International Journal of Computer Vision (2024) 132:5950–5982
https://doi.org/10.1007/s11263-024-02170-z

Multi-source-free Domain Adaptive Object Detection

Sicheng Zhao[1] · **Huizai Yao**[1,2,3] · **Chuang Lin**[4] · **Yue Gao**[1,5] · **Guiguang Ding**[1,5]

Received: 15 December 2023 / Accepted: 28 June 2024 / Published online: 11 July 2024

Abstract

To enhance the transferability of object detection models in real-world scenarios where data is sampled from disparate distributions, considerable attention has been devoted to domain adaptive object detection (DAOD). Researchers have also investigated multi-source DAOD to confront the challenges posed by training samples originating from different source domains. However, existing methods encounter difficulties when source data is unavailable due to privacy preservation policies or transmission cost constraints. To address these issues, we introduce and address the problem of Multi-source-free Domain Adaptive Object Detection (MSFDAOD), which seeks to perform domain adaptation for object detection using multi-source-pretrained models without any source data or target labels. Specifically, we propose a novel Divide-and-Aggregate Contrastive Adaptation (DACA) framework. First, multiple mean-teacher detection models perform effective knowledge distillation and class-wise contrastive learning within each source domain feature space, denoted as "Divide". Meanwhile, DACA integrates proposals, obtains unified pseudo-labels, and assigns dynamic weights to student prediction aggregation, denoted as "Aggregate". The two-step process of "Divide" and "Aggregate" enables our method to efficiently leverage the advantages of multiple source-free models and aggregate their contributions to adaptation in a self-supervised manner. Extensive experiments are conducted on multiple popular benchmark datasets, and the results demonstrate that the proposed DACA framework significantly outperforms state-of-the-art approaches for MSFDAOD tasks.

Keywords Domain adaptive object detection · Multi-source domain adaptation · Source-free domain adaptation · Contrastive learning

Communicated by Hong Liu.

Sicheng Zhao, Huizai Yao and Chuang Lin have contributed equally.

✉ Sicheng Zhao
schzhao@tsinghua.edu.cn

✉ Guiguang Ding
dinggg@tsinghua.edu.cn

Huizai Yao
victoryaohz@gmail.com

Chuang Lin
Chuang.Lin@monash.edu

Yue Gao
gaoyue@tsinghua.edu.cn

[1] BNRist, Tsinghua University, Beijing, China

[2] College of Computer Sciences, Nankai University, Tianjin, China

[3] Shanghai Artificial Intelligence Laboratory, Shanghai, China

[4] Department of Data Science, Monash University, Clayton, Australia

[5] School of Software, Tsinghua University, Beijing, China

1 Introduction

In the last decade, deep learning has significantly advanced computer vision, benefiting various real-world applications. Notably, with the success of deep neural network architectures such as convolutional neural networks (CNNs) (Krizhevsky et al., 2012; He et al., 2016) and vision transformers (ViTs) (Dosovitskiy et al., 2021), recent achievements in deep learning have convincingly demonstrated their effectiveness and substantial potential in various visual tasks, such as image classification (Krizhevsky et al., 2012; Huang et al., 2017; Szegedy et al., 2017; He et al., 2016), object detection (Girshick, 2015; Redmon et al., 2016; Tian et al., 2019), and semantic segmentation (Long et al., 2015; Chen et al., 2017). As a prominent subject of research, object detection (OD) primarily aims to identify visual instances belonging to predefined object categories (Zou et al., 2023). Deep learning innovations have catalyzed the development and implementation of diverse object detectors, such as two-stage detectors (Girshick, 2015; He et al., 2015; Ren et al., 2015; Lin et al., 2017a, 2023) and one-stage detectors (Red-

mon et al., 2016; Liu et al., 2016; Tian et al., 2019; Carion et al., 2020). Recently, researchers have endowed object detection with a novel *set prediction* paradigm, including transformer (Vaswani et al., 2017)-based detection models (Carion et al., 2020; Zhu et al., 2021; Liu et al., 2022) and a latest successful approach DiffusionDet (Chen et al., 2023), which is inspired by the diffusion model (Ho et al., 2020) to detect objects by diffusing bounding boxes. The noteworthy achievements of these object detection models manifest when trained on high-quality datasets equipped with accurate annotations.

Recent developments in complex model designs, increased storage, and enhanced computing power have highlighted the need for large-scale, high-quality data in computer vision. This data, often difficult and costly to gather and label, is crucial for improving model performance and generalization. Particularly in real-world settings, there is a clear risk of affecting model generalization ability due to domain shift (Sun et al., 2016; Amodei et al., 2016; Zhao et al., 2021c, 2022, 2024), i.e., the training and testing data often differ in distribution. To address this challenge, domain adaptation (DA) has become a prominent research task in various computer vision tasks (Zhao et al., 2019a, 2020, 2021b, 2022, 2023). Among various extensively studied DA settings, unsupervised DA (UDA), i.e., DA with no target data labels as supervised information (Wilson & Cook, 2020), has gained substantial attention. With the success of deep neural network architectures, researchers have made crucial attempts in deep visual UDA, encompassing both theoretical analysis (Long et al., 2017; Sun & Saenko, 2016; Li et al., 2017) and algorithm design (Liu & Tuzel, 2016; Ganin & Lempitsky, 2015; Tzeng et al., 2017; Zhao et al., 2019b, 2021a).

Domain adaptive object detection (DAOD), a branch within the realm of deep visual domain adaptation, aiming at addressing domain shift inherent in object detection tasks, has also been the subject of extensive investigation (Chen et al., 2018; Inoue et al., 2018; Saito et al., 2019; Cai et al., 2019; He et al., 2023; Zhang et al., 2023a; Lang et al., 2022; Xu et al., 2023; He and Zhang, 2020). However, these conventional unsupervised DAOD algorithms may encounter several challenges under real-world scenarios. Firstly, primarily designed for single-source scenarios, these algorithms face challenges in accommodating data from multiple source domain distributions (Sun et al., 2015; Zhao et al., 2024; Lin et al., 2021). Secondly, their reliance on high-quality labeled source data for effective knowledge transfer proves impractical in the real-world context where data privacy protection and transmission cost issues come to the forefront (Fang et al., 2022; Yu et al., 2023). In response to these limitations, two distinct DAOD tasks are proposed: multi-source domain adaptive object detection (MSDAOD) and source-free domain adaptive object detection (SFDAOD). MSDAOD (Yao et al., 2021; Wu et al., 2022; Zhang et al., 2022) employs knowledge from multiple source domains to enhance adaptation, whereas SFDAOD (Li et al., 2021; Huang et al., 2021; Li et al., 2022a; Xiong et al., 2021; Vibashan et al., 2023; Chu et al., 2023) operates under constraints where source data and labels are inaccessible during adaptation, relying on a single pretrained model. However, current methods are not yet adept at effectively managing scenarios that require both multi-source and source-free scenarios simultaneously.

Considering the aforementioned limitations, we propose a novel unresolved DA setting: **Multi-Source-Free Domain Adaptive Object Detection (MSFDAOD)**. The goal of MSFDAOD is to adapt the object detection model to target domains for multi-source and source-free conditions under open-world scenarios, which aligns well with the Aim and Scope of https://link.springer.com/journal/11263/updates/25233244. To better illustrate MSFDAOD task, Fig. 1 shows the differences among the four aforementioned settings: DAOD, SFDAOD, MSDAOD, and MSFDAOD. To our knowledge, there is currently no algorithm specifically designed for MSFDAOD tasks. Here we showcase one concrete example of a practical application scenario of MSFDAOD: To train an effective car detector for sandy weather, an organization requests several other institutions for large-scale training data (mostly in non-sandy weather) for superior MSDAOD performance. However, some of the institutions need to protect the citizens' privacy like faces and licenses, and other institutions think it is inconvenient and expensive to transmit data, so they can only provide pretrained models to the organization. Then it is essential for the organization to apply the MSFDAOD approach to leverage multiple pretrained models to perform adaptation, and finally obtain superior performance on the target domain i.e., sandy weather.

To design a particular algorithm for MSFDAOD, an intuitive approach is to directly transfer existing MSFDA approaches focusing primarily on classification or segmentation tasks to detection tasks. However, this intuitive approach may face several challenges: (1) ***Task specificity***. Image classification or semantic segmentation tasks aim to perform accurate classification result at a fixed scale, i.e., image-level or pixel-level, respectively. In contrast, object detection tasks require not only accurate classification outcomes but also precise localization results. (2) ***Cross-model variability***. Object detection models exhibit varying bounding box prediction results for the same image during an iteration, depending on different source models. This variability is not a concern in image classification tasks, where only category representations are needed. (3) ***Object feature discriminability***. MSFDA for image classification or semantic segmentation tasks focuses on learning discriminative image-level or instance-level features, while MSFDAOD tasks demand high-quality discriminative object-level feature learning.

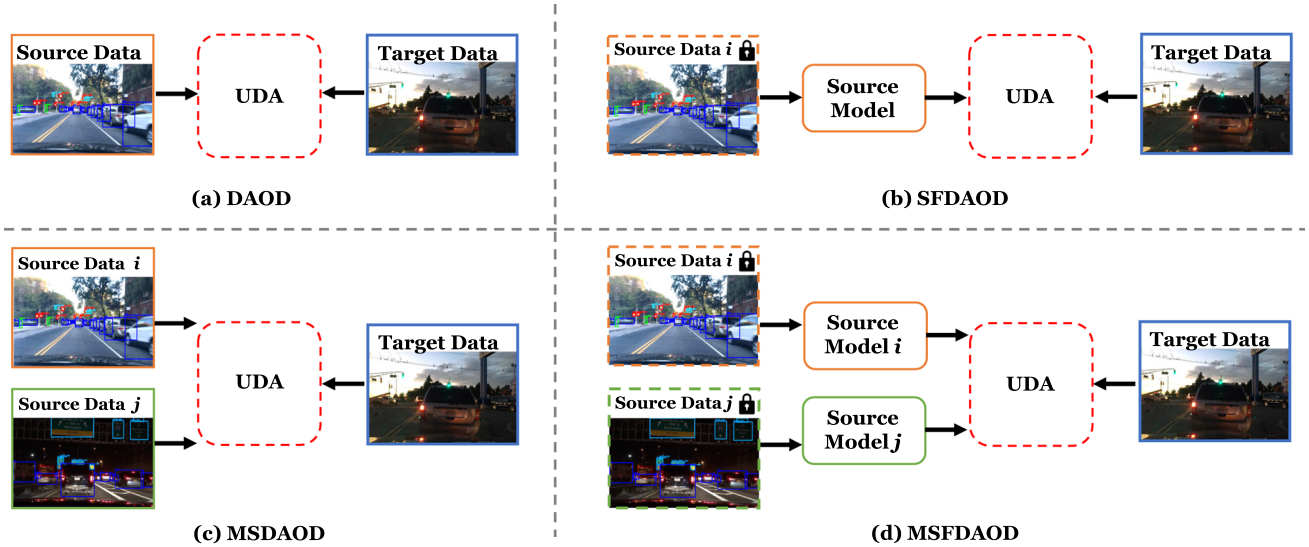

Fig. 1 Illustration of DAOD, SFDAOD, MSDAOD, and the proposed MSFDAOD. In unsupervised DAOD, both source and target images are available, while source labels are available and target labels are unavailable. In SFDAOD, during the adaptation stage, the source data remains unseen considering data transmission costs and ensuring data privacy protection. In MSDAOD, images are sampled from multiple source domains for adaptation purposes. MSFDAOD can be conceptualized as a fusion of MSDAOD and SFDAOD. It lacks access to multiple source domain data and only provides multiple source-pretrained models for adaptation. The image in "Source Data" and "Source Data i" is from Bdd100k (Yu et al., 2020) daytime. The image in "Source Data j" is from Bdd100k night. The image in "Target Data" is from Bdd100k dawn/dusk

In addressing the aforementioned challenges in MSFDAOD, we introduce a novel Divide-and-Aggregate Contrastive Adaptation (DACA) framework, which is built upon a Multi-Source Mean Teacher (MSMT) architecture with a Unified Proposal (UniP) approach. Additionally, two integral components, namely Multi-Source Probabilistic Bounding Box Ensemble (MSPE) and Memory Bank Consensus Contrastive Learning (MBCL), complement the MSMT framework. MSMT, an extension of the mean-teacher framework to MSFDAOD tasks, is designed to handle ***task specificity***. It incorporates a dynamic exponential moving average (EMA) frequency to facilitate stable and effective mutual learning between teacher–student pairs, progressively enhancing classification and localization ability. To tackle ***cross-model variability***, UniP and MSPE are developed to produce unified proposals and pseudo-labels, respectively. UniP generates unified proposals within the same image for distinct source models, providing a consistent training and weighting objective for the MSMT framework. MSPE enhances pseudo-label quality by fusing multi-teacher predictions into a unified pseudo-label set, offering more convincing information through boosted consensus predictions. Additionally, we propose MBCL to address ***object feature discriminability***. By utilizing the popular contrastive learning approach, MBCL systematically learns high-quality object-level features, improving the discriminability of detection model. To fully utilize gradually learned representations, a memory bank is employed to augment contrastive samples across varied image backgrounds. A comprehensive array of experiments is conducted to validate the efficacy of DACA. The results, including both quantitative analyses and visualization results, robustly demonstrate the superior performance of DACA.

In summary, our contributions are threefold:

- We propose to adapt multi-source-pretrained object detection models to the target domain, named Multi-Source-Free Domain Adaptive Object Detection (MSFDAOD). This topic addresses comprehensively the more realistic challenges and constraints of real-world scenarios, particularly addressing issues such as data privacy concerns and the assumption of multi-source distributions. To the best of our knowledge, this is the first work to explore MSFDAOD.
- To address the MSFDAOD problem effectively, we introduce a novel framework termed Divide-and-Aggregate Contrastive Adaptation (DACA), which is constructed within a new Multi-Source Mean Teacher (MSMT) framework. To obtain unified region proposals, perform effective self-supervised learning and learn discriminative object features, DACA also comprises three essential components: a proposal unifying method, a probabilistic bounding box fusion method, and a class-wise contrastive learning method.

– We conduct extensive experiments under various adaptation settings. The results, including quantitative comparisons, ablation studies, visualization results, and extension studies, robustly demonstrate the superior performance of DACA as compared to the state-of-the-art approaches, comprehensively demonstrating the effectiveness of DACA.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 defines the proposed MSFDAOD problem. Section 4 introduces the proposed DACA framework in detail. Section 5 presents experimental settings, results, and corresponding analysis. Section 6 gives the conclusion of our work.

2 Related Work

In this section, we delve into related work that is closely aligned with the MSFDAOD task. This encompasses OD, DAOD, SFDAOD, MSDAOD, and MSFDA. We also make comparisons between our proposed DACA framework and these existing approaches.

2.1 Object Detection

Object detection (OD) involves the localization and classification of existing objects within a given image, sampled from photos, video frames, or live footage. The success of deep learning has yielded numerous effective object detection methodologies grounded in deep networks. These methodologies can be categorized into two main groups: two-stage detectors and one-stage detectors. The inception of two-stage detectors can be traced back to R-CNN (Girshick et al., 2014), the first two-stage detector, and also the first to utilize CNN on object detection tasks. R-CNN frames the localization problem with a proposed *recognition using regions* paradigm together with a *sliding window* paradigm. Subsequent advancements in this family include Fast R-CNN (Girshick, 2015), which improves both R-CNN accuracy and speed by pretrained network, RoI pooling layers, and truncated SVD (Denton et al., 2014). Building upon these improvements, Faster R-CNN (Ren et al., 2015) incorporates region proposal networks (RPNs) to further elevate accuracy and speed.

While two-stage detectors take object detection tasks as "coarse to fine" steps, one-stage detectors attempt to achieve detection in a single step (Zou et al., 2023). SSD (Single Shot MultiBox Detector) (Liu et al., 2016) is a prominent one-stage detector. Without using any region proposals, SSD detects objects by predicting classification and regression results for a fixed set of bounding boxes (Liu et al., 2016). Similarly giving up region proposals, YOLO (You Only Look Once) (Redmon et al., 2016) partitions the input image into a grid of fixed dimensions, generating predictions based on each grid cell. Drawing inspiration from fully connected networks (Long et al., 2015), FCOS (Fully Convolutional One-Stage Object Detector) (Tian et al., 2019) performs pixel-wise prediction across various scales of the feature pyramid, complemented by a center-ness loss to mitigate the influence of some low-quality objects. Introducing a brand new set prediction paradigm for object detection, DETR (DEtection TRansformer) (Carion et al., 2020) designs transformer (Vaswani et al., 2017) encoders, decoders, and prediction feed-forward networks to predict objects based on image features. DiffusionDet (Chen et al., 2023) transfers the popular diffusion model (Ho et al., 2020) to object detection tasks by treating the bounding box prediction as a diffusion process. However, both two-stage and one-stage detectors experience a notable performance drop owing to their limited generalization capacity when the target distribution is different from the source. In contrast to these approaches, our proposed DACA adeptly performs domain adaptation across multiple source domains to the target domain, resulting in commendable generalization ability under domain shift. Furthermore, these OD methods fail when high-quality data with accurate labels is absent. Instead, the proposed DACA is capable of performing effective detection with only source-pretrained models.

2.2 Domain Adaptive Object Detection

Evidently, within real-world applications, conventional object detection models encounter a performance drop attributed to domain shift. Some early works identify the need for single-source domain adaptive object detection (DAOD), also known as cross-domain object detection, and make some attempts. DA-Faster (Chen et al., 2018) mitigates the domain gap by incorporating image and instance-level adaptation, along with a consistency regularization mechanism. DTPL (Inoue et al., 2018) employs a weakly supervised approach for DAOD, involving sequential fine-tuning of fully supervised detectors in two phases. Subsequently, a variety of algorithms and architectures have been applied to DAOD tasks. For instance, the mean teacher (MT) framework (Tarvainen & Valpola, 2017) aggregates information by averaging model parameters using exponential moving average (EMA). MT introduces noise to the model to prevent predictions from exhibiting bias towards specific training targets, aligning with the cross-domain detection setting where training and testing data originate from distinct domains. MTOR (Cai et al., 2019) extends MT to include region-level inter-graph, and region-level intra-graph consistencies, considering graph structures. UMT (Deng et al., 2021) applies MT with augmenting source samples via CycleGAN (Zhu et al., 2017) and employs a teacher–student bias healing pro-

cess. Addressing the challenge of low-quality pseudo-labels, AT (Li et al., 2022b) tackles domain shift by promoting mutual learning and strong-weak augmentations. Similar to teacher–student models, MAF (He et al., 2023) utilizes Paradigm Teacher to overcome source error collapse and guide knowledge distillation for the adaptation process.

While early works primarily rely on diverse public benchmarks, covering a range of situations such as various weather conditions, different art forms, etc., some recent works emphasize specific application scenarios. For instance, DAOD under adverse weather conditions (Sindagi et al., 2020; Li et al., 2023) and lighting conditions (Kennerley et al., 2023) have received significant attention. Several works have also investigated a more challenging cross-domain object detection setting, Domain Generalization for Object Detection, to improve the adapted model's generalization ability to unseen domains (Lin et al., 2021; Xu et al., 2023; Zhang et al., 2023a). While DAOD methods excel in addressing single source-target domain shift, they fall short in handling multi-source or source-free scenarios. In contrast, our proposed DACA framework is specifically engineered to achieve robust adaptation in both of these demanding scenarios.

2.3 Source-Free Domain Adaptive Object Detection

Conventional DAOD always incurs a substantial cost for data transmission (Liang et al., 2020; Fang et al., 2022). Meanwhile, the collected source data may potentially contain sensitive user privacy information and is unavailable due to privacy policy (Yang et al., 2020, 2021a, b; Li et al., 2021; Huang et al., 2021). To address the issues above, researchers make essential efforts on source-free domain adaptive object detection (SFDAOD). SED (Li et al., 2021) represents one of the initial endeavors to address SFDAOD challenges. This method generates high-quality pseudo-labels through self-entropy descent and handles false negatives using mosaic augmentation techniques (Bochkovskiy et al., 2020). HCL (Huang et al., 2021) employs contrastive learning (Kang et al., 2019) on the memorized source hypotheses to acquire instance-discriminative and category-discriminative representations within the target domain. Utilizing pseudo-labeling strategy (Kim et al., 2021; Liang et al., 2020; Li et al., 2021; Huang et al., 2021) on classification loss, LODS (Li et al., 2022a) meanwhile focuses on enhancing target domain style while aligning image-level and instance-level features. This alignment process enhances target model's capability to generalize across various image styles. IRG (Vibashan et al., 2023) draws inspiration from the well-established contrastive representation learning (CRL) framework, SimCLR (Chen et al., 2020b). To enhance the quality of representations and adaptation performance, IRG treats different region proposals surrounding the same object as distinct augmentations of the object. It employs a graph convolution network (Gori et al., 2005) to establish positive–negative pairs and introduces a distillation loss to the training process. In comparison to these methodologies, our proposed DACA harnesses multiple source domain knowledge embedded in multiple pretrained models through the implementation of a pseudo-label ensemble and a source weighting method.

2.4 Multi-source Domain Adaptive Object Detection

Multi-source domain adaptive object detection (MSDAOD) presents the challenge of utilizing source data from multiple distinct domains. Building upon earlier theoretical investigations into multi-source domain adaptation (MSDA) (Mansour et al., 2008; Hoffman et al., 2012; Lin et al., 2020), several early approaches of MSDA can be directly generalized to MSDAOD tasks. For instance, MDAN (Zhao et al., 2018) introduces a novel generalization bound and executes MSDA through adversarial learning within deep neural networks. M3SDA (Peng et al., 2019) aligns the deep feature distribution moments across multiple source domains, emphasizing the advantages of aligning source domains in the context of MSDA.

DMSN (Yao et al., 2021) is the first approach that specifically focuses on MSDAOD tasks. DMSN employs feature alignment strategies on low-level and high-level features, tailored to align the hierarchical features originating from both source and target domains. Recognizing that combined source hypotheses with appropriate weights can effectively represent target hypotheses (Mansour et al., 2008; Peng et al., 2019), DMSN assigns weights to source parameters by an inverse fashion of MT framework, in which multiple source subnets are updated by stochastic gradient descent (Robbins & Monro, 1951) and target subnet is updated by weighted EMA of source parameters. TRKP (Wu et al., 2022) preserves domain-specific knowledge through a disentanglement module during the teacher model pretraining process in an adversarial manner. Additionally, it introduces a target-relevant mining procedure incorporating the k-nearest neighbors (KNN) algorithm (Cover & Hart, 1967) to discover relevant knowledge and alleviate knowledge degradation. MTK (Zhang et al., 2022) proposes a two-stage strategy on low-level and high-level features to address the MSFDAOD problem. With the implementation of the attention mechanism, MTK achieves high-quality information fusion during both the training and testing phases. However, existing methods depend on source data and labels for their foundational training processes. In contrast, our proposed DACA model excels in scenarios where multi-source data is unavailable, with the help of its effective pseudo-labeling method and contrastive learning approach. This capability enhances privacy

protection and contributes to data transmission cost reduction.

2.5 Multi-source-free Domain Adaptation

Multi-source-free domain adaptation (MSFDA) is a more practical and challenging scenario where training data is sampled from multiple source domains and is invisible and unavailable during the adaptation process. DECISION (Ahmed et al., 2021) first focuses on MSFDA and proposes to optimize an information maximization loss to assign proper weights to source hypotheses. These weights are further utilized for weighted pseudo-labeling and clustering processes. CAiDA (Dong et al., 2021) proposes a novel MSFDA generalization bound based on more mild assumptions, proving that multi-source predictors benefit adaptation by high-quality pseudo-labeling. According to the theoretical analysis, CAiDA utilizes a transferability perception to quantify source contributions and a reliable pseudo-labeling module based on a confident anchor to achieve appropriate weights and a reliable pseudo-labeling process. Surrogate (Shen et al., 2023) also gives a novel generalization error bound, which further introduces a bias and variance trade-off, by designing a selective self-training paradigm and assigning several losses to different parts of the framework. All the methods discussed before are based on white-box models, where the network parameters are accessible. To address a more challenging scenario of black-box UDA (Fang et al., 2022), DINE (Liang et al., 2022) introduces an adaptive label smoothing method and structural regularization for self-knowledge distillation (Hinton et al., 2015) on a MT manner, which works well on black-box MSFDA tasks. Notably, US-MSMA (Li et al., 2022c) focuses on MSFDA in semantic segmentation tasks. In the first stage of US-MSMA, backbones and classifiers are randomly combined and trained with cross-model consistency, followed by domain-specific and domain-invariant features learning. In the second stage, an integrated final model is trained with pseudo-labels and mitigated to source models.

Recent works about MSFDA emphasize more on harmonizing discriminability and transferability for MSFDA (Kundu et al., 2022; Han et al., 2023). *Discriminability* stands for the ease of classifying objects into given categories by a pretrained classifier, while *transferability* refers to feature representations invariance between different domains (Chen et al., 2019, 2020a; Kundu et al., 2022). Discriminability and transferability are balanced by utilizing edge-mixup and feature-mixup (Kundu et al., 2022). DATE (Han et al., 2023) gives a novel Bayesian perspective on MSFDA target risk upper bound, which is further explained by discriminability and transferability trade-off by Bayesian formulas. DATE employs a proxy discriminability perception module based on novel sample habitat and habitat density definition, while a source-similarity transferability module is introduced to measure model transferability. The trade-off problem is gradually optimized in a universal decision and optimization process. TGMA (Yang et al., 2023) performs multi-source model selection and pseudo-label correction by domain-level and instance-level transferability matrix, respectively, based on a proposed label-free transferability metric. These algorithms are proven to work well for classification or segmentation tasks. However, given that these studies mainly concentrate on image classification or semantic segmentation tasks, the direct transposition of their methodologies to MSFDAOD tasks proves to be challenging. Fortunately, there exist several works (Lu et al., 2023; Liu et al., 2023b) delving into the domain of federated learning (Li et al., 2020a; Zhang et al., 2021a) on MSDA or MSDAOD task. These works hold a similar setting where training and testing data is stored in different devices with privacy-preserving policies. However, unlike these federated-learning-based approaches, our desired MSFDAOD is more realistic and challenging since we are constrained to accessing solely source-pretrained models, without the capacity to intervene in the detector pretraining process or alternate between source pretraining and aggregation steps.

In summary, previous works have addressed the necessity of multi-source-free domain adaptation in image classification (Ahmed et al., 2021; Dong et al., 2021; Liang et al., 2022; Kundu et al., 2022; Han et al., 2023; Shen et al., 2023) or semantic segmentation (Li et al., 2022c; Yang et al., 2023). Nevertheless, within the constraints given by DAOD, SFDAOD, and MSDAOD in a more confined MSFDAOD scenario, the direct application of any of these methodologies to MSFDAOD tasks proves to be a formidable challenge. For instance, in the context of MSDAOD tasks, the absence of source data or distributions poses a significant challenge, given the unavailability of supervised signals for the adaptation process. For SFDAOD tasks, the straightforward mixture of multiple source domains into a unified domain may lead to substantial knowledge degradation and a performance drop, due to the persistence of domain discrepancy within source domain pairs (Riemer et al., 2019; Yao et al., 2021) and distinct adaptation contributions of source domains to the target domain.

Concentrating on the application of MSFDA in the context of object detection, our proposed method, DACA, demonstrates adept capability in effecting efficient adaptation by leveraging multiple pretrained detection models, thereby attaining superior performance. In our proposed approach, we opt to proficiently acquire visual representations through contrastive learning and efficiently distill knowledge via our novel multi-source mean-teacher framework in a multi-source-free scenario. To the best of our knowledge, we are the first to study the MSFDAOD problem. Compared with all the approaches above, which are not suitable to be directly trans-

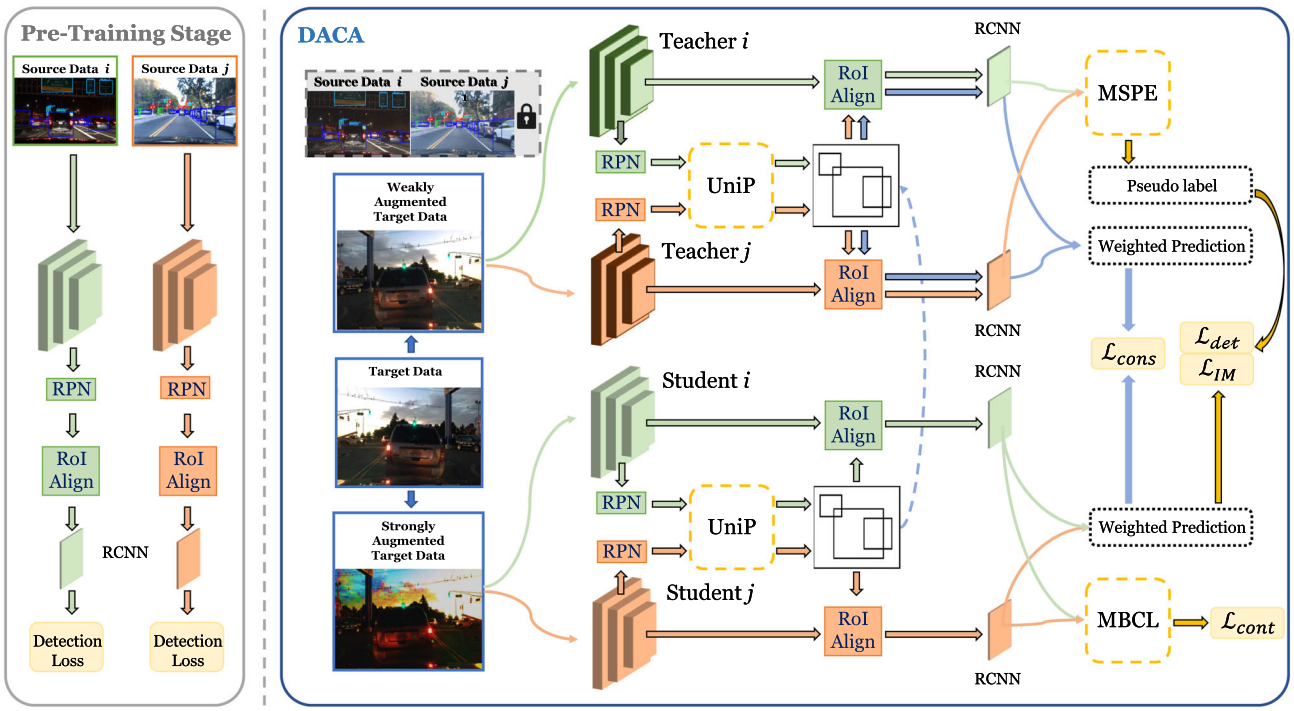

Fig. 2 Detailed framework of the proposed Divide-and-Aggregate Contrastive Adaptation (DACA) network. Source data is only available at source pretraining stage. After pretraining, each one teacher–student pair of multiple teacher–students will be initialized by a corresponding source model, then multiple teacher–student will cooperate in the adaptation stage within DACA. The initialization process is omitted in this figure

ferred to MSFDAOD, our novel approach can proficiently perform multi-source-free adaptation on object detection tasks, leading to superior MSFDAOD performance.

3 Problem Setup

In the MSFDAOD setting, there exists an unlabeled target domain denoted by T and multiple unseen source domains denoted by $S_1, S_2, \ldots, S_m$, where m is the number of source domains. We can only access the target dataset $X_T = \{\mathbf{x}_T^j\}_{j=1}^{N_T}$, where $\mathbf{x}_T$ denotes a single image sampled from the target distribution $p_T(\mathbf{x}_T, \mathbf{y}_T)$. $\mathbf{y}_T$ includes both bounding box label and category label distribution, which are unseen during adaptation. N_T denotes the total number of target samples. As for source information, only a set of source-pretrained models $\theta_S = \{\theta_i\}_{i=1}^m$ is available, where θ_i is the source model pretrained using samples from the i-th source distribution $p_{s_i}(\mathbf{x}_{S_i}, \mathbf{y}_{S_i})$. Source data, label, bounding box distribution, and label distribution are unseen during adaptation. Nevertheless, source models are pretrained with unknown strategies (e.g., supervised, unsupervised, or semi-supervised). Our goal is to transfer detection knowledge from source domains to the target domain and obtain a final target hypothesis $\theta_T : \mathbf{x}_T \rightarrow \mathbf{y}_p$ using only source-pretrained models $\theta_S = \{\theta_i\}_{i=1}^m$ and unlabeled target data $X_T = \{\mathbf{x}_T^j\}_{j=1}^{N_T}$. We implement DACA on Faster R-CNN (Ren et al., 2015) unless otherwise specified.

Similar to MSDAOD and MSFDA methods (Ahmed et al., 2021; Yao et al., 2021; Wu et al., 2022), we hold three assumptions:

(1) Unsupervised, i.e., no supervised information of the target domain is provided;
(2) Homogeneity, i.e., the source and target domains are observed in the identical data space with identical dimensionality;
(3) Closed-set, i.e., the label sets for the source and target domains are identical. Specifically, we denote the label sets for X_T as $\{c_i\}_{i=1}^k$, including k catrgories.

4 Divide-and-Aggregate Contrastive Adaptation

In this section, we present our approach to address the challenges posed by MSFDAOD: *Divide-and-Aggregate Contrastive Adaptation (DACA)* framework. DACA com-

prises a primary framework named Multi-Source Mean Teacher (MSMT) and three components: Unified Proposals (UniP), Multi-Source Probabilistic Bounding Box Ensemble (MSPE), and Memory Bank Consensus Contrastive Learning (MBCL). Throughout the training process, MSMT is employed for source networks to aggregate weighted predictions from source models, thereby distilling knowledge from the source to the target model. This facilitates an enhancement in model discriminability through utilizing stable pseudo-labels and trainable domain weights. To establish a reasonable schema for pseudo-label generation, we utilize MSPE to fuse bounding box predictions from different source models. Given the challenges associated with discriminative object-level features and biased pseudo-labels in source-free settings, MBCL is introduced to learn consensus class-wise high-quality features within each source domain feature space, effectively learning class-wise features. An overall framework of DACA is shown in Fig. 2. In this section, we will introduce the motivation and detailed method for each component, ending with the overall objective.

4.1 Multi-source Mean Teacher Framework

Motivation. To formulate a comprehensive framework for MSFDAOD, we initiate our approach by addressing the challenges associated with the transfer from MSDAOD to MSFDAOD. Given the absence of source data and its supervised signals, conventional MSDAOD algorithms are ineffective. For example, they typically incorporate a branch associated with source image detection loss (Yao et al., 2021; Wu et al., 2022; Zhang et al., 2022). To address this limitation and enhance basic detection performance, we naturally consider a pseudo-label learning pipeline. For previous MSFDA methods employed in classification or segmentation tasks using pseudo-label learning pipeline, possible pseudo-labeling techniques include neighborhood clustering (Yang et al., 2021a, b; Ahmed et al., 2021; Dong et al., 2021; Han et al., 2023) or directly weighting model output (Liang et al., 2022) or both (Shen et al., 2023). However, for detection tasks, these pseudo-labeling methods may exhibit suboptimal performance or be unavilable due to ***cross-model variability*** and ***task specificity***.

Considering that no previous MSFDA pseudo-label learning pipeline is suitable for MSFDAOD tasks, and in pursuit of a robust training process with more reliable pseudo-labels for MSFDAOD tasks, we start with leveraging MT (Tarvainen & Valpola, 2017), which generates pseudo-labels by a teacher model and facilitates teacher–student mutual learning. MT is commonly employed in both SFDAOD (Li et al., 2021, 2022a; Vibashan et al., 2023; Chu et al., 2023) and MSDAOD (Yao et al., 2021; Wu et al., 2022), revealing its potential to excel in MSFDAOD. Furthermore, recognizing that weight assignment for combining source hypotheses plays a crucial role in effectively representing target hypotheses in MSDA, MSDAOD, and MSFDAOD tasks (Mansour et al., 2008; Peng et al., 2019; Yao et al., 2021; Ahmed et al., 2021; Dong et al., 2021; Shen et al., 2023), we opt to optimize a set of weights representing source importance. However, this approach still encounters challenges of weighting source predictions due to ***cross-model variability***. To tackle this issue effectively, we design a strategy for generating unified proposals.

In response to the above observations, we introduce a Multi-Source Mean Teacher (MSMT) framework and a Unified Proposal (UniP) approach. Within the MSMT framework designed for MSFDAOD, students and teachers engage in mutual learning, leading to more stable knowledge distillation and more effective enhancement of model discriminability. Leveraging UniP, MSMT dynamically assigns optimized weights to student model predictions based on consistent proposals. This enables the framework to obtain more appropriate source weights and more accurate predictions by incorporating knowledge from multiple source models.

Method. In the original MT framework (Tarvainen & Valpola, 2017) for semi-supervised learning, a teacher model M_{te} and a student model M_{st} are initialized with identical parameters. The training loss of the student model involves a combination of a supervised loss, L_{sup}, and a consistency regularization term, denoted as L_{con}, designed to enforce consistency between teachers and students. Optimization methods such as stochastic gradient descent are employed for updating the student model, whereas the teacher model updates through the Exponential Moving Average (EMA) of student model parameters.

Within our MSMT framework, there exist m teacher–student pairs. We initialize the i-th teacher θ_i^{te} and the i-th student θ_i^{st} with the pretrained model θ_i. Concurrently, during the initialization stage, we initialize a set of parameterized domain weights $\{\alpha_i\}_{i=1}^m$. Following previous implementations of MT (Tarvainen & Valpola, 2017; Cao et al., 2023; Vibashan et al., 2023; Deng et al., 2023), inputs for teachers and students are weakly augmented and strongly augmented images, respectively. During the training process, we obtain a set of weighted predictions of input images from student models, along with a collection of pseudo-labels generated by teacher models serving as classification and regression targets. We start by discussing the process of obtaining a unified set of weighted predictions from the student models. To address this issue, we introduce UniP, a procedure for obtaining unified proposals.

UniP. At the beginning of the j-th iteration, the process initiates by acquiring the proposal set predicted by each source model, denoted as $\{\mathbf{prop}_i\}_{i=1}^m$. Following the paradigm of Faster R-CNN, the initial proposal set for each source model is established by applying Non-Maximum Suppression (NMS) to eliminate redundant proposals based on RPN

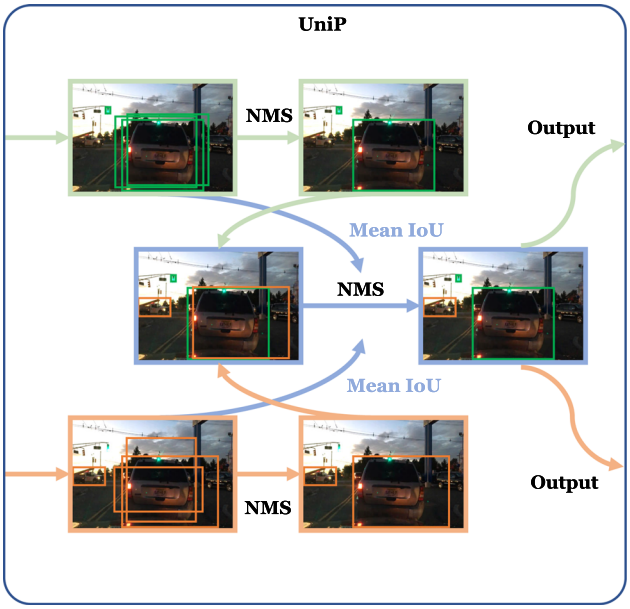

Fig. 3 Proposed Unified Proposal (UniP) approach. Proposals generated by multiple source models will undergo non-maximum suppression (NMS) individually. Subsequently, we combine all the proposals and conduct another round of NMS to eliminate redundant proposals based on the mean Intersection over Union (IoU) scores

scores. After performing NMS, we sort the list of post-NMS proposals within each $\mathbf{prop}_i$ in descending order of RPN prediction scores. We select the top p_t proposals after eliminating redundant boxes in NMS, where p_t is a hyper-parameter.

After performing NMS in all m proposal sets, we simply combine all the proposals into a unified set of proposals. However, simply combining proposals is likely to introduce duplicated proposals, e.g., proposals for easy samples that every RPN can easily detect. Following the observation that localization accuracy can be measured by the variance of bounding box regression (Xu et al., 2021; Zhang et al., 2023b), we implement an additional NMS after combination to eliminate duplicate proposals. In this phase, we calculate the mean Intersection over Union (mIoU) for each proposal, based on its associated proposals within the *original set*. For a single proposal from $\{\mathbf{prop}_i\}_{i=1}^m$, if this proposal is originally from $\mathbf{prop}_i$, we denote $\mathbf{prop}_i$ as its *original set* and the i-th model as its *original model*. Finally, we obtain a high-quality proposal set denoted as $\mathbf{prop}^{fin}$. The process of UniP is shown in Fig. 3.

After obtaining unified proposals for the current image using UniP, we replace each set of original proposals from $\{\theta_i^{st}\}_{i=1}^m$ with $\mathbf{prop}^{fin}$. Naturally, we continue forward propagation within each model detection head. Finally, a set of classification scores $\{\mathbf{cls}_i^{st}\}_{i=1}^m$ and bounding box regression predictions $\{\mathbf{reg}_i^{st}\}_{i=1}^m$ are acquired. For the classification scores, domain weights are applied, and the weighted student classification scores are computed as $\mathbf{cls}^{st} = \sum_{i=1}^m \alpha_i \mathbf{cls}_i^{st}$. As assigning weights to regression predictions is inappropriate, for each proposal in $\mathbf{prop}^{fin}$, its bounding box prediction is retained as the original prediction generated by its *original model*.

After obtaining a set of predictions for the image $\mathbf{x}_T$ using UniP, we can compute loss functions with pseudo labels $\widetilde{\mathbf{y}}_T$ to train the student detection models. To generate high-quality pseudo-labels, motivated by a multimodal bounding box fusion approach ProbEn (Chen et al., 2022), we propose a bounding box fusion method termed Multi-Source Probabilistic Bounding Box Ensemble (MSPE). This method is well-suited for pseudo-labeling scenarios involving predictions from multiple source teachers. Further elaboration on the methodology is provided in Sect. 4.2. After obtaining pseudo-labels, to optimize weight parameters and teacher–student models, we propose three losses within the MSMT framework: detection loss $\mathscr{L}_{det}$, information maximization loss $\mathscr{L}_{\text{IM}}$, and consistency loss $\mathscr{L}_{cons}$.

To construct the basic detection loss, we divide the original detection loss into two parts: RPN loss and R-CNN loss. Normalized weights are assigned to multiple object classification logits to calculate the weighted R-CNN classification loss following DECISION (Ahmed et al., 2021). Meanwhile, RPN loss and R-CNN regression loss are processed separately within each teacher–student pair. There are two reasons for assigning weights to classification scores only: (1) Regression scores treat the bounding box localization task as a regression problem and are defined on a continuous domain, which means it is inappropriate to assign weights to them. (2) Assigning the same weight to classification and regression scores may hurt performance since there exists misalignment between classification and localization due to task independence (Tian et al., 2019; Li et al., 2020b; Wang & Zhang, 2021; Deng et al., 2023). Overall, the detection loss $\mathscr{L}_{det}$ of student models can be given as:

$$\mathscr{L}_{det} = \mathscr{L}_{rpn} + \mathscr{L}_{rcnn}, \tag{1}$$

$$\mathscr{L}_{rpn} = \frac{1}{m}\sum_{i=1}^{m}\left(\mathscr{L}_{rpn}^{cls}\left(\theta_i^{st}(\mathbf{x}_T), \widetilde{\mathbf{y}}_T\right) + \mathscr{L}_{rpn}^{reg}\left(\theta_i^{st}(\mathbf{x}_T), \widetilde{\mathbf{y}}_T\right)\right), \tag{2}$$

$$\mathscr{L}_{rcnn} = \mathscr{L}_{rcnn}^{cls}\left(\sum_{i=1}^{m}\alpha_i\theta_i^{st}(\mathbf{x}_T), \widetilde{\mathbf{y}}_T\right) + \frac{1}{m}\sum_{i=1}^{m}\mathscr{L}_{rcnn}^{reg}\left(\theta_i^{st}(\mathbf{x}_T), \widetilde{\mathbf{y}}_T\right), \tag{3}$$

where $\mathscr{L}_{rpn}$ is the RPN loss and $\mathscr{L}_{rcnn}$ is the R-CNN loss. $\mathscr{L}_{rpn}^{cls}$ and $\mathscr{L}_{rpn}^{reg}$ are the classification and regression loss for RPN, respectively; $\mathscr{L}_{rcnn}^{cls}$ and $\mathscr{L}_{rcnn}^{reg}$ are the classification loss and regression loss for R-CNN, respectively.

To effectively learn source weights, we further implement the information maximization (IM) loss (Hu et al., 2017; Liang et al., 2020; Ahmed et al., 2021; Shen et al., 2023) with weights in our MSMT framework. The IM loss was utilized to promote the network to assign distinct one-hot encodings to the target feature representations and to fit the hypothesis (Hu et al., 2017; Liang et al., 2020). DECISION (Ahmed et al., 2021) first implemented the IM loss on MSFDA tasks for weighted predictions, enhancing the confidence and global diversity of weighted predictions on the target data. Surrogate (Shen et al., 2023) conducts more theoretical analyses on the assumptions about the optimal mixture distribution. Surrogate (Shen et al., 2023) concludes that the optimization goal of the IM loss perfectly matches the attempt to access the optimal mixture distribution. Based on the analyses above, we utilize the IM loss in DACA to optimize source weights and enhance target feature learning. Similar to DECISION and Surrogate, we attempt to optimize a weighted IM loss, i.e., reduce a conditional entropy term as well as increase an empirical label distribution entropy term (Ahmed et al., 2021):

$$\mathscr{L}_{\mathrm{IM}} = \frac{1}{N_T}\sum_{i=1}^{N_T} H\left(\sum_{j=1}^{k}\sigma_j\left(\sum_{p=1}^{m}\alpha_p\theta_p^{st}\left(\mathbf{x}_T^i\right)\right)\right) - H\left(\frac{1}{N_T}\sum_{i=1}^{N_T}\sum_{j=1}^{k}\sigma_j\left(\sum_{p=1}^{m}\alpha_p\theta_p^{st}\left(\mathbf{x}_T^i\right)\right)\right), \quad (4)$$

where σ denotes the softmax operator and H denotes the entropy function.

While MT with strong-weak augmentation can distill knowledge with noisy pseudo-labels to a certain extent, we further ensure more robust mutual learning by facilitating consistency between teacher and student models. We propose to add consistency regularization terms for the MSMT framework. Specifically, we conduct inference on teacher models by getting the teacher proposals replaced by student-generated $\mathbf{prop}^{fin}$. Afterward, we obtain individual teacher predictions and then obtain weighted classification predictions, denoted as $\mathbf{cls}^{te} = \sum_{i=1}^{m}\alpha_i\mathbf{cls}_i^{te}$, where $\mathbf{cls}_i^{te}$ is the classification scores of the i-th teacher. We construct a consistency loss between weighted classification predictions from students and teachers, given as:

$$\mathscr{L}_{cons} = KL(\sigma(\mathbf{cls}^{st}), \sigma(\mathbf{cls}^{te})), \quad (5)$$

where KL denotes the Kullback–Leibler divergence and σ denotes the softmax operator.

4.2 Multi-source Probabilistic Bounding Box Ensemble

Motivation. In DAOD tasks, pseudo-labels, typically consisting of bounding boxes with class logits, are commonly employed as "pseudo" supervisory signals for unsupervised or semi-supervised target domain data, ensuring effective target representation learning (Vibashan et al., 2023). In prior SFDAOD or MSDAOD methods, pseudo-labels are usually generated by a single model, specifically, by the single model that will be optimized for inference (Huang et al., 2021; Li et al., 2021), or by the single teacher model in MT-based methods (Li et al., 2022a; Wu et al., 2022; Vibashan et al., 2023). Through additional operations such as confidence thresholding, the pseudo-labels can be directly employed to construct detection losses. However, in our MSMT framework, each teacher is capable of independently generating a set of pseudo-labels. A natural question arises: *How can we integrate these distinct sets of pseudo-labels into a unified set of pseudo-labels*? Two intuitive approaches involve directly combining, or applying Non-Maximum Suppression (NMS) to all pseudo-labels generated by all teachers. However, simply combining pseudo-labels may result in a significant number of duplicated or inaccurately predicted boxes, introducing substantial bias. Applying NMS with either classification scores or localization metric e.g., mIoU in Unip may also lead to inferior performance as we need to consider classification and localization both to construct more precise pseudo-labels. Borrowing the insightful idea from ProbEn (Chen et al., 2022), a multimodal bounding box fusion method, we extend ProbEn to a methodology for merging detection pseudo-labels from multiple source models for a more reasonable classification and localization ensemble, which is named Multi-Source Probabilistic Bounding Box Ensemble (MSPE).

Method. To formulate the complete MSPE process, we initially establish MSPE for two distinct source models, following the methodology of ProbEn (Chen et al., 2022). We denote the two chosen source models as θ_1 and θ_2. We also denote the current target image and label as $\mathbf{x}_T$, $\mathbf{y}_T$, respectively. An assumption can be made that the teacher models are conditional independent since they are pretrained from independent source domains and are updated from independent student models:

$$p(\theta_1, \theta_2|\mathbf{x}_T, \mathbf{y}_T) = p(\theta_1|\mathbf{x}_T, \mathbf{y}_T) * p(\theta_2|\mathbf{x}_T, \mathbf{y}_T). \quad (6)$$

To obtain the classification prediction results, following the ProbEn (Chen et al., 2022) theory, we begin with:

$$p(\mathbf{y}_T|\theta_1, \theta_2, \mathbf{x}_T) = \frac{p(\mathbf{x}_T, \mathbf{y}_T, \theta_1, \theta_2)}{p(\mathbf{x}_T, \theta_1, \theta_2)} \propto p(\theta_1, \theta_2|\mathbf{x}_T, \mathbf{y}_T)p(\mathbf{x}_T, \mathbf{y}_T)$$

$$
\begin{aligned}
&\propto p(\theta_1|\mathbf{x}_T,\mathbf{y}_T) \propto p(\theta_2|\mathbf{x}_T,\mathbf{y}_T)p(\mathbf{x}_T,\mathbf{y}_T)\\
&\propto \frac{p(\mathbf{y}_T|\theta_1,\mathbf{x}_T)p(\theta_1,\mathbf{x}_T)}{p(\mathbf{x}_T,\mathbf{y}_T)}\\
&\quad\frac{p(\mathbf{y}_T|\theta_2,\mathbf{x}_T)p(\theta_2,\mathbf{x}_T)}{p(\mathbf{x}_T,\mathbf{y}_T)}p(\mathbf{x}_T,\mathbf{y}_T)\\
&\propto \frac{p(\mathbf{y}_T|\theta_1,\mathbf{x}_T)p(\mathbf{y}_T|\theta_2,\mathbf{x}_T)}{p(\mathbf{x}_T,\mathbf{y}_T)}.
\end{aligned}
\tag{7}
$$

Then we can simply extend Eq. (7) to m source domains:

$$
p\left(\mathbf{y}_T|\{\theta_i\}_{i=1}^m,\mathbf{x}_T\right) \propto \frac{\Pi_{i=1}^m p(\mathbf{y}_T|\theta_i,\mathbf{x}_T)}{p(\mathbf{x}_T,\mathbf{y}_T)^{m-1}}. \tag{8}
$$

Following the fact introduced by ProbEn that the softmax denominator's partition function is not the current class label's function (Chen et al., 2022), by utilizing softmax posterior as an activate function, the softmax posterior for class-k for a single source model can be derived as:

$$
\begin{aligned}
p(\mathbf{y}_T{=}k|\theta_i,\mathbf{x}_T) &= \frac{\exp(\theta_i(\mathbf{x}_T)[k])}{\sum_j \exp(\theta_i(\mathbf{x}_T)[j])}\\
&\propto \exp(\theta_i(\mathbf{x}_T)[k]).
\end{aligned}
\tag{9}
$$

Then MSPE for classification is derived as:

$$
\begin{aligned}
p\left(\mathbf{y}_T=k|\{\theta_i\}_{i=1}^m,\mathbf{x}_T\right) &\propto \frac{\Pi_{i=1}^m \exp(\theta_i(\mathbf{x}_T)[k])}{p^{m-1}(\mathbf{x}_T,\mathbf{y}_T)}\\
&\propto \frac{exp(\sum_{i=1}^m \theta_i(\mathbf{x}_T)[k])}{p^{m-1}(\mathbf{x}_T,\mathbf{y}_T)}.
\end{aligned}
\tag{10}
$$

This implies that the class-wise prediction is related to summing class logits of correlated predictions from different source models. After summing logits, softmax is applied, and the results are divided by a joint distribution of $\mathbf{x}_T$ and $\mathbf{y}_T$. According to this result, we can derive our MSPE algorithm. In our algorithm, we neglect the influence of $p^{m-1}(\mathbf{x}_T,\mathbf{y}_T)$ and solely sum class logits. This decision is based on our empirical observation that approximating $p^{m-1}(\mathbf{x}_T,\mathbf{y}_T)$ is challenging, and simply summing logits yields optimal performance.

For bounding box fusion, we define $\mathbf{z}_T$ for the bounding box continuous random variable for a target image, following ProbEn (Chen et al., 2022). Except for the assumption of a uniform prior on $p(\mathbf{z}_T)$ as a bounding box can lie anywhere within the image with the same probability as stated in ProbEn, we also assume a uniform prior on $p(\mathbf{x}_T,\mathbf{z}_T)$ implying that all combinations of an image and a bounding box are of the same probability without any preferences if there exists no prior knowledge. We denote σ_i^2 as the Gaussian distribution variance and μ_i as the box coordinates from the i-th source model prediction following ProbEn (Chen et al., 2022). Then we can derive MSPE bounding box coordination fusion beginning with two source domains:

$$
\begin{aligned}
p(\mathbf{z}_T|\theta_1,\theta_2,\mathbf{x}_T) &= \frac{p(\theta_1,\theta_2|\mathbf{x}_T,\mathbf{z}_T)p(\mathbf{x}_T,\mathbf{z}_T)}{p(\mathbf{x}_T,\theta_1,\theta_2)}\\
&\propto p(\theta_1|\mathbf{x}_T,\mathbf{z}_T)p(\theta_2|\mathbf{x}_T,\mathbf{z}_T)p(\mathbf{x}_T,\mathbf{z}_T)\\
&\propto \frac{p(\mathbf{z}_T|\theta_1,\mathbf{x}_T)p(\theta_1,\mathbf{x}_T)}{p(\mathbf{x}_T,\mathbf{z}_T)}\frac{p(\mathbf{z}_T|\theta_2,\mathbf{x}_T)p(\theta_2,\mathbf{x}_T)}{p(\mathbf{x}_T,\mathbf{z}_T)}p(\mathbf{x}_T,\mathbf{z}_T)\\
&\propto p(\mathbf{z}_T|\theta_1,\mathbf{x}_T)p(\mathbf{z}_T|\theta_2,\mathbf{x}_T)\frac{p(\theta_1,\mathbf{x}_T)p(\theta_2,\mathbf{x}_T)}{p(\mathbf{x}_T,\mathbf{z}_T)}\\
&\propto p(\mathbf{z}_T|\theta_1,\mathbf{x}_T)p(\mathbf{z}_T|\theta_2,\mathbf{x}_T)\\
&\propto \exp\left(\frac{\|\mathbf{z}_T-\mu_1\|^2}{-2\sigma_1^2}\right)\exp\left(\frac{\|\mathbf{z}_T-\mu_2\|^2}{-2\sigma_2^2}\right).
\end{aligned}
\tag{11}
$$

Afterward, the derivation process is the same as that of Eq. (7) in ProbEn (Chen et al., 2022). We refer readers to the remaining proof in ProbEn (Chen et al., 2022) paper and ProbEn appendix.

Considering the theoretical proof above, we can easily design our MSPE procedure following the ProbEn procedure, as shown in Algorithm 1 and Fig. 4. In our implementation, we also utilize UniP for multiple teachers to obtain unified proposals before MSPE, as shown in Fig. 2. Since our MSFDAOD setting cannot access any supervised signals, a vanilla implementation of MSPE may introduce bias. Rationales for this are discussed in Sect. 5.3.

Algorithm 1 MSPE

Input: Bounding boxes $\{bbox_i\}_{i=1}^m$ from m teacher models, with corresponding logits. IoU threshold τ_{iou}. confidence threshold τ_{con}.
Output: Final bounding boxes with logits

Initialization: empty bounding box sets $bbox_{ens}$, $bbox_{fin}$, $\{bbox_i^o\}_{i=1}^m$.
repeat
  Find the single bounding box $bbox_{cur}$ in $\{bbox_i\}_{i=1}^m$ with largest class posterior.
  for $i = 1$ to m **do**
    Find all bounding boxes from $bbox_i$ with IoU $\geq \tau_{iou}$ with $bbox_{cur}$, add eligible bounding boxes to $bbox_i^o$.
  end for
  for $i = 1$ to m **do**
    if $bbox_i^o$ is not empty **then**
      In $bbox_i^o$ find the bbox having the largest IoU with $bbox_{cur}$, add it to $bbox_{ens}$.
    end if
  end for
  if $bbox_{ens}$ is not empty **then**
    Sum class logits and average box coordinates of all bounding boxes in $bbox_{ens}$, add the ensembled bounding box to $bbox_{fin}$.
  end if
  Remove $\{bbox_i^o\}_{i=1}^m$ from $\{bbox_i\}_{i=1}^m$.
  Empty $bbox_{ens}$ and $\{bbox_i^o\}_{i=1}^m$.
until $\{bbox_i\}_{i=1}^m$ are all empty
Filter confident bounding boxes in $bbox_{fin}$ by τ_{con}.
return Filtered $bbox_{fin}$.

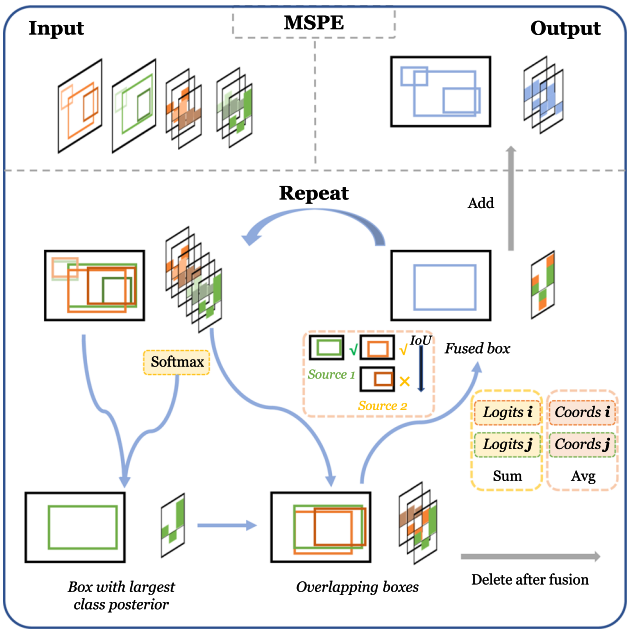

Fig. 4 Proposed Multi-source Probabilistic Bounding Box Ensemble (MSPE) approach. For multiple teacher inputs, including both bounding box predictions and class logits, we combine them into a unified input set. To begin MSPE, we identify the bounding box with the highest class posterior probability, denoting it as the current box. Subsequently, we identify all bounding boxes with an IoU $\geq \tau_{iou}$. From the bounding box set generated by each teacher model, we select the box having the highest IoU with the current box. We then merge this selected box with the current box to form a singular bounding box prediction, which is appended to the output set. The current box and overlapping boxes are then removed. This process iterates until the input set is empty

Discussions. We provide some discussions about the MSPE design and methodology by answering several potential questions as follows:

– ***What's the difference between ProbEn*** (Chen et al., 2022) ***and MSPE?*** Firstly, MSPE and ProbEn are implemented on different tasks. ProbEn is designed for multimodal image object detection tasks, while MSPE is adopted in MSFDAOD tasks. Models can access images of the same view in different modalities such as RGB, thermal, etc., in ProbEn tasks. However, models can only access unsupervised target domain images in MSFDAOD tasks. Secondly, the theoretical derivation processes of ProbEn and MSPE are different. The prerequisites of ProbEn include multimodal images, while those of MSPE include target images and multiple source models. In Eq. (6) of ProbEn (Chen et al., 2022), the conditional confidence distribution is inversely proportional to $p(y = k)^{(M-1)}$, in which y is the object distribution and M is the number of modalities. In Eq. (10) of DACA, the conditional confidence distribution is inversely proportional to $p^{m-1}(\mathbf{x}_T, \mathbf{y}_T)$, which is completely different to $p(y = k)^{(M-1)}$. Considering that the target domain is unsupervised in the MSFDA setting, it is extremely challenging to approximate the joint distribution $p^{m-1}(\mathbf{x}_T, \mathbf{y}_T)$. Thus, we borrow the solution from ProbEn which simply neglects the denominator distribution. In that case, we can simply modify the original ProbEn algorithm to ensemble bounding boxes from multiple source models, extending ProbEn to MSFDAOD tasks. This leads to a high similarity of methodologies between ProbEn and MSPE. Generally, the differences between ProbEn and MSPE are mainly (1) Application tasks and (2) Theoretical derivation. Moreover, since we implement UniP before MSPE, there is also a slight difference between ProbEn and MSPE methodologies.

– ***Can MSPE be extended to student prediction ensemble?*** Given that the proposed MSPE is essentially a bounding box ensemble algorithm, it is also capable of integrating student predictions. However, we do not recommend this approach, since multi-source weights will not be effectively optimized, as no source weights will be assigned to student models.

4.3 Memory Bank Consensus Contrastive Learning

Motivation. In MSMT, where consensus pseudo-labels are constructed, each source model is empowered to perform effective knowledge distillation. However, the quality of feature representation remains suboptimal under the raw mean-teacher framework (Vibashan et al., 2023). Previous research has also sought solutions to this challenge. IRG (Vibashan et al., 2023) obtains object pair-wise relations by constructing a Graph Convolutional Network (GCN) (Kipf & Welling, 2016). Nevertheless, graph convolution operations may prove time-consuming in real-world scenarios (Yu & Qin, 2020). CMT (Cao et al., 2023) has introduced a contrastive learning strategy involving teacher and student features at multiple scales, yet it may learn on incorrectly annotated features since no supervisory signals of the target domain are provided. Considering these concerns, to address the challenge of ***object feature discriminability***, we propose a Memory Bank Consensus Contrastive Learning (MBCL) approach within the MSMT framework. This approach aims to lead the network to contrastively learn consensus class-wise features, enhance the network's ability to acquire high-quality feature representations and improve its discriminability.

Method. The InfoNCE loss (Oord et al., 2018), a widely employed unsupervised contrastive learning loss function, is designed to enhance the maximization of mutual information between input and context. This improves the model's capability to distinguish between positive and negative pairs. The crucial component for effective contrastive learning lies in the construction of positive and negative pairs. Exploiting the InfoNCE characteristic to ensure the proximity of posi-

tive features to the cluster centroid and the clarity of decision boundaries, it is instinctive to assign instances sharing the same category label to the positive cluster and vice versa. Nevertheless, considering the existence of noise in pseudo-labels, particularly the occurrence of incorrect predictions in source-free and unsupervised scenarios, we employ the following strategy to discern high-quality feature representations as keys.

During the training phase of the i-th student network, we acquire object-level features by applying RoI-Alignment with both image-level features and corresponding proposals of the current image. It is noteworthy that distinct sets of RoI features are obtained with image-level features from different students, and distinct sets of RoI features are processed separately to ensure consistent learning within the same feature space. To construct memory banks with high-quality features, we employ three criteria for sample selection: (1) A class probability exceeding a predefined threshold fixed at 0.7; (2) Prediction entropy below the average entropy in the current minibatch; (3) Consensus predictions, where "consensus" is defined as (a) the category predictions of the same instance from the current source model and the ensemble model are consistent, and (b) the final prediction, i.e., the weighted prediction and the pseudo-labels are consistent. Criteria (1) and (2) aim to incorporate representative class-wise information into the key set while excluding potentially challenging samples. Criterion (3) serves to safeguard the key set against pollution by biased or erroneous predictions.

Following the above filtering process, we categorize current object-level features into consensus features and non-consensus features. Aiming to perform class-wise contrastive learning within each source domain feature space, we establish $m \cdot (k+1)$ class-wise memory banks, as we also consider background features for abundant samples and better generalization ability. Here, we denote the i-th student model's j-th category memory bank as $\mathbf{MB}_i^j$. The memory bank size is fixed at l_{mb} for all source domains and categories, where l_{mb} is a hyperparameter. Upon acquiring the features for each potential category, we adopt a First-In-First-Out (FIFO) approach to update the respective memory bank. It is important to note that we update memory banks each iteration. Through these steps, we assemble a memory bank containing high-quality consensus features for each source domain and category.

In each iteration, after updating the feature memory bank, we construct and back-propagate the contrastive loss. We denote the input query features as $\{\{\mathcal{Q}_i^j\}_{j=1}^{k+1}\}_{i=1}^{m}$, where $\mathcal{Q}_i^j$ represents the feature belonging to the j-th category extracted by the i-th student model. For a given set of query features from a source domain with the same predicted category, we take memory bank features of the same category as positive keys, while memory bank features from the remaining categories serve as negative keys. According to the principle that features belonging to the same category as the query are positive keys and vice versa, the final contrastive loss is given by:

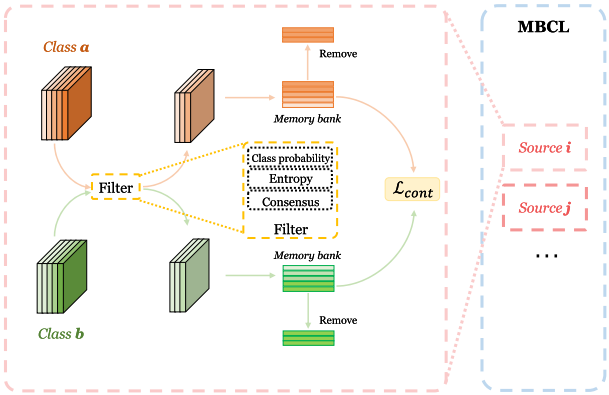

Fig. 5 Proposed Memory Bank Consensus Contrastive Learning (MBCL) approach. For the class-wise input features derived from each student model, we extract consensus features and subsequently update the memory bank with these selected features. Then we formulate a class-wise contrastive loss by the consensus features and the memory bank features within each source feature space

$$\mathcal{L}_{cont} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{c_i}\sum_{j=1}^{k+1}\mathcal{L}_{cont}^{ij}, \tag{12}$$

$$\mathcal{L}_{cont}^{ij} = \frac{1}{|\mathcal{Q}_i^j|}\sum_{q\in\mathcal{Q}_i^j}\frac{-1}{|\mathcal{K}_{pos}|} \sum_{K^+\in\mathcal{K}_{pos}}\log\frac{\exp(q\cdot K^+/\tau)}{\sum_{K\in\mathcal{K}_{all}}\exp(q\cdot K/\tau)}, \tag{13}$$

$$\mathcal{K}_{pos} = \mathbf{MB}_i^j, \tag{14}$$

$$\mathcal{K}_{all} = \{\mathbf{MB}_i^{j'}\}_{j'=1}^{k+1}, \tag{15}$$

where $\mathcal{K}_{pos}$ represents positive keys, and $\mathcal{K}_{all}$ represents all keys, including both positive and negative keys for $\mathcal{Q}_i^j$, the query for the i-th source domain and j-th category, to conduct contrastive learning. Here, c_i represents the total number of predicted categories in the current iteration, and τ denotes the temperature of contrastive learning, which is a hyperparameter. For better readability, we use lowercase k to denote the total number of categories and capital K to denote the key in a key set. The process of MBCL is illustrated in Fig. 5.

4.4 Overall Objective

In this section, we introduce the training and testing process in an overall objective.

Training. At the start of each training iteration, a minibatch of target images is sampled from the target dataset as input. Weakly augmented images are fed to the teacher models, while strongly augmented images are fed to the student models. Once forward propagation and loss calculation are completed across all components, we combine the final loss, denoted as $\mathscr{L}_{fin}$, using the proposed detection loss, information maximization loss, consistency loss, and contrastive loss with weighting hyperparameters ω_1 to ω_4:

$$\mathscr{L}_{fin} = \omega_1 \mathscr{L}_{det} + \omega_2 \mathscr{L}_{IM} + \omega_3 \mathscr{L}_{cons} + \omega_4 \mathscr{L}_{cont}. \tag{16}$$

Subsequently, we perform backpropagation through $\mathscr{L}_{fin}$ to update model parameters. The backpropagation process is applied only to the student models. Once the parameters of the student models are updated, we update each corresponding teacher model using the exponential moving average (EMA) of its associated student model. The update process for the student models and weights, as well as the EMA process, are denoted by Eqs. (17), (18) and (19), respectively:

$$\theta_i^{st} \leftarrow \theta_i^{st} + \gamma \frac{\partial(\mathscr{L}_{fin})}{\partial \theta_i^{st}}, \tag{17}$$

$$\{\alpha_i\}_{i=1}^m \leftarrow \{\alpha_i\}_{i=1}^m + \gamma \frac{\partial(\mathscr{L}_{fin})}{\partial\left(\{\alpha_i\}_{i=1}^m\right)}, \tag{18}$$

$$\theta_i^{te} \leftarrow \eta \theta_i^{te} + (1-\eta)\theta_i^{st}, \tag{19}$$

where η is a predefined coefficient controlling the teacher updating rate, and γ represents the student learning rate. During the iteration for executing EMA, we perform EMA for the m teacher–student pairs $\{\theta_i^{te}, \theta_i^{st}\}_{i=1}^m$. Instead of executing EMA every iteration, we set the EMA frequency to a dynamic number of iterations for smoother and more stable updates. The empirical equation for determining the EMA frequency is:

$$EMA\_iter = k * N_T // 500. \tag{20}$$

Testing. To fully leverage the finely adjusted domain weights and ensure consistency with the training objective during inference, we assign weights to the final predictions from multiple source models using UniP-generated proposals to obtain the final testing results.

5 Experiments

In this section, we conduct a series of experiments, including comparisons between DACA and other state-of-the-art approaches, ablation studies, visualizations, a model integration study, and an extension study of DACA. The objective is to demonstrate the effectiveness of DACA through both quantitative comparisons and visualization results. The source code will be released at https://github.com/HuizaiVictorYao/DACA.

5.1 Experimental Settings

We introduce the datasets, baselines, and implementation details in this section.

5.1.1 Datasets

Cityscapes. Cityscapes (Cordts et al., 2016) is a dataset for semantic urban scene understanding, featuring images captured in diverse urban environments. The dataset comprises 2,975 training images and 500 validation images, each annotated at the pixel level.

Bdd100k. Bdd100k (Yu et al., 2020) is a large-scale dataset with a total of 100,000 images, with 80,000 of them annotated with bounding boxes. Among the accurately annotated images, 70,000 are divided into a training set and 10,000 are divided into a validation set. The annotated images are distributed across three distinct periods: daytime, night, and dawn/dusk. In our partitioning, daytime has 36,728 training and 5258 validation images; night has 27,971 training and 3929 validation images; dawn/dusk has 5027 training and 778 validation images.

KITTI. KITTI (Geiger et al., 2012), specifically designed for autonomous driving systems, provides a benchmark dataset with images captured across diverse scenes, including cities, highways, and rural areas. In our configuration, we utilize all 7481 images in the original training set as source domain images.

SHIFT. SHIFT (Sun et al., 2022) emerges as a comprehensive synthetic dataset for autonomous driving, featuring various domain shifts such as different times of the day (sun altitude angle) and adverse weather conditions. In our experimental settings, we only employ images under daytime and specific weather conditions including clear, heavy cloudy, heavy foggy, and heavy rainy. In our partitioning, clear has 18,906 training and 3152 validation images; heavy cloudy has 13,308 training and 2218 validation images; heavy foggy has 12,543 training and 2091 validation images; heavy rainy has 19,191 training and 3199 validation images.

Sim10k. Sim10k (Johnson-Roberson et al., 2017) is a synthetic dataset derived from the video game Grand Theft Auto V (GTA V), simulating real-world scenarios with a total of 10,000 images. In our experimental setting, we utilize all 9000 training images and 1000 testing images from Sim10k.

5.1.2 Baselines

To conduct comprehensive comparisons, we benchmark DACA against the following baselines:

(1) *Source-only*: This involves training on the source domain and testing on the target domain without any adaptation, serving as a lower bound for domain adaptation performance.
(2) *SFDAOD*: This involves using a single source domain or constructing a unified source domain for SFDAOD methods.
(3) *MSDAOD*: This involves the direct utilization of existing MSDA approaches with annotated source domains. Corresponding methods include vanilla MSDAOD and transferring MSDA for classification to detection.
(4) *MSFDAOD*: As no pioneers in MSFDA research focus on object detection tasks, we have adapted several MSFDA for classification or segmentation methods to object detection as MSFDAOD baselines.
(5) *Oracle*: This intends training and testing directly on target domain with its supervisory signals. This represents the upper bound of domain adaptation in some extent.

To obtain a single domain from multiple source domains in Source-only and SFDAOD experiments, we employ two strategies:

(1) *Single-source*. In this approach, we only select one source domain. In scenarios with two source domains, we report evaluation results for both. In cases with more than two source domains, we focus on *Single-best* and *Single-worst*, meaning we perform adaptation on each single domain and report the corresponding best or worst results. The division of *Single-best* and *Single-worst* is for illustrating different contributions to adaptation performance due to different degrees of domain gap.
(2) *Source-combined*, where we combine all source domain datasets into a single dataset, creating a *combined* domain.

For source-only experiments, we contend that comparing our method with single-source source-available DAOD approaches is meaningless due to significant differences in experimental settings. Consequently, we only employ Faster R-CNN (Ren et al., 2015) as the source-only object detection baseline.

For SFDAOD experiments, we compare against SOAP (Xiong et al., 2021), LODS (Li et al., 2022a), and IRG (Vibashan et al., 2023). These approaches are selected because they have fully runnable open-source codebases.

For MSDAOD experiments, we compare against DMSN (Yao et al., 2021), TRKP (Wu et al., 2022), MDAN (Zhao et al., 2018), and M3SDA (Peng et al., 2019). MDAN and M3SDA are originally designed for classification tasks, so they are modified to object detection tasks for a comprehensive comparison, following DMSN. It is worth noting that MSDAOD approaches are likely to outperform MSFDAOD methods due to the availability of abundant source data and labels. However, we can still gain insights from these results to analyze the capacity of MSFDAOD methods to learn without any supervisory information.

For MSFDAOD experiments, we adapt DECISION (Ahmed et al., 2021) and US-MSMA (Li et al., 2022c), originally designed for classification and segmentation tasks, respectively, to object detection tasks, considering factors such as training time consumption and modification feasibility.

5.1.3 Implementation Details

Task-specific details. Here, we introduce some task-specific details and necessary modifications to demonstrate how we conduct fair and comprehensive comparisons:

– *SFDAOD*. For SFDAOD experiments, all experiments are conducted using corresponding open-source implementations. It is important to note that we implement VGG16 (Simonyan & Zisserman, 2014) as the backbone for IRG (Vibashan et al., 2023), which is originally implemented with a ResNet (He et al., 2016) backbone, to ensure fair comparisons. We employ *Single-source* and *Source-combined* strategies to construct the source domain as previously stated.
– *MSDAOD*. For DMSN (Yao et al., 2021) and TRKP (Wu et al., 2022) experiments, as they are not open-sourced, we only refer to existing results in the original paper. For MDAN (Zhao et al., 2018) and M3SDA (Peng et al., 2019), it is feasible to directly apply them to object detection tasks. For MDAN, we treat the RPN and R-CNN detection head in Faster R-CNN as "Desired Task" in the original paper (Zhao et al., 2018), with the VGG16 backbone serving as "Feature Extractor". For M3SDA, we take the VGG16 backbone, RPN, and R-CNN convolutional layers before the R-CNN classifier as the "Feature Extractor" mentioned in the original paper (Peng et al., 2019), considering that only object-level classification predictions are provided by the last classification layers of R-CNN head.
– *MSFDA to MSFDAOD*. For MSFDA to MSFDAOD experiments, we make necessary modifications based on the original MSFDA methods. It is important to note that, due to ***task specificity*** and ***cross-model variability*** mentioned above, these two methods cannot be directly applied to MSFDAOD. For instance, the weighted information maximization and weighted pseudo-labeling in DECISION cannot be directly utilized for object detection tasks, as Regions of Interest (RoIs) vary from model to model, i.e., ***cross-model variability***. To address this challenge, we employ UniP to obtain a unified proposal set for each minibatch of images. Nevertheless, as the MT framework is not present in DECISION and US-MSMA,

Table 1 Comparisons with state-of-the-art SFDAOD, MSDAOD, and MSFDAOD methods under **cross-camera** setting measured by mAP (%)

Settings and methods				Results (%)
Settings	Methods	Original task	Source domain	mAP
Source-Only	Faster R-CNN (Ren et al., 2015)	Detection	C	**44.60**
			K	28.60
			C+K combined	43.20
SFDAOD	LODS (Li et al., 2022a)	Detection	C	42.07
			K	34.11
			C+K combined	45.67
	SOAP (Xiong et al., 2021)	Detection	C	39.55
			K	34.26
			C+K combined	41.47
	IRG (Vibashan et al., 2023)	Detection	C	**47.73**
			K	34.46
			C+K combined	45.87
MSDAOD	MDAN (Zhao et al., 2018)	Classification	C+K	43.20
	M3SDA (Peng et al., 2019)	Classification	C+K	44.10
	DMSN (Yao et al., 2021)	Detection	C+K	49.20
	TRKP (Wu et al., 2022)	Detection	C+K	**58.40**
MSFDAOD	DECISION (Ahmed et al., 2021)	Classification	C+K	44.89
	US-MSMA (Li et al., 2022c)	Segmentation	C+K	40.08
	DACA (Ours)	Detection	C+K	**51.45**
Oracle	Faster R-CNN (Ren et al., 2015)	Detection	Bdd100k	**59.20**

The best method within each setting is emphasized in bold. To save space, we denote Cityscapes as "C" and KITTI as "K"

we initially try to apply UniP in every iteration. However, this straightforward solution could lead to severe performance degradation. This is because the quality of the localization ability cannot be assured in classification/segmentation approaches due to ***task specificity***. To address this problem, we choose to generate proposals for all target images at the initialization process of the adaptation stage using UniP, denoted as *pre-adaptation proposals*. In the training process of the adaptation stage, we replace every set of RPN-generated proposals with corresponding *pre-adaptation proposals*, ensuring reasonable and comparable results.

Overall details. Unless otherwise specified, our primary focus is on the two-stage model, Faster R-CNN (Ren et al., 2015), with a VGG16 (Simonyan & Zisserman, 2014) backbone, serving as the foundational detection model in all experiments following (Li et al., 2021; Yao et al., 2021; Wu et al., 2022; Li et al., 2022a). Following (Li et al., 2021; Ren et al., 2015; Saito et al., 2019), we resize the shorter side of all images to 600 pixels. The initial learning rate γ for all experiments is set to 0.0025, and the MSMT framework is trained with cosine learning rate decay. To maintain consistency with previous SFDAOD approaches (Li et al., 2021, 2022a; Vibashan et al., 2023), we maintain a fixed batch size of 1 in all experiments.

Without intricate adjustment of weighting hyperparameters, we assign a weight of 0.4 to the contrastive loss weight ω_4, while ω_1 to ω_3 are simply set to 1. We employ Smooth L1 Loss for all regression losses. RPN classification loss is addressed with Cross-Entropy Loss, and R-CNN classification loss is handled using Focal Loss (Lin et al., 2017b; Yao et al., 2021). The number of top-rated proposals p_t in UniP is consistently set to 300, and memory bank size l_{mb} in MBCL is fixed at 256 for all categories. The temperature parameter τ for contrastive loss in Eq. (12) is fixed at 0.1. The IoU threshold τ_{iou} and confidence threshold τ_{con} for MSPE in Algorithm 1 are set to 0.7 and 0.4, respectively. We use the stochastic gradient descent (SGD) optimizer. For each teacher–student pair, the EMA momentum coefficient η is set to 0.999, and the EMA frequency follows results outlined in Eq. (20) for a smooth and stable teacher update.

In the testing phase, for fair comparisons, we employ the same evaluation metric: mean average precision (mAP) with a 0.5 Intersection over Union (IoU) threshold. Following some previous methods utilizing MT, we utilize teacher models for testing, ensuring more reliable performance (Vibashan et al., 2023; Cao et al., 2023). All code implementations are

Table 2 Comparisons with state-of-the-art SFDAOD, MSDAOD, and MSFDAOD methods under **cross-time** setting measured by mAP (%)

Settings and methods				Results (%)										
Settings	Methods	Original task	Source domain	Bike	Bus	Car	Motor	Person	Rider	Light	Sign	Train	Truck	mAP
Source-only	Faster R-CNN (Ren et al., 2015)	Detection	D	35.10	51.70	52.60	9.90	31.90	17.80	21.60	36.30	0	47.10	**30.40**
			N	27.90	32.50	49.40	15.00	28.70	21.80	14.00	30.50	0	30.70	25.00
			D+N combined	31.50	46.90	52.90	8.40	29.50	21.60	21.70	34.30	0	42.20	28.90
SFDAOD	LODS (Li et al., 2022a)	Detection	D	41.53	53.17	58.37	25.00	38.35	24.53	28.93	41.75	0	49.07	36.07
			N	37.44	37.90	52.52	12.50	35.19	21.52	21.07	34.70	0	39.30	29.21
			D+N combined	29.23	42.81	55.96	9.65	33.25	18.91	27.44	36.52	0	38.89	29.27
	SOAP (Xiong et al., 2021)	Detection	D	35.38	55.45	53.05	27.45	40.00	23.56	34.61	44.59	0	53.48	**36.76**
			N	31.79	35.30	51.97	14.83	36.51	24.95	27.56	36.97	0	38.39	29.83
			D+N combined	33.36	48.06	51.07	24.26	35.39	29.18	24.74	35.97	0	46.04	32.81
	IRG (Vibashan et al., 2023)	Detection	D	31.84	47.98	65.97	13.41	39.14	22.51	44.15	48.39	0	42.09	35.55
			N	27.50	32.75	58.64	18.76	35.77	20.06	37.47	42.58	0	32.56	30.61
			D+N combined	34.79	40.86	64.86	11.11	39.24	25.76	46.12	47.71	0	39.04	34.95
MSDAOD	MDAN (Zhao et al., 2018)	Classification	D+N	37.10	29.90	52.80	15.80	35.10	21.60	24.70	38.80	0	20.10	27.60
	M3SDA (Peng et al., 2019)	Classification	D+N	36.90	25.90	51.90	15.10	35.70	20.50	24.70	38.10	0	15.90	26.50
	DMSN (Yao et al., 2021)	Detection	D+N	36.50	54.30	55.50	20.40	36.90	27.70	26.40	41.60	0	50.80	35.00
	TRKP (Wu et al., 2022)	Detection	D+N	-	-	-	-	-	-	-	-	-	-	**39.80**
MSFDAOD	DECISION (Ahmed et al., 2021)	Classification	D+N	39.47	48.80	53.09	27.49	39.54	23.87	29.43	39.49	0	45.62	34.68
	US-MSMA (Li et al., 2022c)	Segmentation	D+N	38.09	46.68	53.25	24.30	39.45	25.85	27.46	37.59	0	50.61	34.33
	DACA (Ours)	Detection	D+N	45.87	53.91	60.05	27.84	44.73	31.20	36.79	45.99	0	53.05	**39.94**
Oracle	Faster R-CNN (Ren et al., 2015)	Detection	Dawn/Dusk	27.20	39.60	51.90	12.70	29.00	15.20	20.00	33.10	0	37.50	**26.60**

The best method within each setting is emphasized in bold. To save space, we denote Bdd100k daytime as "D" and Bdd100k night as "N"

Table 3 Comparisons with state-of-the-art SFDAOD, MSDAOD, and MSFDAOD methods under **synthetic-to-real** setting measured by mAP (%)

Settings and methods				Results (%)
Settings	Methods	Original task	Source domain	mAP
Source-only	Faster R-CNN (Ren et al., 2015)	Detection	Sim10k	**39.07**
			SHIFT	32.06
			Sim10k+SHIFT combined	34.94
SFDAOD	LODS (Li et al., 2022a)	Detection	Sim10k	44.41
			SHIFT	36.90
			Sim10k+SHIFT combined	45.10
	SOAP (Xiong et al., 2021)	Detection	Sim10k	42.22
			SHIFT	32.50
			Sim10k+SHIFT combined	44.23
	IRG (Vibashan et al., 2023)	Detection	Sim10k	46.73
			SHIFT	38.83
			Sim10k+SHIFT combined	**47.14**
MSDAOD	MDAN (Zhao et al., 2018)	Classification	Sim10k+SHIFT	44.18
	M3SDA (Peng et al., 2019)	Classification	Sim10k+SHIFT	**45.38**
MSFDAOD	DECISION (Ahmed et al., 2021)	Classification	Sim10k+SHIFT	43.64
	US-MSMA (Li et al., 2022c)	Segmentation	Sim10k+SHIFT	41.69
	DACA (Ours)	Detection	Sim10k+SHIFT	**55.34**
Oracle	Faster R-CNN (Ren et al., 2015)	Detection	Cityscapes	**53.86**

The best method within each setting is emphasized in bold

built with PyTorch (Paszke et al., 2019). Each experiment is conducted on one NVIDIA GeForce RTX 3090 GPU.

5.2 Comparison with State-of-the-art

Tables 1, 2, 3 and 4 present the experimental results of the proposed DACA framework compared with (1) **Source-only**, (2) **SFDAOD**, (3) **MSDAOD**, and (4) **MSFDAOD**. All the datasets mentioned above are included in the experiments. These tables present four MSDAOD settings respectively, which are categorized as:

- *Cross-camera*. Following Yao et al. (2021), we designate Cityscapes and KITTI as source domains, with Bdd100k daytime as the target domain for cross-camera adaptation, focusing exclusively on the *car* category. Within the cross-camera setting, source and target images are captured across various city landscapes and filming angles, serving as the primary domain gap.
- *Cross-time*. Following Yao et al. (2021), we designate Bdd100k daytime and night as source domains, with Bdd100k dawn/dusk as the target domain. This setting encompasses all 10 categories, including bike (bicycle), car, bus, motor (motorcycle), person, rider, light (traffic light), sign (traffic sign), train, and truck. This cross-time setting mirrors a real-world scenario where images are collected at different times of the day, indicating different natural or artificial light conditions as the main domain gap.
- *Synthetic-to-real*. In the synthetic-to-real experimental setting, Sim10k and SHIFT daytime clear are designated as source domains, while Cityscapes is utilized as the target domain, concentrating solely on the *car* category. This setting explores the potential usage of synthetic images for adaptation to real-world images, particularly beneficial for real-world applications in which preprocessing and annotating real-world datasets are challenging and time-consuming. The primary domain gap lies in distinct appearances of real-world and synthetic objects under different environments, such as lighting conditions and texture.
- *Cross-weather*. In the cross-weather setting, SHIFT daytime images under clear, heavy cloudy, and heavy foggy weather conditions are taken as source domains, while SHIFT daytime images under heavy rainy weather conditions serve as the target domain. The primary domain gap in the cross-weather setting stems from the different appearances of objects under various adverse weather conditions.

From the experimental results, several observations can be made:

Table 4 Comparisons with state-of-the-art SFDAOD, MSDAOD, and MSFDAOD methods under **cross-weather** setting measured by mAP (%)

Settings and methods				Results (%)						
				Clear, cloudy, foggy to rainy						
Settings	Methods	Original task	Source domain	Car	Pedestrian	Truck	Bus	Motorcycle	Bicycle	mAP
Source-only	Faster R-CNN (Ren et al., 2015)	Detection	Single-best	51.78	46.00	48.27	42.64	44.92	38.96	**45.43**
			Single-worst	47.95	45.24	43.31	39.65	43.27	36.25	42.62
			Combined	49.75	42.58	50.01	43.77	43.84	39.16	44.85
SFDAOD	LODS (Li et al., 2022a)	Detection	Single-best	48.49	36.19	49.63	43.03	38.47	40.38	42.70
			Single-worst	47.60	32.45	43.00	50.23	40.57	37.54	41.90
			Combined	50.05	32.46	52.30	46.55	43.55	43.33	44.71
	SOAP (Xiong et al., 2021)	Detection	Single-best	46.33	39.82	49.95	42.37	39.34	40.80	43.10
			Single-worst	45.30	39.28	47.72	43.82	41.08	40.46	42.94
			Combined	46.86	38.80	45.58	42.83	39.92	39.53	42.25
	IRG (Vibashan et al., 2023)	Detection	Single-best	50.29	43.08	52.64	47.29	44.04	44.19	**46.92**
			Single-worst	49.42	45.13	43.10	39.32	42.74	36.15	42.64
			Combined	50.84	45.29	48.94	44.05	45.21	39.93	45.71
MSDAOD	MDAN (Zhao et al., 2018)	Classification	Multi-source	50.41	42.50	49.46	45.68	42.98	41.25	45.38
	M3SDA (Peng et al., 2019)	Classification	Multi-source	51.81	42.78	51.08	47.78	43.61	42.50	**46.59**
MSFDAOD	DECISION (Ahmed et al., 2021)	Classification	Multi-source	51.96	44.26	52.74	47.73	46.58	43.11	47.73
	US-MSMA (Li et al., 2022c)	Segmentation	Multi-source	50.81	46.15	52.54	46.32	46.12	42.08	47.34
	DACA (Ours)	Detection	Multi-source	51.01	48.16	54.13	48.87	46.59	43.99	**48.79**
Oracle	Faster R-CNN (Ren et al., 2015)	Detection	Rainy	50.36	43.22	51.59	46.60	46.74	41.63	**46.69**

The best method within each setting is emphasized in bold

(1) *Comparing source-only with others*: Source-only consistently exhibits the poorest performance when compared with other adaptation approaches. The presence of domain shift, both between the source and target domains and among source domains in the *Source-combined* settings, results in a significant performance drop when source-pretrained models are directly applied to the target domain without adaptation. This performance drop underscores the evident difference in joint probability distributions of image and label spaces between source-source and source-target domain pairs.

(2) *Comparing source-only with SFDAOD under single-source setting*: This comparison reveals that almost all methods with adaptation outperform Source-only methods. For instance, in Sim10k to Cityscapes adaptation in synthetic-to-real setting, LODS, SOAP, and IRG exhibit performance improvements of 5.37%, 3.15% and 7.66%, respectively. This demonstrates that adapting from a single source to a target domain significantly enhances detection performance. It is noteworthy that some single-source adaptation results are different from the recorded results in the original paper. For example, KITTI to Cityscapes adaptation records a 43.90% mAP in LODS (Li et al., 2022a), while our result is 34.11% mAP. The main reason is that we adopt the source-only baseline in DMSN (Yao et al., 2021) for all experiments for fair comparisons of adaptation performance. For example, KITTI to Cityscapes Source-only baseline in DMSN is 28.6% mAP, which is much lower than that in LODS (Li et al., 2022a) which is 39.2% mAP.

(3) *Comparing source-only with SFDAOD under Source-combined setting*: When comparing Source-only with SFDAOD results using a combined source domain, it becomes evident that *Source-combined* results are sometimes superior to single-source results. This is mainly because of the benefit of rich domain-invariant features from multiple source domains. However, in most cases, directly combining source domains leads to similar or inferior performance compared to single-source adaptation. This results from distribution shifts between source-source and source-target domain pairs. On the other hand, while *Source-combined* results may occasionally exhibit marginal improvements, this comes at the expense of additional pre-processing and training time. Both observations underscore the necessity of MSDA and MSFDA to effectively leverage information from multiple source domains.

(4) *Comparing MSDAOD with SFDAOD and MSFDAOD*: When source data is available, MSDAOD almost consistently outperforms SFDAOD and MSFDAOD results. For example, TRKP (Wu et al., 2022) reaches 58.40% mAP, which is 6.95% mAP higher than DACA and 10.67% mAP higher than SFDAOD. This superior performance is attributed to multi-source adaptation and abundant supervisory information, providing supervision for detection models and leveraging knowledge from multiple source domains.

(5) *Comparing MSFDA for classification or segmentation with MSDAOD and DACA*: For both DECISION and US-MSMA, detection performance is generally lower than that of MSDAOD and DACA. In comparison to DACA, the inferior performance of MSFDA to detection results is mainly associated with ***task specificity***, ***cross-model variability***, and ***object feature discriminability***, as mentioned earlier. The fixed scales for classification or segmentation tasks lead to a lack of localization training objectives when MSFDA is applied straightforwardly to detection tasks. This emphasizes the need to consider object-level discriminative features and localization ability in MSFDAOD tasks.

(6) *Comparing DACA with others*: Generally, DACA outperforms all MSFDAOD and SFDAOD methods and some results of MSDAOD methods. This demonstrates the effective utilization of multi-source detection knowledge of DACA, even with only source-pretrained models under source-free scenarios. DACA's superior performance results from robust mutual learning and knowledge distillation of MSMT and UniP, high-quality pseudo-labels from MSPE, and discriminative feature learning with MBCL. Notably, DACA also outperforms some MSDAOD and Oracle results even under source-free constraints. For example, DACA gains a 39.94% mAP in the cross-time adaptation scenario, outperforming DMSN (Yao et al., 2021) by 4.94% mAP and TRKP (Wu et al., 2022) by marginal 0.14% mAP. These detection results are all much higher than the 26.60% mAP of Oracle as recorded in DMSN. This emphasizes that Oracle results may not always be the upper bound of adaptation performance (Yao et al., 2021), since multiple source domain data or pretrained models may contain rich domain-invariant information, which is beneficial for adaptation. This comparison highlights that our DACA framework can leverage knowledge stored in multiple source-pretrained models, resulting in considerable performance. Additionally, this comparison suggests the potential for MSDAOD and MSFDAOD approaches to fully exploit the domain knowledge hidden in model parameters.

5.3 Ablation Study

In this section, we conduct ablation studies to illustrate the effectiveness of different components for detection and adaptation performance in DACA. Note that in the rest of this section, we only showcase results under the Synthetic-to-

Table 5 Ablation study on the effectiveness of MSMT framework and MBCL methodology

MSPE	MSMT	MBCL	mAP
✓			48.41
✓		✓	50.42
✓	✓		53.38
✓	✓	✓	**55.34**

The best result is emphasized in bold

Table 6 Ablation study on multi-source adaptation performance compared with single-source adaptation performance

Settings	mAP
Sim10k to Cityscapes	47.95
SHIFT to Cityscapes	48.28
Sim10k + SHIFT to Cityscapes	**55.34**

The best result is emphasized in bold

real setting unless otherwise specified. Details of this setting are described in Sect. 5.2.

To begin with, as the DACA framework comprises three main components: MSMT, MSPE, and MBCL, it is intuitive to consider keeping or eliminating these components to conduct a comprehensive comparison. However, it is not feasible to directly eliminate MSPE, as the entire framework relies on available pseudo-labels. Therefore, we first evaluate the effectiveness of MSMT and MBCL while keeping MSPE. For MSPE effectiveness, we evaluate it by comparing it with several alternative pseudo-labeling approaches, which will be discussed later.

MSMT and MBCL effectiveness. We demonstrate the effectiveness of MSMT and MBCL in Table 5. We start by showcasing how we "eliminate" these components. If MSMT is eliminated, no teacher models are used in the adaptation process, and correlated components such as consistency loss and teacher–student consensus in MBCL are also removed. If MBCL is eliminated, no memory bank or contrastive loss is constructed. Results in Table 5 show that without MSMT, the performance drops by 6.93% or 4.92% mAP, depending on whether MBCL is eliminated. Similarly, without MBCL, the performance will drop by 6.93% or 1.96% mAP, depending on whether MSMT is eliminated. These results highlight the beneficial roles of MSMT and MBCL in MSFDAOD performance by providing robust mutual learning and high-quality representation learning processes.

Multi-source adaptation effectiveness. Additionally, an experiment is conducted to explore the scenario when only a single source is provided to DACA. This experiment is also able to reveal the adaptation effectiveness of MSMT and MBCL components. Since only a single source-pretrained model is given, we can only initialize one teacher–student pair.

Table 7 Ablation study on the effectiveness of each loss, including detection loss $\mathscr{L}_{det}$, information maximization loss $\mathscr{L}_{im}$, consistency loss $\mathscr{L}_{cons}$ in MSMT, and contrastive loss $\mathscr{L}_{cont}$ in MBCL

$\mathscr{L}_{det}$	$\mathscr{L}_{IM}$	$\mathscr{L}_{cons}$	$\mathscr{L}_{cont}$	mAP
✓				52.35
✓	✓			53.51
✓		✓		53.82
✓			✓	54.39
✓	✓	✓		53.38
✓	✓		✓	54.45
✓		✓	✓	54.06
✓	✓	✓	✓	**55.34**

The best result is emphasized in bold

UniP and MSPE are also eliminated since we can directly take proposals and pseudo-labels generated by the single teacher as final proposals and pseudo-labels without any fusion or ensemble operations. Note that the pseudo-label generation process for a single model is simply filtering predictions by confidence threshold τ_{con}. Additionally, no source weights will be initialized or trained. The results in Table 6 show that multi-source full DACA can significantly improve single-source performance by about 6–7% mAP compared with single-source incomplete DACA, emphasizing DACA's capability of leveraging multi-source domain knowledge and exploring multi-source complementarity.

Loss fuction effectiveness. We also assess the effectiveness of each loss function as shown in Table 7. DACA only yields a 52.35% mAP when solely employing weighted detection loss. The inclusion of any one of the information maximization loss, consistency loss, or contrastive loss results in an overall mAP improvement, thereby substantiating the positive impact of each loss. Conversely, the elimination of any one of these three losses leads to a decline in mAP by 1.96%, 0.89%, and 1.28% mAP, relative to the full performance of 55.34% mAP, respectively. This observation underscores the contribution of each loss function within the DACA framework to proficiently execute the MSFDAOD task.

Pseudo-labeling method comparisons. We evaluate the effectiveness of MSPE in Table 8. This is to supplement the analysis in Table 5 where MSPE is retained. Here, we present several alternative pseudo-label generation methods to obtain a unified pseudo-label set: (1) **All combined**: This method involves collecting and combining all pseudo-labels generated by single teacher models. (2) **UniP+Weighted**: In this approach, a unified proposal set is obtained using UniP, and classification scores are obtained by weighting classification scores using $\{\alpha_i\}_{i=1}^{m}$. The bounding box coordinate predictions of a proposal are provided by the teacher model that generated it. (3) **WBF+Weighted.** WBF (Solovyev et al., 2021) is a novel method to combine bounding boxes pre-

Table 8 Results of different pseudo-labeling methods

Pseudo-labeling methods	mAP
All combined	49.22
UniP + weighted	54.49
WBF + weighted	53.51
Naïve MSPE	52.87
WBF + MSPE	54.99
UniP + MSPE	**55.34**

The final results of adapted models trained from each method are presented for comparison.
The best result is emphasized in bold

dicted by multiple object detection models, which has been used by some SFDAOD approaches (Liu et al., 2023a). However, in practice, we find that directly using WBF to fuse teacher prediction results as pseudo-labels in MSMT leads to poor performance. One possible reason is that bounding box predictions are unstable due to significant domain shift at the beginning of the adaptation process. Thus, we only implement WBF to obtain a unified proposal set, i.e., replacing UniP. The scores in WBF are class posteriors by R-CNN (Ren et al., 2015) head. After obtaining unified proposals with WBF, the weighting process is the same as in (2). (4) **Naïve MSPE**: This method directly uses MSPE to fuse bounding box predictions from multiple teachers, where single-teacher model predictions are generated by corresponding independent proposals instead of unified proposals by UniP. (5) **UniP+MSPE** and **WBF+MSPE**: These approaches involve replacing individual teacher proposal sets with the unified proposal set generated by UniP or WBF, and fusing bounding box predictions from multiple teachers using MSPE.

Results in Table 8 show that combining UniP+MSPE achieves the best performance. We can easily conclude that MSPE outperforms simple combination or weighted predictions due to MSPE's ability to boost consensus predictions and reasonably fuse localization results. The reasons why utilizing unified proposals for pseudo labeling helps are twofold: (1) Some source models may have inferior localization ability on the target domain due to domain shift, resulting in inaccurate bounding box coordinates from individual proposals. Utilizing a unified set of consensus proposals inhibits this inaccuracy. (2) Domain shift may cause overconfident incorrect predictions towards background or irrelevant objects. This kind of bias may be boosted in the MSPE process without unified proposal constraint, as MSPE always chooses bounding boxes with the highest class posterior to be fused. Thus, a set of unified proposals may eliminate this bias and improve performance.

Parameter sensitivity. To demonstrate the robustness of DACA under parameter fluctuations and further analyze the impact of each loss function on the overall design, we conduct parameter sensitivity experiments in the Synthetic-to-real setting. We simply adjust one of ω_1 to ω_4 in Eq. (16) while keeping the other three fixed. Experiment results are shown in Fig. 6. Generally, when these parameters are adjusted within acceptable ranges, detection performance mainly fluctuates between 55.5% and 54.5% mAP, only within a small range of 1% mAP. The results demonstrate that DACA is robust to slight parameter perturbations in real-world applications. We can also analyze the impact of each loss function individually.

(1) For the detection loss weight ω_1, assigning too small a weight will impair basic detection performance, while assigning too large a weight will lead to the ignorance of other loss functions for adaptation.
(2) Regarding the information maximization loss weight ω_2, as it is mainly responsible for domain weight optimization, loss weights below 0.5 lead to about a 1% mAP performance drop due to inferior prediction weighting ability. Huge weights cause the model to excessively concentrate on entropy functions, ignoring basic detection ability.
(3) Generally, both too large and too small consistency loss weights ω_3 lead to worse performance. Student models may deviate from teacher models after optimization if we assign too small a weight to the consistency loss, while forcing consistency too much can hurt the effectiveness of student model exploration.
(4) From a general perspective, neither too large nor too small contrastive loss weight ω_4 leads to good performance. This highlights the effectiveness of high-quality object-feature learning. However, as adjusting the contrastive loss weight leads to more frequent performance fluctuations, it is necessary to explore how to make memory bank and filter-based contrastive learning more robust in the future.

EMA frequency sensitivity. We also compare different EMA frequencies, deviating from the common practice of performing EMA in each iteration, as observed in previous works (Li et al., 2021, 2022a; Vibashan et al., 2023). We conduct this experiment in the Synthetic-to-real setting. It is noteworthy that AASFOD (Chu et al., 2023) also employs an empirical EMA frequency set to 2500 iterations. To assess the effectiveness of the EMA frequency calculated by Eq. (20), we conduct experiments encompassing a wide range of EMA frequencies. Specifically, a frequency of 0 iterations implies that the teacher model is never updated.

Results presented in Table 9 demonstrate that applying our dynamic EMA frequency in MSMT yields better results than performing EMA updates every iteration. This is because MSMT learns multiple source domain knowledge without any supervised information, leading to occasional instability in student updates in a single iteration. Adopting EMA

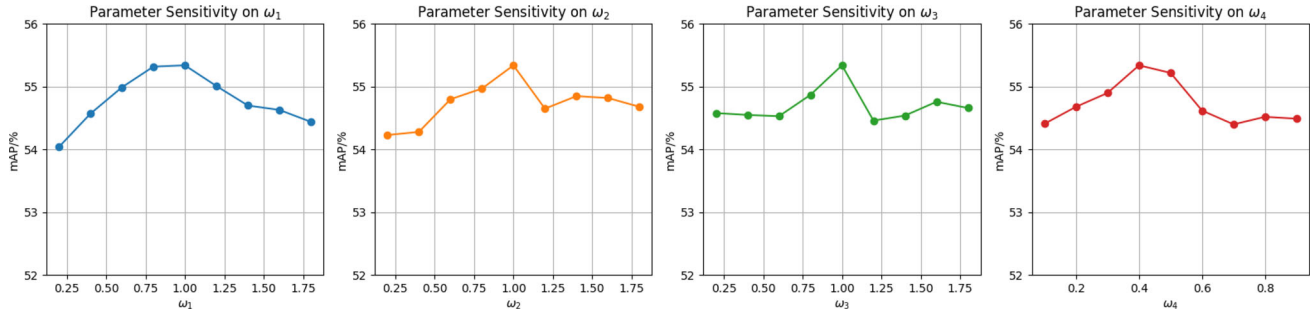

Fig. 6 Ablation study on parameter sensitivity of each loss function. Weighting parameters ω_1 to ω_4 correspond to the detection loss $\mathscr{L}_{det}$, information maximization loss $\mathscr{L}_{im}$, consistency loss $\mathscr{L}_{cons}$, and contrastive loss $\mathscr{L}_{cont}$ in Eq. (16), respectively. While we adjust one of the weighting parameters, we fix the other parameters. Original data points are highlighted. We compare the performance using mAP (%)

Table 9 Comparisons of various EMA updating frequencies in the synthetic-to-real setting

EMA update iterations	mAP
0	47.82
1	39.16
Dynamic = 5	**55.34**
10	53.68
20	50.09
50	45.53
100	43.61
500	40.76

In the synthetic-to-real setting, the dynamic EMA updating frequency, calculated by Eq. (20), is set to 5. An EMA update iteration of 0 implies no updating of teacher models. The best result is emphasized in bold

after a suitable number of iterations enhances the robustness of parameter updates. However, excessively slow updates are also harmful, as the MT framework necessitates a high enough optimization frequency to perform effective knowledge distillation.

5.4 Visualization Results

This section shows the visualization results to further demonstrate the effectiveness of DACA.

Detection visualization. Figure 7 showcases predictions of the same image generated by single-source models before adaptation, i.e., source-only, and by DACA. The top two lines represent predictions from two different source models, while the bottom line represents predictions from DACA. Obviously, model performance is improved after adaptation, providing evidence of the DACA model's effectiveness. Moreover, our model harnesses high-quality information from multiple sources. For instance, in column (b), DACA not only utilizes predictions of cars on the right side of the top image to complement missing predictions in the middle image but also eliminates duplicate predictions in the top image through acquired high-quality knowledge. Furthermore, DACA employs learned knowledge to explore and identify additional challenging objects, such as the car at the end of the road in the image in column (b). Generally, DACA is able to utilize multi-source knowledge proficiently and improve the generalization ability on the target domain.

Pseudo-label visualization. According to the analysis in previous sections, one of the main reasons for DACA's superior performance on MSFDAOD tasks is that MSMT and MSPE cooperate to generate and utilize reliable high-quality pseudo-labels. Thus, we visualize pseudo-labels generated by our method and several alternative methods described in Sect. 5.3 for comparison, including Naïve MSPE, WBF+MSPE, UniP+Weighted, and UniP+MSPE. Figure 8 shows the visualization results of these pseudo-labeling methods across two MSDAOD settings: cross-time and synthetic-to-real. Both dataset settings are described in Sect. 5.2 in detail.

To begin with, we compare column (c), which lacks MSPE, with (a), (b), and (d). It becomes apparent that MSPE uncovers and highlights additional objects by enhancing the consistency of prediction scores through logit summation. For instance, in the first line, a truck is situated on the right side of the image. Without MSPE, predictions around this object in (c) are suppressed due to passive-weighted predictions. Conversely, (a), (b), and (d) all precisely localize this object with MSPE, although (a) and (b) inaccurately classify it as a car category, which shares similar features with the truck category.

In addition, applying UniP+MSPE proves to enhance pseudo-label quality. This becomes evident when comparing UniP+MSPE with WBF+MSPE or Naïve MSPE. For example, UniP+MSPE gives the most high-quality pseudo-labels among all pseudo-labels for the image in the third line, accurately localizing the row of cars. These observations further substantiate the analyses presented in Sect. 5.3.

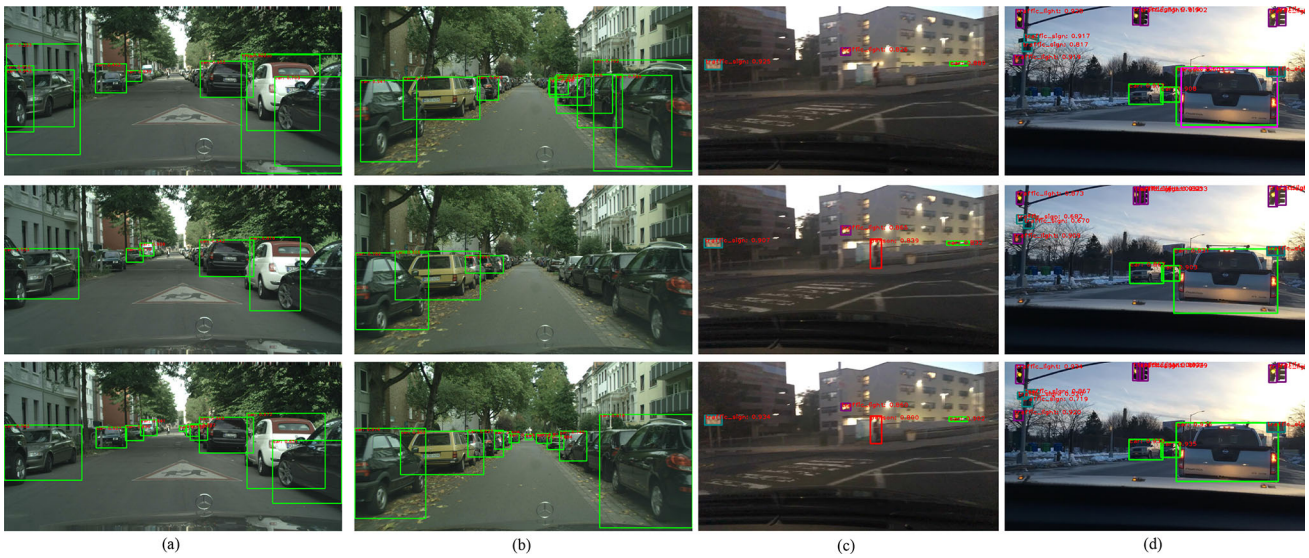

Fig. 7 Comparison of multi-source adapted predictions with source-only predictions. Predictions in columns **a**, **b**, **c** and **d** correspond to different target images. Images in columns **a** and **b** are from the synthetic-to-real setting, while the rest are from the cross-time setting. Predictions in the top and middle lines are generated by distinct source-only models, and predictions in the bottom line are generated by DACA. Specifically, the top line of **a** and **b** is from Sim10k to Cityscapes; the top line of **c** and **d** is from Bdd100k daytime to dawn/dusk; the middle line of **a** and **b** is from SHIFT daytime clear to Cityscapes; the middle line of **c** and **d** is from Bdd100k night to dawn/dusk. Generally, green boxes represent "car", cyan-blue boxes represent "traffic sign", deep purple boxes represent "traffic light", pink boxes represent "truck", and red boxes represent "person"

5.5 Model Integration

As model ensemble inevitably increases computational complexity and inference time, it is natural to ask, "***Is model integration a better choice?***". We argue that it's evident that multiple teacher models are necessary since we need to consider preserving source domain-specific knowledge, providing abundant domain-invariant knowledge, and the feasibility of initialization. However, it seems that the decision to adopt multiple students is still worth considering, given computational resources and concerns about complexity. Considering these factors, in this section, we conduct analyses and experiments to address the rationale for our decision to adopt multiple students from two perspectives: performance and time efficiency.

5.5.1 Performance Analysis

In this section, we begin by outlining our rationale for using multiple students from a performance perspective, followed by an empirical study to further support our reasoning.

- Simply aggregating all domain knowledge into a single model may lead to inferior performance due to domain conflicts. This can be empirically demonstrated in the following experiment.
- Efficiently initializing a single student model from multiple models remains a challenge. In the MSFDA scenario, where there are multiple source pretrained models, it is natural and convenient to create multiple student models initialized by their corresponding source models. However, initializing a single model is more complex. To the best of our knowledge, there still does not exist an effective method to directly integrate multiple sets of source model parameters into one set of model parameters without any training or fine-tuning. Simply initializing the single student with an ImageNet (Krizhevsky et al., 2012) pretrained model or a randomly chosen source pretrained model leads to inferior performance.

Let's also review several Multi-Source-Free Domain Adaptation for classification or segmentation approaches from previous studies, including DECISION (Ahmed et al., 2021), CAiDA (Dong et al., 2021), DINE (Liang et al., 2022), US-MSMA (Stage I) (Li et al., 2022c), DATE (Han et al., 2023), Surrogate (Shen et al., 2023), etc. Except for DINE (Liang et al., 2022), which is designed for black-box domain adaptation without access to even pretrained model parameters, we can observe that almost all these approaches adopt multiple students or multiple models, or at least initialize multiple students with pretrained models in part of their training procedure. Based on the rationales and investigations discussed above, it appears that implementing multiple

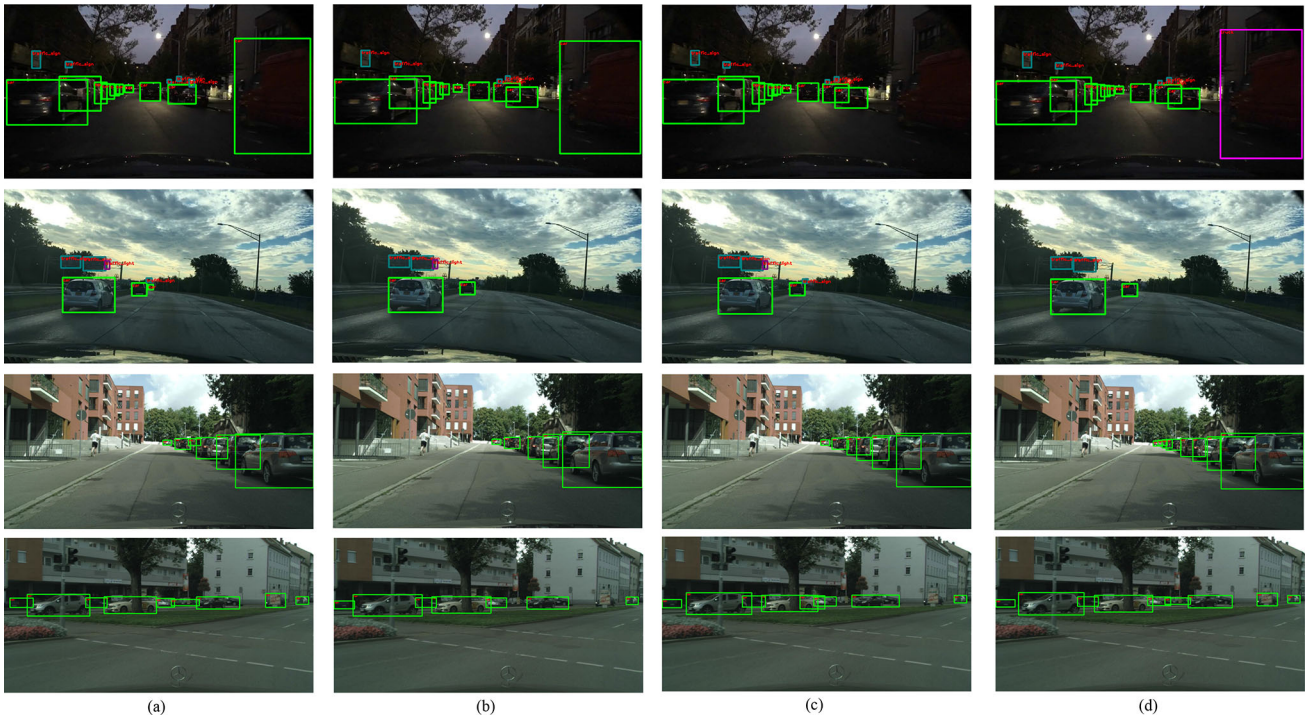

Fig. 8 Comparison of visualization results for pseudo-labeling methods. Predictions in columns **a**, **b**, **c** and **d** are generated by: **a** Naïve MSPE, **b** WBF+MSPE, **c** UniP+Weighted, and **d** UniP+MSPE, respectively. Predictions and images from the top two lines are from the cross-time setting, while the rest are from the synthetic-to-real setting. Similar to Table 7, green boxes represent "car", cyan-blue boxes represent "traffic sign", deep purple boxes represent "traffic light", and pink boxes represent "truck"

students is still the most applicable method for addressing MSFDAOD tasks.

Since our reason for adopting a multiple-student framework in DACA is to address the challenges of initializing a single model and the inferior performance of aggregating multiple models, we conduct an empirical study to compare the performance of multiple students versus a single student. For a comprehensive comparison, we conduct experiments with only a single model in DACA, as well as delve into previous MSFDA research and identify some possible methods for integrating multi-source models into the training procedure.

- Shared Student. This setting simply makes all student models in DACA share parameters. As we conduct the experiment in the Synthetic-to-real setting, we initialize student models with (1)Sim10k (Johnson-Roberson et al., 2017) pretrained, (2)SHIFT (Sun et al., 2022) daytime clear pretrained, and (3)ImageNet (Krizhevsky et al., 2012) pretrained, respectively in three separate experiments. Note that this setting is different from the setting in Table 6. We utilize all teacher models to generate pseudo-labels here. In contrast, in Table 6, we only initialize and optimize one teacher model to conduct single-source adaptation.
- Knowledge Distillation, which is proposed by DECISION (Ahmed et al., 2021) and KD (Hinton et al., 2015). Following DECISION, after original DACA training, we obtain weighted predictions by multiple teachers as pseudo labels to train a single model, distilling multiple source knowledge from multiple models into a single model.
- Stage II of US-MSMA (Li et al., 2022c). As US-MSMA is originally designed for semantic segmentation tasks, which have only pixel-level classification branches, we add regression branches processed similarly to classification branches. We use MSE (Mean Squared Error) loss in regression branches to reduce regression prediction discrepancy, similar to the original approach in US-MSMA Stage II, which used Kullback–Leibler divergence to reduce classification prediction discrepancy.
- Weighted EMA, which was proposed by DMSN (Yao et al., 2021). It involves replacing the updated single student parameters in EMA in Mean-Teacher with weighted updated multiple student parameters. Considering that "Pseudo-Subnet Learning" in DMSN is introduced to the training pipeline after 10 burn-in epochs, which means multiple source subnets have already performed 10-epoch adaptation to the target domain, we introduce the single model and weighted EMA after the original

training steps of DACA and add a few epochs to train the single model. In the training steps with weighted EMA, the training procedure of other components is the same as that of DACA, while the single model is updated by weighted EMA using continually updated student models and source weights $\{\alpha_i\}_{i=1}^{m}$.

These experiments are all conducted in the Synthetic-to-real setting. Results are shown in Table 10. Note that for KD, US-MSMA Stage II, and Weighted EMA we only report the best result among the three initialization methods. We can easily make some observations:

- Shared Student: Initializing shared student with a source pretrained model leads to about 4–5% less mAP in the Synthetic-to-real setting. This is mainly due to the bias towards the source domain from which the shared student is initialized. In this case, knowledge from other source domains in other source models is weakened, and the introduced domain shift between the chosen source domain and other source domains from teacher models will lead to worse performance. Initializing the shared student with an ImageNet pretrained model causes a severe performance drop since the ImageNet pretrained model has none of the source domain knowledge. Furthermore, the Shared Student is not able to optimize the source weights effectively.
- Knowledge Distillation: By providing weighted teacher predictions as pseudo labels, KD achieves 51.11% mAP, 4.23% mAP lower than Multiple Students. This performance drop is due to the loss of domain information in the distillation process. For example, this lost information may contain domain-specific knowledge, which is beneficial for detection performance.
- US-MSMA Stage II: This results in a significant performance drop. As US-MSMA is originally designed for semantic segmentation without the need for careful regression branch design, directly transferring it to detection tasks hurts localization performance.
- Weighted EMA: By updating the final model by weighted EMA of student models, the performance reaches 51.37% mAP, which is 3.97% mAP lower than Multiple Students. This performance drop also comes from domain conflict introduced by weighted parameters.

In general, simply implementing multiple students results in superior performance compared to the possible alternative model integration methods mentioned above. However, it is obvious that multiple students introduce more training and inference time than a single student. To better analyze the time efficiency between multiple students and a single student, we conduct a time efficiency analysis in Sect. 5.5.2.

5.5.2 Time Efficiency Analysis

Furthermore, we analyze the training and inference time of DACA and model integration methods.

Computational complexity. We start by analyzing the computational complexity of the training phase in DACA. To analyze the computational cost in training time, we first review the training process of DACA. Initially, m sets of image-level features are extracted by m backbones, where m represents the number of source domains. Subsequently, m RPNs generate m sets of region proposals respectively. To consolidate these proposals into unified ones, UniP is employed to process the m sets of region proposals. Following this, proposals and features are forwarded to the detection head. Pseudo-labels are generated using MSPE for the detection loss, while MBCL is utilized for the contrastive loss.

In UniP, we need to conduct NMS within m models respectively, then an additional NMS using mIoU scores. According to analyses in Soft-NMS (Bodla et al., 2017), the computational complexity for traditional greedy-NMS is $O(N_b^2)$, in which N_b is the number of detection boxes. After NMS in each Faster R-CNN, only the top-N proposals are retained, typically set to 256 or 300, which is significantly lower than N_b. Consequently, the computational complexity of NMS with mIoU scores can be simplified to $O(mN_b^2)$.

Similarly, MSPE follows a procedure akin to NMS, aiming to identify one box and all overlapping boxes, followed by the deletion of these boxes upon obtaining a single box. Given that the total Region of Interest (RoI) number is typically significantly lower than the proposal number N_b in UniP, we can neglect the computational complexity of MSPE compared to UniP.

For MBCL analysis, we begin with InfoNCE (Oord et al., 2018) analysis. In InfoNCE, the fundamental concept is to promote the model to enhance the similarity between positive samples while reducing the similarity between negative samples, thereby facilitating the improvement of representations. Typically, in a contrastive learning setup, the query size is usually equivalent to the positive key size (and they are totally equivalent in all our settings), denoted as N_p, while the negative key size is denoted as N_n. The InfoNCE calculation process can be segmented into three steps: (1) Calculating the cosine similarity between queries and positive keys; (2) Determining the cosine similarity between queries and negative keys; and (3) Summing these similarities. We can easily derive InfoNCE complexity as $O(N_p^2 + N_pN_n + N_p + N_n)$, which can be simplified to $O(N_p^2 + N_pN_n)$. In MBCL, assuming a worst-case scenario where every memory bank is full, we can substitute N_p with l_{mb} and N_n with kl_{mb}. Note that since we consider the background category, there are totally $k + 1$ category-wise memory banks, as elaborated in Sect. 4.3. Since contrastive loss is computed for each cate-

Table 10 Training time and performance comparisons between several model integration or single student methods

Integration setting	mAP (%)	Training time (%)
Shared student (Sim10k)	50.44	**86.12**
Shared student (SHIFT)	50.97	
Shared student (ImageNet)	28.18	
KD (Hinton et al., 2015; Ahmed et al., 2021)	51.11	125.34
US-MSMA (Li et al., 2022c) Stage II	38.91	167.33
Weighted EMA (Yao et al., 2021)	51.37	153.09
Multiple students	**55.34**	100.00

The best performance result and the fewest training time are emphasized in bold

gory in each student model, we derive the MBCL complexity as $O(m(k+1)(l_{mb}^2 + kl_{mb}^2))$, simplified to $O(mk^2 l_{mb}^2)$.

Then a set of unified proposals is fed to the RoI Align layer and then m detection heads. We denote the computational complexity of a Faster R-CNN without considering NMS is $O(P)$, where P is determined by factors such as the height and width of a feature map, etc. Then the computational cost of DACA is $O(mN_b^2 + mk^2 l_{mb}^2 + mP)$, while that of a single full Faster R-CNN is $O(N_b^2 + P)$ considering the original NMS in Faster R-CNN. Here, m, representing the number of source domains, is considered constant and independent of N_b^2, k^2, l_{mb}^2, and P. Therefore, both complexities, $O(mN_b^2 + mk^2 l_{mb}^2 + mP)$ and $O(N_b^2 + P)$, are of the same order of magnitude. This occurs because, as m is a constant independent of other factors, regardless of the increase in input size, these two terms of computational complexity both exhibit the same growth rate. Given that DACA's computational complexity is of the same order of magnitude as that of a single Faster R-CNN, DACA's computational complexity can be considered acceptable.

For the computational complexity during the inference phase in DACA, we can similarly derive it as $O(mN_b^2 + mP)$ since only UniP is utilized during inference. Consequently, the computational complexity of inference time for DACA and Faster R-CNN are also of the same order of magnitude.

In conclusion, based on the theoretical analysis of computational complexity, we infer that both the training and inference time of DACA are of the same order of magnitude compared to that of a single model Faster R-CNN. While multiple students inevitably introduce extra training and inference time, the increased time remains within an acceptable range. To further explore the training and inference time consumption in practical applications, we conduct several experiments to measure training and inference time, providing corresponding analyses.

Training and inference time. In fact, components such as UniP and MBCL contribute only a small fraction to the additional training time. To provide a more empirical understanding of computational costs, we measured the average training time of each method relative to DACA training time. For clarity, we denote the average training time of DACA as 100% and report the relative proportion of other methods. Results are presented in Table 10. From training time results in Table 10, we observe that Shared Student reduces training time by only 13.88% by eliminating UniP and performing MBCL within a single feature space. This finding underscores that UniP and MBCL contribute minimally to additional computational costs, with the primary cost stemming from the original Faster R-CNNs and DACA components in multiple teachers. Additionally, it is evident that additional model integration techniques, such as KD after the original DACA training steps, substantially increase training time.

Table 11 Inference time comparisons between single student and multiple students with 2 or 3 source domains

Settings/m	$m = 2$	$m = 3$
Single student	0.2169 ± 0.0034	0.1688 ± 0.0081
Multiple students	0.2955 ± 0.0059	0.3021 ± 0.0034

We choose models with similar performance and measured the average time for inferring one image. Results are presented with second(s) and m stands for the number of source domains

For a comprehensive analysis, we also measured the average inference time. To ensure a fair comparison, we selected inference models from the Single Student and Multiple Student settings with similar performance. The experiments were conducted in the Synthetic-to-real setting with 2 source domains ($m = 2$ in Table 11) and the Cross-weather setting with 3 source domains ($m = 3$ in Table 11). Results are presented in Table 11. From the results, we observe that Multiple Students do not introduce m times inference time in practice, despite having multiple students. At the same time, Multiple Students consistently achieve superior performance as presented in Table 10. Therefore, we continue to favor the Multiple Students approach for our model design. It is worth noting that the comparison between $m = 2$ and $m = 3$ was not conducted due to the significant performance gap between the two dataset settings.

Future work. Since we currently conduct training and inference sequentially, introducing parallel programming in the future could significantly reduce both training and infer-

ence times. Components other than UniP can all be easily parallelized since multiple teachers (or students) share the same forward and backward propagation procedures individually. Theoretically, reasonable parallelization could bring the training or inference time of Multiple Students to a similar level as that of Single Student, but this remains an area for future work. Furthermore, developing a more efficient model integration algorithm for multiple source models also remains a future task.

5.6 Extension to Other Detection Frameworks

Although DACA is mainly designed on Faster R-CNN, DACA is *framework-invariant*, i.e., it can be modified to successfully perform MSFDAOD based on different detection frameworks. Compared with some *framework-specific* DAOD approaches like GPA (Xu et al., 2020) and RPA (Zhang et al., 2021b) which rely on RPN network in Faster R-CNN, our framework is not limited to object detector types. We demonstrate the effectiveness of DACA across different detection frameworks by implementing DACA on a well-known anchor-free one-stage object detector FCOS (Tian et al., 2019). Without using any predefined anchor boxes, FCOS extracts pixel-level features on feature maps of Fully Convolutional Network (FCN) (Long et al., 2015). FCOS gives multi-level prediction on multi-level feature maps and inhibits far-away bounding boxes with center-ness loss. We conclude the main difference between FCOS and Faster R-CNN affecting our experiment as follows: (1) Faster R-CNN generates anchors and Regions of Interest (RoI) and then gives bounding box prediction. FCOS directly predicts bounding boxes on feature maps. (2) Faster R-CNN extracts object-level features with RoI-Alignment (He et al., 2017) on the last feature map of the backbone network, while FCOS extracts multi-level feature maps with Feature Pyramid Network (Lin et al., 2017a). (3) Faster R-CNN performs object-level prediction based on predefined anchors, leading to ***cross-model variability***. FCOS performs bounding box prediction directly on pixels from multi-level feature maps, avoiding ***cross-model variability*** in region proposals.

Due to the difference mentioned above, we made several modifications to implement DACA with FCOS: (1) With pixel-level predictions on multi-level feature maps, ***cross-model variability*** in region proposals does not exist. Thus, we do not need to implement UniP in MSMT. This means we can directly assign weights to pixel-level classification predictions. (2) For center-ness losses, considering that it is calculated based on regression targets and is designed to inhibit far-away boxes from single predictions, we calculate the center-ness loss for each student independently, without assigning any weights. (3) Since FCOS only extracts feature maps with FPN, we cannot extract object-level features directly. Instead of utilizing RoI-Alignment on multi-level feature maps following SoCo (Wei et al., 2021) which still relies on Faster R-CNN components, we directly perform pixel-level contrastive learning for modified MBCL.

Table 12 Results of DACA extension to FCOS using Resnet-101 (He et al., 2016) as backbone

Method	Setting	mAP
Source-only	Sim-only	41.80
	SHIFT-only	**44.87**
EPM (Hsu et al., 2020)	Sim to Cityscapes	51.20
SSAL (Munir et al., 2021)		51.80
MGA-DA (Zhou et al., 2022)		**54.10**
DACA	Sim + SHIFT	**51.98**
Oracle	Cityscapes	**70.40**

The Oracle result is cited from MGA-DA (Zhou et al., 2022). We use Sim to represent Sim10k dataset. The best result within each setting is emphasized in bold

Results in Table 12 show that after adaptation, the final performance on the target domain is improved by 10.18% and 7.11% mAP as compared with Sim10k-only result and SHIFT-only result, respectively. This proves that DACA is able to leverage multi-source information and explore multi-source complementarity for target performance improvement even under other detection frameworks. We also compared DACA with several FCOS-based single-source source-available DAOD approaches including EPM (Hsu et al., 2020), SSAL (Munir et al., 2021), and MGA-DA (Zhou et al., 2022). Although DACA cannot access source data, it still outperforms SSAL by 0.18% mAP by leveraging multiple domain knowledge, which further proves DACA's strong multi-source-free adaptation ability. This result proves that DACA is *framework-invariant* and DACA can be utilized as a universal solution for MSFDAOD tasks.

The most important key for transferring DACA to another framework is ensuring that multiple source detection heads i.e., classification, and regression layers perform prediction on the same set of targets, such as a unified proposal for Faster R-CNN and pixel-level prediction for FCOS. DACA can also be extended to more modern detection frameworks with a few modifications. Take Deformable DETR (Zhu et al., 2021) and YOLO (Redmon et al., 2016) as examples. For Deformable DETR, we can apply UniP in a two-stage Deformable DETR to obtain a unified set of proposals as queries, and then feed these unified queries to multiple detection heads. As for the YOLO series, we can simply weight grid-level predictions for each image. To implement MBCL, we can utilize RoI Align or other techniques for object-level features.

In general, extending DACA to Deformable DETR is similar to the process for Faster R-CNN, while extending it to YOLO is similar to the approach for FCOS. Although

we have already demonstrated the effectiveness of DACA on Faster R-CNN and FCOS, extending DACA to YOLOs, DETRs, or other possible frameworks remains a task for future work due to limited computational resources and time.

6 Conclusion

In this paper, we introduced a novel research focus within the Domain Adaptive Object Detection (DAOD) area, namely Multi-Source-Free Domain Adaptive Object Detection (MSFDAOD). To tackle the associated challenges, we proposed a novel framework termed Divide-and-Aggregate Contrastive Adaption (DACA). Following the pretraining phase for each source domain, we initialized each corresponding teacher–student pair in MSMT using the relevant pretrained model. Subsequently, within the MSMT framework, we employed strongly augmented images for students and weakly augmented images for teachers during the feed-forward process. To aggregate predictions from multiple source models, we designed the UniP mechanism to aggregate and eliminate redundant proposals. We formulated weighted detection loss and information maximization loss based on unified predictions utilizing UniP proposals. The bounding boxes and categories, generated by MSPE, served as pseudo-labels. To further learn target discriminative features, we leveraged MBCL to acquire consensus high-quality representations. Experimental results underscored the superior performance of DACA, achieving mAP values of 54.45%, 39.94%, 55.34% and 48.79% in cross-camera, cross-time, synthetic-to-real, and cross-weather MSDAOD tasks, respectively, outperforming both SFDAOD and MSFDA to OD approaches. For a comprehensive demonstration, we conducted ablation, visualization studies, and necessary analysis to demonstrate the effectiveness of DACA components. Additionally, we extended DACA to other detection frameworks, which revealed DACA's scalability.

Limitations. Nevertheless, our proposed method still has limitations. Although multiple students achieve superior performance, a single model is obviously less time-consuming and more beneficial for real-time detection and deployment. The problem of how to effectively aggregate multiple models and improve computational efficiency in MSFDAOD remains as future work. Moreover, there is a limitation in MSPE where we simply neglect the target joint distribution. Considering that the target joint distribution may enhance pseudo-label quality, we aim to explore methods for approximating this distribution without supervisory signals in the future. Furthermore, we primarily optimize the classification branch through source weighting, consistency regularization, and contrastive learning. For the localization branch, we only compute a regression loss in the detection loss. Exploring how to effectively optimize the regression branch, obtain regression weights, and learn localization features, among other aspects, are valuable topics for future work.

Future work. To broaden the scope of our research, we express interest in extending MSFDAOD approaches to multi-target adaptation. Moreover, we also plan to investigate domain generalization, aligning with real-world scenarios and contributing to the future exploration of more generalized domain adaptation and object detection approaches.

Societal impact. This work is applicable for adapting object detection networks from multiple source domains to a target domain, even when source data and labels are not accessible during the adaptation stage. The proposed method is capable of reducing the burden of collecting and organizing large-scale supervised data in open-world scenarios. Our work is extremely helpful when training data and annotations are not available due to privacy preservation policies and data transmission constraints in real-world scenarios. By evaluating the superior performance in the four dataset settings, we can also demonstrate that our method is applicable to various kinds of domain shifts in real-world urban environments, facilitating further development of real-world urban detection systems. Although we have achieved state-of-the-art performance, negative impacts still exist. Since we have not yet developed a proper way to integrate multiple student models, our proposed method adopts a multi-student framework. Therefore, the proposed method should not be used in some systems with extremely high detection latency requirements.

Acknowledgements This work is supported by CCF-DiDi GAIA Collaborative Research Funds for Young Scholars and the National Natural Science Foundation of China (Nos. 61925107, 62021002).

References

Ahmed, S. M., Raychaudhuri, D. S., Paul, S., Oymak, S., & Roy-Chowdhury, A. K. (2021). Unsupervised multi-source domain adaptation without access to source data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10103–10112).

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv:1606.06565

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv:2004.10934

Bodla, N., Singh, B., Chellappa, R., & Davis, L. S. (2017). Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision* (pp. 5561–5569).

Cai, Q., Pan, Y., Ngo, C. W., Tian, X., Duan, L., Yao, T. (2019). Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11457–11466).

Cao, S., Joshi, D., Gui, L. Y., & Wang, Y. X. (2023). Contrastive mean teacher for domain adaptive object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 23839–23848).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Proceedings of the European conference on computer vision* (pp. 213–229).

Chen, C., Zheng, Z., Ding, X., Huang, Y., & Dou, Q. (2020a). Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8869–8878).

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(4), 834–848.

Chen, S., Sun, P., Song, Y., & Luo, P. (2023). Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 19830–19843).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020b). A simple framework for contrastive learning of visual representations. In *Proceedings of the international conference on machine learning* (pp. 1597–1607).

Chen, X., Wang, S., Long, M., & Wang, J. (2019). Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *Proceedings of the international conference on machine learning* (pp. 1081–1090).

Chen, Y., Li, W., Sakaridis, C., Dai, D., & Van Gool, L. (2018). Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3339–3348).

Chen, Y. T., Shi, J., Ye, Z., Mertz, C., Ramanan, D., & Kong, S. (2022). Multimodal object detection via probabilistic ensembling. In *Proceedings of the European conference on computer vision* (pp. 139–158).

Chu, Q., Li, S., Chen, G., Li, K., & Li, X. (2023). Adversarial alignment for source free object detection. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 452–460).

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3213–3223).

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27.

Deng, J., Li, W., Chen, Y., & Duan, L. (2021). Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4091–4101).

Deng, J., Xu, D., Li, W., & Duan, L. (2023). Harmonious teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 23829–23838).

Denton, E. L., Zaremba, W., Bruna, J., LeCun, Y., & Fergus, R. (2014). Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in neural information processing systems* (pp. 1269–1277).

Dong, J., Fang, Z., Liu, A., Sun, G., & Liu, T. (2021). Confident anchor-induced multi-source free domain adaptation. In *Advances in neural information processing systems* (pp. 2848–2860).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., & Uszkoreit, J. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the international conference on learning representations*.

He, Z., Zhang, L. (2020). Domain adaptive object detection via asymmetric tri-way faster-rcnn. Proceedings of the European (pp. 309–324). Conference on Computer Vision

Fang, Y., Yap, P. T., Lin, W., Zhu, H., & Liu, M. (2022). Source-free unsupervised domain adaptation: A survey. arXiv:2301.00265

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *Proceedings of the international conference on machine learning* (pp. 1180–1189).

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The kitti vision benchmark suite. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3354–3361).

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1440–1448).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 580–587).

Gori, M., Monfardini, G., & Scarselli, F. (2005). A new model for learning in graph domains. In *Proceedings of the IEEE international joint conference on neural networks* (pp. 729–734).

Han, Z., Zhang, Z., Wang, F., He, R., Su, W., Xi, X., & Yin, Y. (2023). Discriminability and transferability estimation: A Bayesian source importance estimation approach for multi-source-free domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 7811–7820)

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(9), 1904–1916.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 770–778).

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2961–2969).

He, Z., Zhang, L., Gao, X., & Zhang, D. (2023). Multi-adversarial faster-RCNN with paradigm teacher for unrestricted object detection. *International Journal of Computer Vision, 131*(3), 680–700.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv:1503.02531

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in neural information processing systems* (pp. 6840–6851).

Hoffman, J., Kulis, B., Darrell, T., & Saenko, K. (2012). Discovering latent domains for multisource domain adaptation. In *Proceedings of the European conference on computer vision* (pp. 702–715).

Hsu, C. C., Tsai, Y. H., Lin, Y. Y., & Yang, M. H. (2020). Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *Proceedings of the European conference on computer vision* (pp. 733–748).

Hu, W., Miyato, T., Tokui, S., Matsumoto, E., & Sugiyama, M. (2017). Learning discrete representations via information maximizing self-augmented training. In *Proceedings of the international conference on machine learning* (pp. 1558–1567).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4700–4708).

Huang, J., Guan, D., Xiao, A., & Lu, S. (2021). Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. In *Advances in neural information processing systems* (pp. 3635–3649).

Inoue, N., Furuta, R., Yamasaki, T., & Aizawa, K. (2018). Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5001–5009).

Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S. N., Rosaen, K., & Vasudevan, R. (2017). Driving in the matrix: Can virtual

worlds replace human-generated annotations for real world tasks? In *Proceedings of the IEEE international conference on robotics and automation* (pp. 746–753).

Kang, G., Jiang, L., Yang, Y., & Hauptmann, A. G. (2019). Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4893–4902).

Kennerley, M., Wang, J. G., Veeravalli, B., & Tan, R. T. (2023). 2pcnet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11484–11493).

Kim, Y., Cho, D., Han, K., Panda, P., & Hong, S. (2021). Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence, 2*(6), 508–518.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv:1609.02907

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* Vol. 25.

Kundu, J. N., Kulkarni, A. R., Bhambri, S., Mehta, D., Kulkarni, S. A., Jampani, V., & Radhakrishnan, V. B. (2022). Balancing discriminability and transferability for source-free domain adaptation. In *Proceedings of the international conference on machine learning* (pp. 11710–11728).

Lang, Q., Zhang, L., Shi, W., Chen, W., & Pu, S. (2022). Exploring implicit domain-invariant features for domain adaptive object detection. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(4), 1816–1826.

Li, J., Xu, R., Ma, J., Zou, Q., Ma, J., & Yu, H. (2023). Domain adaptive object detection for autonomous driving under foggy weather. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 612–622).

Li, S., Ye, M., Zhu, X., Zhou, L., & Xiong, L. (2022a). Source-free object detection by learning to overlook domain style. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8014–8023).

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine, 37*(3), 50–60.

Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., & Yang, J. (2020b). Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Advances in neural information processing systems* (pp. 21002–21012).

Li, X., Chen, W., Xie, D., Yang, S., Yuan, P., Pu, S., & Zhuang, Y. (2021). A free lunch for unsupervised domain adaptive object detection without source data. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8474–8481).

Li, Y., Wang, N., Shi, J., Liu, J., & Hou, X. (2017). Revisiting batch normalization for practical domain adaptation. In *Proceedings of the international conference on learning representations workshops*.

Li, Y. J., Dai, X., Ma, C. Y., Liu, Y. C., Chen, K., Wu, B., He, Z., Kitani, K., & Vajda, P. (2022b). Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7581–7590).

Li, Z., Togo, R., Ogawa, T., & Haseyama, M. (2022c). Union-set multi-source model adaptation for semantic segmentation. In *Proceedings of the European conference on computer vision* (pp. 579–595).

Liang, J., Hu, D., & Feng, J. (2020). Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the international conference on machine learning* (pp. 6028–6039).

Liang, J., Hu, D., Feng, J., & He, R. (2022). Dine: Domain adaptation from single and multiple black-box predictors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8003–8013).

Lin, C., Zhao, S., Meng, L., & Chua, T. S. (2020). Multi-source domain adaptation for visual sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2661–2668).

Lin, C., Yuan, Z., Zhao, S., Sun, P., Wang, C., & Cai, J. (2021). Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8771–8780).

Lin, C., Sun, P., Jiang, Y., Luo, P., Qu, L., Haffari, G., Yuan, Z., Cai, J. (2023). Learning object-language alignments for open-vocabulary object detection. In *Proceedings of the international conference on learning representations*.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017a). Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2117–2125).

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017b). Focal loss for dense object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2980–2988).

Liu, M. Y., & Tuzel, O. (2016). Coupled generative adversarial networks. In *Advances in neural information processing systems* (pp. 469–477).

Liu, Q., Lin, L., Shen, Z., & Yang, Z. (2023a). Periodically exchange teacher–student for source-free object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6414–6424).

Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., & Zhang, L. (2022). Dab-detr: Dynamic anchor boxes are better queries for detr. In *Proceedings of the international conference on learning representations*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Proceedings of the European conference on computer vision* (pp. 21–37).

Liu, X., Xi, W., Li, W., Xu, D., Bai, G., & Zhao, J. (2023). Co-MDA: Federated multi-source domain adaptation on black-box models. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(12), 7658–7670.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3431–3440).

Long, M., Zhu, H., Wang, J., & Jordan, M. I. (2017). Deep transfer learning with joint adaptation networks. In *Proceedings of the international conference on machine learning* (pp. 2208–2217).

Lu, P. J., Jui, C. Y., & Chuang, J. H. (2023). A privacy-preserving approach for multi-source domain adaptive object detection. In *Proceedings of the IEEE international conference on image processing* (pp. 1075–1079).

Mansour, Y., Mohri, M., & Rostamizadeh, A. (2008). Domain adaptation with multiple sources. In *Advances in neural information processing systems* (pp. 1041–1048).

Munir, M. A., Khan, M. H., Sarfraz, M., & Ali, M. (2021). Ssal: Synergizing between self-training and adversarial learning for domain adaptive object detection. In *Advances in neural information processing systems* (pp. 22770–22782).

Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv:1807.03748

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., & Desmaison, A. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems* (pp. 8024–8035).

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., & Wang, B. (2019). Moment matching for multi-source domain adaptation. In *Pro-*

ceedings of the IEEE/CVF international conference on computer vision (pp. 1406–1415).

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 779–788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).

Riemer, M., Cases, I., Ajemian, R., Liu, M., Rish, I., Tu, Y., & Tesauro, G. (2019). Learning to learn without forgetting by maximizing transfer and minimizing interference. In *Proceedings of the international conference on learning representations*.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 400–407.

Saito, K., Ushiku, Y., Harada, T., & Saenko, K. (2019). Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6956–6965).

Shen, M., Bu, Y., & Wornell, G. W. (2023). On balancing bias and variance in unsupervised multi-source-free domain adaptation. In *Proceedings of the international conference on machine learning* (pp. 30976–30991).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the international conference on learning representations*.

Sindagi, V. A., Oza, P., Yasarla, R., & Patel, V. M. (2020). Prior-based domain adaptive object detection for hazy and rainy conditions. In *Proceedings of the European conference on computer vision* (pp. 763–780).

Solovyev, R., Wang, W., & Gabruseva, T. (2021). Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing, 107*, 104117.

Sun, B., & Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation. In *Proceedings of the European conference on computer vision* (pp. 443–450).

Sun, B., Feng, J., & Saenko, K. (2016). Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2058–2065).

Sun, S., Shi, H., & Wu, Y. (2015). A survey of multi-source domain adaptation. *Information Fusion, 24*, 84–92.

Sun, T., Segu, M., Postels, J., Wang, Y., Van Gool, L., Schiele, B., Tombari, F., & Yu, F. (2022). Shift: A synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 21371–21382).

Szegedy, C., Ioffe, S., Vanhoucke, V.,& Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 4278–4284).

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems* (pp. 1195–1204).

Tian, Z., Shen, C., Chen, H.,& He, T. (2019). Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9627–9636).

Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7167–7176).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser Ł, Polosukhin, I. (2017) Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Vibashan, V., Oza, P., & Patel, V. M. (2023). Instance relation graph guided source-free domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3520–3530).

Wang, K., & Zhang, L. (2021). Reconcile prediction consistency for balanced object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3631–3640).

Wei, F., Gao, Y., Wu, Z., Hu, H., & Lin, S. (2021). Aligning pretraining for detection via object-level contrastive learning. In *Advances in neural information processing systems* (pp. 22682–22694).

Wilson, G., & Cook, D. J .(2020). A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology*, *11*(5), 51:1–51:46

Wu, J., Chen, J., He, M., Wang, Y., Li, B., Ma, B., Gan, W., Wu, W., Wang, Y., & Huang, D. (2022). Target-relevant knowledge preservation for multi-source domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5301–5310).

Xiong, L., Ye, M., Zhang, D., Gan, Y., Li, X., & Zhu, Y. (2021). Source data-free domain adaptation of object detector through domain-specific perturbation. *International Journal of Intelligent Systems, 36*(8), 3746–3766.

Xu, M., Wang, H., Ni, B., Tian, Q., & Zhang, W. (2020). Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12355–12364).

Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., & Liu, Z. (2021). End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3060–3069).

Xu, M., Qin, L., Chen, W., Pu, S., & Zhang, L. (2023). Multi-view adversarial discriminator: Mine the non-causal factors for object detection in unseen domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8103–8112).

Yang, C., Liu, Y., & Yuan, Y. (2023). Transferability-guided multi-source model adaptation for medical image segmentation. In *Proceedings of the international conference on medical image computing and computer-assisted intervention* (pp. 703–712).

Yang, S., Wang, Y., Van De Weijer, J., Herranz, L., & Jui, S. (2020). Unsupervised domain adaptation without source data by casting a bait. *1*(2), 5. arXiv:2010.12427

Yang, S., Wang, Y., Van De Weijer, J., Herranz, L., & Jui, S. (2021a). Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8978–8987).

Yang, S., van de Weijer, J., Herranz, L., & Jui, S. (2021). Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in Neural information processing systems, 34*, 29393–29405.

Yao, X., Zhao, S., Xu, P., & Yang, J. (2021). Multi-source domain adaptation for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp 3273–3282).

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan ,V., & Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2633–2642).

Yu, W., & Qin, Z. (2020). Graph convolutional network for recommendation with low-pass collaborative filters. In *Proceedings of the international conference on machine learning* (pp. 10936–10945).

Yu, Z., Li, J., Du, Z., Zhu, L., & Shen, H. T. (2023). A comprehensive survey on source-free domain adaptation. arXiv:2302.11803

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. *Knowledge-Based Systems, 216*, 106775.

Zhang, D., Ye, M., Liu, Y., Xiong, L., & Zhou, L. (2022). Multi-source unsupervised domain adaptation for object detection. *Information Fusion, 78*, 138–148.

Zhang, L., Qin, L., Xu, M., Chen, W., Pu, S., & Zhang, W. (2023). Randomized spectrum transformations for adapting object detector in unseen domains. *IEEE Transactions on Image Processing, 32*, 4868–4879.

Zhang, S., Zhang, L., & Liu, Z. (2023b). Refined pseudo labeling for source-free domain adaptive object detection. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing* (pp. 1–5).

Zhang, Y., Wang, Z., Mao, Y. (2021b). Rpn prototype alignment for domain adaptive object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12425–12434).

Zhao, H., Zhang, S., Wu, G., Moura, J. M., Costeira, J. P., & Gordon, G. J. (2018). Adversarial multiple source domain adaptation. In *Advances in neural information processing systems* (pp. 8568–8579).

Zhao, S., Li, B., Yue, X., Gu, Y., Xu, P., Hu, R., Chai, H., & Keutzer, K. (2019a). Multi-source domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems* (pp. 7285–7298).

Zhao, S., Lin, C., Xu, P., Zhao, S., Guo, Y., Krishna, R., Ding, G., & Keutzer, K. (2019b). Cycleemotiongan: Emotional semantic consistency preserved cyclegan for adapting image emotions. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2620–2627).

Zhao, S., Wang, G., Zhang, S., Gu, Y., Li, Y., Song, Z., Xu, P., Hu, R., Chai, H., & Keutzer, K. (2020). Multi-source distilling domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 12975–12983).

Zhao, S., Chen, X., Yue, X., Lin, C., Xu, P., Krishna, R., Yang, J., Ding, G., Sangiovanni-Vincentelli, A. L., & Keutzer, K. (2021). Emotional semantics-preserved and feature-aligned cyclegan for visual emotion adaptation. *IEEE Transactions on Cybernetics, 52*(10), 10000–10013.

Zhao, S., Li, B., Xu, P., Yue, X., Ding, G., & Keutzer, K. (2021). Madan: Multi-source adversarial domain aggregation network for domain adaptation. *International Journal of Computer Vision, 129*(8), 2399–2424.

Zhao, S., Xiao, Y., Guo, J., Yue, X., Yang, J., Krishna, R., Xu, P., & Keutzer, K. (2021c). Curriculum cyclegan for textual sentiment domain adaptation with multiple sources. In *Proceedings of the the web conference* (pp. 541–552).

Zhao, S., Yue, X., Zhang, S., Li, B., Zhao, H., Wu, B., Krishna, R., Gonzalez, J. E., Sangiovanni-Vincentelli, A. L., Seshia, S. A., et al. (2022). A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems, 33*(2), 473–493.

Zhao, S., Hong, X., Yang, J., Zhao, Y., & Ding, G. (2023). Toward label-efficient emotion and sentiment analysis. *Proceedings of the IEEE, 111*(10), 1159–1197.

Zhao, S., Chen, H., Huang, H., Xu, P., & Ding, G. (2024). More is better: Deep domain adaptation with multiple sources. In *Proceedings of the international joint conference on artificial intelligence*.

Zhou, W., Du, D., Zhang, L., Luo, T., & Wu, Y. (2022). Multi-granularity alignment domain adaptation for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9581–9590).

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2223–2232).

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2021). Deformable detr: Deformable transformers for end-to-end object detection. In: *Proceedings of the international conference on learning representations*.

Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE, 111*(3), 257–276.