CSDA: Learning Category-Scale Joint Feature for Domain Adaptive Object Detection

Changlong Gao[1*], Chengxu Liu[1,2*], Yujie Dun[1], Xueming Qian[1,2]
[1]Xi'an Jiaotong University
[2]Shaanxi Yulan Jiuzhou Intelligent Optoelectronic Technology Co., Ltd
{gaochanglong,chengxuliu}@stu.xjtu.edu.cn, {dunyj,qianxm}@mail.xjtu.edu.cn

Abstract

Domain Adaptive Object Detection (DAOD) aims to improve the detection performance of target domains by minimizing the feature distribution between the source and target domain. Recent approaches usually align such distributions in terms of categories through adversarial learning and some progress has been made. However, when objects are non-uniformly distributed at different scales, such category-level alignment causes imbalanced object feature learning, refer as the inconsistency of category alignment at different scales. For better category-level feature alignment, we propose a novel DAOD framework of joint category and scale information, dubbed CSDA, such a design enables effective object learning for different scales. Specifically, our framework is implemented by two closely-related modules: 1) SGFF (Scale-Guided Feature Fusion) fuses the category representations of different domains to learn category-specific features, where the features are aligned by discriminators at three scales. 2) SAFE (Scale-Auxiliary Feature Enhancement) encodes scale coordinates into a group of tokens and enhances the representation of category-specific features at different scales by self-attention. Based on the anchor-based Faster-RCNN and anchor-free FCOS detectors, experiments show that our method achieves state-of-the-art results on three DAOD benchmarks.

1. Introduction

Benefiting from the training of large numbers of high-quality labeled data, object detection algorithms have been significantly advanced in recent years [15, 20, 24, 25, 33]. However, when dealing with new scenarios with unlabeled data, the detector suffers from severe performance degradation. To tackle this problem, the domain adaptive object detection is proposed, where the trained detector is transferred from a scenario with labeled data (*i.e.*, source domain) to a new scenario without labeled data (*i.e.*, target domain).

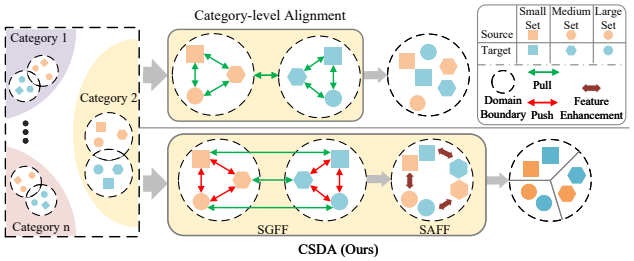

Figure 1. Illustration of the difference between our method and the category-level alignment methods. It is worth noting that we only depict one category in the figure for brevity. Our method learns category-level features jointly with scales for more effective feature alignment. SGFF: scale-guided feature fusion module. SAFE: scale-auxiliary feature enhancement module.

The large difference in data distribution between the source and target domain is an essential factor affecting the performance. To improve the generalization ability of detectors, most existing methods align feature distribution from three aspects: 1) DAF [3] uses adversarial learning [8] to align the feature distribution of pixel- and instance-level with the help of the gradient reverse layer (GRL). 2) EPM [11] and CFFA [43] weaken the negative impact from the background and focus on the cross-domain alignment of the foreground. 3) KTNet [32], GPA [37], DBGL [1], and SIGMA [17] achieve adaptation at category-level by aligning cross-domain class-conditional distributions and make significant progress. However, these works above ignore the impact of object scale in transferring and aligning category-level features, as shown in Fig. 1, which prevents them from satisfactory results.

In our view, there are still two challenges existing in current category-level DAOD works. The first challenge is the large variance in category features at different scales, which negatively affects the category-level feature alignment. Commonly, category features of small objects lack texture details or even category-specific features with dis-

*Equal contribution.

criminative ability. This makes it difficult to directly align features of small and large objects in the same category. Recent US-DAF [28] separately aligns features at different scales by assigning more fine-grained scale labels (small, medium, and large) for each category. OADA [38] introduces the bounding box offsets to conditionally align the feature distributions based on FCOS [33]. Although these algorithms above optimize feature learning for objects at different scales to some extent, they ignore the category-level feature consistency intra-/inter-domain with the same scale, resulting in inefficient learning of domain-invariant features for each category.

The second challenge is the weak perception of small objects. For humans, if we know what a "near car" looks like, then we can easily recognize the "far car". That is, when the detector can detect a large object, it also has the potential to detect the small one. Inspired by this, it is necessary to facilitate learning between objects at different scales. In DAOD, although MGADA [44] proposes an omni-scale gated fusion module consisting of convolution with different sizes to learn multi-scale features, it cannot be removed during inference, increasing the parameters.

To overcome these challenges, we propose a DAOD framework of joint category and scale information (CSDA) for more advanced cross-domain feature alignment and enhancement. In particular, as shown in Fig. 1, we proposed a Scale-Guided Feature Fusion module (SGFF) to promote feature learning and alignment for each category. It pulls objects of the same scale closer while pushing objects of different scales away in an adversarial learning manner. Such a design significantly alleviates the model degradation due to the large variance between features at different scales. Then, to improve the perceptive capability in multi-scale scenarios, especially for small objects, we design a Scale-Auxiliary Feature Enhancement module (SAFE) to enhance feature interactions at different scales. It encodes the category features and scale information of objects as a set of tokens to learn the relations between different scales through self-attention [34].

To be summarized, our contributions are as follows.

- We propose a joint category and scale feature framework for DAOD, dubbed CSDA, which achieves more effective cross-domain feature alignment and feature enhancement for different scale objects.
- We propose a scale-guided feature fusion module (SGFF), which eliminates the negative impact of excessive differences in features at different scales, and a scale-auxiliary feature enhancement module (SAFE), which enhances the perceptive capability of small objects in multi-scale scenarios.
- Extensive experiments demonstrate that the proposed CSDA can significantly outperform existing SOTA methods in three widely-used benchmarks.

2. Related Work

2.1. Object Detection

CNN-based object detection methods can be grouped into anchor-based and anchor-free frameworks. Anchor-based object detectors set anchor boxes with different scales and aspect ratios on each cell, and then obtain the category and location of each object by regression. Faster-RCNN [25] is a classical anchor-based detector. It utilizes RPN to obtain the region proposals by filtering out and adjusting the pre-defined anchors and then uses the classification and regression branch to obtain the categories of proposals and make further adjustments for their locations. YOLO [24] and SSD [22] are also anchor-based detectors with faster speeds. Anchor-free object detectors directly use full convolutional layers to complete object localization and classification without relying on anchor boxes. As a typical anchor-free object detection method, FCOS [33] directly predicts the category and offset of each pixel on the feature map. Based on different regression targets, other related detectors are CornerNet [15] and CenterNet [7]. In this paper, we use the Faster-RCNN [25] and FCOS [33] to build our DAOD framework for more representation.

2.2. Domain Adaptive Object Detection

DAOD trains a model with labeled source domain data and unlabeled target domain data, so as to achieve satisfactory performance on the target domain. Recently, many works solve the DAOD task by Mean Teacher framework [2, 6, 18, 40] and domain-alignment [3, 11, 26, 42, 44]. The methods based on Mean Teacher [31] guide the training of the student model by the high-quality pseudo labels generated from the teacher model and update the teacher model by temporally copying the weights of the student model [2, 6, 18, 40]. Domain-alignment methods are applied in three levels, including pixel-level [11, 26, 44], instance-level [3, 42], and category-level [17, 32, 37, 43, 44]. Typically, DAF [3] firstly proposes pixel- and instance-level domain discriminators to implement feature alignment in an adversarial learning [8] manner. SWDA [26] proposes a strong-weak alignment strategy that pays more attention to images that are globally similar while ignoring dissimilar ones. MGADA [44] incorporates multi-scale information and proposes a unified multi-granularity alignment-based object detection framework, including pixel-, instance-, and category-levels. PT [2] treats the anchors as learnable ones for more suitable toward the target domain and introduces strong augmentation to alleviate the intra-domain gap from a data augmentation perspective. SIGMA [17] aligns the category-level features in

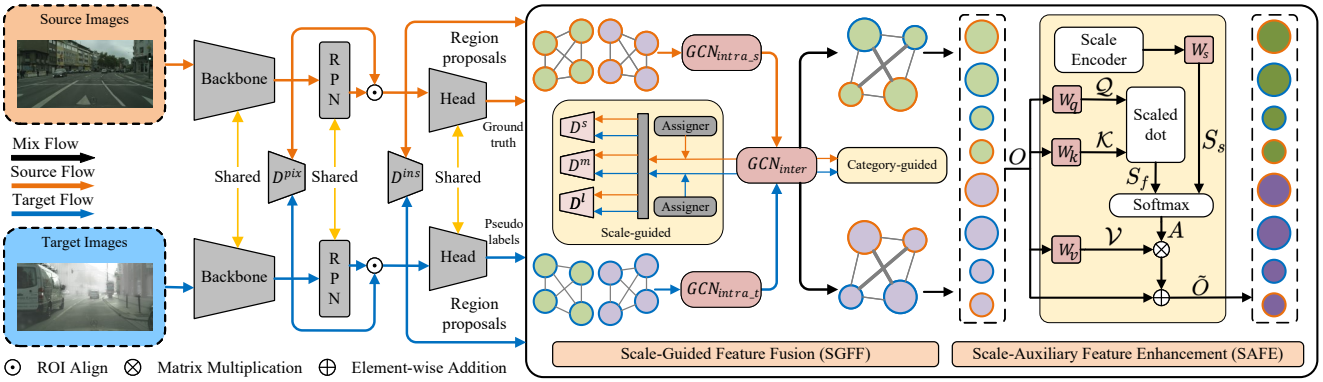

Figure 2. **Overview of our proposed DAOD framework (CSDA) based on Faster-RCNN detector.** Images from the source and target domain are fed into a shared backbone and RPN to extract domain-invariant features on pixel- and instance-level. Then, based on the region proposals, SGFF conducts the category relation graph and performs scale-separated feature alignment. Finally, SAFE is used to enhance the category-specific feature representation at different scales through the self-attention mechanism. Notably, for source and target input, we use ground truth and pseudo label from a shared detection head as supervision of our modules, respectively.

a graph-matching perspective. TIA [42] uses multiple auxiliary classifiers and localizers to align features separately and effectively guarantees the accuracy of detectors.

However, most recent DAOD works [17, 32, 37, 43, 44] mainly explore cross-domain alignment between category-level features, while ignoring the challenges caused by large variances in object scales for category-level alignment and intra-/inter-domain knowledge transfer. In this paper, we propose a domain adaptive framework that unifies category and scale information to achieve better cross-domain feature alignment and enhancement.

3. Method

It is worth noting that our method can be applied in different detectors, such as anchor-based Faster-RCNN [25] and anchor-free FCOS [33]. Without loss of generality, we mainly take Faster-RCNN as an example and annotate the difference with the implementation of FCOS.

As shown in Fig. 2, images from source domain s and target domain t are firstly fed into the backbone to compute pixel-level feature maps, followed by RPN with ROI Align to generate instance-level features. Then, the instance-level features are passed to the scale-guided feature fusion module (SGFF) to facilitate the learning of features at each scale. Benefiting from the robust spatial modeling capability of GCN [14], we model the category relations intra-/inter-domain while guaranteeing the feature consistency on each scale through three discriminators. Finally, the features are delivered to the scale-auxiliary feature enhancement module (SAFE) for enhancing the feature representation by transferring category-specific knowledge at different scales. Specifically, we encode the category features and scale information as tokens. Then the features are learned by self-attention mechanism [34].

3.1. Baseline Model

We use VGG16 [29] as our backbone. For Faster-RCNN [25], we extract the last stage C5 as the pixel-level feature. For FCOS [33], we extract the last three stages of VGG16 [29] and combine them into multi-level feature maps as pixel-level features including P3, P4, P5, P6, and P7 by FPN. Similar to the previous works [3, 11], pixel- and instance-level discriminators, $D^{pix}(\cdot)$ and $D^{ins}(\cdot)$, are used to perform pixel- and instance-level alignment of feature maps, respectively. Their loss function L_{pix} and L_{ins} are similar, which can be defined as:

$$L_{pix} = -\sum_{(i,j)} [L(D^{pix}(F^s(i,j)), y^s_{i,j}) + L(D^{pix}(F^t(i,j)), y^t_{i,j})], \quad (1)$$

where L is a binary cross-entropy function. $F^s(i,j)$ and $F^t(i,j)$ denote the feature at pixel (i,j) in the pixel-level feature map from the source and target domain, respectively. $y^s_{i,j} = 1$ and $y^t_{i,j} = 0$ are the ground truth labeled according to which domain the pixel comes from.

$$L_{ins} = -\sum_{i}^{n} [L(D^{ins}(F^s_i), y^s_i) + L(D^{ins}(F^t_i), y^t_i)], \quad (2)$$

where n is the number of region proposals. Similarly, F_i denotes the instance-level feature map from the i^{th} region proposal, and y_i is the ground truth labeled according to which domain the i^{th} proposal comes from.

Since we have the label of the source domain, we can use it to supervise the training of the detectors. About the detection loss, we exactly follow existing works [25, 33]. Specifically, for Faster-RCNN [25], there are two stages of output, *i.e.*, the RPN and RoI, each with a cross-entropy loss and a smooth L1 loss for classification L_{cls} and localization

L_{loc}, respectively. The detection loss can be defined as:

$$L_{det} = L_{cls}^{rpn} + L_{loc}^{rpn} + L_{cls}^{roi} + L_{loc}^{roi}. \tag{3}$$

In summary, the total loss of the baseline model in Faster-RCNN [25] can be defined as:

$$L_{baseline} = L_{det} + L_{pix} + L_{ins}. \tag{4}$$

For FCOS [33], the detection heads consist of classification, centerness, and regression branches, which are supervised by the focal loss [20] L_{cls}, cross-entropy loss L_{ctr}, and IoU loss [39] L_{iou}, respectively. The object detection loss can be defined as:

$$L_{det} = L_{cls} + L_{ctr} + L_{iou}. \tag{5}$$

Notably, since FCOS [33] belongs to the anchor free-based detection method, the total loss of the baseline model does not include the instance-level loss, which can be defined as:

$$L_{baseline} = L_{det} + L_{pix}. \tag{6}$$

3.2. SGFF

Most of the existing domain adaptation methods [1, 32, 36, 37, 43] focus on the alignment of category-level, ignoring the effect of scales. However, feature variance at different scales within categories may be larger than that at similar scales across categories, which fails to achieve advanced alignment [30]. To solve this problem, as shown in Fig. 2, we use GCN [14] to model intra-/inter-domain relations and align the inter-domain category-level features at the same scale by discriminators.

Graph-based feature fusion. We construct the proposals outputted from the RPN [1] as a graph $G^k = (V^k, E^k), k \in \{s, t\}$, where $V^k = \{v_i^k\}_{i=1}^{n^k}$ and $E^k = \{e_{i,j}^k\}_{i,j=1}^{n^k}$ represent the set of proposals and edges (relations among proposals), respectively. $n^k, k \in \{s, t\}$ is the number of sampling proposals. We obtain the category labels $C^k = \{c_i^k\}_{i=1}^{n^k}$ of the proposals according to the ground truth of the source domain and the pseudo-label of the target domain. Each element $e_{i,j}^k$ in the adjacency matrix E^k can be obtained by:

$$e_{i,j}^k = \begin{cases} 1 & \text{if } c_i^k = c_j^k \\ CosineSim\left\langle v_i^k, v_j^k \right\rangle & \text{if } c_i^k \neq c_j^k \end{cases}, \tag{7}$$

where $k \in \{s, t\}$, v_i^k and v_j^k is defined as the i^{th} and j^{th} proposal, c_i^k and c_j^k is the category label of the i^{th} and j^{th} proposal ($1 \leq i, j \leq n^k$).

In terms of formula, the proposals of source and target domain are fed into GCN_{intra_s} and GCN_{intra_t} to structure the relations intra-domain separately. Then, GCN_{inter} is applied to achieve knowledge transfer inter-domain. This process can be formulated as:

$$\begin{aligned} O = GCN_{inter}\,[&GCN_{intra\_s}(V^s, E^s) \\ &\copyright GCN_{intra\_t}(V^t, E^t), E\,], \end{aligned} \tag{8}$$

where, similar as the E^s and E^t, the adjacency matrix E of all proposals can be obtained from Eq.(7) and © represents concatenation operation. $O = \{o_i\}_{i=1}^{n^s+n^t}$ is the generated fusion proposals.

Category-guided alignment. To ensure the GCN models the correct intra-/inter-domain relations, we configure a classifier $f_{cls}(\cdot)$ consisting of two fully connected layers to supervise its training equipped with cross-entropy. The loss function L_{cg} of it can be formulated as:

$$L_{cg} = -\sum_{i=1}^{n^s+n^t} y_i \log\left(softmax(f_{cls}(o_i))\right), \tag{9}$$

where $y_i, i \in \{1, 2, \ldots, n^s + n^t\}$ denotes the label of output o_i, which comes from the truth-valued label of $v_i^s, i \in \{1, 2, \ldots, n^s\}$ in source domain and the pseudo-label of $v_i^t, i \in \{1, 2, \ldots, n^t\}$ in target domain.

Scale-guided alignment. As mentioned above, in addition to using GCN [14] to sufficiently learn intra-/inter-domain knowledge, we further align features at different scales. Following COCO [21], the size of proposals can be divided into three categories: small ($[0, 32^2]$), medium($(32^2, 96^2]$), and large($(96^2, +\infty)$), which is defined as the "Assigner" in Fig. 2. Based on this, we divide the O above into different scale ranges, represented as $o_{i,j}$, $i \in \{s, m, l\}$ and j is the j^{th} feature in corresponding scales.

Specifically, inspired by the adversarial learning [8] mechanism, we use three discriminators $D^s(\cdot)$, $D^m(\cdot)$, and

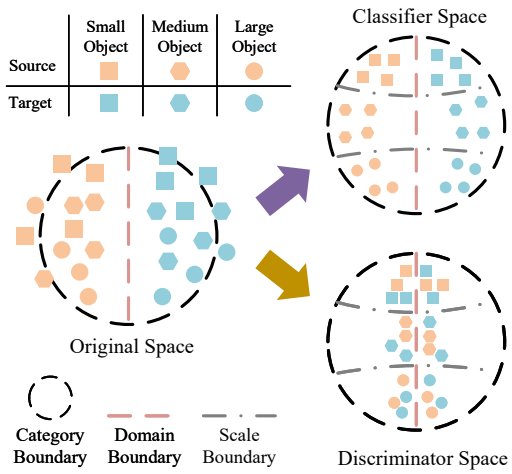

Figure 3. Illustration of the difference of scale alignment between classification [28] and discrimination method on a specific category. The purple arrow and brown arrow represent classification and discrimination methods respectively.

[1] We sample the category-level features from the output of FPN [19] for the implementation of FCOS [33]

$D^l(\cdot)$ (corresponding to small, medium, and large) to pull together inter-domain features with the same scales, which helps the GCN pay more attention on the feature fusion between the same scales. Therefore, the loss function L_{sg} of scale alignment can be defined as:

$$L_{sg} = \sum_{i \in \{s,m,l\}} \sum_{j} \lambda_i L(D^i(o_{i,j}), y_{i,j}), \tag{10}$$

where L denotes the binary cross-entropy function, $D^i(\cdot)$ denotes the discriminator of different scales, λ_i is the balancing factors, and $y_{i,j}$ is the ground truth according to which domain the $o_{i,j}$ comes from, $i \in \{s, m, l\}$.

Notably, our scale alignment method is distinct from the US-DAF [28], which aligns proposal features at different scales within a category by a multi-label classification. Our proposed discriminator-based scale alignment method outperforms the above multi-label classification method [28]. In particular, as shown in Fig. 3, without scale alignment, there will be a large gap between proposal features of the same category across domains. Scale classifiers can divide features into different scale spaces, but fail to distinguish features of different domains. Unlike classifiers, the discriminator makes the network generate more similar features in scale through GRL [8]. Therefore, the distribution of features with the same scale is more concentrated near the domain boundaries. Meanwhile, the features with different scales are also distinguishable.

3.3. SAFE

To further explore the network's ability to perceive objects at different scales, inspired by [12], we design the SAFE for further feature representation enhancement through interaction learning between different scale features. As shown in Fig. 2, it consists of a scale encoder for encoding region proposal scales into high-dimensional features and a feature enhancement module for interactively learning features across scales.

Scale encoder. Large-size objects contain rich and fine-grained category information. Therefore, for two groups of objects with different scales, if the scale difference within each group is the same, the difference between features in the group with the smaller scale is larger than the group with the larger scale. In brief, the difference between objects of size 10×10 and 20×20 is greater than the difference between objects of size 200×200 and 210×210, despite the fact that their scale difference has the same 10 pixels. Based on the above considerations, we propose a new scale mapping function $f_{sm}(\cdot)$ to encode object scales as vectors that can be embedded in high-dimensional features, which is formulated as:

$$f_{sm}(x) = \alpha \frac{1 - e^{-\frac{x}{\beta}}}{1 + e^{-\frac{x}{\beta}}}, \tag{11}$$

where x is the width or height of proposals, α and β are adjustable factors. We adjust the value of β, then encode the centroid offset by logarithmic operation and width or length by $f_{sm}(\cdot)$ to generate a 4-d feature. Following the method in [34], which computes cosine and sine functions of different wavelengths, this 4-d feature is embedded in a high-dimensional representation and encoded to scale similarity matrix, denoted as S_s, via the FC layers W_s.

Feature enhancement. We use transformer-based [34] self-attention mechanism to achieve interaction learning among features. Specifically, we first use FC layers W_q, W_k, and W_v to encode all proposals as query $\mathcal{Q} = \{q_i\}_{i=1}^{n^s+n^t}$, key $\mathcal{K} = \{k_i\}_{i=1}^{n^s+n^t}$, and value $\mathcal{V} = \{v_i\}_{i=1}^{n^s+n^t}$. Notably, to avoid confusion, the symbols q, k, and v in this section refer to the tokens in query, key, and value. n^s and n^t are the number of proposals in the source domain and target domain, respectively.

Then, based on $\mathcal{Q}$ and $\mathcal{K}$, we can calculate the feature similarity matrix, denoted as S_f, among all proposals. We combine the feature similarity matrix $S_f = \{q_i k_j^T\}_{i,j=1}^{n^s+n^t}$ and scale similarity matrix $S_s = \{s_{i,j}\}_{i,j=1}^{n^s+n^t}$ obtained above to generate the potential relations $A = \{a_{i,j}\}_{i,j=1}^{n^s+n^t}$ among proposal features by a weighted formulation of softmax, formulated as:

$$a_{i,j} = \frac{s_{i,j} \cdot exp(q_i k_j^T)}{\sum_{m=1}^{n^s+n^t} s_{i,m} \cdot exp(q_i k_m^T)}, \tag{12}$$

where $a_{i,j}$ refers to affinity score between i^{th} proposal and j^{th} proposal.

Finally, we obtain the residuals by weighting the values $\mathcal{V}$ according to the affinity score A, and sum the residuals with the output $O = \{o_i\}_{i=1}^{n^s+n^t}$ of SGFF as the output of SAFE $\tilde{O} = \{\tilde{o}_i\}_{i=1}^{n^s+n^t}$, formulated as:

$$\tilde{o}_i = o_i + \sum_{j=1}^{n^s+n^t} a_{i,j} v_j. \tag{13}$$

Similarly, the loss function L_{fe} of this module can be calculated with the same as Eq. (9).

3.4. Overall Training Loss Function

As discussed above, the loss of the baseline part is supervised by $L_{baseline}$. Besides, the SGFF is optimized in category and scale, indicated as L_{cg} and L_{sg}, respectively, and the SAFE is constrained by L_{fe}. In general, the overall training loss function is defined as:

$$L = L_{baseline} + \omega_1 \cdot (L_{cg} + L_{sg}) + \omega_2 \cdot L_{fe}, \tag{14}$$

where ω_1 and ω_2 are the balancing factors.

Method	Arch.	person	rider	car	truck	bus	train	mbike	bicycle	mAP
Source Only	Faster-RCNN+VGG16	17.8	23.6	27.1	11.9	23.8	9.1	14.4	22.8	18.8
DAF [3]		25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
SW-DA [26]		29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.2
SW-Faster-ICR-CCR [36]		32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
CFFA [43]		43.2	37.4	52.1	34.7	34.0	**46.9**	29.9	30.8	38.6
RPNPA [41]		33.6	43.8	49.6	32.9	45.5	46.0	35.7	36.8	40.5
UMT [6]		33.0	46.7	48.6	34.1	**56.5**	46.8	30.4	37.4	41.7
MeGA-CDA [35]		37.7	49.0	52.4	25.4	49.2	**46.9**	34.5	39.0	41.8
TIA [42]		**52.1**	38.1	49.7	37.7	34.8	46.3	**48.6**	31.1	42.3
PT [2]		40.2	48.8	59.7	30.7	51.8	30.6	35.4	44.5	42.7
TDD [10]		39.6	47.5	55.7	33.8	47.6	42.1	37.0	41.4	43.1
MGADA [44]		43.9	49.6	**60.6**	29.6	50.7	39.0	38.3	42.8	44.3
CSDA(Ours)		43.1	**58.4**	44.7	**50.0**	33.3	37.3	42.9	**50.5**	**45.0**
Oracle	Faster-RCNN+VGG16	44.7	63.9	37.3	47.6	32.1	37.2	40.8	48.2	44.0
Source Only	FCOS+VGG16	29.6	26.3	37.1	7.9	14.1	6.3	12.9	28.1	20.3
EPM [11]		41.9	38.7	56.7	22.6	41.5	26.8	24.6	35.5	36.0
KTNet [32]		46.4	43.2	60.6	25.8	41.2	40.4	30.7	38.8	40.9
SSAL [23]		45.1	47.4	59.4	24.5	50.0	25.7	26.0	38.7	39.6
MGADA [44]		45.7	47.5	60.6	**31.0**	52.9	44.5	29.0	38.0	43.6
SIGMA [17]		46.9	**48.4**	**63.7**	27.1	50.7	35.9	**34.7**	**41.4**	43.5
OADA [38]		**47.3**	45.6	62.8	30.7	48.0	49.4	34.6	39.5	44.8
CSDA(Ours)		46.6	46.3	63.1	28.1	**56.3**	**53.7**	33.1	39.1	**45.8**
Oracle	FCOS+VGG16	48.3	44.6	66.9	33.4	50.2	43.6	32.3	38.2	44.7

Table 1. Experimental results(%) on Cityscapes to Foggy Cityscapes.

4. Experiments

4.1. Evaluation and Datasets

Following the same DAOD setting in [3, 11], we conduct extensive experiments on three benchmarks. We use the mean average precisions (mAP) at the IoU threshold of 0.5 to compare with state-of-the-art methods.

Cityscapes→Foggy Cityscapes. The Cityscapes [4] is a street scene dataset collected from 50 cities, which contains eight categories: person, rider, car, truck, bus, train, motorcycle, and bicycle. It consists of 2,975 images for training and 500 images for validation. Foggy Cityscapes [27] is a synthesized dataset based on the Cityscapes with three levels of foggy weather (0.05, 0.01, 0.02) and shares the same annotations with it. For the adaptation scenario from normal to foggy, we set Cityscapes as the source domain and Foggy Cityscapes (0.02 level) as the target domain.

Sim10k→Cityscapes. Sim10k [13] is a synthesized dataset collected from the computer game Grand Theft Auto V (GTA5). It covers 10,000 images and their annotated bounding boxes in the car category. For the adaptation scenario from synthetic- to real-world, we set Sim10k as the source domain and Cityscapes as the target domain.

KITTI→Cityscapes. KITTI [9] is a real-world traffic scene dataset collected from an autonomous driving platform, which covers 7,481 images in the training set. In addition to the difference in camera setup, it's similar to Cityscapes [4]. For cross-camera adaptation, we set KITTI as the source domain and Cityscapes as the target domain with only the car category.

4.2. Implementation Details

We apply VGG16 [29] pre-trained on ImageNet [5] as our backbone and conduct our DAOD framework based on two different detectors (*i.e.*, Faster-RCNN [25] and FCOS [33]). For the input size, we scale all images by resizing the shorter side of the image to 600 for Faster-RCNN and 800 for FCOS, following the default settings in [3, 11]. We only use random horizontal flip as data augmentation for FCOS and add additional multi-scale augmentation for Faster-RCNN. Our model is trained with the SGD optimizer with a 0.0025 learning rate, 4 batch-size, momentum of 0.9, and weight decay of 5×10^{-4}. The balancing factors in Eq. (10) are set as $\lambda_s = 0.01$, $\lambda_m = 0.05$, $\lambda_l = 0.1$, β in Eq. (11) is set as 20, and $\omega_1 = 1.0$, $\omega_2 = 0.2$ in Eq. (14). All experiments are implemented in PyTorch.

4.3. Comparison with SOTA

As presented in Tab. 1 and Tab. 2, we compare 18 state-of-the-art domain adaptive detectors on three widely-used benchmarks. Besides, we also train the detectors without DAOD methods using only source domain data, as well as the annotated target data, and their performance is referred to as Source Only and Oracle, respectively.

Cityscapes→Foggy Cityscapes. In Tab. 1, we achieve 45.0% and 45.8% based on Faster-RCNN [25] and

Method	Arch.	Sim10k	KITTI
		mAP(car)	mAP(car)
Source Only	FR+VGG16	30.1	30.2
DAF [3]		39.0	38.5
SW-DA [26]		40.1	37.9
SC-DA [45]		43.0	42.5
CFFA [43]		43.8	-
SAPNet [16]		44.9	43.4
RPNPA [41]		45.7	-
MeGA-CDA [35]		44.8	43.0
TIA [42]		-	44.0
MGADA [44]		49.8	45.2
TDD [10]		53.4	47.4
CSDA(Ours)		**56.9**	**48.6**
Oracle	FR+VGG16	66.9	66.9
Source Only	FCOS+VGG16	39.8	34.4
EPM [11]		49.0	43.2
KTNet [32]		50.7	45.6
SSAL [23]		51.8	45.6
MGADA [44]		54.6	48.5
SIGMA [17]		53.7	45.8
OADA [38]		56.6	46.3
CSDA(Ours)		**57.8**	**48.6**
Oracle	FCOS+VGG16	73.0	73.0

Table 2. Experimental results(%) on Sim10k/KITTI to Cityscapes.

FCOS [33], and we obtain the highest mAP overall compared methods. By using Faster-RCNN, we achieve 0.7% gain over the second-best MGADA [44]. Taking into account that we don't add additional parameters during model inference, the performance improvement (+0.7%) we achieve is quite considerable. Besides, we surpass MGADA [44], SIGMA [17] , and OADA [38] with 2.2%, 2.3%, and 1.0% mAP using the same FCOS detector. Notably, our method performs better (+1.0% for Faster-RCNN and +1.1% for FCOS) than the Oracle results.

Sim10k→Cityscapes. As shown in the left part of Tab. 2, our method achieves the best mAP of 56.9% and 57.8%. Compared with the same Faster-RCNN [25] detector, it can be observed that our method outperforms the previous SOTA by 3.5%. Besides, our method acquires 3.2%, 4.1%, and 1.2% gains compared with MGADA [44], SIGMA [17], and OADA [38] on FCOS [33] detector.

KITTI→Cityscapes. The comparison results are shown in the right part of Tab. 2. Based on Faster-RCNN [25], our method achieves the best result of 48.6%, which obtains 3.4% and 1.2% gains compared with MGADA [44] and TDD [10]. Compared to recent methods SIGMA [17] and OADA [38], our method achieves 2.8% and 2.3% gains separately. Even for the FCOS [33] implementation of MGADA [44] which adds additional multi-scale data augmentation during training, our performance is still better.

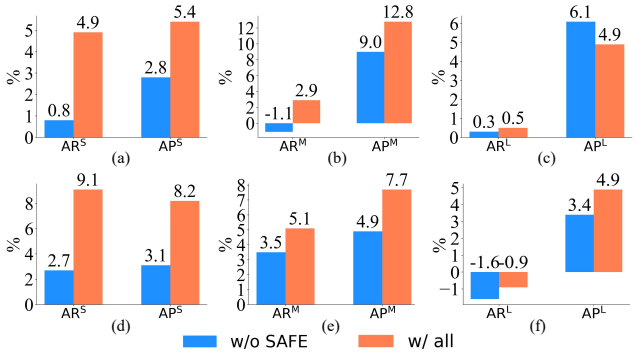

Figure 4. Ablation studies on Sim10k (Row 1) / KITTI (Row 2) to Cityscapes (%). (a) to (c) and (d) to (f) represent the AR/AP of small, medium, and large objects, respectively. For each subplot, the vertical axis represents the relative value (AR and AP) with baseline.

4.4. Ablation Study

To verify the effectiveness of each proposed module and the superiority of discriminator-based scale alignment in multi-scale scenarios, we constructed ablation studies on the three benchmarks mentioned above in Tab. 3, Tab. 4, Tab. 5, and Fig. 4. Following MGADA [44] and COCO [21] metrics, we adapt AP^S/AR^S, AP^M/AR^M, and AP^L/AR^L which denotes the mAP/mAR of the object area in the range $[0, 32^2]$, $(32^2, 96^2]$, and $(96^2, +\infty)$ respectively. All the ablation studies are applied on FCOS [33].

As shown in Fig. 5, we also visualize the results of Source only, EPM [11], CSDA, and the ablation studies on Cityscapes to Foggy Cityscapes to demonstrate the superiority of CSDA and the effectiveness of SGFF and SAFE.

Effectiveness of SGFF. To analyze the effectiveness of SGFF and the superiority of discriminator-based scale alignment over the classifier-based [28], we conduct two experiments "+SGFF (w/ scale discriminator)" and "+SGFF (w/ scale classifier)" in Tab. 3. Specifically, "+SGFF (w/ scale discriminator)" can achieve 44.6% with 5.7% gains compared with the baseline. Compared with "+SGFF (w/ scale discriminator)", the performance of "+SGFF (w/ scale classifier)" is reduced by 1.1%. As shown in Tab. 4, we design more detailed ablation studies to validate the effectiveness of each component of SGFF. Each component can improve the performance of the detector. And the addition of the scale-guided constraint promotes a significant improvement in detector performance (2.3%) compared to GFF only. Moreover, we also conduct experiments on Sim10k/KITTI to Cityscapes to validate the generalization ability of the SGFF in different scenarios. As shown in Fig. 4, we can find that the SGFF can significantly improve mAP at different scales regardless of the scenario.

Effectiveness of SAFE. As shown in Tab. 3, our CSDA further improves the mAP (+1.2%) with the help of SAFE

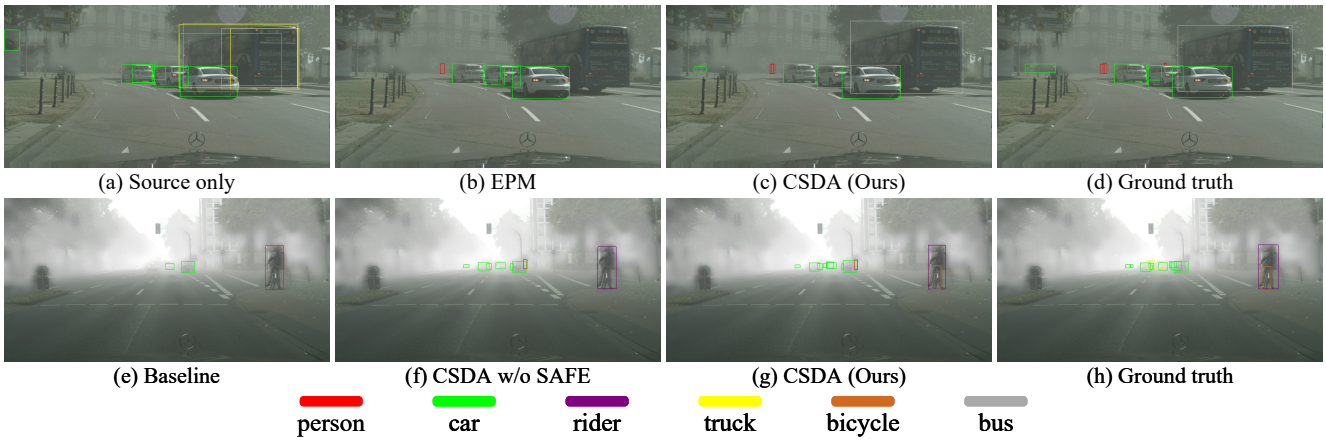

Figure 5. Visual comparison of the detection results on Cityscapes to Foggy Cityscapes scenario among (a) source only, (b) EPM [11], (c) CSDA(Ours), and (d) Ground truth in 1^{st} row and (e) Baseline, (f) CSDA w/o SAFE, (g) CSDA, and (h) Ground truth in the 2^{nd} row.

Method	mAP	AP^S	AP^M	AP^L
EPM [11]	36.0	8.3	36.7	61.6
MGADA [44]	43.6	10.1	43.1	72.5
Baseline	38.9	8.1	39.1	64.8
+SGFF (w/ scale classifier)	43.5	10.8	41.8	71.2
+SGFF (w/ scale discriminator)	44.6	11.1	43.5	71.8
+SAFE	41.4	11.0	40.7	68.4
CSDA (+SGFF+SAFE)	**45.8**	**11.3**	**45.0**	**72.8**
Oracle	44.7	12.3	42.6	70.9

Table 3. **Ablation studies of CSDA on Cityscapes to Foggy Cityscapes.** Both SGFF and SAFE can bring performance gains. Cascading SAFE after SGFF leads to further performance gains.

Method	SGFF				AP^S	AP^M	AP^L	mAP
	s	m	l	GFF				
Baseline	-	-	-	-	8.1	39.1	64.8	38.9
Proposed	✓				10.3	40.8	65.5	40.1
		✓			9.9	42.1	66.3	41.0
			✓		9.7	40.5	68.5	40.7
				✓	10.1	41.7	68.1	42.3
	✓	✓	✓		10.5	41.1	69.0	41.8
+SGFF			✓		**11.1**	**43.5**	**71.8**	**44.6**

Table 4. **Effectiveness of each component in our SGFF.** "s", "m", and "l" correspond to three discriminators: $D^s(\cdot)$, $D^m(\cdot)$, and $D^l(\cdot)$. "GFF" represents the graph-based feature fusion.

and performs better than Oracle. Besides, As shown in Fig. 4, our model significantly improves the perception of small and medium objects in both Sim10k and KITTI to Cityscapes scenarios. In particular, the SAFE has more significant effects on small objects than on medium objects on AP and AR. It fully validates that SAFE can effectively facilitate feature learning and improve the perception of objects at different scales.

Besides, we also notice that our method has more advanced performance than EPM [11] and MGADA [44] at all scales. Although MGADA [44] are reaching the upper bound of the FCOS [33] detector performance, our method can also acquire 2.2%, 1.2%, 1.9%, 0.3% gains on mAP, AP^S, AP^M, and AP^L.

β	5	10	15	20	25	30	40
mAP	44.1	44.9	45.4	45.8	45.0	45.0	44.6

Table 5. Sensitivity analysis of β on Cityscapes to Foggy Cityscapes.

Sensitivity Analysis of β. As shown in Tab. 5, we compare the different values of β in Eq. (11). The results for different β are all above 44.0%, which demonstrates that our CSDA can also achieve satisfactory results regardless of the effect of β. In particular, as the value of β increases, the performance of our model achieves the maximum mAP of 45.8% when $\beta = 20$, and then it starts to degrade.

5. Conclusions

In this work, we investigate the impact of object scale in existing category-level alignment-based DAOD methods and then analyze the challenges of large feature variance in different scales as well as the weak perception of small objects. To avoid feature degradation due to the effect of object scales, we propose a novel DAOD framework of joint category and scale feature learning, dubbed CSDA. It includes 1) a scale-guided feature fusion module (SGFF) for learning scale-separated category-specific features, which eliminates the negative impact of excessive differences in features at different scales, and 2) a scale-auxiliary feature enhancement module (SAFE) for facilitating the feature interaction between different scales, avoiding the weak perception for small objects. On three widely-used benchmarks, experimental results show significant superiority of the proposed CSDA compared with the existing SOTA methods on different detectors.

Acknowledgement. This work was supported in part by the NSFC under Grant 62272380 and 62103317, the Science and Technology Program of Xi'an, China under Grant 21RGZN0017.

References

[1] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *ICCV*, pages 2703–2712, 2021. 1, 4

[2] Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, and Shiliang Pu. Learning domain adaptive object detection with probabilistic teacher. In *ICML*, pages 3040–3055, 2022. 2, 6

[3] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive Faster R-CNN for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 1, 2, 3, 6, 7

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

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 6

[6] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, 2021. 2, 6

[7] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. CenterNet: Keypoint triplets for object detection. In *ICCV*, pages 6569–6578, 2019. 2

[8] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189. PMLR, 2015. 1, 2, 4, 5

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

[10] Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao. Cross domain object detection by target-perceived dual branch distillation. In *CVPR*, pages 9570–9580, 2022. 6, 7

[11] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *ECCV*, pages 733–748. Springer, 2020. 1, 2, 3, 6, 7, 8

[12] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, pages 3588–3597, 2018. 5

[13] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016. 6

[14] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 3, 4

[15] Hei Law and Jia Deng. CornerNet: Detecting objects as paired keypoints. In *ECCV*, pages 734–750, 2018. 1, 2

[16] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *ECCV*, pages 481–497. Springer, 2020. 7

[17] Wuyang Li, Xinyu Liu, and Yixuan Yuan. SIGMA: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, pages 5291–5300, 2022. 1, 2, 3, 6, 7

[18] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *CVPR*, pages 7581–7590, 2022. 2

[19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 4

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 1, 4

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 4, 7

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 2

[23] Muhammad Akhtar Munir, Muhammad Haris Khan, M Sarfraz, and Mohsen Ali. SSAL: Synergizing between self-training and adversarial learning for domain adaptive object detection. *NeurIPS*, 34:22770–22782, 2021. 6, 7

[24] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 1, 2

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 1, 2, 3, 4, 6, 7

[26] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 2, 6, 7

[27] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126:973–992, 2018. 6

[28] Wenxu Shi, Lei Zhang, Weijie Chen, and Shiliang Pu. Universal domain adaptive object detector. In *ACM MM*, pages 2258–2266, 2022. 2, 4, 5, 7

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 6

[30] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection snip. In *CVPR*, pages 3578–3587, 2018. 4

[31] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30, 2017. 2

[32] Kun Tian, Chenghao Zhang, Ying Wang, Shiming Xiang, and Chunhong Pan. Knowledge mining and transferring for domain adaptive object detection. In *ICCV*, pages 9133–9142, 2021. 1, 2, 3, 4, 6, 7

[33] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019. 1, 2, 3, 4, 6, 7, 8

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2, 3, 5

[35] Vibashan Vs, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. MeGA-CDA: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *CVPR*, pages 4516–4526, 2021. 6, 7

[36] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, pages 11724–11733, 2020. 4, 6

[37] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, pages 12355–12364, 2020. 1, 2, 3, 4

[38] Jayeon Yoo, Inseop Chung, and Nojun Kwak. Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In *ECCV*, pages 691–708. Springer, 2022. 2, 6, 7

[39] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *ACM MM*, pages 516–520, 2016. 4

[40] Jinze Yu, Jiaming Liu, Xiaobao Wei, Haoyi Zhou, Yohei Nakata, Denis Gudovskiy, Tomoyuki Okuno, Jianxin Li, Kurt Keutzer, and Shanghang Zhang. Mttrans: Cross-domain object detection with mean teacher transformer. In *ECCV*, pages 629–645. Springer, 2022. 2

[41] Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *CVPR*, pages 12425–12434, 2021. 6, 7

[42] Liang Zhao and Limin Wang. Task-specific inconsistency alignment for domain adaptive object detection. In *CVPR*, pages 14217–14226, 2022. 2, 3, 6, 7

[43] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *CVPR*, pages 13766–13775, 2020. 1, 2, 3, 4, 6, 7

[44] Wenzhang Zhou, Dawei Du, Libo Zhang, Tiejian Luo, and Yanjun Wu. Multi-granularity alignment domain adaptation for object detection. In *CVPR*, pages 9581–9590, 2022. 2, 3, 6, 7, 8

[45] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *CVPR*, pages 687–696, 2019. 7