Open-Scenario Domain Adaptive Object Detection in Autonomous Driving

Zeyu Ma
University of Electronic Science and Technology of China
cnzeyuma@hotmail.com

Ziqiang Zheng
University of Electronic Science and Technology of China
zhengziqiang1@gmail.com

Jiwei Wei
University of Electronic Science and Technology of China
mathematic6@gmail.com

Xiaoyong Wei
Sichuan University,
Peng Cheng Laboratory
cswei@scu.edu.cn

Yang Yang*
University of Electronic Science and Technology of China
yang.yang@uestc.edu.cn

Heng Tao Shen
University of Electronic Science and Technology of China
shenhengtao@hotmail.com

ABSTRACT

Existing domain adaptive object detection algorithms (DAOD) have demonstrated their effectiveness in discriminating and localizing objects across scenarios. However, these algorithms typically assume a single source and target domain for adaptation, which is not representative of the more complex data distributions in practice. To address this issue, we propose a novel **O**pen-**S**cenario **D**omain **A**daptive Object Detection (**OSDA**), which leverages multiple source and target domains for more practical and effective domain adaptation. We are the first to increase the granularity of the background category by building the foundation model using contrastive vision-language pre-training in an open-scenario setting for better distinguishing foreground and background, which is under-explored in previous studies. The performance gains by introducing the pre-training have been observed and have validated the model's ability to detect objects across domains. To further fine-tune the model for domain-specific object detection, we propose a hierarchical feature alignment strategy to obtain a better common feature space among the various source and target domains. In the case of multi-source domains, the cross-reconstruction framework is introduced for learning more domain invariances. The proposed method is able to alleviate knowledge forgetting without any additional computational costs. Extensive experiments across different scenarios demonstrate the effectiveness of the proposed model.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision problems**.

KEYWORDS

object detection, domain adaptation, autonomous driving

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611854

ACM Reference Format:
Zeyu Ma, Ziqiang Zheng, Jiwei Wei, Xiaoyong Wei, Yang Yang, and Heng Tao Shen. 2023. Open-Scenario Domain Adaptive Object Detection in Autonomous Driving. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611854

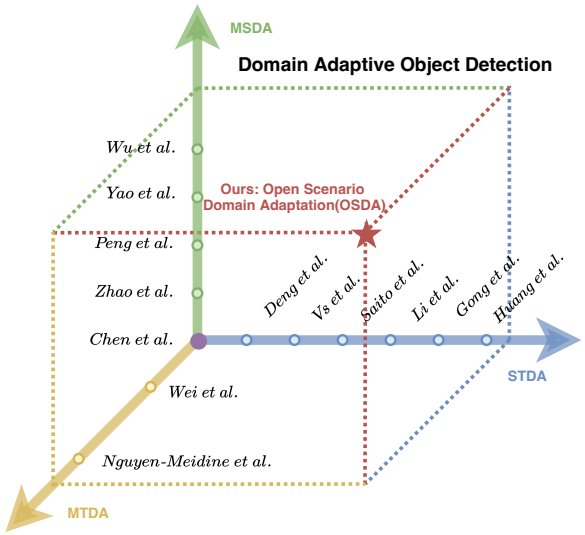

Figure 1: Open-Scenario Domain Adaptive Object Detection (★) is a novel problem that has not been formally defined and addressed so far. Though related to the STDA, MSDA and MTDA, OSDA offers its own more practical challenges, which when addressed, improve the practicality of object detectors.

1 INTRODUCTION

Object detection [25, 31, 38], a fundamental and essential problem in autonomous driving, is to identify and localize the objects of interest based on visual perception in deep learning [42–45]. Unfortunately, the optimized detectors in the closed-domain or closed-scenario visual data would suffer from a severe decrease in performance when applied to unseen scenarios due to the domain shift [3, 6, 17]. The autonomous driving system is desired to work in various environments with different lighting, weather, and cityscapes.

Collecting annotated data in such an environment can be challenging, while obtaining large amounts of unlabeled data is easier.

Domain adaptive object detection (DAOD) has thus been proposed to detect the common objects in novel target domains using unlabeled data. The Domain Adaptive Faster R-CNN [3] (shown as • in Fig. 1) is a pioneering work to address this issue, in which the feature alignment and gradient reversal layer [8] are proposed to conduct effective cross-domain object detection. The existing domain-adaptive detection algorithms [6, 10, 17, 35] have demonstrated promising performance but are hinged on the assumption that there is only a single labeled source domain and a single unlabeled target domain, which is a highly idealized setting. For simplicity, let us denote this group of methods that follows *Single-Source-Single-Target* domain adaptation setting as **STDA** hereafter.

To improve the fidelity to real scenarios, *Multi-Source-Single-Target* domain adaptation methods [28, 48, 53, 59] (**MSDA**) (Fig. 1) have been introduced, which involve more training data from different data distributions (various sources) for better representation learning towards domain-invariant features. The most commonly adopted approach for combining data from multiple sources is to treat all sources as one source. However, this line of approach ignores the difference among various sources. In many situations, conflicts between multiple source domains can lead to negative transfer. Moreover, both STDA and MSDA assume that only one target domain is being adapted, which is infeasible for the goal of learning a single model that can adapt to multiple target domains simultaneously. The *One-Source-Multi-Target domain adaptation* methods [19, 46] (**MTDA**) (Fig. 1) proposed to perform the adaptation from the single source domain to the multiple target domains progressively. However, MTDA methods still suffer from knowledge forgetting in the successive multi-target domain adaptation.

In this paper, we propose an *Open-Scenario Domain Adaptation* task (denoted as ★ in Fig. 1), which is conducted in combination with the above three tasks. OSDA tasks conduct knowledge transfer from multi-source domains to multi-target domains and are thus more challenging and practical. The main challenges of OSDA include 1) achieving a comparable/better performance than domain-specific STDA methods in an open-domain setting, 2) bridging the domain gap between multiple source domains when adapting to the target domains, and 3) alleviating knowledge forgetting when performing successively multi-target domain adaptation.

We argue that existing domain adaptation methods [3, 6, 17] are constrained by the narrow definition of "domain" as a closed set of concepts in the autonomous driving scene (e.g., car, pedestrian, lane). As a result, the learned adaptivity mainly reflects the model's ability to discriminate among the closed vocabulary with the assumption that concepts are distinct enough from one another. It focuses too much on the separability of concepts of interest and pays less attention to the clutter within the background. The "background" in existing settings is in fact an ill-defined concept consisting of a great variety of hidden concepts (e.g., buildings, rubbish bins, statues) which can bring significant confusion into the RPN(Region Proposal Network) learning, and is a dominating category with the greatest variance in the feature space which makes RPN harder to distinguish the foreground and background. The most straightforward way to address this issue is to increase the granularity of the background category, but this is limited by the high cost of labeling. Fortunately, recent advances in large multi-modal models have enabled learning vision and category relations through pre-training in a label-free manner. We are inspired by the idea and propose to pre-train the model prior to the actual learning. This paves the way for a comparably unbiased understanding of the relationship between the background and foreground concepts, and implicitly refines the granularity of the "background" category.

Similar to the open vocabulary learning [56, 62], the pre-training is using image caption datasets without specific region-category annotation and builds the foundation for adapting a model from multi-sources to multi-targets. The model after the pre-training procedure possesses the generalized category feature distribution, which helps downstream tasks to better distinguish foreground from background and get a stable start in further fine-tuning. More importantly, the general knowledge learned in the foundation model could also enrich the recognition ability to discriminate objects under open scenarios. The proposed cross-reconstruction module based on the feature representations from the multi-source domains as shown in Fig. 2 can effectively narrow the gap between multi-source domains, and the fixed low-level parts in the visual backbone can alleviate the problem of knowledge forgetting. Therefore, we call it *Open-Scenario Domain Adaptation* method (**OSDA**).

The main contributions of our paper are summarized as follows:

- We revisit the domain adaptive object detection and propose a more challenging and practical open-scenario object detection setting, which could better approach real-world cross-domain object detection. The proposed setting also includes the existing multi-source domain adaptation (MSDA) and multi-target domain adaptation (MTDA).
- We have investigated that diversity within "background" categories leads to difficulty in distinguishing the foreground categories. We propose to increase the granularity of the background categories by building the foundation model, which could better serve for DAOD.
- We introduce a cross-construction module for extracting domain invariance in MSDA and alleviating knowledge forgetting in MTDA without any extra computational cost. The results of extensive experiments demonstrate the generalization and effectiveness of our method.

2 RELATED WORK

2.1 Pre-training for Foundation Model

Pre-training is an important and effective technique to obtain a powerful foundation model, which could be extended to various downstream vision tasks. The learned general knowledge has shown a strong few-shot and zero-shot learning ability to perform open-vocabulary object detection. CLIP [30] was optimized by a huge amount of text-image pairs, and has achieved impressive open-vocabulary object recognition results on various downstream tasks. The visual encoder and the text encoder could project the text and image inputs to a common feature space. The CLIP model had been extended to image generation, image-text retrieval [52], image classification [30], and image captioning [1] due to its strong generalization ability to recognize various object categories from diverse languages. [62] proposed RegionCLIP, which aligns regional features and corresponding descriptions to perform open-vocabulary object detection, The vision-language pre-training [36, 56, 62] is beneficial for downstream visual tasks.

Open-vocabulary object detection (OVD) aims to detect novel object categories beyond the training data. [56] proposed to disentangle object detection into recognition and localization and then optimize the parallel tasks separately based on specially designed supervisions for each corresponding task. The recognition is optimized based on the language captions designed for open-vocabulary recognition while the localization is optimized from bounding box annotations. [56] believed that open vocabulary tasks can be extended to other downstream tasks, just as a knowledgeable person can learn knowledge faster than an illiterate person. The generalized category features distribution in pre-training can help OSDA better distinguish foreground from background and get a stable start in further fine-tuning. We introduce the pre-training method to obtain a strong foundation model for our OSDA task, which has been under-explored in the existing methods [10, 17].

2.2 Domain Adaptive Object Detection

Source-Target Domain Adaptation (STDA) aims to transfer the object detector trained on a labeled source domain to a novel target domain with different data distribution. STDA is important and practical since the annotation of the target domain data may be scarce, expensive, or even inaccessible. The feature alignment [3], gradient reversal [3], maximum classifier discrepancy [20], cycle consistency [57], self-supervised learning [55] and knowledge distillation [14] have been introduced for domain adaptation. DAOD [3] was the pioneering work that introduced the domain classifier and gradient reversal into feature alignment for STDA. The following works [10, 11, 17, 47, 63] further improved the detection performance in the target domain. However, these methods only assume that there is a single source and target domain meanwhile.

Multi-Source Domain Adaptation (MSDA) considers a more generalized setting, in which multiple source domains are available for training [2, 28, 53, 59]. It is beneficial to model generalization ability as more diverse data are included. [26, 37] proposed to handle this task through a weighted combination of multiple source domains to achieve target-relevant prediction with rigorous theoretical analysis. Recent attempts have conducted this reweighting process in adversarial adaptation [29, 39] for dynamically aligning moments of feature distributions, which consist of pairs of source and target domains and those of source domains. Rather than explicit feature alignment, DMSN [53] was the first work to introduce MSDA into object detection. DMSN develops feature alignment among sources and pseudo-subnet learning for a more effective combination. TRKP [48] aimed at preserving more target-relevant knowledge from different source domains to facilitate multi-source DAOD. Our method OSDA innovatively extracts domain invariance by introducing the feature cross-reconstruction.

Multi-Target Domain Adaptation (MTDA) involves large domain shifts between complex source and target distributions compared with STDA. STDA for detection can be applied to MTDA by adapting one model per target, or one common model with a mixture of data from target domains. However, these approaches are either costly or inaccurate. MTDA remains largely unexplored in object detection, despite its practical relevance in several real-world applications, such as cross-time autonomous driving, because there is knowledge forgetting of the middle target domain in successive multi-target domain adaptation. MTDA-KD [46] is the first to introduce MTDA into object detection and proposes a novel multi-domain adaptive method for object detection based on incremental learning. MTDA-DTM [19] introduces a more efficient approach based on MTDA-KD that generalizes well to multiple target domains, and leverages domain discriminators to train a novel Domain Transfer Module (DTM), which incurs an extra overhead and depends on the order of targets selected. Our method OSDA effectively alleviates knowledge forgetting by freezing low-level layers in the backbone without any extra computing cost.

3 METHODS

3.1 Overview

We first define Open-Scenario Domain Adaptive (OSDA) object detection. Specifically, OSDA refers to recognizing and localizing objects under open-scenario domains including multi-source domains (MSDA) [28, 48, 53, 59] and multi-target domains (MTDA) [19, 46], for which most existing STDA methods [3, 6, 10, 11, 17] that focus on the source-target scenario cannot be applied. We are not only the first to propose OSDA, but also the first to introduce the two-stage training for solving this more realistic detection setting. We propose to formulate solutions to OSDA by defining a foundation model [56, 62] in advance, which covers the pairs of visual-region and language concepts as much as possible. Motivated by this, a novel framework is designed as illustrated in Fig. 2 upper part. The foundation model is created by exploring the grounding model to align latent concept distribution generated from both visual patterns and language guidance. In this way, the scale of visual region [56, 62] and language concept [7, 62] pairs can be made up by expanding the coverage of language guidance, which can be easily achieved by leveraging a caption dataset [23]. We believe that the pre-training procedure learns the generalized knowledge that is not limited to a closed scenario, and increases the granularity of the "background" category to help downstream specific tasks to better distinguish foreground from background and get a stable start in further fine-tuning. In Fig. 2 lower part, the vision and language knowledge are then transferred by loading parameters of the foundation model to better use the relationship between visual and language to guide downstream task learning. Furthermore, various source inputs have never been considered by existing domain adaptive object detection frameworks [10, 17], the OSDA makes full use of data from the multi-source domain and leverages to enable the learning of domain-agnostic representation within the multi-source domain. In addition, we explore alleviating knowledge forgetting in successive multi-target domain adaptation without any extra computational cost. In this way, a novel framework OSDA is formulated to detect objects under the open-scenario domains. In Sec. 4, we conduct our experiments under STDA, MSDA, and MTDA compared with existing methods [3, 6, 17, 19, 28, 46, 48, 53, 59]. The comprehensive experiments and superior results have demonstrated the effectiveness of the proposed method.

3.2 Building the Foundation

In this section, we will introduce the process of building the "Foundation" model. Recently, with the popularity of CLIP [30], the multimedia foundation model has demonstrated its potential in various downstream task applications. Inspired by them, we introduce

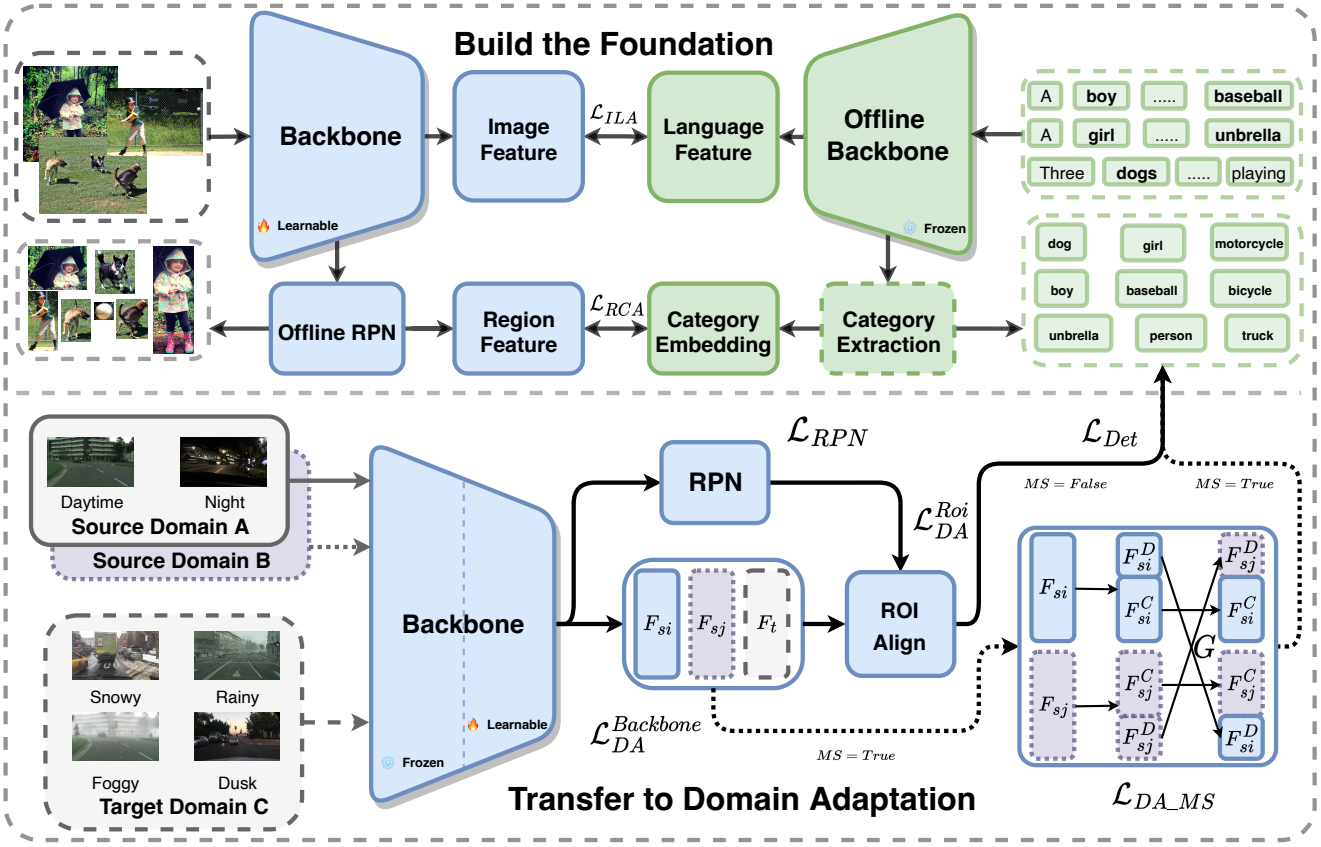

Figure 2: The architecture of our method includes pre-training (upper) and training (lower). In pre-training, we perform contrastive learning in the feature space for building the foundation model by introducing $\mathcal{L}_{ILA}$ and $\mathcal{L}_{RCA}$. In training, images from the source domain and target domain are fed into the model at the same time, but only source domain images will be trained for object detection. If there are multi-source datasets, MS will be True and the model will start multi-source feature cross-reconstruction. Moreover, we replace the classification head in Faster R-CNN with the fixed category embedding to retain the mapping relationship between vision and language in the pre-training.

multi-level visual and language aligning models [5, 30, 62], which apply the contrastive learning to pre-train a model on a large image captioning dataset [23]. Thanks to that the captioning corpus includes a rich vocabulary and semantic structure, it is possible to learn the meaning of various category concepts for constructing a generalized "Foundation" model. We employ a weakly supervised grounding technique to learn potential semantic relationships between visuals and language because the specific correspondence of words and regions in caption datasets is not given. Specifically, hierarchical tasks are involved including Image-Language Aligning (ILA) and Region-Category Aligning (RCA).

During pre-training, we follow RegionCLIP [62] to use a pre-trained ResNet-50 [13, 30] as a visual backbone and use a frozen language encoder of clip [30, 62] as language backbone which takes a whole sentence as input. We take image-caption pairs as input, and get visual features and language features, respectively. Then we introduce global contrastive learning loss for aligning global feature pairs to pre-train our visual backbone. We design a linear layer named I2L (Image to Language) to map visual representation f_i^I into the language embedding space f_j^L for the ILA task. The global aligning score for image-caption pairs is defined as follows:

$$\langle I_i, L_j \rangle_A = \left\langle f_i^I, f_j^L \right\rangle_{Cosine}, \tag{1}$$

$\langle ., . \rangle$ denotes cosine similarity for calculating the distance of two vectors.

The optimization goal is to maximize the global aligning score for a matching image-caption pair while minimizing the score for a non-matching pair. As a result, with a sampling batch containing different caption and image pairs for contrastive learning, the grounding objective functions are defined as follows:

$$\mathcal{L}_{ILA} = -\frac{1}{N} \sum_i \log \frac{\exp \langle I_i, L_j \rangle_A}{\sum_{L' \in B_L} \exp \left(\langle I_i, L' \rangle_A / \tau \right)} \tag{2}$$

where $\langle I_i, L_j \rangle_A$ is the best match pair and B_L is the batch for image and language modality respectively.

Optimizing the aligning objectives results in a learned visual backbone and V2L layer that can map an image into a sentence that best describe them for learning multimodality feature.

Table 1: The experimental result comparison under the STDA setting for Cityscapes → Foggy Cityscapes.

Dataset	Method	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	mAP
Cityscapes→Foggy Cityscapes	*Pre-training Baseline*	43.4	46.3	53.8	26.6	36.1	31.6	31.9	41.8	38.9
	DAF [3]	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0
	SCDA [65]	33.8	42.1	52.1	26.8	42.5	26.5	29.2	34.5	35.9
	CR-SW [49]	34.1	44.3	53.5	24.4	44.8	38.1	26.8	34.9	37.6
	GPA [51]	32.9	46.7	54.1	24.7	45.7	41.1	32.4	38.7	39.5
	MEGA [40]	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8
	SCAN [21]	41.7	43.9	57.3	28.7	48.6	48.7	31.0	37.3	42.1
	TIA [61]	34.8	46.3	49.7	31.1	52.1	48.6	37.7	38.1	42.3
	SCFA [58]	**51.4**	51.7	64.1	26.7	48.5	13.1	**38.1**	**49.5**	42.8
	O^2Net [10]	48.7	51.5	63.6	31.1	47.6	47.8	38.0	45.9	46.8
	AQT [17]	49.3	52.3	**64.4**	27.7	53.7	46.5	36.0	46.4	47.1
	OSDA-ST	48.7	**53.8**	61.1	**36.1**	**55.4**	**50.5**	37.8	47.0	**48.8**
	Oracle	55.7	58.5	71.9	46.3	63.6	55.9	43.0	53.0	56.0

To learn local features, we employ a Region-Category Aligning (RCA) strategy. Specifically, we introduce offline object extraction RPN [31, 62], which is pre-trained on foreground-background annotation datasets without using specific category annotation. We extract image regions and get region visual feature f_i^r. Moreover, to better represent the local region features with language description, we use the new word embedding representation of categories [62]. Meanwhile, to build the mapping between category word embedding and visual region feature, we also design a linear layer named R2C (Region to Concept) to map visual representation f_i^R into the language embedding space f_j^C for the RCA task. The Local grounding objective functions $\mathcal{L}_{RCA}$ is same as $\mathcal{L}_{ILA}$, but replaced f_i^I, f_j^L with f_i^R, f_j^C.

To optimize the overall pre-training process consisting of two complementary tasks, the total loss is as follows:

$$\mathcal{L}_{Pre-train} = \mathcal{L}_{ILA} + \mathcal{L}_{RCA}. \tag{3}$$

3.3 Transfer to Domain Adaptation

In this section, the process of leveraging the foundation model for detecting objects under the open-scenario is described in detail. We try to solve a common issue that is caused by domain appearance changes in the open world, i.e. applying a model trained in normal weather to carry out perception tasks in the open-scenario including cross-weather and cross-time. Open-scenario domain adaptation contains source-target domain adaptation (STDA), multi-source domain adaptation (MSDA), and multi-target domain adaptation (MTDA). We simplify the domain adaptive loss function in the network for deployment, which is different from other domain adaptive methods. And a lightweight feature alignment network can achieve very nice results in OSDA tasks. To bridge the cross-domain gap, we design a global domain adaptive component by aligning the global feature representation distribution across domains. The optimization goal is to maximize the model's ability to extract domain-agnostic feature representations by introducing the gradient reversal layer. As a result, with a sampling batch containing different domains feature for domain invariance learning, the global domain adaptive discrimination loss is defined as follows:

$$\mathcal{L}_{DA}^{Backbone} = -D_i \log P(F_i^{Backbone}) - (1-D_i) \log(1-P(F_i^{Backbone})), \tag{4}$$

where D_i denotes the domain label of the i-th training sample in batch size, with $D_i = 0$ for the source domain and $D_i = 1$ for the target domain, $F_i^{Backbone}$ denotes the global feature of the i-th training sample.

Meanwhile, to align the local domain distributions, we also optimize the ROIAlign's parameters of the detection network to maximize local domain classification loss aimed at extracting local domain-agnostic feature representations. The local domain adaptive discrimination loss $\mathcal{L}_{DA}^{Roi}$ is same as $\mathcal{L}_{DA}^{Backbone}$, but replaced $F_i^{Backbone}$ with F_i^{Roi}.

In addition, we develop a novel OSDA framework to introduce various source subsets to learn more domain invariances. This network supports single and multi-source domain annotation datasets as input. It is worth mentioning that if there are multi-source annotation datasets as input, we also need narrow the gap between source domains. Similar to source and target domain alignment, the multi-source domain discrimination loss is defined as follows:

$$\mathcal{L}_{DA\_MS} = -D_{s_i} \log P(F_{s_i}^{Backbone}) - (1-D_{s_i}) \log(1-P(F_{s_i}^{Backbone})), \tag{5}$$

where $D_{s_i} = 0$ for the source A domain and $D_{s_i} = 1$ for the source B domain.

Inspired by image fusion [50], we disentangle source image feature $F_{s_i}^{Backbone}$ into domain component $F_{s_i}^D$ and concept component $F_{s_i}^C$. G performs features mean coupling. The coupled features are used for normal object detection training with the supervision of $F_{s_i}^C$, loss function are defined as follows:

$$\mathcal{L}_{Det\_MS} = L_{Det}(G(F_{s_j}^D, F_{s_i}^C)), \tag{6}$$

where G performs features mean coupling, L_{Det} is object detection loss.

To retain the mapping relationship between vision and language in the foundation model, we replace the classification head in Faster R-CNN with different category embedding in pre-training. We use similarity [22] for calculating the matching probability between region feature F_i^{Roi} and category embedding F_{Cj} of annotated categories in the training dataset. It can be calculated as follows:

$$P(Cj|F_i^{Roi}) = \frac{\exp\left(Sim(F_i^{Roi}, F_{Cj})\right)}{\sum_{Ci \in C} \exp\left(Sim(F_i^{Roi}, F_{Ci})\right)}. \tag{7}$$

Table 2: The experimental result comparison under the STDA setting for Cityscapes → BDD100K Daytime.

Dataset	Method	Person	Rider	Car	Truck	Bus	Motor	Bike	mAP
Cityscapes→BDD100K Daytime	*Pre-training Baseline*	39.0	30.9	53.9	13.8	13.2	16.8	25.7	27.6
	DAF [3]	28.9	25.4	44.1	17.9	16.1	14.2	22.4	24.9
	SCDA [65]	29.3	29.2	44.4	20.3	19.6	14.8	23.2	25.8
	CR-DA [49]	30.8	29.0	44.8	20.5	19.8	14.1	22.8	26.0
	CR-SW [49]	32.8	29.3	45.8	22.7	20.6	14.9	25.5	27.4
	SFA [41]	40.2	27.6	57.5	19.1	23.4	15.4	19.2	28.9
	CACO [16]	32.7	32.2	50.6	20.2	23.5	19.4	25.0	29.1
	AQT [17]	38.2	33.0	58.4	17.3	18.4	16.9	23.5	29.4
	O^2Net [10]	40.4	31.2	58.6	20.4	**25.0**	14.9	22.7	30.5
	OSDA-ST	**50.3**	**37.4**	**66.1**	**28.6**	24.8	**26.0**	**33.6**	**38.1**
	Oracle	55.6	40.5	68.9	46.6	50.0	39.7	43.7	49.3

where $Sim()$ is similarity function, and Ci is specific category in C. Other object detection losses L_{Det} is same with the standard Faster R-CNN [31].

3.4 Total Loss

Through formulating the detection across domains into a unified framework for more practical open-scenario domain adaptive object detection, the final training loss of our methods is as follows:

$$\mathcal{L}_{Train} = \alpha\mathcal{L}_{DA} + \mathcal{L}_{Det} + (\beta\mathcal{L}_{DA\_MS} + \gamma\mathcal{L}_{Det\_MS})_{MS=True}. \quad (8)$$

where α, β, γ are hyper-parameters to balance the contribution of different terms, $\mathcal{L}_{DA}$ is the sum of $\mathcal{L}_{DA}^{Backbone}$ and $\mathcal{L}_{DA}^{ROI}$ and $\mathcal{L}_{Det}$ is Faster R-CNN supervision loss.

It should be noted that we do not design a loss function or any additional modules specifically for the MTDA task. We choose to freeze low-level parameters in the backbone to get the balance between final target domain learning and middle domain knowledge forgetting. This is essentially different from previous methods [19, 46] where the issue is addressed by specifically designing extra modules and the training order of targets. We will discuss this in Sec. 4.2.

4 EXPERIMENTS

4.1 Implementation Details

Pre-training Details. We used a pre-trained and frozen CLIP language encoder [30, 62] as the language backbone and a ResNet-50 [13, 30] as the visual backbone. The RPN used in pretraining was trained with the base categories of COCO [23] dataset. The other details of contrastive learning are the same as [56, 62].

Training Details. For training, we use the category embedding vectors generated by offline CLIP language encoder [62], to replace categories classifier weights and fix them. We start the training process by setting the learning rate as 0.005 and then decreasing it to 0.0005 and 0.00005 when appropriate.

Dataset Pre-processing. Before inference, we noticed a lack of samples for the category "train" in the RainCityscapes evaluation subset. Thus, a set of 100 randomly selected images with the category "train" was extracted out of 500 images in the training set and transferred to evaluation [19].

Table 3: The experimental result comparison under the STDA setting for BDD100K Daytime → BDD100K Night, and Sim10K → Cityscapes(Car).

Dataset	Method	mAP
Day.→Night	CycleGAN [64]	25.9
	DCLGAN [12]	26.2
	CUT [27]	26.3
	OA-FSUI2IT [60]	30.5
	OSDA-ST	**37.3**
Sim10K→Cityscapes(Car)	DAF [3]	41.9
	SWDA [33]	44.6
	GPA [51]	47.6
	ViSGA [32]	49.3
	SFA [41]	52.6
	AQT [17]	53.4
	O^2Net [10]	54.1
	TDD [14]	63.3
	OSDA-ST	**64.8**

4.2 Further Analyses and Discussion

Considering that our proposed setting of detecting objects from open-scenario domains is a completely all-around task including STDA, MSDA and MTDA. We compare the mainstream methods under three tasks for a fair comparison.

STDA Quantitative Evaluation.

As mentioned, our model can detect objects in the open-scenario setting. To demonstrate the effectiveness of our method, we first conduct a comparison under the common STDA setting. As reported in Table 1, our model produces the best results among all the methods in the cross-weather setting and outperforms the previous SOTA AQT [17] by **1.7**. With the cross-camera setting, our model could also outperform the existing O^2Net [10] by an obvious increase (**7.6**). Meanwhile, we also consider the cross-time scenario (*i.e.* BDD100K [54] Daytime-to-Night) in Table 3. Our method has achieved more gains. Finally, we also performed the common STDA task (i.e. Sim10K [34]-to-Cityscapes [4]) in Table 3. We outperformed the best TDD [14] by **1.5**.

According to the quantitative results, we can draw two important conclusions: 1) Our method has a strong ability to resist the

Table 4: The experimental result comparison under the MSDA setting for BDD100K Daytime/Night → Dawn/Dusk and Cityscapes and KITTI → BDD100K Daytime(Car)

Dataset	Method	mAP
D.+N.→D.+D.	MDAN [59]	27.6
	M^3SDA [28]	26.5
	DMSN [53]	35.0
	TRKP [48]	39.8
	OSDA-MS	**42.2**
	Oracle	37.9
C.+K.→D.(Car)	MDAN [59]	43.2
	M^3SDA [28]	44.1
	DMSN [53]	49.2
	TRKP [48]	58.4
	OSDA-MS	**63.3**
	Oracle	64.7

long-tail data distribution, and the experimental results on "train" and "truck" are significantly better than the existing methods. The existing methods usually initialize the classifier randomly and the model would consistently overfit the dominant categories. In contrast, we replace the classifier with fixed category embedding and retain the mapping relationship between vision and language extracted from the foundation model. In this way, we could get a balance between mAP in total categories and AP in tail categories, and alleviate the negative effects of long-tail distribution. 2) As for the cross-camera setting in Table 2, our method outperforms all SOTA methods in nearly all the categories without any data augmentation or image-to-image translation [60]. We attribute the improvement to the foundation model, which has aligned visual features and language representations in advance. The introduction of our foundation model fully exploits the knowledge contained in the semantic descriptions for the discrimination task to get a stable start. We will include more discussions about the effects of fixed classifiers and the pre-training model in Sec. 4.3.

MSDA Quantitative Evaluation. We then perform the MSDA experiments and report the results in Table 4. Our method outperforms the previous MSDA algorithms by **2.4** and **4.9** on (BDD100K daytime/night→BDD100Kdawn/dusk) and (Cityscapes/KITTI [9]→ BDD100K Daytime), respectively. We investigate that the main challenge of MSDA is the knowledge degradation caused by domain shifts inside the multiple source domains. Inspired by image fusion [18, 24], we propose the process of cross-reconstructing features inside the multi-source domains. The reconstructed features are used for the final object detection training with localization supervision and domain supervision. The lower performance of the Oracle model in Table 4 can be attributed to the limited number of training images available in the target domain [48].

MTDA Quantitative Evaluation. As shown in Table 5, our method outperformed the previous MTDA methods by **7.3** and **10.2** on FoggyCityscapes [34] and RainCityscapes [15] dataset, respectively. The forgetting rate (%, compared with the respective baselines) is used as a metric to quantitatively evaluate anti-forgetting. As reported, our method has a stronger anti-forgetting ability compared with other algorithms. The difficulty of MTDA is knowledge forgetting during the multi-target domain adaptation procedure. Due to the uncertainty of the target domain, traditional methods either choose to store the previously learned knowledge or additionally design a knowledge distillation module and a specific learning order of targets. Either way, extra additional training resources are required. We investigate that the main reason leading to the knowledge forgetting is that all the parameters of the backbone are constantly updated as the target domain changes. In contrast, we only update a part of the whole network backbone to achieve a better trade-off between knowledge forgetting and target domain learning. The experimental results show that our method could effectively alleviate knowledge forgetting, and achieve 6.1% forgetting rate on Foggy Cityscapes dataset compared with the STDA-ST method in Table 1, though the proposed method is not specially designed for this task. We also achieve the best results on Rain Cityscapes dataset.

Table 5: The experimental result comparison under the MTDA setting for Cityscapes(S) → Foggy Cityscapes(T1) → Rain Cityscapes(T2), Rate denotes forgetting rate(↓ is better).

Dataset	Method	mAP	Rate%
Cityscapes→Foggy C.→ Rain C.	Train: S → T1 → T2 → Test: T1		
	HTCN(Base) [2]	39.8	-
	MTDA-KD [46]	36.0	9.5%
	MTDA-DTM [19]	37.5	**5.8%**
	OSDA-ST(Base)	48.8	-
	OSDA-MT	**45.8**	6.1%
	Train: S → T1 → T2 → Test: T2		
	MTDA-KD [46]	40.2	-
	MTDA-DTM [19]	40.8	-
	OSDA-MT	**51.0**	-

4.3 Ablation Study

To explore the effectiveness of each component proposed in our method, comprehensive ablation studies are carried out and the corresponding results are illustrated in Table 6. Under the ST setting, without transferring the pre-training model, there is a noticeable performance degradation, indicating that the foundation model is essentially important. If we do not replace the classifier and initialize it randomly, the model will ignore the mapping relationships between the visual and language representations from the foundation model. Our research reveals that the existing methods underestimate the power of the transfer of language knowledge in the pre-training model. We believe that it is essential to transfer both visual and language knowledge for retaining the mapping relationship between these two counterparts. Then we remove the cross-reconstruction module and the detection results under the MS setting also decreased significantly, demonstrating that cross-reconstruction could better narrow the gap between the multiple source datasets. Thus, to effectively alleviate the knowledge forgetting in the MT setting, we do not freeze the backbone, and in

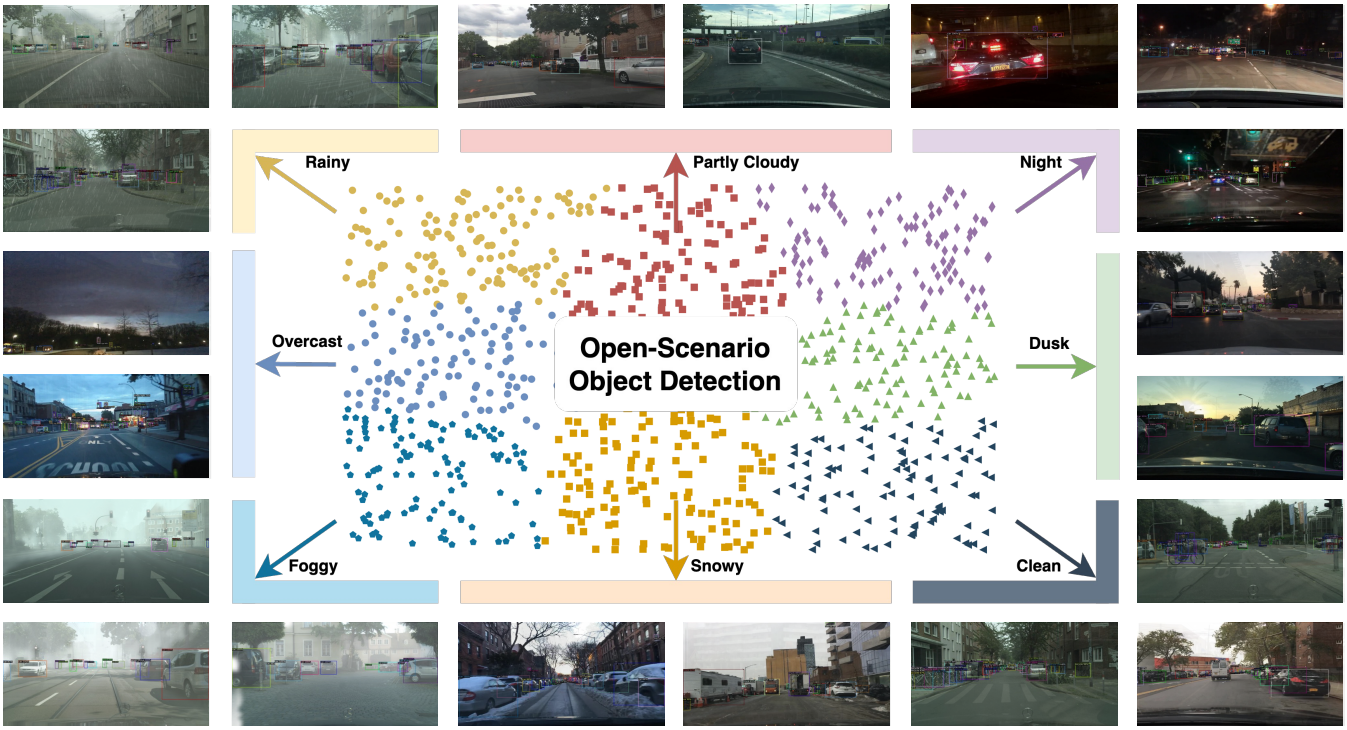

Figure 3: The object detection visualization of images from 8 different domains including *Rainy, Night,* etc, demonstrates that the proposed method could benefit the open-scenario object detection.

contrast, we only freeze the very shallow layer (C1 -C2 layers). Experimental results demonstrate that freezing only the shallow parts of the backbone can effectively alleviate knowledge forgetting. We consider the knowledge forgetting as an important problem in continual domain adaptation.

Table 6: The ablation studies of the proposed method under different settings: single-target (ST), multi-source (MS) and multi-target (MT) domain adataption settings. "all" denotes pre-training and replace classifier.

Task	Target Domain	Ablation	mAP
ST	FoggyCityscapes	w/o pre-training	42.4
		w/o replace classifier	44.1
		w/ all	**48.8**
MS	BDD100K Daytime(Car)	w/o cross-reconstruction	62.1
		w/ $\mathcal{L}_{DA\_MS}$	62.8
		w/ $\mathcal{L}_{Det\_MS}$	62.4
		w all	**63.3**
MT	FoggyCityscapes	w/o freeze	34.8
		w/ freeze C1	45.4
		w/ freeze C1+C2	**45.8**

4.4 Visualization

In Fig. 3, we visualize object detection results of the images from eight different domains under the open-scenario setting. In general, our method demonstrates that open-scenario domain adaptive object detection should be considered and solved in an integrated manner to help learn a more realistic object detection model.

5 CONCLUSION

In this paper, we have proposed to address a novel open-scenario domain adaptive object detection task, which targets to resolve the more realistic cross-domain object detection in autonomous driving at once. We have built a powerful foundation model to increase the granularity of the "background" category and enrich the ability of the foundation model to distinguish the foreground categories. We have demonstrated the potential of introducing the powerful foundation model to downstream cross-domain object detection. Moreover, we introduced the multi-source feature cross-construction module for extracting domain invariance between multiple source domains and froze shallow layers of the network backbone to alleviate the knowledge forgetting in successive multi-target domain adaptation without introducing any extra computational cost. The experimental results have demonstrated that the proposed method could achieve better object detection results than the existing methods under a large range of settings. We believe our work will help the community rethink the domain adaptive object detection in autonomous driving.

ACKNOWLEDGMENTS

This work was partially supported by the Key-Area Research and Development Program of Guangdong Province(2021B0101400002), the National Natural Science Foundation of China under grant U20B2063, 62220106008 and China Postdoctoral Science Foundation with Project 2022M720660.

REFERENCES

[1] Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. 2022. The unreasonable effectiveness of CLIP features for image captioning: an experimental analysis. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4662–4670.

[2] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. 2020. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8869–8878.

[3] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2018. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3339–3348.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3213–3223.

[5] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. 2019. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2601–2610.

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

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*. PMLR, 1180–1189.

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research* 32, 11 (2013), 1231–1237.

[10] Kaixiong Gong, Shuang Li, Shugang Li, Rui Zhang, Chi Harold Liu, and Qiang Chen. 2022. Improving Transferability for Domain Adaptive Detection Transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1543–1551.

[11] Lijun Gou, Jinrong Yang, Hangcheng Yu, Pan Wang, Xiaoping Li, and Chao Deng. 2022. A Semantic Consistency Feature Alignment Object Detection Model Based on Mixed-Class Distribution Metrics. *arXiv preprint arXiv:2206.05765* (2022).

[12] Junlin Han, Mehrdad Shoeiby, Lars Petersson, and Mohammad Ali Armin. 2021. Dual contrastive learning for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 746–755.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[14] Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao. 2022. Cross domain object detection by target-perceived dual branch distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9570–9580.

[15] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. 2019. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8022–8031.

[16] Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu, and Ling Shao. 2022. Category contrast for unsupervised domain adaptation in visual tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1203–1214.

[17] Wei-Jie Huang, Yu-Lin Lu, Shih-Yao Lin, Yusheng Xie, and Yen-Yu Lin. 2022. AQT: Adversarial Query Transformers for Domain Adaptive Object Detection. In *31st International Joint Conference on Artificial Intelligence, IJCAI 2022*. International Joint Conferences on Artificial Intelligence, 972–979.

[18] Zi-Rong Jin, Liang-Jian Deng, Tian-Jing Zhang, and Xiao-Xu Jin. 2021. BAM: Bilateral activation mechanism for image fusion. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4315–4323.

[19] Madhu Kiran, Marco Pedersoli, Jose Dolz, Louis-Antoine Blais-Morin, Eric Granger, et al. 2022. Incremental multi-target domain adaptation for object detection with efficient domain transfer. *Pattern Recognition* 129 (2022), 108771.

[20] Shuang Li, Chi Harold Liu, Binhui Xie, Limin Su, Zhengming Ding, and Gao Huang. 2019. Joint adversarial domain adaptation. In *Proceedings of the 27th ACM International Conference on Multimedia*. 729–737.

[21] Wuyang Li, Xinyu Liu, Xiwen Yao, and Yixuan Yuan. 2022. SCAN: Cross Domain Object Detection with Semantic Conditioned Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 6. 7.

[22] Xiaoxu Li, Jijie Wu, Zhuo Sun, Zhanyu Ma, Jie Cao, and Jing-Hao Xue. 2020. BSNet: Bi-similarity network for few-shot fine-grained image classification. *IEEE Transactions on Image Processing* 30 (2020), 1318–1331.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. Springer, 740–755.

[24] Risheng Liu, Zhu Liu, Jinyuan Liu, and Xin Fan. 2021. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1600–1608.

[25] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*. Springer, 21–37.

[26] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2008. Domain adaptation with multiple sources. *Advances in Neural Information Processing Systems* 21 (2008).

[27] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. 2020. Contrastive learning for unpaired image-to-image translation. In *Proceedings of the European Conference on Computer Vision*. Springer, 319–345.

[28] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1406–1415.

[29] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1406–1415.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems* 28 (2015).

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

[33] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. 2019. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6956–6965.

[34] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2018. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision* 126, 9 (2018), 973–992.

[35] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. 2019. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint arXiv:1911.02559* (2019).

[36] Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei Cai. 2022. ProposalCLIP: unsupervised open-category object proposal generation via exploiting clip cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9611–9620.

[37] Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. 2011. A two-stage weighting framework for multi-source domain adaptation. *Advances in Neural Information Processing Systems* 24 (2011).

[38] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9627–9636.

[39] Naveen Venkat, Jogendra Nath Kundu, Durgesh Singh, Ambareesh Revanur, et al. 2020. Your classifier can secretly suffice multi-source domain adaptation. *Advances in Neural Information Processing Systems* 33 (2020), 4647–4659.

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

[41] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. 2021. Exploring Sequence Feature Alignment for Domain Adaptive Detection Transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1730–1738.

[42] Jiwei Wei, Xing Xu, Zheng Wang, and Guoqing Wang. 2021. Meta self-paced learning for cross-modal matching. In *Proceedings of the 29th ACM international conference on multimedia*. 3835–3843.

[43] Jiwei Wei, Xing Xu, Yang Yang, Yanli Ji, Zheng Wang, and Heng Tao Shen. 2020. Universal weighting metric learning for cross-modal matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13005–13014.

[44] Jiwei Wei, Yang Yang, Xing Xu, Jingkuan Song, Guoqing Wang, and Heng Tao Shen. 2023. Less is Better: Exponential Loss for Cross-Modal Matching. *IEEE Transactions on Circuits and Systems for Video Technology* (2023). https://doi.org/10.1109/TCSVT.2023.3249754

[45] Jiwei Wei, Yang Yang, Xing Xu, Xiaofeng Zhu, and Heng Tao Shen. 2022. Universal Weighting Metric Learning for Cross-Modal Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 10 (2022), 6534–6545. https://doi.org/10.1109/TPAMI.2021.3088863

[46] Xing Wei, Shaofan Liu, Yaoci Xiang, Zhangling Duan, Chong Zhao, and Yang Lu. 2020. Incremental learning based multi-domain adaptation for object detection. *Knowledge-Based Systems* 210 (2020), 106420.

[47] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. 2021. Instance-invariant domain adaptive object detection via progressive disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 8 (2021), 4178–4193.

[48] Jiaxi Wu, Jiaxin Chen, Mengzhe He, Yiru Wang, Bo Li, Bingqi Ma, Weihao Gan, Wei Wu, Yali Wang, and Di Huang. 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*. 5301–5310.

[49] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. 2020. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11724–11733.

[50] Han Xu, Xinya Wang, and Jiayi Ma. 2021. DRF: Disentangled representation for visible and infrared image fusion. *IEEE Transactions on Instrumentation and Measurement* 70 (2021), 1–13.

[51] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. 2020. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12355–12364.

[52] Yahui Xu, Yi Bin, Jiwei Wei, Yang Yang, Guoqing Wang, and Heng Tao Shen. 2023. Multi-Modal Transformer with Global-Local Alignment for Composed Query Image Retrieval. *IEEE Transactions on Multimedia* (2023).

[53] Xingxu Yao, Sicheng Zhao, Pengfei Xu, and Jufeng Yang. 2021. Multi-source domain adaptation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3273–3282.

[54] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2636–2645.

[55] Jin Yuan, Feng Hou, Yangzhou Du, Zhongchao Shi, Xin Geng, Jianping Fan, and Yong Rui. 2022. Self-supervised graph neural network for multi-source domain adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3907–3916.

[56] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. 2021. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14393–14402.

[57] Hui Zhang, Junkun Tang, Yihong Cao, Yurong Chen, Yaonan Wang, and QM Jonathan Wu. 2022. Cycle Consistency Based Pseudo Label and Fine Alignment for Unsupervised Domain Adaptation. *IEEE Transactions on Multimedia* (2022).

[58] Jingyi Zhang, Jiaxing Huang, Zichen Tian, and Shijian Lu. 2022. Spectral unsupervised domain adaptation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9829–9840.

[59] Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and GeoffreyJ Gordon. 2018. Adversarial multiple source domain adaptation. *Advances in Neural Information Processing Systems* 31 (2018).

[60] Lifan Zhao, Yunlong Meng, and Lin Xu. 2022. OA-FSUI2IT: A Novel Few-Shot Cross Domain Object Detection Framework with Object-Aware Few-Shot Unsupervised Image-to-Image Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3426–3435.

[61] Liang Zhao and Limin Wang. 2022. Task-specific Inconsistency Alignment for Domain Adaptive Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14217–14226.

[62] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16793–16803.

[63] Qianyu Zhou, Qiqi Gu, Jiangmiao Pang, Xuequan Lu, and Lizhuang Ma. 2023. Self-adversarial disentangling for specific domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2223–2232.

[65] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. 2019. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 687–696.