Target-Relevant Knowledge Preservation for Multi-Source Domain Adaptive Object Detection

Jiaxi Wu[1,2], Jiaxin Chen[2]*, Mengzhe He[3], Yiru Wang[4], Bo Li[4],
Bingqi Ma[4], Weihao Gan[4,5], Wei Wu[4,5], Yali Wang[3], Di Huang[1,2]
[1]State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Science
[4]SenseTime Research [5]Shanghai AI Laboratory
{wujiaxi,jiaxinchen,dhuang}@buaa.edu.cn, {hemz, yl.wang}@siat.ac.cn,
{libo, mabingqi, wuwei}@senseauto.com, {wangyiru, ganweihao}@sensetime.com

Abstract

Domain adaptive object detection (DAOD) is a promising way to alleviate performance drop of detectors in new scenes. Albeit great effort made in single source domain adaptation, a more generalized task with multiple source domains remains not being well explored, due to knowledge degradation during their combination. To address this issue, we propose a novel approach, namely target-relevant knowledge preservation (TRKP), to unsupervised multi-source DAOD. Specifically, TRKP adopts the teacher-student framework, where the multi-head teacher network is built to extract knowledge from labeled source domains and guide the student network to learn detectors in unlabeled target domain. The teacher network is further equipped with an adversarial multi-source disentanglement (AMSD) module to preserve source domain-specific knowledge and simultaneously perform cross-domain alignment. Besides, a holistic target-relevant mining (HTRM) scheme is developed to re-weight the source images according to the source-target relevance. By this means, the teacher network is enforced to capture target-relevant knowledge, thus benefiting decreasing domain shift when mentoring object detection in the target domain. Extensive experiments are conducted on various widely used benchmarks with new state-of-the-art scores reported, highlighting the effectiveness.

1. Introduction

In the past decade, convolutional neural networks [11,26,37] (CNNs) have achieved great progress and delivered significant improvement in visual object detection [18,20,24]. Unfortunately, the well-built detectors suffer from remarkable performance drop when applied to unseen scenes due to domain shift [39,47]. Because it is rather expensive and time-consuming to annotate newly collected data, domain adaptive object detection (DAOD) [3,5,47] has been receiving increasing attention. It originates from unsupervised domain adaptation (UDA) [1,6,30], which proves effective in transferring knowledge from the learned domain (known as source domain) to a novel domain (known as target domain) with only unlabeled image for classification. Compared to UDA, DAOD is even more challenging as it simultaneously locates and classifies all instances of different objects in images with domain shift, requiring generating domain-invariant representations to reduce such a discrepancy in the presence of complex foreground and background variations.

Many efforts have been made on DAOD in the literature, and the methods mainly address it in the paradigm of adversarial feature alignment [15,25,39,47] or semi-supervised learning [2,5,43]. The former directly aligns the features in the source and target domains through adversarial discriminator confused by gradient reversal layer [25,47], and it can be fulfilled at the image-level [3,15], instance-level [3,25] or/and category-level [39,47]. The latter predicts pseudo labels according to the model trained in the source domain and adopts them as guidance to the target domain [2,5], and the domain gap can be bridged through enforcing the model consistency. Both the two types of methods show promising results in DAOD for a single pair of source and target.

Multi-source domain adaptation (MSDA) is considered as a more practical scenario in UDA since it assumes that various sources are available for better adaptation to the target domain [12,23,45]. In addition to the gap between the source and target domains [12,40,44], MSDA also deals with the discrepancy among different sources to avoid neg-

*Corresponding author.

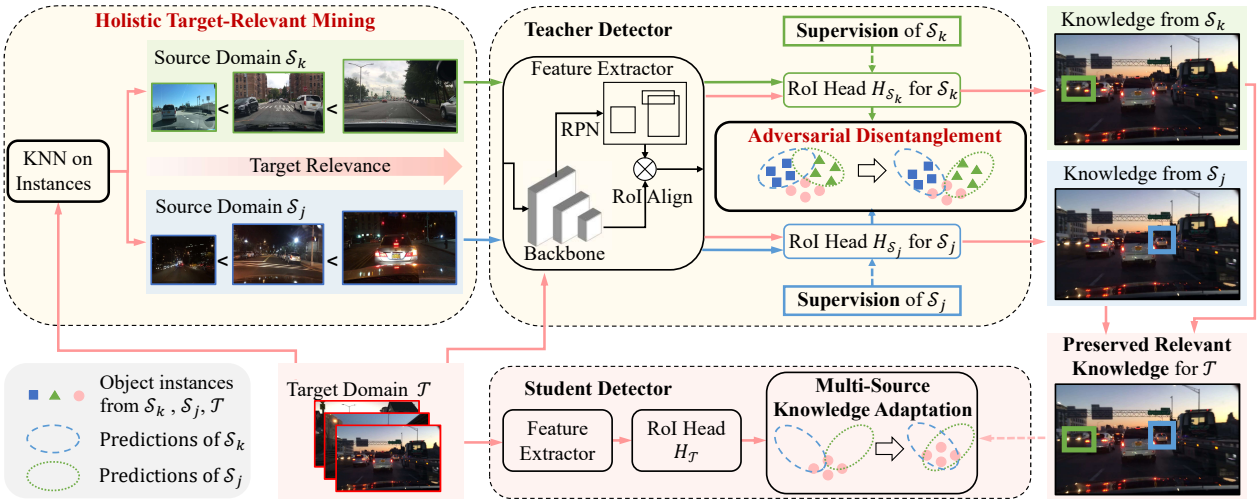


Figure 1. Framework overview of the proposed TRKP approach. The solid arrows refer to forward propagation and the dashed ones denote supervision. The teacher detector is trained on labeled source images and generates pseudo labels for unlabeled images in the target domain, which mentors the student detector. TRKP leverages the adversarial multi-source disentanglement (AMSD) module to preserve source domain-specific knowledge and the holistic target-relevant mining (HTRM) scheme to strengthen encoding target relevance knowledge, which significantly facilitates adapting multi-source knowledge to the target domain.

ative transfer [23,32]. Albeit its prevalence in classification, the multi-source problem has seldom been investigated in detection. To the best of our knowledge, the only attempt is recently given by DMSN [41]. It follows the pipeline that primarily assigns dynamic weights to multiple sources for alignment and then adapts the compound source to the target in MSDA [23, 46], and illustrates the necessity of knowledge of different domains to facilitate DAOD. However, there exist two major limitations: (1) the divide-and-merge spindle network conducts early alignment of multiple sources, which often incurs degradation of domain knowledge learned in individual sources for their gaps; (2) the loss memory bank measures target-relevant knowledge in source domains by a temporary discrepancy, leading to a local optimum. Both the facts suggest much room for amelioration.

To tackle the issues aforementioned, this study proposes a novel target-relevant knowledge preservation (TRKP) approach to multi-source DAOD, aiming at enhancing target-relevant knowledge learning from different sources and reducing domain knowledge degradation in adaptation to the target. Specifically, TRKP performs multi-source DAOD in the teacher-student framework, where a multi-head teacher network is constructed to extract knowledge from individual labeled source domains and mentor the student network on detector building in the unlabeled target domain (refer to Fig. 1 for an overview). To restrain knowledge degradation, the teacher network embeds an adversarial multi-source disentanglement (AMSD) module to preserve source domain-specific knowledge acquired by corresponding independent detection heads as much as possible during cross-domain alignment. Further, a holistic target-relevant mining (HTRM) scheme is developed to re-weight source images according to source-target relevance. By this means, the teacher network is enforced to capture and highlight target-relevant knowledge at the global level, thus benefiting domain gap decreasing for detector adaptation in the target domain. Extensive experiments are carried out on public benchmarks with state of the art performance reported, demonstrating the advantages of TRKP.

The contributions of this study are three-fold:

1) We propose a novel teacher-student network for multi-source DAOD, which alleviates target-relevant source domain knowledge degradation for alignment through a multi-head teacher structure along with an adversarial source disentanglement module.

2) We propose a target-relevant mining procedure to measure relevance between the source and target domains at the global-level, substantially strengthening target-relevant knowledge acquiring from different sources.

3) We not only outperform the top counterpart by a large margin in existing protocols, but also achieve a good baseline on a harder scenario with more sources.

2. Related Work

Domain Adaptive Object Detection. As a well-tuned detector suffers performance degradation when applied to new scenes, unsupervised domain adaptation (UDA) is a promising solution to this dilemma. Domain adaptive object detection (DAOD) addresses the problem by diminishing the domain shift between seen and unseen scenes [3, 13, 47]. Most of recent studies can be grouped into two cate-

gories: (1) feature alignment based methods that tackle the domain shift by aligning discrepant features in detectors [3, 25, 33, 39, 47]; and (2) semi-supervised learning based methods that directly formulate UDA as a semi-supervised learning problem [2, 5, 14, 43]. However, these studies are designed on the single-source assumption and fail to deal with multiple source domains. Here we propose a novel semi-supervised learning based approach specially for multi-source DAOD.

Multi-Source Domain Adaptation. The studies on UDA generally focus on alignment between a single pair of source and target domains. Multi-source domain adaptation (MSDA) considers a more generalized case that multiple source domains are available [12, 23, 45]. It is beneficial to model generalization ability as more diverse data included but more challenging since domain shift also exists among source domains. There are several early studies [12, 22, 27, 28] handling this problem through a weighted source combination to achieve target-relevant prediction with rigorous theoretical analysis. Recent attempts conduct this re-weighting process in adversarial adaptation [35, 40, 44]. Besides, many investigations aim to diminish domain shifts between multiple sources [9, 23, 32]. [23] dynamically aligns moments of feature distributions, which consist of pairs of source and target domains and those of source domains. Rather than explicit feature alignment, [32] uses pseudo-labeled target samples for implicit alignment. All the methods above focus on classification, and to the best of our knowledge, DMSN [41] is the first to introduce MSDA into object detection. In addition to general DAOD approaches, it develops feature alignment among sources and pseudo subnet learning for their weighted combination. However, its alignment is limited by knowledge degradation and its temporary domain discrepancy measurement leads to a local optimum. By contrast, our TRKP aims at preserving more target-relevant knowledge from different source domains to facilitate multi-source DAOD.

3. Method

3.1. Framework Overview

We firstly describe the problem setting of unsupervised multi-source DAOD and subsequently overview the framework of the proposed approach.

Similar to the general MSDA [23, 32, 40] task, we consider K label-rich source domains $\{\mathcal{S}_1, \cdots, \mathcal{S}_K\}$ and an unlabeled target domain $\mathcal{T}$. Formally, we assume that there exist $N_{\mathcal{S}_k}$ labeled images $D_{\mathcal{S}_k} = \{(I_i^{\mathcal{S}_k}, \boldsymbol{y}_i^{\mathcal{S}_k})\}_{i=1}^{N_{\mathcal{S}_k}}$ in $\mathcal{S}_k$ ($k = 1, \cdots, K$), and $N_{\mathcal{T}}$ unlabeled images $D_{\mathcal{T}} = \{I_i^{\mathcal{T}}\}_{i=1}^{N_{\mathcal{T}}}$ in $\mathcal{T}$, where $I_i^{\mathcal{S}_k}$ is the i-th image from the k-th source domain $\mathcal{S}_k$ and $\boldsymbol{y}_i^{\mathcal{S}_k}$ refers to the corresponding label including the bounding boxes and their classes.

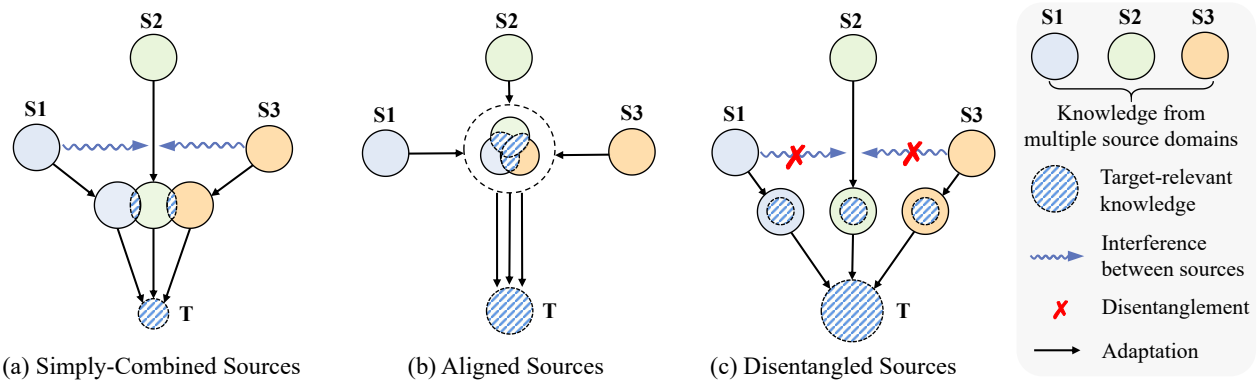
In MSDA, the unsupervised DAOD aims to learn a detector delivering high performance in the unlabeled target domain, by transferring knowledge for detection in $\{\mathcal{S}_k\}_{k=1}^K$ to $\mathcal{T}$ based on $\{D_{\mathcal{S}_k}\}_{k=1}^K \cup D_{\mathcal{T}}$. To achieve this goal, we propose a novel approach, namely target-relevant knowledge preservation (TRKP). Inspired by the success of semi-supervised learning in single source DAOD [2, 5], TRKP adopts the teacher-student framework, which proves effective in transferring domain knowledge and bridging the source-to-target gap [2, 5]. Specifically, as shown in Fig. 1, TRKP mainly consists of a teacher detector TeDet$(\cdot)$ and a student detector StDet$(\cdot)$, which encodes the knowledge for detection from the source domains and performs object detection in the target domain, respectively. As in [21], StDet$(\cdot)$ adopts the same architecture as TeDet$(\cdot)$. Usually, the 'teacher' TeDet$(\cdot)$ is applied to encode knowledge in the source domains by training on $\{D_{\mathcal{S}_k}\}_{k=1}^K$, and subsequently generate a pseudo label $\hat{\boldsymbol{y}}_j^{\mathcal{T}}$ for each unlabeled image $I_j^{\mathcal{T}}$, which is finally utilized to mentor the 'student' StDet$(\cdot)$, *i.e.* training StDet$(\cdot)$ on $\{(I_j^{\mathcal{T}}, \hat{\boldsymbol{y}}_j^{\mathcal{T}})\}_{j=1}^{N_{\mathcal{T}}}$.

As pointed out in [23, 41], both the multi-source domain shifts and the source-to-target domain gap notably affect the multi-source adaptation to the target domain. DMSN [41] deals with these problems by employing an early multi-source alignment and a local memory bank, which however incurs degradation of knowledge in source domains, thus only reaching a local optimum. To overcome the issues above, we develop an adversarial multi-source disentanglement (AMSD) module together with a holistic target-relevant mining (HTRM) scheme as shown in Fig. 1, which are further incorporated into the teacher-student framework. AMSD enables TeDet$(\cdot)$ to disentangle the single-source knowledge from multiple sources and prevent their mutual interference via adversarial learning, thus fulfilling domain-specific knowledge preservation. HTRM re-weights images from the sources $\{D_{\mathcal{S}_k}\}_{k=1}^K$ according to their relevance with those from the target $D_{\mathcal{T}}$ in a holistic manner, further facilitating TeDet$(\cdot)$ to encode globally refined target-relevant knowledge. By leveraging both the advantages of AMSD and HTRM, TRKP remarkably alleviates the knowledge degradation, therefore significantly boosting the overall performance. We describe the details of AMSD in Sec. 3.2 and HTRM in Sec. 3.3, respectively.

3.2. Adversarial Multi-Source Disentanglement

3.2.1 Knowledge Degradation in MSDA

Current approaches for MSDA typically deal with the domain gaps by multi-source combination or alignment. As shown in Fig. 2 (a), the combination based methods bridge the source-target domain gap by taking all the sources as a whole, regardless of their discrepancies. As a consequence, the target-relevant knowledge extracted from one source (*e.g.* S1) may be negatively interfered by another (*e.g.* S2). This kind of knowledge degradation deteriorates the qual-

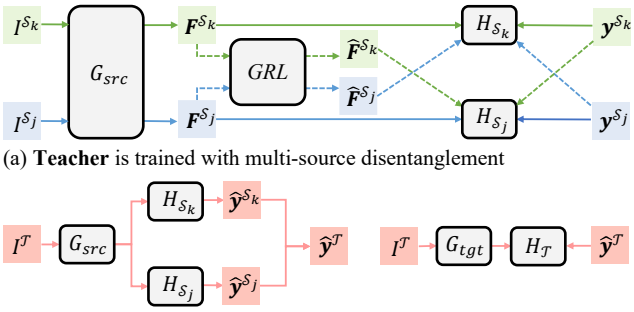


(a) Simply-Combined Sources (b) Aligned Sources (c) Disentangled Sources

Figure 2. Illustration of different strategies for multi-source adaptation. The area of a circular displays the amount of knowledge. (a) Simply-combined sources probably incurs mutual interference, due to domain shifts among sources. (b) Multi-source alignment reduces the domain shift, but degrades the target-relevant knowledge when performing alignment without the guidance of the target domain. (c) Our method preserves domain-specific target-relevant knowledge by disentangling multiple sources and preventing their mutual interference.

ity of transferred multi-source knowledge. In contrast, as illustrated in Fig. 2 (b), the alignment based approaches pay more attention to removing domain shifts among distinct sources, but probably incur severe loss of knowledge related to the target without the guidance of the target domain, leading to another kind of knowledge degradation.

As we aim to explore target-relevant knowledge from multiple label-rich sources to train detectors in the unlabeled target domain, both two kinds of knowledge degradation aforementioned should be reduced. There exist several studies emphasizing domain-specific knowledge preservation in heterogeneous domain adaptation [16, 17, 31] or face recognition under various domain biases [8, 34], yet not directly applicable to MSDA. This motivates us to present a solution that can jointly preserve domain-specific knowledge and align the source and target domains as in Fig. 2 (c). We elaborate the details of our solution in Sec. 3.2.2.



(a) **Teacher** is trained with multi-source disentanglement

(b) **Teacher** generates supervison (c) **Student** adapts to target

Figure 3. Illustration of the entire training pipeline based on AMSD. The solid arrows refer to teacher-student training and the dashed ones denote disentanglement. (a) The teacher detector is trained on multiple sources with disentanglement. (b) The teacher detector generates pseudo labels for images from the target domain. (c) The student detector adopts pseudo labels for training, and thus accomplishes the multi-source domain adaptation.

3.2.2 Knowledge Preservation via Disentanglement

In order to alleviate the knowledge degradation, we present AMSD during training TeDet$(\cdot)$ as shown in Fig. 3, by encoding the domain-specific knowledge from multiple sources without mutual interference.

Particularly, we employ the multi-head structure as in [41] in TeDet$(\cdot)$, where each source domain $\mathcal{S}_k$ has an individual RoI detection head $H_{\mathcal{S}_k}(\cdot)$, but shares the same base network $G_{src}(\cdot)$ (including the backbone and Region Proposal Network known as RPN) with the other source domains. This structure proves effective for its strong generalization ability [23, 32, 41]. Besides, it also facilitates the implementation of multi-source disentanglement and knowledge preservation, since the multiple heads $\{H_{\mathcal{S}_k}(\cdot)\}$ have separated parameters for distinct source domains. The student detector StDet$(\cdot)$ adopts the same multi-head architecture as TeDet$(\cdot)$, which is constituted of a base network $G_{tgt}(\cdot)$ and a detection head $H_{\mathcal{T}}(\cdot)$.

Inspired by [8], we disentangle multiple sources by correlation minimization via adversarial learning. Instead of employing additional domain discriminators, we impose constraints on the heads $\{H_{\mathcal{S}_k}\}$ and features across source domains, without increasing the model complexity. Specifically, given labeled images $\{(I_i^{\mathcal{S}_k}, \boldsymbol{y}_i^{\mathcal{S}_k})\}$ from multiple sources, the corresponding deep features are fetched by G_{src}, denoted as $\{\boldsymbol{F}_i^{\mathcal{S}_k} = G_{src}(I_i^{\mathcal{S}_k})\}$. A gradient reverse layer $GRL(\cdot)$ is introduced between the feature extractor G_{src} and heads $\{H_{\mathcal{S}_k}\}$ to implement adversarial learning. In the forward propagation of GRL, an adversarial feature $\hat{\boldsymbol{F}}_i^{\mathcal{S}_k} = GRL(\boldsymbol{F}_i^{\mathcal{S}_k})$ is generated for an input $\boldsymbol{F}_i^{\mathcal{S}_k}$. In the back propagation of GRL, the sign of the input gradient is simply reversed and multiplied by a factor μ. To facilitate

learning domain-specific knowledge from the k-th source domain $\mathcal{S}_k$, we formulate the following loss w.r.t. the k-th detection head $H_{\mathcal{S}_k}$:

$$\mathcal{L}_i^{H_{\mathcal{S}_k}} = l[H_{\mathcal{S}_k}(\boldsymbol{F}_i^{\mathcal{S}_k})] + \frac{\lambda}{K-1}\sum_{j\neq k}^{K} l[H_{\mathcal{S}_j}(\hat{\boldsymbol{F}}_i^{\mathcal{S}_k})], \quad (1)$$

where $l[\cdot]$ is the conventional detection loss (*e.g.*, the focal loss and smooth L_1 loss), and λ is a trade-off parameter. The label $\boldsymbol{y}_i^{\mathcal{S}_k}$ is simply omitted here for succinctness.

As observed from Eq. (1), the standard detection loss $l[H_{\mathcal{S}_k}(\boldsymbol{F}_i^{\mathcal{S}_k})]$ trains $H_{\mathcal{S}_k}$ by using the feature from $\mathcal{S}_k$, thus encoding knowledge from $\mathcal{S}_k$. The additional loss $l[H_{\mathcal{S}_j}(\hat{\boldsymbol{F}}_i^{\mathcal{S}_k})]$ measures the discrepancy between the ground-truth label and the prediction by the head $H_{\mathcal{S}_j}$ using the adversarial feature $\hat{\boldsymbol{F}}_i^{\mathcal{S}_k}$ from a distinct source domain $\mathcal{S}_j$ ($j \neq i$). Recall that the gradient w.r.t. $\hat{\boldsymbol{F}}_i^{\mathcal{S}_k}$ is reversed via GRL in back propagation. Therefore, minimizing $l[H_{\mathcal{S}_j}(\hat{\boldsymbol{F}}_i^{\mathcal{S}_k})]$ will increase the prediction error made by $H_{\mathcal{S}_j}$ on $\boldsymbol{F}_i^{\mathcal{S}_k}$. In other words, the loss $\mathcal{L}_i^{H_{\mathcal{S}_k}}$ in Eq. (1) enforces $H_{\mathcal{S}_k}$ to encode domain-specific knowledge from $\mathcal{S}_k$ and simultaneously puzzles the other heads $H_{\mathcal{S}_j}$ ($j \neq i$) by forcing them to yield distinct predictions.

Based on Eq. (1), the teacher detector is trained as below:

$$\min_{G_{src},\{H_{\mathcal{S}_k}\}_{k=1}^{K}} \sum_{k=1}^{K}\sum_{i=1}^{N_{\mathcal{S}_k}} \mathcal{L}_i^{H_{\mathcal{S}_k}}. \quad (2)$$

As being optimized in Eq. (2), each head $H_{\mathcal{S}_k}$ is disentangled from the other sources, thus encoding domain-specific knowledge. By this means, the mutual interference between sources can be mitigated, benefiting decreasing knowledge degradation.

3.2.3 Multi-Source Knowledge Adaptation

After training the teacher detector TeDet($\cdot$) by AMSD, the domain-specific knowledge encoded in each head is subsequently adapted to the target domain via training the student detector StDet($\cdot$). Concretely, given an unlabeled image $I_j^{\mathcal{T}}$ from the target domain, each head $H_{\mathcal{S}_k}$ separately generates a prediction $\hat{\boldsymbol{y}}_j^{\mathcal{T},\mathcal{S}_k}$, and the averaged one (conducted on RoI) $\hat{\boldsymbol{y}}_j^{\mathcal{T}} = \frac{1}{K}\sum_k^K \hat{\boldsymbol{y}}_j^{\mathcal{T},\mathcal{S}_k}$ is utilized as the pseudo label. Finally, the 'student' StDet($\cdot$) is mentored by TeDet($\cdot$) via the following optimization process based on $\{(I_j^{\mathcal{T}}, \hat{\boldsymbol{y}}_j^{\mathcal{T}})\}_{j=1}^{N_{\mathcal{T}}}$:

$$\min_{G_{tgt},H_{\mathcal{T}}} \sum_{j=1}^{N_{\mathcal{T}}} l[H_{\mathcal{T}}(G_{tgt}(I_j^{\mathcal{T}}))]. \quad (3)$$

During training StDet($\cdot$) based on Eq. (3), the multi-source domains and the target domain are implicitly aligned. However, training the 'student' StDet($\cdot$) with a fixed 'teacher' TeDet($\cdot$) tends to incur overfitting [29]. The Exponential Moving Average (EMA) [21] mechanism addresses this issue by regularizing the learning of TeDet($\cdot$) with the gradient of StDet($\cdot$). We therefore employ it in our framework to fulfill the multi-source knowledge adaptation in a more effective way.

3.3. Holistic Target-Relevant Mining

As observed in Eq. (2), images from multiple sources are treated equally when training TeDet($\cdot$). Due to the lack of guidance of the target, images that are less relevant to the target domain are given the same importance as more relevant ones, which deteriorates the quality of knowledge adaption. Previous works in MSDA [12, 27, 28] tackle this problem by using a distribution-weighted combination specially designed for classification, which is not fully suitable for object detection. DMSN [41] makes the first attempt in detection by proposing a dynamic loss memory bank to measure the discrepancy between the source and target domains. Nevertheless, it only captures local relevance information in mini-batches, leading to a local optimal solution.

To address the issue above, we develop HTRM to guarantee that the teacher detector encodes target-relevant knowledge at the global level, by assigning each source image $I_i^{\mathcal{S}_k}$ a target-relevant weight $\alpha_i^{\mathcal{S}_k}$. To achieve this goal, we first extract the deep feature $\boldsymbol{F}_i^{\mathcal{S}_k}$ via $G_{src}(\cdot)$ for each image $I_i^{\mathcal{S}_k}$. To avoid the interference from massive backgrounds, we only select the RoI features locating in the object area according to the label $\boldsymbol{y}_i^{\mathcal{S}_k}$, which are further pooled as a set of features denoted by $\{\boldsymbol{f}_{i,j}^{\mathcal{S}_k}\}_{j=1}^{|\boldsymbol{y}_i^{\mathcal{S}_k}|}$. Here, $|\boldsymbol{y}_i^{\mathcal{S}_k}|$ stands for the number of annotated bounding boxes in the i-th image $I_i^{\mathcal{S}k}$. By repeating this procedure, we finally obtain the instance-level feature set for all the images from the multi-source domains, denoted by $\mathcal{G} = \{\{\{\boldsymbol{f}_{i,j}^{\mathcal{S}_k}\}_{j=1}^{|\boldsymbol{y}_i^{\mathcal{S}_k}|}\}_{i=1}^{N_{\mathcal{S}_k}}\}_{k=1}^{K}$. Similarly, based on the pseudo labels $\{\hat{\boldsymbol{y}}_m^{\mathcal{T}}\}$ and $G_{tgt}(\cdot)$ of the student detector, we extract the instance-level feature set from the target domain, denoted by $\mathcal{Q} = \{\{\boldsymbol{f}_{n,m}^{\mathcal{T}}\}_{m=1}^{|\hat{\boldsymbol{y}}_n^{\mathcal{T}}|}\}_{n=1}^{N_{\mathcal{T}}}$.

We follow [28] by applying the nearest neighbor algorithm to mine cross-domain relevance $\{\alpha_i^{\mathcal{S}_k}\}$. As summarized in Algorithm 1, the mining process mainly consists of two steps: 1) for each feature $\boldsymbol{f}_{n,m}^{\mathcal{T}} \in \mathcal{Q}$ from the target domain, we search its K' nearest neighbors $\mathcal{N}_{\boldsymbol{f}_{n,m}^{\mathcal{T}}}$ in $\mathcal{G}$ from the source domains, where the cosine distance is used as the similarity metric; 2) for the i-th image $I_i^{\mathcal{S}_k}$ from the k-th source domain represented by $\{\boldsymbol{f}_{i,j}^{\mathcal{S}_k}\}_{j=1}^{|\boldsymbol{y}_i^{\mathcal{S}_k}|}$, we compute the frequency $w_i^{\mathcal{S}_k}$ by counting the number of elements in $\mathcal{Q}$ that include at least one member in $\{\boldsymbol{f}_{i,j}^{\mathcal{S}_k}\}_{j=1}^{|\boldsymbol{y}_i^{\mathcal{S}_k}|}$ as K' nearest neighbors. Note that $w_i^{\mathcal{S}_k}$ in step 2) is computed by using the holistic feature set from the target domain, thus mining

Algorithm 1 Holistic Target-Relevant Mining

Input: The object-level feature set $\mathcal{G}$ from multiple source domains and the feature set $\mathcal{Q}$ from the target domain; the hyper-parameter K'.

Output: The relevance weights $\{\alpha_i^{\mathcal{S}_k}\}$ of the source images w.r.t the target domain.

Initialize: $w_i^{\mathcal{S}_k} := 0$.

for $\boldsymbol{f}^{\mathcal{T}}$ in $\mathcal{Q}$ **do**
  Find the K'-nearest neighbors of $\boldsymbol{f}^{\mathcal{T}}$ in $\mathcal{G}$ as $\mathcal{N}_{\boldsymbol{f}^{\mathcal{T}}}$
  for $\boldsymbol{f}_{i,j}^{\mathcal{S}_k}$ in the neighborhood $\mathcal{N}_{\boldsymbol{f}^{\mathcal{T}}}$ **do**
    $w_i^{\mathcal{S}_k} := w_i^{\mathcal{S}_k} + 1$
  end for
end for
Compute the weight $\{\alpha_i^{\mathcal{S}_k}\}$ based on $\{w_i^{\mathcal{S}_k}\}$ and Eq. (4)

the target-relevance in a global view. Based on $w_i^{\mathcal{S}_k}$, the relevance weight $\alpha_i^{\mathcal{S}_k}$ is formulated as below:

$$\alpha_i^{\mathcal{S}_k} = \begin{cases} \gamma \log(\frac{w_i^{\mathcal{S}_k}}{K'}) + \beta, & w_i^{\mathcal{S}_k} > K', \\ 0, & w_i^{\mathcal{S}_k} \leq K', \end{cases} \tag{4}$$

where γ and β control the magnitude of $\alpha_i^{\mathcal{S}_k}$. From Eq. (4), we can observe that $\alpha_i^{\mathcal{S}_k}$ becomes large if the source image $I_i^{\mathcal{S}_k}$ is closely relevant to the target, and turns to 0 otherwise.

Based on $\{\alpha_i^{\mathcal{S}_k}\}$, we can re-weight the importance of images from multiple sources as illustrated in Fig. 1, and apply it to train a target-relevant teacher detector by reformulating the loss function in Eq. (2) as the following:

$$\min_{G_{src}, \{H_{\mathcal{S}_k}\}_{k=1}^{K}} \sum_{k=1}^{K} \sum_{i=1}^{N_{\mathcal{S}_i}} \alpha_i^{\mathcal{S}_k} \mathcal{L}_i^{H_{\mathcal{S}_k}}. \tag{5}$$

Based on Eq. (5), TeDet$(\cdot)$ is explicitly enforced to learn from target-relevant samples, and thus restrains from the interference from the information irrelevant to the target.

4. Experiments

In this section, we evaluate the performance of TRKP by following the settings in [41], including the cross camera adaptation in Sec. 4.1 and the cross time adaptation in Sec. 4.2. In addition, we present a new setting, which contains more sources with mixed domain gaps in Sec. 4.3. We also conduct ablation studies as summarized in Sec. 4.4

Implementation Details. Similar to [41, 47], we adopt Faster R-CNN [24] with RoI Align [10] and VGG16 [26] backbone as the basic detector to make fair comparisons. All the input images are resized such that the shorter lengths have 600 pixels. As for the teacher-student learning framework, we adopt the same settings as in UBT [21], which is a representative of semi-supervised object detection. Concretely, the confidence threshold for pseudo labeling is set to 0.7. The smoothing coefficient in EMA is set as 0.9999. For AMSD, the hyper-parameters λ and μ are fixed to 0.2 and 0.01, respectively. For HTRM, the number of nearest neighbors K' is set to 5. The scaling factors γ and β in Eq. (4) are fixed as 1.0 and 0.5 by default. The learning rate is 0.01 with the batch size at 16. We utilize 20 epochs in training, where the teacher detector is trained individually for the first 10 epochs, after which HTRM is conducted to re-weight source images, followed by training StDet$(\cdot)$ for domain adaptation. All the experiments are carried out on 8 NVIDIA 1080Ti GPUs.

Setting	Source	Method	AP
Source Only	C	FRCNN [24]	44.6
	K		28.6
	C+K		43.2
Single Source	C	SW [25]	45.5
		CRDA [38]	46.5
		UMT [5]	47.5
		UBT [21] (Baseline)	48.4
Single Source	K	SW [25]	29.6
		CRDA [38]	30.8
		UMT [5]	35.4
		UBT [21] (Baseline)	33.8
Source Combined	C+K	SW [25]	41.9
		CRDA [38]	43.6
		UMT [5]	47.0
		UBT [21] (Baseline)	47.6
MSDA	C+K	MDAN [44]	43.2
		M^{3}SDA [23]	44.1
		DMSN [41]	49.2
		HTRM (Ours)	52.9
		AMSD (Ours)	56.8
		TRKP (Ours)	**58.4**
Oracle	BDD100K	FRCNN [24]	60.2

Table 1. Results on cross camera adaptation. 'C' and 'K' indicate Cityscapes and KITTI respectively, which constitute source domains. BDD100K is the target domain. AP (%) of *car* is reported.

Comparative Approaches. We compare TRKP to the following state-of-the-art approaches: (1) Source-only method which applies the basic Faster R-CNN [24] detector without adaptation to the target domain; (2) Single-Source & Source-Combined methods including SW [25], GPA [39], UMT [5] and UBT [21], which conduct DAOD with the single-source assumption; (3) MSDA methods including MDAN [44], M^{3}SDA and DMSN [41]. We also report the performance of Oracle trained by fully labeled target images, as an estimated upper bound.

4.1. Cross Camera Adaptation

Settings. The images captured by different cameras incur the domain shift problem due to various settings of camera parameters, viewpoints and scenes during data collection. To address this concern, we evaluate our method in the setting of cross camera adaptation. By following [41], we select Cityscapes [4] and KITTI [7] as the source do-

mains and BDD100K [42] as the target domain, and meanwhile only use the images from the *car* category for training and evaluation. Cityscapes [4] is a benchmark for semantic urban scene understanding and KITTI [7] is a widely used dataset for autonomous driving, containing 2,975 and 7,481 annotated training images, respectively. BDD100K is a large-scale dataset for autonomous driving, where only the *daytime* subset is adopted, including 36,728 unlabeled images for training and 5,258 validation images for evaluation. The widely used average precision (AP) is adopted as the evaluation metric.

Results. As shown in Table 1, the previous DAOD methods, which simply combine Cityscapes and KITTI (see the row in *"Source Combined"*) during training, generally report worse performance compared to those only adopt Cityscapes (see the row in *"Single Source"*). The reason lies in that knowledge transferred from Cityscapes to BDD100K is probably interfered by the domain shift between Cityscapes and KITTI, resulting in severe knowledge degradation during adaptation. Despite of increasing amount of data in multiple sources, most existing MSDA based methods only achieve minor gains or perform even worse, compared to the source combined approaches. By contrast, our method improves the accuracy by a large margin. For instance, the AP by applying TRKP is 9.2% higher than the second best, *i.e*. DMSN. It is worth noting that our method is based on the UBT baseline. When separately applying the proposed AMSD and HTRM modules to UBT, the gains are 5.3% and 9.2%, respectively, clearly showing their effectiveness. By combining AMSD and HTRM, TRKP achieves an AP of 58.4%, reaching a new state-of-the-art, which reduces the gap with Oracle (full supervision) to 1.8%.

4.2. Cross Time Adaptation

Settings. In real-world applications, a detector is often deployed at different time, where changes in illumination and scene can be extremely large. To evaluate the performance of our method against such a factor, we follow the setting in [41] to adapt knowledge learned in the daytime and nighttime to corner cases, *i.e*. at dawn or dusk. Concretely, BDD100K [42] is divided into three subsets by time, including *daytime*, *night*, *dawn/dusk*. 36,728 images in the *daytime* and 27,971 images at *night* constitute two source domains. Images collected by excluding the ones in the daytime and nighttime are relatively few, where 5,027 unlabeled images are used for training and 778 validation images for evaluation at *dawn/dusk* as the target domain. The mean average precision (mAP) over 10 categories is reported for comparison.

Results. The results on cross time adaptation are summarized in Table 2, where more detailed comparisons are provided in the *supplementary material* due to space limit. As shown in Table 2, previous DAOD methods fail to boost the performance when using images from both the *daytime* and *night* subsets, due to the interference of the large discrepancy between the two domains. By multi-source disentanglement, our TRKP improves the performance by large margins, *e.g*. 4.8% higher than the second best based on DMSN. The HTRM and AMSD modules also achieve remarkable gains in performance. Specifically, AMSD disentangles multiple sources and prevents the interference among them, thus improving the UBT baseline by 4.9%. HTRM performs re-weighting at the global level, yielding better performance than DMSN [41] that adopts the dynamic weighting strategy. Besides, it is worth noting that TRKP exceeds Oracle significantly and boosts the detection accuracy to 39.8% in mAP. The relatively poor performance of Oracle is owing to insufficient training images in the target domain, and our remarkable performance improvement shows the effectiveness of transfer learning in such situations by target-relevant knowledge adaptation.

Setting	Source	Method	mAP
Source Only	D	FRCNN [24]	30.4
	N		25.0
	D+N		28.9
Single Source	D	SW [25]	31.4
		GPA [39]	31.8
		CRDA [38]	31.2
		UMT [5]	33.8
		UBT [21] (Baseline)	33.2
Single Source	N	SW [25]	26.9
		GPA [39]	27.6
		CRDA [38]	28.4
		UMT [5]	21.6
		UBT [21] (Baseline)	24.2
Source Combined	D+N	SW [25]	29.9
		GPA [39]	30.6
		CRDA [38]	30.2
		UMT [5]	33.5
		UBT [21] (Baseline)	33.1
MSDA	D+N	MDAN [44]	27.6
		M^3SDA [23]	26.5
		DMSN [41]	35.0
		HTRM (Ours)	35.5
		AMSD (Ours)	38.0
		TRKP (Ours)	**39.8**
Oracle	BDD100K	FRCNN [24]	26.6

Table 2. Results on cross time adaptation. 'D' and 'N' indicate the *daytime* and *night* subsets of BDD100K, respectively. mAP (%) over 10 categories on BDD100K *dawn/dusk* is reported.

4.3. Extension to Mixed Domain Adaptation

Settings. As there always exist more than one factors leading to domain shift in practice, we extend existing settings of cross camera/time adaptions with only two source domains, and present a new setting by considering a more complex case with mixed domain gaps. Specifically, based on the scene adaptation scenario in [38] that chooses

Setting	Source	Method	mAP
Source Only	C	FRCNN [24]	23.4
Single Source	C	UBT [21] (Baseline)	29.7
Source Only	C+M	FRCNN [24]	29.7
Source Combined	C+M	UBT [21] (Baseline)	18.5
MSDA	C+M	**TRKP (Ours)**	**35.3**
Source Only	C+M+S	FRCNN [24]	30.9
Source Combined	C+M+S	UBT [21] (Baseline)	25.1
MSDA	C+M+S	**TRKP (ours)**	**37.1**
Oracle	BDD100K	FRCNN [24]	38.6

Table 3. Results on mixed domain adaptation. 'C'/'M'/'S' indicate Cityscapes/MS COCO/Synscapes, respectively.

Cityscapes [4] as the source and BDD100K [42] as the target, we employ MS COCO [19] and Synscapes [36] as two extra sources. MS COCO contains common scenes distinct from street views and Synscapes is a synthetic dataset, both of which enlarge the data scale and bring in more kinds of domain gaps and category shifts. 2,975/71,749/25,000 images from Cityscapes/MS COCO/Synscapes are used for training. 36,728 images in the *daytime* subset from BDD100K are used as unlabeled target data. 5,258 images from BDD100K in the *daytime* subset are used for evaluation. mAP over 7 classes is reported.

Results. As summarized in Table 3, by adopting more sources, the performance of the source only detector, *i.e.* FRCNN, is consistently improved. However, the source combined method, *i.e.* UBT, performs poorly due to severe negative transfers caused by mixed domain gaps. In contrast, TRKP achieves a significant performance gain, *e.g.* 5.6% in mAP when using two sources, and 6.2% in mAP for three sources, demonstrating its effectiveness when applying to mixed source domains.

4.4. Ablation Study

We detailedly analyze the modules and hyper-parameters of TRKP in the setting of Cross Time Adaptation.

On Disentanglement. As displayed in Table 4, training a separated detector for each source domain performs much worse than training a common backbone with combined sources, showing the necessity of a shared feature extractor. The multi-head structure also contributes, improving the mAP by 1.3%. When performing AMSD on the classification head and regression head, mAPs are boosted by 2.1% and 1.5% respectively, highlighting the advantage of using adversarial disentanglement. A combination of them further promotes the accuracy.

On Hyper-Parameters. As described in Sec. 3.2, μ and λ control the magnitude of AMSD . As shown in Table 5, TRKP achieves the best result when $\mu = 0.01$ and $\lambda = 0.2$. As for HTRM, we study the effect of the number of neighbors K', where the best result is reached when $K' = 5$. Moreover, HTRM focuses on mining source-target relevance at the instance level, rather than at the image-level as in most existing MSDA approaches [12, 46]. To validate the impact of instance-level relevance, we report the mAPs by performing HTRM at different levels. As in Table 5, HTRM clearly performs better at the instance level, which makes sense as object detection is an instance-aware task.

Shared Feature	Multi-head	Cls	Reg	mAP
	✓	✓	✓	24.7
✓				33.1
✓	✓			34.4
✓	✓	✓		36.5
✓	✓		✓	35.9
✓	✓	✓	✓	**38.0**

Table 4. mAP (%) by performing AMSD on different structures. Shared Feature refers to training with shared backbone and RPN. Multi-head indicates assigning each source an independent RoI head. Cls/Reg refer to applying disentanglement on the classification/regression heads.

AMSD			HTRM		
μ	λ	mAP	Features	K'	mAP
0.05	0.2	36.8	Image-level	5	32.6
0.002	0.2	37.0	Instance-level	3	34.9
0.01	0.2	**38.0**	Instance-level	5	**35.5**
0.01	1.0	37.2	Instance-level	10	35.2
0.01	0.04	36.7	Instance-level	30	34.6

Table 5. mAP (%) of ablation studies on AMSD and HTRM.

5. Conclusion

In this paper, we present a novel multi-source domain adaptation approach for object detection. To avoid knowledge degradation, we propose an adversarial multi-source disentanglement module and a holistic target-relevant mining scheme to preserve target-relevant knowledge during adaption. Extensive experiments clearly show the effectiveness of our method compared to the state-of-the-art. Besides, we apply our method to a harder scenario with mixed sources and provide a competitive baseline.

Acknowledgement

This work is partly supported by the National Natural Science Foundation of China (No. 62022011, No. 61876176 and U1813218), the Guangdong NSF Project (2020B1515120085), the Shanghai Committee of Science and Technology, China (21DZ1100100), the Shenzhen Research Program (RCJC20200714114557087), the Research Program of State Key Laboratory of Software Development Environment (SKLSDE-2021ZX-04), the Fundamental Research Funds for the Central Universities, and the Joint Lab of CASHK.

References

[1] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, pages 95–104, 2017. 1

[2] 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 *CVPR*, pages 11457–11466, 2019. 1, 3

[3] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster R-CNN for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 1, 2, 3

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 6, 7, 8

[5] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, 2021. 1, 3, 6, 7

[6] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015. 1

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 6, 7

[8] Sixue Gong, Xiaoming Liu, and Anil K. Jain. Jointly debiasing face recognition and demographic attribute estimation. In *ECCV*, pages 330–347, 2020. 4

[9] Jianzhong He, Xu Jia, Shuaijun Chen, and Jianzhuang Liu. Multi-source domain adaptation with collaborative learning for semantic segmentation. In *CVPR*, pages 11008–11017, 2021. 3

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988, 2017. 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[12] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In *NeurIPS*, pages 8256–8266, 2018. 1, 3, 5, 8

[13] 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 *ECCV*, pages 733–748, 2020. 2

[14] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *ICCV*, pages 6091–6100, 2019. 3

[15] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *CVPR*, pages 12456–12465, 2019. 1

[16] Haoliang Li, Sinno Jialin Pan, Renjie Wan, and Alex C. Kot. Heterogeneous transfer learning via deep matrix completion with adversarial kernel embedding. In *AAAI*, pages 8602–8609, 2019. 4

[17] Jingjing Li, Mengmeng Jing, Ke Lu, Lei Zhu, and Heng Tao Shen. Locality preserving joint transfer for domain adaptation. *IEEE TIP*, 28(12):6103–6115, 2019. 4

[18] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2999–3007, 2017. 1

[19] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014. 8

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *ECCV*, 2016. 1

[21] 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. In *ICLR*, 2021. 3, 5, 6, 7, 8

[22] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *NeurIPS*, pages 1041–1048, 2008. 3

[23] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415, 2019. 1, 2, 3, 4, 6, 7

[24] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 1, 6, 7, 8

[25] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 1, 3, 6, 7

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1, 6

[27] Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. A two-stage weighting framework for multi-source domain adaptation. In *NeurIPS*, pages 505–513, 2011. 3, 5

[28] Shi-Liang Sun and Hong-Lei Shi. Bayesian multi-source domain adaptation. In *ICML*, pages 24–28, 2013. 3, 5

[29] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pages 1195–1204, 2017. 5

[30] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, July 2017. 1

[31] Laurens van der Maaten and Geoffrey Hinton. Knowledge preserving and distribution alignment for heterogeneous domain adaptation. *TOIS*, 40(16):1–29, 2021. 4

[32] Naveen Venkat, Jogendra Nath Kundu, Durgesh Kumar Singh, Ambareesh Revanur, and Venkatesh Babu R. Your classifier can secretly suffice multi-source domain adaptation. In *NeurIPS*, 2020. 2, 3, 4

[33] 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 *CVPR*, pages 4516–4526, 2021. 3

[34] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *CVPR*, pages 9319–9328, 2020. 4

[35] Junfeng Wen, Russell Greiner, and Dale Schuurmans. Domain aggregation networks for multi-source domain adaptation. In *ICML*, pages 10214–10224, 2020. 3

[36] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint*, 1810.08705, 2018. 8

[37] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 5987–5995, 2017. 1

[38] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, pages 11721–11730, 2020. 6, 7

[39] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, pages 12352–12361, 2020. 1, 3, 6, 7

[40] Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *CVPR*, pages 3964–3973, 2018. 1, 3

[41] Xingxu Yao, Sicheng Zhao, Pengfei Xu, and Jufeng Yang. Multi-source domain adaptation for object detection. In *ICCV*, 2021. 2, 3, 4, 5, 6, 7

[42] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving video database with scalable annotation tooling. *arXiv preprint*, 1805.04687, 2018. 7, 8

[43] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *ECCV*, pages 86–102, 2020. 1, 3

[44] Han Zhao, Shanghang Zhang, Guanhang Wu, José M. F. Moura, João Paulo Costeira, and Geoffrey J. Gordon. Adversarial multiple source domain adaptation. In *NeurIPS*, pages 8568–8579, 2018. 1, 3, 6, 7

[45] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source domain adaptation for semantic segmentation. In *NeurIPS*, pages 7285–7298, 2019. 1, 3

[46] Sicheng Zhao, Guangzhi Wang, Shanghang Zhang, Yang Gu, Yaxian Li, Zhichao Song, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source distilling domain adaptation. In *AAAI*, pages 12975–12983, 2020. 2, 8

[47] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *CVPR*, pages 13763–13772, 2020. 1, 2, 3, 6