Cross Domain Object Detection by Target-Perceived Dual Branch Distillation

Mengzhe He[1,3], Yali Wang*[1,6], Jiaxi Wu[5], Yiru Wang[2],
Hanqing Li[2], Bo Li[2], Weihao Gan[2,4], Wei Wu[2,4], Yu Qiao†[1,4]

[1] ShenZhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2] SenseTime Research [3] University of Chinese Academy of Science [4] Shanghai AI Laboratory, Shanghai, China
[5] Beihang University [6] SIAT Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society
{hemz,yl.wang,yu.qiao}@siat.ac.cn, wujiaxi@buaa.edu.cn
{lihanqing,libo,wuwei}@senseauto.com, {wangyiru,ganweihao}@sensetime.com

Abstract

Cross domain object detection is a realistic and challenging task in the wild. It suffers from performance degradation due to large shift of data distributions and lack of instance-level annotations in the target domain. Existing approaches mainly focus on either of these two difficulties, even though they are closely coupled in cross domain object detection. To solve this problem, we propose a novel Target-perceived Dual-branch Distillation (TDD) framework. By integrating detection branches of both source and target domains in a unified teacher-student learning scheme, it can reduce domain shift and generate reliable supervision effectively. In particular, we first introduce a distinct Target Proposal Perceiver between two domains. It can adaptively enhance source detector to perceive objects in a target image, by leveraging target proposal contexts from iterative cross-attention. Afterwards, we design a concise Dual Branch Self Distillation strategy for model training, which can progressively integrate complementary object knowledge from different domains via self-distillation in two branches. Finally, we conduct extensive experiments on a number of widely-used scenarios in cross domain object detection. The results show that our TDD significantly outperforms the state-of-the-art methods on all the benchmarks. The codes and models will be released afterwards.

1. Introduction

Object detection has achieved remarkable success with the help of advanced deep neural networks [2, 12–14, 26, 28–31, 36]. However, it still faces challenges in realistic applications such as autonomous driving and mobile robots, where data variance is often large due to various conditions of weather, illumination, object appearance, etc. Hence, cross-domain object detection has attracted lots of attention in recent years. In general, there are two difficulties in this problem. First, object detection is more vulnerable to domain shift. The main reason is that, object detection focuses on instance-level prediction, which is more sensitive to object variance in various image styles and contents. Second, object annotations are more expensive and labor-intensive to get, causing the scarcity of discriminative object supervision in a new domain. Both of them inevitably deteriorate the detection performance in target domain.

Figure 1. Two typical examples of detection results on the adverse weather conditions adaptation experiments with different methods. Semi-supervised method UBT [27] lacks awareness of objects in the fog. Adversarial based GPA [49] attempts to exploit the objects in the fog but gives some wrong predictions, such as the motorcycle in the first row and the person in the second row. Our methods can predict the boxes and categories more accurately.

Recently, several approaches have been proposed for cross-domain object detection [5, 24, 32, 34, 49]. Unfortunately, most of them focus on either domain shift or label deficiency, which limits their power in cross domain object detection. For example, domain adaption approaches [5, 34, 49] propose to reduce domain shift via adversarial training. Besides of unstable model optimization, the discrimination ability of the network is limited in such adversarial design. As shown in Figure 1, adversarial based GPA [49] tends to produce wrong predictions on the regions

* Equal contribution. † Corresponding author.

where the target domain characteristics are significant. Alternatively, self-training based approaches [1, 16, 22, 23, 55] study the problem from the viewpoint of semi-supervised learning, and propose to generate pseudo object supervision via label distillation. In this way, many advanced semi-supervised methods can be transferred to this task. However, these approaches are often insufficient to deal with the complex domain shifts. In Figure 1, it is difficult for a semi-supervised method like UBT [27] to aware objects in the target domain. Hence, both types of solutions are unsatisfactory in cross domain object detection.

Based on these discussions, we propose a novel Target-perceived Dual-branch Distillation (TDD) framework, which can effectively tackle domain shift and label deficiency via object perception and knowledge distillation in a concise dual-branch detection network. Specifically, our network consists of a source-adaptive branch and a target-like branch, both of which are elaborately designed to be target-oriented for domain shift reduction. For the source-adaptive branch, we introduce a distinct Target Proposal Perceiver, which leverages iterative cross-attention to discover target-domain contexts for each proposal. As a result, it can adaptively enhance source branch to perceive objects in the target domain image. For the target-like branch, we transfer source images into target-like images. Via training this branch with these labeled images, we can learn discriminative object knowledge of target domain reliably. Finally, we design a concise Dual Branch Self Distillation strategy for network training. It is a tailored mean-teacher style framework to generate pseudo annotations of target images from both source-adaptive and target-like branches. Through three well-designed training steps, namely joint-domain pretraining, cross-domain distillation and dual-teacher refinement, we can progressively integrate complementary object knowledge from different domains to boost cross domain object detection.

In summary, this paper has the following contributions. First, we develop a novel Target-perceived Dual-branch Distillation (TDD) framework, which leverages two distinct detection branches to address both domain shift and label deficiency in a unified teacher-student learning manner. Second, we introduce a smart Target Proposal Perceiver module, which can adaptively guide source detection branch to perceive target domain objects, via cross-attention-style transformer on proposal contexts. Finally, we conduct extensive experiments on a number of widely-used benchmarks and our TDD outperforms the state-of-the-art methods with a large margin.

2. Related Work

Object detection. Object detection is one of the fundamental tasks in computer vision. Boosted by the strong representation ability of deep neural network, object detection has obtained a promising performance in recent years. Previous work can be roughly categorized into two-stage [2, 12–14, 31] and one-stage [28–30, 36] detectors. Recently, some anchor-free [10, 40, 51, 53] and transformer [3, 45, 58] based methods also stand out in the detection task.

Cross domain object detection. [5]first propose image and instance level domain classifiers to implement feature alignment in an adversarial manner. Following this, [34]impose a strong-weak alignment strategy to the local and global features respectively. [15] and [47] employ multi level domain feature alignment. [48]exploit the categorical consistency between image-level and instance-level prediction with the help of a multi-label classification model. [17] propose a center-aware feature alignment method to allow the discriminator to focus on features coming from the object region. Some other works [16, 24, 32, 38, 57] add additional constraint during the adversarial learning stage. [54, 56] emphasis the different strategies to deal with foreground and background features.

Another mainstream method [1, 16, 22, 23, 55] is dedicated to solving the problem of inaccurate label in target domain. [22]retrain the object detector using the original labeled data and the refined machine-generated annotations in the target domain. [1] study the problem from the viewpoint of semi-supervised learning and integrate the object relations into the measure of consistency cost between teacher and student modules. [9]propose a cross-domain distillation method which utilizes both the source-like and target-like images. It uses soft label and instance selection to heal the model bias in Mean-Teacher. Different from [9], our method proposes a dual-branch framework with a cross-domain perceiver for teacher-student mutual learning.

Semi-supervised object detection. Semi-supervised object detection attempts to solve the problem when there are only a part of annotations for the train set. In this setting [20] propose a consistency-based method, enforcing the predictions consistency between an input image and its flipped version. [37] pre-train a detector using a small amount of labeled data and generate pseudo-labels on unlabeled data to fine-tune the pre-trained detector. [27] propose to use strong and weak augmentations to improve the mean-teacher method and can get more accurate pseudo labels by EMA training. Those methods can be easily applied to the cross domain object detection problem owing to the similar data setting. But they did not take the domain difference into consideration, which limited their detection performance unavoidably.

3. Proposed Methods

3.1. Overview

As shown in Figure 2, we propose a novel Target-perceived Dual-branch Distillation framework (TDD),

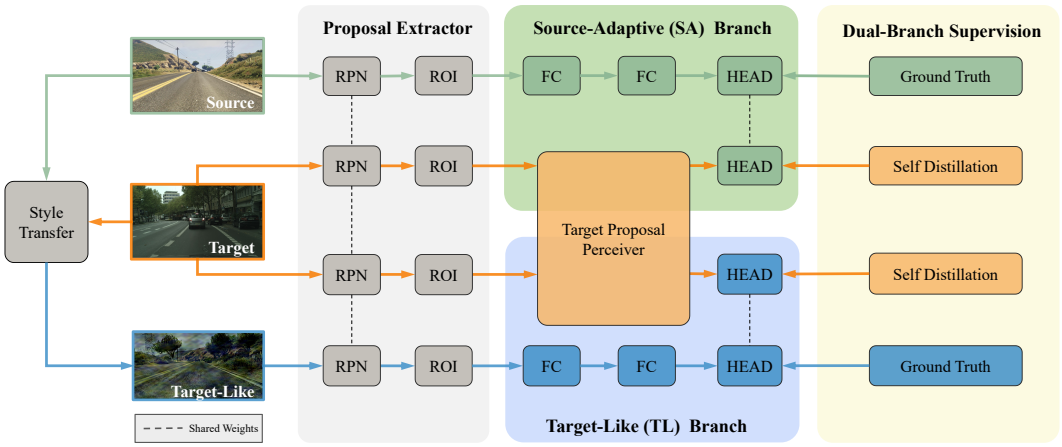


Figure 2. An overview of Target-perceived Dual-branch Distillation framework. To simplify the description, ROI refers to the operation to get proposal features of each image. First, a source domain image is transferred into target-like domain. All of the images from three domains are fed to a shared proposal extractor to get proposals and proposal features. Then, the proposal features of source and target-like images are used to train corresponding branches with supervision of ground truth. Moreover, we feed the proposal features of a real target domain image into both branches, for learning object knowledge from both source and target-like domains. As the images from target domain are not annotated, the model is optimized by self-distillation.

which tackles domain shift and label deficiency together in cross domain object detection task.

First, we introduce a style transfer module from the aspect of input image. It is used to transfer source images into style that is close to target domain. In this case, we can bridge the domain gap by such target-like domain. Moreover, since target-like images inherit label annotations from the corresponding source images, they can be used as extra object supervision in the target-like domain. In this paper, we mainly use a concise and effective Fourier transform [50] method as this module.

Second, we design a novel dual branch detection network from the aspect of model architecture. Via such design, we can effectively extract complementary object knowledge from different domains to boost object detection on the target images. Basically, our network consists of a shared proposal extractor and two individual detection branches. The former allows us to construct domain-invariant feature space of all the images for domain generalization, while the latter preserves domain-specific object characteristics of each image for domain discrimination. Specifically, two detection branches are Source-Adaptive (SA) and Target-Like (TL) branch respectively. We feed the proposals of source images to train the SA branch, while feeding the proposals of target-like images to train the TL branch. Moreover, the proposals of a real target image are sent into both branches, for learning object knowledge from both source and target-like domains. However, source domain may be significantly different from target domain. In this case, the proposals of a target image cannot be detected accurately in the SA branch, without any target-oriented guidance. To tackle this problem, we design a novel Target Proposal Perceiver. Inspired by perceiver in [19], it smartly uses iterative cross attention between proposal features in two branches. In this case, we leverage contextual proposals of TL branch as guidance, which can effectively guide SA branch to perceive object proposals in the target domain. We will explain the details of this module in 3.2.

Finally, we introduce a concise dual-branch self-distillation approach from the aspect of supervision. As introduced before, all the images do not have any annotations in target domain. Hence, it is critical to generate reliable supervision in this domain. Thanks to our dual-branch network, we can construct discriminative pseudo labels of each target image from the cooperative SA and TL branches. To effectively leverage these pseudo labels, our self-distillation is based on teacher-student mutual learning, which can dynamically adjust teacher in the training procedure to progressively boost target-domain supervision of our two branches. We will explain the details in 3.3.

3.2. Target Proposal Perceiver

As discussed in our TDD framework, we feed proposal features of each target-domain image respectively into SA and TL branches, for learning object knowledge from both domains. However, SA branch is not good at exploiting objects from these features due to the large shift between the source and real target domain. To guide SA branch to discover target domain objects, we propose a novel Target Proposal Perceiver between SA and TL branches, which can progressively exploit object contexts in the TL branch to enhance proposal features in the SA branch.

Note that, we inherit the name of Perceiver from [19], since our motivation is also to mimic humans and other animals to take in data from many sources and integrate it seamlessly. But different from the generic Perceiver

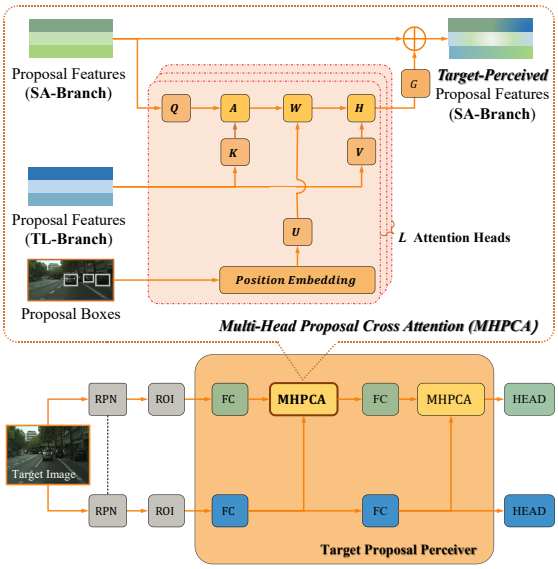


Figure 3. The structure of our Target Proposal Perceiver. The cross attention between SA and TL is explored to help source branch perceive target domain objects.

[19] architecture, our Target Proposal Perceiver is elaborately tailored for cross-domain object detection, by using Transformer-style cross attention iteratively to reduce domain shift in the instance level.

As shown in Figure 3, we feed a target-domain image $\mathbf{X}^t$ into proposal extractor, and generate its proposal features $\mathbf{P}^t$. Subsequently, we put these proposal features respectively into SA and TL branches, where Target Proposal Perceiver leverages cross attention to process them in the following,

$$\mathbf{\Phi}_{SA} = \mathcal{F}_{SA}(\mathbf{P}^t), \tag{1}$$

$$\mathbf{\Phi}_{TL} = \mathcal{F}_{TL}(\mathbf{P}^t), \tag{2}$$

$$\mathbf{\Psi}_{SA} = \text{MHPCA}(\mathbf{\Phi}_{SA}, \mathbf{\Phi}_{TL}). \tag{3}$$

First, to extract object knowledge from both SA and TL branches, we use the FC layer $\mathcal{F}_{SA}(.)$ and $\mathcal{F}_{TL}(.)$ to encode $\mathbf{P}^t$ as source features $\mathbf{\Phi}_{SA}$ and target-like features $\mathbf{\Phi}_{TL}$ in Eq. (1)-(2). Second, we introduce a novel Multi-Head Proposal Cross Attention (MHPCA) between $\mathbf{\Phi}_{SA}$ and $\mathbf{\Phi}_{TL}$ in Eq. (3). This allows us to leverage target-like proposal features $\mathbf{\Phi}_{TL}$ as context guidance, for enhancing source proposal features $\mathbf{\Phi}_{SA}$ to perceive objects in the target image.

Proposal Cross Attention. Specifically, our MHPCA is a concise Transformer style with Query-Key-Value. In each cross attention head, we use FC layers to encode $\mathbf{\Phi}_{SA}$ as Query, and encode $\mathbf{\Phi}_{TL}$ as Key and Value. The similarity between Key and Query is used to discover affinity between $\mathbf{\Phi}_{SA}$ and $\mathbf{\Phi}_{TL}$. Then, we use such affinity as guidance to aggregate target-like features $\mathcal{V}(\mathbf{\Phi}_{TL})$ as cross-domain contexts for the SA branch.

$$\mathbf{H}_{TL} = \mathcal{W}(\mathcal{Q}(\mathbf{\Phi}_{SA}), \mathcal{K}(\mathbf{\Phi}_{TL})) \cdot \mathcal{V}(\mathbf{\Phi}_{TL}), \tag{4}$$

where Query, Key and Value are respectively $\mathcal{Q}(\mathbf{\Phi}_{SA})$, $\mathcal{K}(\mathbf{\Phi}_{TL})$ and $\mathcal{V}(\mathbf{\Phi}_{TL})$. The affinity function is $\mathcal{W}$. Typically, scaled dot-product is used as $\mathcal{W}$ in transformer [41], $\mathbf{A}_{i,j} = \mathcal{Q}_i(\mathbf{\Phi}_{SA}) \cdot \mathcal{K}_j^\top(\mathbf{\Phi}_{TL})/\sigma$, where σ is a scale parameter that is root square of the dimension of a query feature vector. However, we consider an object detection problem, where spatial position information can be important to describe similarity between proposals. In this work, the geometry weight in [18] is used to describe positional similarity between any two proposal boxes. We use this geometry weight $\mathbf{U}$ to enhance feature similarity $\mathbf{A}$ and describe proposal affinity in Eq. (4) via a weighted formulation of softmax, i.e., $\mathcal{W}(\mathcal{Q}(\mathbf{\Phi}_{SA}), \mathcal{K}(\mathbf{\Phi}_{TL})) = \mathbf{W}$,

$$\mathbf{W}_{i,j} = \frac{\mathbf{U}_{i,j} \cdot \exp(\mathbf{A}_{i,j})}{\sum_{k=1}^{K} \mathbf{U}_{i,k} \cdot \exp(\mathbf{A}_{i,k})}, \tag{5}$$

where $\mathbf{W}_{i,j}$ refers to affinity score between proposal i in SA branch and proposal j in TL branch.

Iterative MHPCA. After obtaining target-like contexts $\mathbf{H}_{TL}$ from each cross attention head, we use FC layer $\mathcal{G}(.)$ to summarize all these contexts from L attention heads to construct MHPCA, denoted as $\mathbf{\Psi}_{SA} = \mathbf{\Phi}_{SA} + \mathcal{G}([\mathbf{H}_{TL}^{(1)}, ..., \mathbf{H}_{TL}^{(L)}])$. In this case, we enhance source proposal features $\mathbf{\Phi}_{SA}$ into target-perceived ones $\mathbf{\Psi}_{SA}$, which allows SA branch to be aware of related object contexts in the target image. Additionally, we perform MHPCA in an iterative manner, by which our Target Proposal Perceiver can progressively exploit target-like proposal contexts from TL branch to boost learning capacity of SA branch. Typically, there are two FC layers to encode proposal features in Faster RCNN. Hence, we iteratively use MHPCA twice in our design, as shown in Figure 3.

3.3. Dual-Branch Self Distillation

After introducing our network, we explain how to train it for cross domain object detection. As mentioned before, the images are unlabeled in the target domain. Hence, it is critical to generate reliable pseudo annotations of these images for effective training. To achieve this goal, we design a generic Dual-Branch Self Distillation approach, which can generate pseudo labels from both SA and TL branches to cooperatively boost our detection network via self-training. Specifically, it is based on the general procedure of teacher-student mutual learning [27, 39], but with elaborate designs for cross domain object detection. As shown in Figure 4, it consists of three key stages, i.e., Joint-Domain Pretraining, Cross-Domain Distillation, and Dual-Teacher Refinement.

Joint-Domain Pretraining. This stage is to generate reliable initialization of dual-branch network. As mentioned before, target-like images have same annotations inheriting from source images. Hence, we pretrain our dual-branch network jointly, by multi-task learning on the labeled images of both source and target-like domains. Specifically,

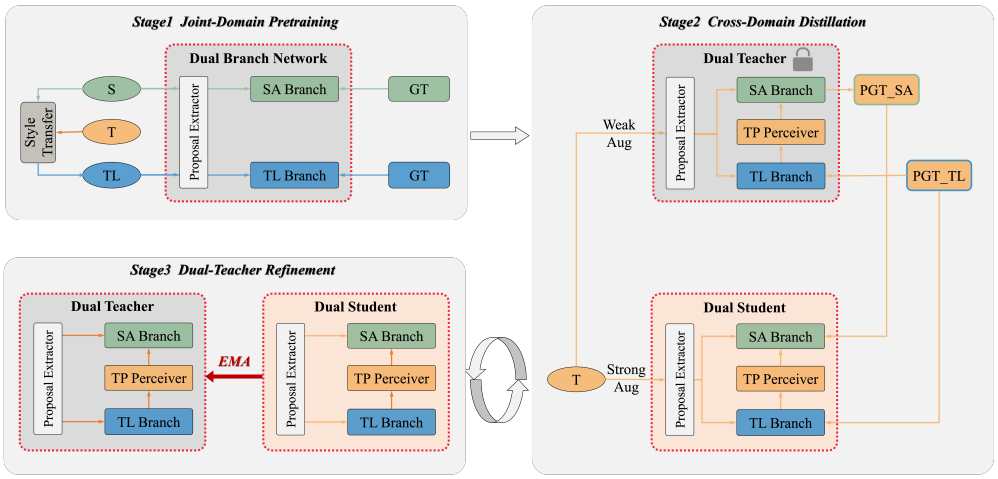


Figure 4. The whole training process of our Dual-Branch Self Distillation models. First, in the joint-domain pretraining stage, we pretrain our dual-branch network jointly, by multi-task learning on the labeled images of both source and target-like domains. Second, in the cross-domain distillation stage, we feed a target-domain image into the fixed and well-trained teacher, which can generate pseudo object annotations from both SA and TL branches. Finally, to generate more stable pseudo annotations, we refine teacher gradually from student via Exponential Moving Average (EMA).

the training loss in this stage consists of three terms.

$$\mathcal{L}_{JDP} = \mathcal{L}_{RPN}^{(\mathcal{S}+\mathcal{TL})} + \mathcal{L}_{SA}^{(\mathcal{S})} + \mathcal{L}_{TL}^{(\mathcal{TL})}. \tag{6}$$

First, RPN is shared among all the domains to generate domain-invariant feature. We use both source and target-like data to train this module, i.e., $\mathcal{L}_{RPN}^{(\mathcal{S}+\mathcal{TL})} = \mathcal{L}_{RPN}(\mathbf{X}^s, \mathbf{Y}^s) + \mathcal{L}_{RPN}(\mathbf{X}^{tl}, \mathbf{Y}^{tl})$, where RPN loss contains the RPN classification and regression losses in Faster RCNN [31]. Then, different detection branches are used to learn different domain-specific object knowledge. Hence, we use source and target-like data respectively to train SA and TL branches, i.e., $\mathcal{L}_{SA}^{(\mathcal{S})} = \mathcal{L}_{SA}(\mathbf{X}^s, \mathbf{Y}^s)$ and $\mathcal{L}_{TL}^{(\mathcal{TL})} = \mathcal{L}_{TL}(\mathbf{X}^{tl}, \mathbf{Y}^{tl})$ where each branch loss contains the ROI classification and regression losses in Faster RCNN.

Cross-Domain Distillation. After joint-domain pretraining, we leverage the well-initialized network to generate pseudo annotations of unlabeled images in the target domain. In this case, we can further adjust our network without target-domain ground truth labels. As shown in Figure 4, this stage is a concise self distillation procedure, where both teacher and student are based on dual branch network. Specifically, we feed a target-domain image into the fixed and well-trained teacher, which can generate pseudo object annotations from both SA and TL branches. We use NMS to remove the duplicated boxes and then set a threshold to obtain confident box predictions as object annotations of this target image in each branch. Subsequently, we also feed this target image into the learnable student, and train student by pseudo annotations from teacher.

$$\mathcal{L}_{CDD} = \mathcal{L}_{RPN}^{(\mathcal{T})} + \mathcal{L}_{SA}^{(\mathcal{T})} + \mathcal{L}_{TL}^{(\mathcal{T})}. \tag{7}$$

Since pseudo labels $\hat{\mathbf{Y}}_{SA}^t$ and $\hat{\mathbf{Y}}_{TL}^t$ are from SA and TL branches, the RPN loss contains two terms $\mathcal{L}_{RPN}^{(\mathcal{T})} = \mathcal{L}_{RPN}(\mathbf{X}^t, \hat{\mathbf{Y}}_{SA}^t) + \mathcal{L}_{RPN}(\mathbf{X}^t, \hat{\mathbf{Y}}_{TL}^t)$. Moreover, both SA and TL branches are also trained with pseudo labels of target-domain images, i.e., $\mathcal{L}_{SA}^{(\mathcal{T})} = \mathcal{L}_{SA}(\mathbf{X}^t, \hat{\mathbf{Y}}_{SA}^t)$ and $\mathcal{L}_{TL}^{(\mathcal{T})} = \mathcal{L}_{TL}(\mathbf{X}^t, \hat{\mathbf{Y}}_{TL}^t)$. Additionally, it is important to increase diversity of student to refine teacher afterwards. As suggested in [27], for each target image, we use its strong augmentation as input of student to predict object boxes, while using its weak augmentation as input of teacher to provide reliable pseudo annotations. Finally, we also use Eq. (6) to train student network with source and target-like images in this stage, to reduce learning difficulties in two detection branches.

Dual-Teacher Refinement. To generate more stable pseudo annotations, we refine teacher gradually from student via Exponential Moving Average (EMA) [27, 39],

$$\Theta_{teacher} \leftarrow \alpha\Theta_{teacher} + (1-\alpha)\Theta_{student} \tag{8}$$

where $\Theta_{teacher}$ and $\Theta_{student}$ are the learnable parameters in teacher and student models. Note that, we perform distillation and refinement in an iterative manner, which can boost cross domain object detection by mutual learning, i.e., teacher generates pseudo labels to train student, and student passes what it learns to update teacher.

Finally, we explain how to train Target Proposal Perceiver in this procedure. We only train it in the last two stages. In the cross-domain distillation stage, we use the pretrained network as teacher, and use the pretrained network with randomly-initialized Target Proposal Perceiver as student. After a number of training iterations in this stage, we can obtain well-trained Target Proposal Perceiver. Subsequently, in the refinement stage, we update teacher from

the entire student network where all the modules are fully trained. From then on, distillation and refinement can be iteratively performed without any difficulties. Moreover, TPR is just used in training stage to guide SA branch. With the dual-branch framework, we only use the SA branch teacher to get the detection results during inference. As it has been well refined by student and TL branch.

4. Experiments

In this section, we conduct experiments on popular cross domain object detection benchmarks with distinct domain shift, including Adverse Weather Conditions Adaptation, Synthetic to Real Adaptation and Cross Camera Adaptation.

4.1. Implementation details

We adopt Faster R-CNN with the VGG16 and Res50 pretrained on ImageNet [8] as the backbone network. The shorter edge of each input image is resized to 600 pixels following the implementation of Faster RCNN with ROI-alignment [13]. The network is trained by SGD [33] optimizer with 0.0005 weight decay and 0.9 momentum. The learning rate and maximum training iterations are set as 0.01 and 25000 for all experiments, with 9000 iterations for the joint-domain pretraining stage and 16000 iterations for cross-domain distillation and dual-teacher refinement stage. Follow the [27], we use Focal-loss as the classification loss and the strong-weak augmentations are used during our whole training stage. For our proposal cross attention, we set the attention head number L=16. The proposal features are encoded to Key-Query-Value by three FC layers with output dimension=1024. We set the frequency parameter$\beta = 0.1$ for the Fourier transform module. The threshold to obtain pseudo annotations of target image is set to 0.7. During the dual-teacher refinement stage, we set the EMA ratio $\alpha = 0.9996$ to update teacher model. We use 8 NVIDIA GeForce 1080 Ti GPUs for training in our experiments. Each mini-batch contains 2 images per GPU, one from the source domain and the other from target domain.

4.2. Adverse Weather Conditions Adaptation

Datesets. In this experiment, we use Cityscapes as source domain and Foggy Cityscapes as target domain to implement adaptation under adverse weather conditions (C→F). Cityscapes [6] is a dataset of real urban scenes containing 3,475 images. 2,975 images are used for training and the remaining 500 for validation. Foggy Cityscapes [35] is a synthetic dataset generated from the Cityscapes. We use the fog level ($\beta = 0.02$) with highest intensity in our experiments. The Cityscapes train set and unlabeled Foggy Cityscapes train set are used for training and the validation set of Foggy Cityscapes is used for evaluation.

Results. The detection results are demonstrated in Table 1. Source only denotes the Faster RCNN model trained with only source domain data. Oracle(tgt) model is trained with labeled target domain data. Oracle(src+tgt) model is trained with labeled data from both source and target domain. Same augmentations are also used for training the oracle models. We compare with the methods implemented with same backbone for fair comparison. For the VGG-based methods, the state-of-the-art MeGA [42] has achieved 41.8% mAP, while our results show a significant +1.3% improvement. For the Res50-based methods, we outperform all prior works and get a significant mAP gain of +1.4%. It is worth noting that our methods show a competitive performance with two oracle models. It proves that our model can perceive target domain knowledge while retaining the useful information of the source domain for discrimination.

Table 1. The mean Average Precision (mAP) of different models on Foggy Cityscapes validation set for C → F transfer.

method	Arch	person	rider	car	truck	bus	train	motor	bike	mAP
DA-Faster [5]	V16	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
SCDA [57]	V16	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
D&Match [24]	V16	30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
SWDA [34]	V16	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
ICR-CCR [48]	V16	32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
HTCN [4]	V16	33.2	47.5	47.9	31.6	47.4	40.9	32.3	37.1	39.8
SAPNet [25]	V16	40.8	46.7	59.8	24.3	46.8	37.5	30.4	40.7	40.9
ATF [16]	V16	34.6	47.0	50.0	23.7	43.3	38.7	33.4	38.8	38.7
CDN [38]	V16	35.8	45.7	50.9	30.1	42.5	29.8	30.8	36.5	36.6
UMT [9]	V16	33.0	46.7	48.6	34.1	56.5	46.8	30.4	37.3	41.7
MeGA [42]	V16	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8
RPA [54]	V16	33.4	44.3	50.1	29.9	44.8	39.1	29.9	36.3	38.5
source only	V16	28.5	34.2	39.9	14.7	26.3	11.4	23.4	28.3	25.8
TDD(ours)	V16	39.6	47.5	55.7	33.8	47.6	42.1	37.0	41.4	**43.1**
ocacle(tgt)	V16	39.1	44.9	56.7	33.3	50.4	34.8	32.3	39.0	41.3
ocacle(src+tgt)	V16	39.5	47.5	58.1	34.2	49.3	41.9	36.4	41.0	43.5
DA-Faster [5]	R50	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0
D&Match [24]	R50	31.8	40.5	51.0	20.9	41.8	34.3	26.6	32.4	34.9
SW-DA [34]	R50	31.8	44.3	48.9	21.0	43.8	28.0	28.9	35.8	35.3
SC-DA [57]	R50	33.8	42.1	52.1	26.8	42.5	26.5	29.2	34.5	35.9
MTOR [1]	R50	30.6	41.4	44.0	21.9	38.6	40.6	28.3	35.6	35.1
AFAN [43]	R50	42.5	44.6	57.0	26.4	48.0	28.3	33.2	37.1	39.6
GPA [49]	R50	32.9	46.7	54.1	24.7	45.7	41.1	32.4	38.7	39.5
ViSGA [32]	R50	38.8	45.9	57.2	29.9	50.2	51.9	31.9	40.9	43.3
SFA [44]	R50	46.5	48.6	62.6	25.1	46.2	29.4	28.3	44.0	41.3
DSS [46]	R50	50.9	57.6	61.1	35.4	50.9	36.6	38.4	51.1	47.8
MKT [7]	R50	43.5	52.0	63.2	34.7	52.7	45.8	37.1	49.4	47.3
source only	R50	36.9	36.1	44.5	21.7	32.3	9.2	21.5	32.4	28.3
TDD(ours)	R50	50.7	53.7	68.2	35.1	53.0	45.1	38.9	49.1	**49.2**
oracle(tgt)	R50	50.1	51.7	70.1	33.4	49.5	42.8	37.6	44.3	47.4
oracle(src+tgt)	R50	50.0	50.2	69.9	35.6	56.3	47.4	41.0	43.4	49.2

4.3. Synthetic to Real Adaptation

Datesets. In this experiment, the model is adapted from synthetic data to real world examples. Sim10k is used as source domain dataset and Cityscapes represents target domain (S→ C). SIM10K [21] is a simulated dataset containing 10,000 images. We train the detector only on the common class "car". The whole dataset Sim10k and unlabeled train set of Cityscapes is used for training and the validation set of Cityscapes is used for evaluation.

Table 2. The car Precision (mAP) of different models on Cityscapes validation set for S→ C and K → C adaptaion.

method	Arch	S→ C	K → C	method	Arch	S→ C	K → C
DA-Faster [5]	V16	39.0	38.5	DA-Faster [5]	R50	41.9	41.8
SCDA [57]	V16	43.0	42.5	SCDA [57]	R50	45.1	43.6
SWDA [34]	V16	47.7	37.9	SWDA [34]	R50	44.6	43.2
CoT [55]	V16	44.5	43.6	GPA [49]	R50	47.6	47.9
SAPNet [25]	V16	44.9	43.4	ViSGA [32]	R50	49.3	47.6
EPM [17]	V16	49.0	43.2	SFA [44]	R50	52.6	41.3
ATF [16]	V16	42.8	42.1	D&Match [24]	R50	43.9	42.7
MeGA [42]	V16	44.8	43.0	DSS [46]	R50	44.5	42.7
RPA [54]	V16	45.7	-	MKT [7]	R50	50.2	44.3
C2F [56]	V16	43.8	-	AFAN [43]	R50	45.5	-
UMT [9]	V16	43.1	-	MTOR [1]	R50	46.6	-
source only	V16	37.8	30.2	source only	R50	42.8	32.5
TDD(ours)	V16	**53.4**	**47.4**	TDD(ours)	R50	**63.3**	**49.8**
oracle(tgt)	V16	60.0	60.0	oracle(tgt)	R50	75.9	75.9
oracle(src+tgt)	V16	60.1	62.5	oracle(src+tgt)	R50	76.4	75.8

Table 3. The mean Average Precision (mAP) of different models on BDD100k daytime validation set for C → B transfer.

method	Arch	person	rider	car	truck	bus	motor	bicycle	mAP
DA-Faster [5]	V16	26.9	22.1	44.7	17.4	16.7	17.1	18.8	23.4
SWDA [34]	V16	30.2	29.5	45.7	15.2	18.4	17.1	21.2	25.3
ICR-CCR [48]	V16	31.4	31.3	46.3	19.5	18.9	17.3	23.8	26.9
source only	V16	29.3	28.2	45.7	15.5	16.6	16.0	22.1	24.8
TDD(ours)	V16	39.6	38.9	53.9	24.1	25.5	24.5	28.8	**33.6**
oracle(tgt)	V16	39.7	35.9	57.9	47.1	48.0	32.3	33.0	42.0
oracle(src+tgt)	V16	39.6	39.2	59.4	45.6	48.0	31.0	33.8	42.4
source only	R50	50.4	33.3	67.4	18.1	20.8	19.6	28.9	34.1
TDD(ours)	R50	57.9	47.4	74.5	31.5	27.5	32.0	36.5	**43.9**
oracle(tgt)	R50	68.0	52.0	83.7	61.2	61.6	44.9	49.9	60.2
oracle(src+tgt)	R50	69.5	54.1	84.4	61.1	61.5	43.8	53.2	61.1

Results. The results of car AP are reported in Table 2. We can see our proposed TDD methods can achieve the state-of-the-art performance between two dissimilar domains. It outperforms VGG-based EPM [17] by +4.4% and Res50-based SFA [44] +10.7%, which shows a stable ability of our methods to tackle domain adaptation problems.

4.4. Cross Camera Adaption

Datesets. We conduct on two cross camera adapataion experiments involving KITTI [11], Cityscapes and BDD100k [52] datasets. In the first experiment, we adapt from KITTI to Cityscapes, where only the category car is used for evaluation (K→C). KITTI is a similar scene dataset to Cityscapes except that KITTI has different camera setup. It consists of 7,481 labeled images for training. In the second experiment, we adapt from Cityscapes to BDD100K (C→B), which is a more challenging setting with more categories and scenes. The daytime subset of BDD100k are used as our target domain, including 36,278 training and 5,258 validation images.

Results. The KITTI adaptation results are shown in Table 2. We outperform the sota VGG-based approach by 3.8% and R50-based approach by 1.9%. Meanwhile, the results on BDD100K are summarized in Table 3. Our method surpasses all the previous works with a large margin. This demonstrates that our method performs well under more complex situation. We also observe an obvious improvement with R50 backbone, increasing the source only results by 9.8%. It further verifies the robustness of our methods.

Table 4. Dual Branch Structure

Structure	S	T	TL	C→ F	S→ C	C→ B
Single	✓			34.8	48.3	34.3
	✓	✓		41.2	59.0	38.9
	✓	✓	✓	47.4	61.1	39.4
Dual	✓	✓	✓	**48.3**	**62.6**	**42.2**

Table 5. Multi Head Proposal Cross Attention

Target Proposal Perceiver	C→ F	S→ C	C→ B
without	48.3	62.6	42.2
with	**49.2**	**63.3**	**43.9**
Self-Attention	46.8	61.0	40.6
Sym Cross-Attention	48.1	62.4	43.7
Asym Cross-Attention	**49.2**	**63.3**	**43.9**

Table 6. Dual-Branch Self Distillation Procedure

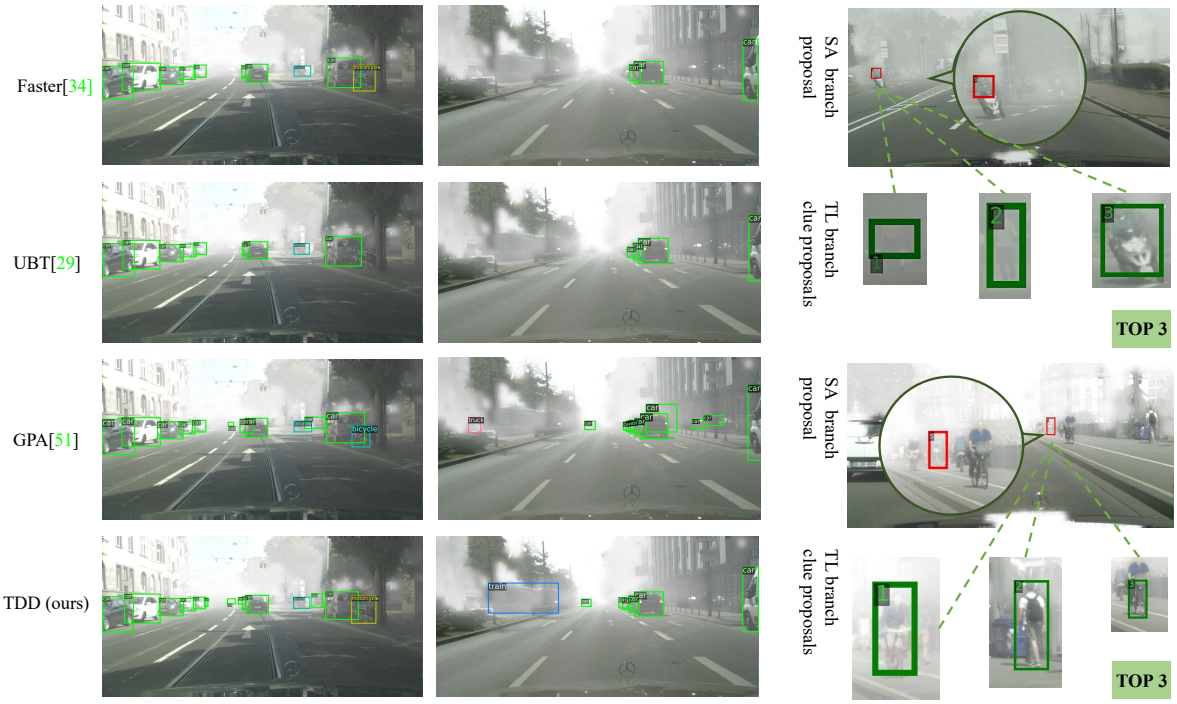
Dual-Branch Self Distillation	C→ F	S→ C	C→ B
JDP	37.4	56.7	37.5
JDP+CDD	44.1	62.1	42.7
JDP+CDD+DTR	**49.2**	**63.3**	**43.9**
Refine α=0.96	39.3	59.1	28.7
Refine α=0.996	48.4	**63.6**	41.5
Refine α=0.9996	**49.2**	63.3	**43.9**

4.5. Ablation Studies and Analysis

To verify designs in our network, we conduct a set of ablation studies on the Res50 backbone.

Dual branch. To validate the effectiveness of our dual branch structure, we conduct a set of ablation studies with images from different domains. Table 4 shows the results of different experiments. When implemented with a single branch, target domain images are supervised by the pseudo annotations generated by the single teacher branch, while the target-like images were feed to the network paired with source images. The proposed Target-proposal-perceiver is not used in this dual-branch structure to fairly compare with the single-branch experiments. We can observed that the model performance improved step by step with the target and target-like images participating in the training. This verifies our motivation that data from each domain is useful. The dual-branch experiment outperforms all the single branch methods, which demonstrates that our dual-branch distillation framework can effectively retain the useful source domain knowledge and explore target domain information simultaneously.

Multi Head Proposal Cross Attention. We implement the MHPCA to guide the source adaptive branch to learn knowledge closer to the target domain with the help of target-like domain branch. Table 5 shows the effectiveness of our MHPCA module. First, we can see a significant improvement with the MHPCA module added. Moreover, to



(a) Qualitative results. (b) Visualization for MHPCA head

Figure 5. (a): Qualitative detection results of on the $C \rightarrow F$ scenario for different models. We set confidence thresh=0.6 for visualization. (b): Top 3 most relevant clue proposals in TL branch found by our MHPCA.

validate that domain difference between two branches matters for our attention model, we also experiment with self-attention. Besides, to explore the guide manner between two branch, we also add a cross attention head in the target-like branch. Sym Cross-Attention means that a same cross-attention module is added on both the source-adaptive and target-like branch. While Asym Cross-Attention refers to our TDD methods which equip the SA branch with the cross attention module. Our asymmetric TDD performs best in these three manners. It also confirms that in our framework, the cross attention manner is needed for the SA branch due to the lack of target domain knowledge.

Dual-Branch Self Distillation. We do ablation studies to verify the effectiveness of our dual-branch self distillation procedure, which is composed of Joint-Domain Pretraining (JDP), Cross-Domain Distillation (CDD), and Dual Teacher Refinement (DTR) steps. We see from Table 6 that all of the three steps in our method improve former step results. We also experiment with different EMA rate α in dual-teacher refinement stage. The smaller the value of α, the more information teacher receives from the target image during the refine stage. When α is set to be small (e.g., 0.96), the model performance drops significantly. Additionally, when $\alpha = 1$, the teacher is not refined which is JDP+CDD in Table 6. All these show the teacher model should be updated gradually. A reasonable rate is needed to impart the target domain knowledge learned by student to teacher.

Qualitative results. We show the detection results of Faster [31], GPA [49], UBT [27] and our TDD in Figure 5 (a). We can see that many objects can not be detected by the Faster RCNN and UBT due to the heavy fog, while the GPA attempts to capture objects in the fog but gives wrong prediction. Our TDD can localize and classify objects more accurately. We also visualize the working mechanism for our cross-domain MHPCA module. For a SA branch proposal, our attention head can discover useful contextual proposal features in TL branch as clues for detection. As the top image shown in Figure 5(b), a rider is classified with the guidance of a motorcycle and two person proposals.

5. Conclusion

In this work, we propose a novel Target-perceived Dual branch Distillation framework. Through a target proposal perceiver and our dual-branch self distillation procedure, we tackle domain shift and label deficiency together in cross domain object detection. Extensive experiments are conducted on multiple benchmarks, and the results clearly show that our TDD surpasses the existing state-of-the-art models. **Acknowledgement:** This work is partially supported by the National Natural Science Foundation of China (61876176,U1813218), the Joint Lab of CASHK, Guangdong NSF Project (No. 2020B1515120085,the Shenzhen Research Program(RCJC20200714114557087), the Shanghai Committee of Science and Technology, China (Grant No. 21DZ1100100).

References

[1] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11449–11458, 2019. 2, 6, 7

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 1, 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2

[4] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8866–8875, 2020. 6

[5] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2018. 1, 2, 6, 7

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[7] Botos Csaba, Xiaojuan Qi, Arslan Chaudhry, Puneet Dokania, and Philip Torr. Multilevel knowledge transfer for cross-domain object detection. *arXiv preprint arXiv:2108.00977*, 2021. 6, 7

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 6

[9] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4091–4101, June 2021. 2, 6, 7

[10] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6569–6578, 2019. 2

[11] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision Pattern Recognition*, 2012. 7

[12] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 1, 2

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 1, 2, 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 1, 2

[15] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6667–6676, 2019. 2

[16] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 309–324, Cham, 2020. Springer International Publishing. 2, 6, 7

[17] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 733–748, Cham, 2020. Springer International Publishing. 2, 7

[18] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3588–3597, 2018. 4

[19] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. *CoRR*, abs/2103.03206, 2021. 3, 4

[20] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32:10759–10768, 2019. 2

[21] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, and R. Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *IEEE International Conference on Robotics Automation*, 2017. 6

[22] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William Macready. A robust learning approach to domain adaptive object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 480–490, 2019. 2

[23] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6091–6100, 2019. 2

[24] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12448–12457, 2019. 1, 2, 6, 7

[25] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 481–497, Cham, 2020. Springer International Publishing. 6, 7

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

[27] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021. 1, 2, 4, 5, 6, 8

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 1, 2

[29] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 1, 2

[30] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1, 2

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 1, 2, 5, 8

[32] Farzaneh Rezaeianaran, Rakshith Shetty, Rahaf Aljundi, Daniel Olmeda Reino, Shanshan Zhang, and Bernt Schiele. Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9204–9213, October 2021. 1, 2, 6, 7

[33] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951. 6

[34] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6949–6958, 2019. 1, 2, 6, 7

[35] Sakaridis, Christos, Dai, Dengxin, Van, Gool, and Luc. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 2018. 6

[36] Zhiqiang Shen, Zhuang Liu, Jianguo Li, Yu-Gang Jiang, Yurong Chen, and Xiangyang Xue. Dsod: Learning deeply supervised object detectors from scratch. In *Proceedings of the IEEE international conference on computer vision*, pages 1919–1927, 2017. 1, 2

[37] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 2

[38] Peng Su, Kun Wang, Xingyu Zeng, Shixiang Tang, Dapeng Chen, Di Qiu, and Xiaogang Wang. Adapting object detectors with conditional domain normalization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 403–419, Cham, 2020. Springer International Publishing. 2, 6

[39] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. 4, 5

[40] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 2

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4

[42] Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A. Sindagi, and Vishal M. Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4516–4526, June 2021. 6, 7

[43] Hongsong Wang, Shengcai Liao, and Ling Shao. Afan: Augmented feature alignment network for cross-domain object detection. *IEEE Transactions on Image Processing*, 30:4046–4056, 2021. 6, 7

[44] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. Exploring sequence feature alignment for domain adaptive detection transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1730–1738, 2021. 6, 7

[45] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021. 2

[46] Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, YangYang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9603–9612, 2021. 6, 7

[47] Rongchang Xie, Fei Yu, Jiachao Wang, Yizhou Wang, and Li Zhang. Multi-level domain adaptive learning for cross-domain detection. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3213–3219, 2019. 2

[48] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11721–11730, 2020. 2, 6, 7

[49] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12352–12361, 2020. 1, 6, 7, 8

[50] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4084–4094, 2020. 3

[51] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9657–9666, 2019. 2

[52] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Dar-

rell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2633–2642, 2020. 7

[53] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020. 2

[54] Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12425–12434, June 2021. 2, 6, 7

[55] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 86–102, Cham, 2020. Springer International Publishing. 2, 7

[56] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13763–13772, 2020. 2, 7

[57] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 687–696, 2019. 2, 6, 7

[58] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2