Differential Alignment for Domain Adaptive Object Detection

Xinyu He, Xinhui Li, Xiaojie Guo*

College of Intelligence and Computing, Tianjin University, Tianjin, China
xy_he68@tju.edu.cn, lixinhui@tju.edu.cn, xj.max.guo@gmail.com

Abstract

Domain adaptive object detection (DAOD) aims to generalize an object detector trained on labeled source-domain data to a target domain without annotations, the core principle of which is *source-target feature alignment*. Typically, existing approaches employ adversarial learning to align the distributions of the source and target domains as a whole, barely considering the varying significance of distinct regions, say instances under different circumstances and foreground *vs* background areas, during feature alignment. To overcome the shortcoming, we investigate a differential feature alignment strategy. Specifically, a prediction-discrepancy feedback instance alignment module (dubbed PDFA) is designed to adaptively assign higher weights to instances of higher teacher-student detection discrepancy, effectively handling heavier domain-specific information. Additionally, an uncertainty-based foreground-oriented image alignment module (UFOA) is proposed to explicitly guide the model to focus more on regions of interest. Extensive experiments on widely-used DAOD datasets together with ablation studies are conducted to demonstrate the efficacy of our proposed method and reveal its superiority over other SOTA alternatives.

Code — https://github.com/EstrellaXyu/Differential-Alignment-for-DAOD

Introduction

As one of the fundamental tasks in computer vision, object detection plays a vital role in a wide spectrum of applications, such as autonomous driving (Li, Chen, and Shen 2019), video surveillance (Nascimento and Marques 2006), and person re-identification (Ye et al. 2021), to name a few. With the significant development of deep learning and the availability of large-scale annotated datasets, it has experienced remarkable performance improvements in recent years (Girshick 2015; Ren et al. 2017; Redmon et al. 2016; Carion et al. 2020; Zhao et al. 2024). Despite such advancements, the detection environment is not always as expected in practice, leading to performance discrepancies between the training and testing domains, *a.k.a. the domain shift issue* (Chen et al. 2018; Wang et al. 2021). This shift, driven by variations in illumination, weather, background, and other factors, frequently degrades the performance of detection. To mitigate this issue, a straightforward way is to acquire and annotate ample and diverse real-world data for training. Nevertheless, even if this manner were possible, it would be extremely laborious and time-consuming.

As a more appealing option, domain adaptive object de-

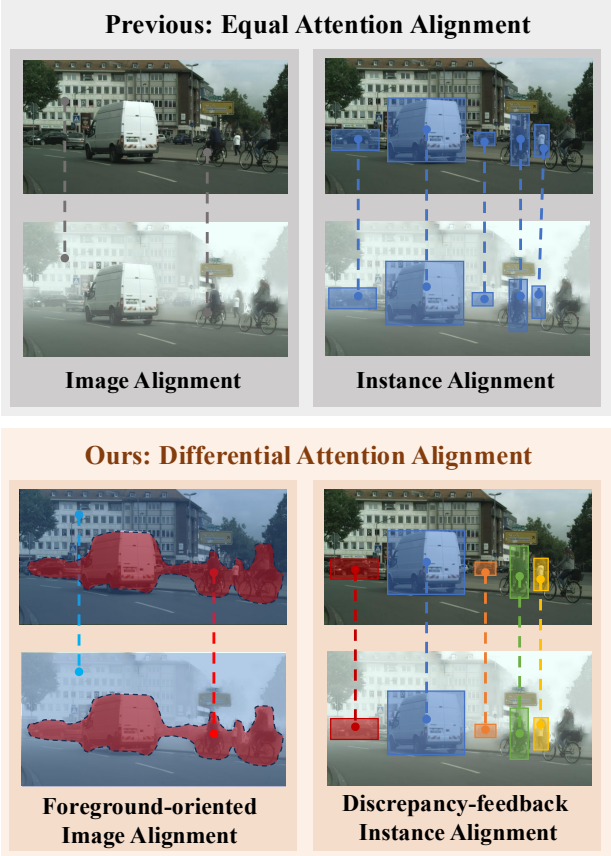

Figure 1: Different from previous methods adopting equal attention feature alignment (**upper part**), our design manipulates features from the backbone and ROI head with differential attention (**lower part**). Different colors represent different alignment weights/attentions.

*Corresponding author

tection (DAOD) (Li et al. 2022b; Cai et al. 2019; Chen et al. 2022; Wang et al. 2021; Huang et al. 2022; Zhao et al. 2023; Chen et al. 2018) has emerged to solve the challenge by following the principle of source-target feature alignment. The goal is to generalize the model trained on labeled source-domain data to the unlabeled target domain, thus alleviating the reliance on target-domain annotations. In recent years, several methods (Chen et al. 2018; Li et al. 2022b; Wang et al. 2021; Huang et al. 2022; Zhao et al. 2023) have been proposed in the literature, which employ adversarial learning to align the distributions of the source and target domains as a whole. Despite having made advancements, they hardly account for *the varying importance of different regions* in feature alignment. For example, foreground objects like cars and persons should be prior to background elements such as roads and sky; and instances in different haze densities should be treated unequally during alignment. Thus, it is desired to design a *differential alignment* mechanism that can effectively adjust attention to regions of different importance, instead of treating all the regions uniformly.

To achieve the goal, we delve into the differential alignment strategy via investigating cues to prioritize the treatment of critical features during the adaptation. This study presents two modules to respectively cope with (1) instances under different circumstances, and (2) foreground *vs* background areas. More concretely, a Prediction-Discrepancy Feedback instance-level Alignment module (PDFA), is proposed to dynamically adjust the model's attention on different instances. It can automatically assess the amount of domain-specific information contained in a given instance, which in turn determines the level of alignment effort required for that instance. In addition, an Uncertainty-based Foreground-Oriented image-level Alignment module (UFOA) is customized to explicitly guide the model to concentrate more on foreground regions than background ones. In other words, UFOA can sense regions of higher interest/importance and concern more on such regions than the others to overcome the deficiency of traditional feature alignment, *i.e.*, identical attention on foreground and background features. Please see Fig. 1 for illustration. Through these two designs, our method can address the varying importance of different regions in feature alignment with significant performance gains over other SOTA alternatives.

Our primary contributions can be summarized as follows:

- We propose an instance-level alignment module (PDFA) to dynamically adjust the focus of model based on the prediction-discrepancy feedback, which enables differential alignment of different instances according to their richness in domain-specific information.
- We develop an image-level alignment module (UFOA) to adaptively prioritize foreground-object areas, which mitigates the limitation of traditional equal feature alignment by balancing foreground and background through an uncertainty factor.
- Extensive experiments are conducted to verify the efficacy of our design, and show that our proposed method remarkably outperforms existing approaches by more than 4% in AP_{50} on three DAOD benchmarks.

Related Work

In recent years, various DAOD schemes have been devised to alleviate the performance degradation when applying models trained on a source domain to data from another domain with different distributions. Existing methods can be roughly grouped into three classes, *i.e.*, domain translation, self-training, and domain alignment approaches. Domain translation methods realize the adaption purpose by transforming the source-domain data to resemble the target domain, typically adopting generative models. Although valid in aligning distributions to some extent, these schemes are fundamentally constrained by information loss, and substantial computational requirement, limiting their applicability and performance in comparison with approaches in the other two categories. Given the focus of this work, we will primarily review representative methods in self-training and domain alignment families.

Self-Training with Pseudo Labels. Methods in this group, typically built upon teacher-student architectures, utilize pseudo labels for unlabeled data, whose effectiveness has been witnessed by several works. To be specific, AT (Li et al. 2022b) arms the self-training paradigm with weak-strong data augmentation, yielding noticeable increase in accuracy. While PT (Chen et al. 2022) further leverage the uncertainty of pseudo-labels to promote adaptation during training. Alternatively, new advances in general object detection have also promoted the development of DAOD. As a consequence, MTTrans (Yu et al. 2022) constructs an end-to-end cross-domain detection Transformer based on the teacher-student framework. From the perspective of capturing context relations within target-domain images, MIC (Hoyer et al. 2023) and MRT (Zhao et al. 2023) introduce the masked image consistency into the self-training framework to generate high-quality pseudo-labels. Besides, HT (Deng et al. 2023) enhances the quality of pseudo labels via regularizing the consistency of classification and localization scores. ALDI (Kay et al. 2024) introduces a unified benchmarking and implementation framework built upon the teacher-student framework. Although significant progress has been made, the aforementioned methods overlook meaningful information conveyed by different degrees of discrepancy between predictions. That is to say, greater discrepancy of prediction shall reflect heavier domain-specific information existed in corresponding regions. Our approach harnesses this potential by designing a prediction-discrepancy feedback mechanism.

Domain Alignment via Adversarial Learning. Domain alignment approaches advocate the use of adversarial learning to align distributions across domains. As two classic works, DA-Faster (Chen et al. 2018) and SADA (Chen et al. 2021) propose image-level and instance-level alignment modules to alleviate the domain discrepancy. In addition to strong data augmentation, AT (Li et al. 2022b) also leverages image-level domain adaptive learning. MGA (Zhang et al. 2024) introduces a unified multi-granularity alignment-based detection framework to learn domain-invariant representations and explore the relationship between features of different granularities in alignment. More recently, REACT (Li et al. 2024) adaptively compensates the extracted

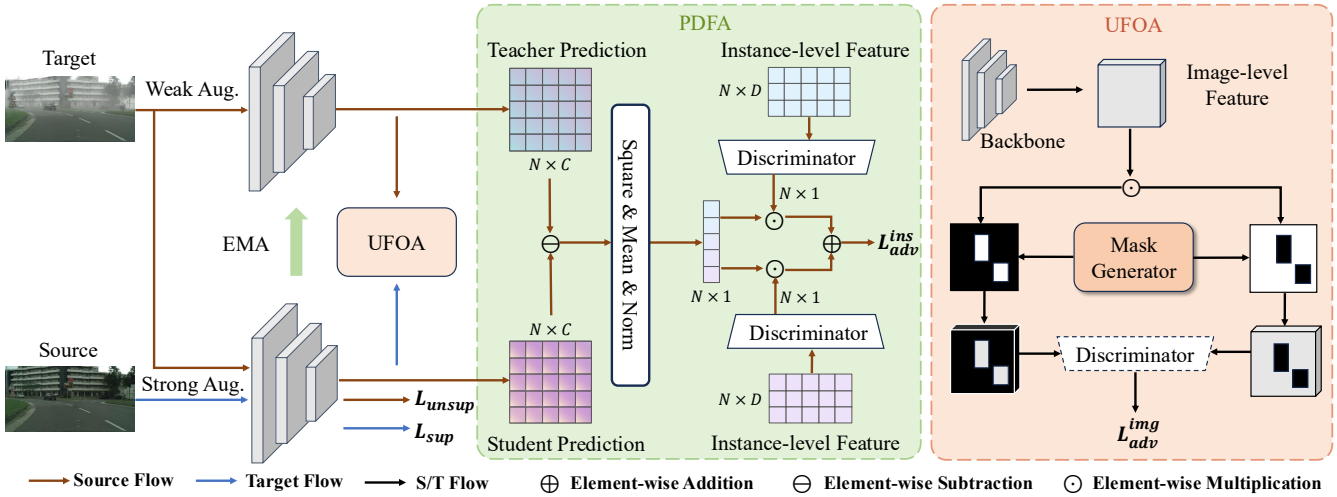

Figure 2: **Overview of our method.** Our approach is built upon the adaptive teacher-student framework. PDFA adjusts weights to different instances with respect to the discrepancy between predictions of the teacher and the student, while UFOA consists of a mask generator and an image-level discriminator. The mask generator produces a foreground-indicating mask to roughly separate the features of the last stage of the FPN into foreground and background parts.

features with the remainder features for generating task-relevant features. Among Def DETR-based (Zhu et al. 2020) attempts, SFA (Wang et al. 2021), O^{2}net (Gong et al. 2022) and MRT (Zhao et al. 2023) adopt domain query-based feature alignment, which is specially designed for domain adaptation of detection transformers. O^{2}net introduces object-aware alignment to emphasize foreground regions containing objects. Similar but different, our proposed strategy takes into account not only foreground but also background information with varying weights, as the alignment of background information is also indispensable.

Motivated by the above, this work designs a differential feature alignment strategy to adjust the treatment of features in light of the varying importance of different regions.

Methodology

Schematic Overview

In the context of DAOD, suppose we have a set of N_s images with category and object bounding-box labels, *i.e.*, $\mathcal{D}_s = \{(\mathbf{X}_1, \mathbf{L}_1), ..., (\mathbf{X}_{N_s}, \mathbf{L}_{N_s})\}$, from the source domain, and another set of N_t unlabeled images, say $\mathcal{D}_t = \{\mathbf{Y}_1, ..., \mathbf{Y}_{N_t}\}$, from the target domain. The goal of DAOD is to enhance the performance of detection on the target domain by leveraging labeled $\mathcal{D}_s$ and unlabeled $\mathcal{D}_t$. Due to the rationality, the adaptive teacher-student paradigm (Chen et al. 2018) with several successful follow-ups (Li et al. 2024; Zhao et al. 2023) has become popular in the field, which mainly adopts the self-training framework in conjunction with adversarial learning. Our proposed method also follows this technical route, as schematically illustrated in Fig. 2. Before launching our contributions, the core components of the paradigm shall be briefed for clarity.

Self-training framework. As can be seen from Fig. 2, the teacher $\mathcal{T}$ and student networks $\mathcal{S}$ share the backbone and detector structures. The teacher processes each weakly augmented target image to generate pseudo-labels, while the student is optimized based on the supervision by the ground-truth labels of source-domain data and the pseudo labels of target-domain samples. The teacher is then updated by Exponential Moving Average (EMA) (Arpit et al. 2022) from the student in the following manner:

$$\Theta_{\mathcal{T}} \leftarrow \alpha\Theta_{\mathcal{T}} + (1-\alpha)\Theta_{\mathcal{S}}, \tag{1}$$

where $\Theta_{\mathcal{T}}$ and $\Theta_{\mathcal{S}}$ denote the learnable parameters of $\mathcal{T}$ and $\mathcal{S}$, respectively. In addition, the hyper-parameter $\alpha \in [0, 1]$ designates the smoothing factor of EMA. In our experiments, setting α to 0.9996 works sufficiently well.

Discriminators for adversarial alignment. In addition to the teacher-student framework, domain discriminators are utilized to facilitate the alignment of feature distributions between the source and target domains. Specifically, the domain discriminators are strategically positioned after certain components to distinguish whether the feature is from the source domain or the target, which executes an adversarial learning process. In our implementation, we adopt Faster R-CNN as the detector, and set up two discriminators for the backbone and ROI head to assist with image-level and instance-level alignment, respectively.

Although previous approaches have incorporated the aforementioned strategies, they usually overlook that different regions within an image contain varying amounts of domain-specific information. The failure to consider this aspect results in suboptimal alignment, particularly when domain-specific features differ across various spatial locations within an image. To address this limitation, this study builds two key modules to flexibly alter the alignment attention at two different levels, including an adaptive prediction discrepancy feedback instance-level alignment (PDFA), and an uncertainty-based foreground-guided image-level align-

ment (UFOA), as depicted in Fig. 2. The subsequent sections will detail these two designs.

Prediction-Discrepancy Feedback Alignment

Since the instance features generated by the ROI head of Faster R-CNN detector are the most proximal image features to final detection results, aligning these features seems to be direct and rational. Again, we emphasize that different instance regions may contain varying amounts of domain-specific information. Figure 3 offers two examples, from which we can see that the fog density surrounding different cars in the images considerably differs. It is reasonable to deem that different instances should receive varying strengths of alignment. To this end, a prediction-discrepancy feedback module is introduced to automatically identify how rich domain-specific information appears in a certain instance region, and determine how much alignment attention to pay on this instance accordingly.

Prediction-discrepancy feedback. To automatically recognize instances with rich domain-specific information, we attempt to measure the prediction discrepancy between the teacher and student models on the same instances. Let us take a closer look at the right column of Fig. 3. The red proposals stand for those of heavy prediction discrepancy between the teacher and student models, while the blue ones are of light discrepancy. As can be observed, those areas embracing more inconsistently predicted proposals tend to be of richer domain-specific information (denser fog in these cases). In the sequel, we guide the instance-level alignment by the measurement of instance-wise prediction discrepancy, making the model concentrate more on the alignment of regions with greater prediction discrepancies. Owing to the simplicity, this work directly employs the classification map as the prediction map. Given N instance candidates and C classes in total, the prediction discrepancy matrix $\mathbf{P}_{\text{div}} \in \mathbb{R}^{N\times C}$ can be simply calculated as follows:

$$\mathbf{P}_{\text{div}} = \text{Square}(\mathbf{P}_{\mathcal{T}} - \mathbf{P}_{\mathcal{S}}), \tag{2}$$

where $\mathbf{P}_{\mathcal{T}} \in \mathbb{R}^{N\times C}$ and $\mathbf{P}_{\mathcal{S}} \in \mathbb{R}^{N\times C}$ represent the prediction classification maps derived from the teacher and student models, respectively.

Weighted instance-level alignment. Having $\mathbf{P}_{\text{div}}$ computed, we come to the construction of instance-wise alignment attention. The initial weight $\mathbf{w}_{\text{ins}} \in \mathbb{R}^{N\times 1}$ can be simply obtained by:

$$\mathbf{w}_{\text{ins}} = \frac{1}{C}\sum_{c}^{C} \mathbf{P}_{\text{div}}(:, c). \tag{3}$$

By further applying the min-max normalization on $\mathbf{w}_{\text{ins}}$, the weight is restricted into the range of $[0, 1]$ as follows:

$$\tilde{\mathbf{w}}_{\text{ins}} = \frac{\mathbf{w}_{\text{ins}} - \min(\mathbf{w}_{\text{ins}})}{\max(\mathbf{w}_{\text{ins}}) - \min(\mathbf{w}_{\text{ins}})}. \tag{4}$$

Moreover, the instance-wise adversarial loss is computed by:

$$\mathbf{f}_{\text{ins}} = -\mathbf{d} \odot \log(\mathfrak{D}_1(\mathbf{F}_{\text{ins}})) - \bar{\mathbf{d}} \odot \log(1 - \mathfrak{D}_1(\mathbf{F}_{\text{ins}})), \tag{5}$$

where $\odot$ means Hadamard product. Moreover, $\mathbf{d} \in \{0, 1\}^{N\times 1}$ is the domain flag vector, and $\bar{\mathbf{d}}$ is the complement version of $\mathbf{d}$. $\mathfrak{D}_1(\cdot)$ represents the instance discriminator whose input is the collection of N instance candidates

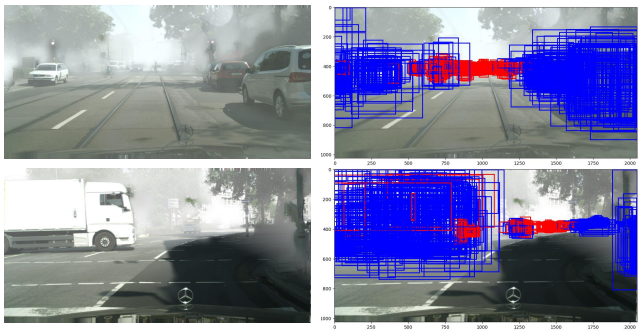

Figure 3: The proposals with top 2% prediction discrepancies are marked in red, while the rest are colored in blue.

$\mathbf{F}_{\text{ins}} \in \mathbb{R}^{N\times D}$. Its outputs the domain discriminating results $\mathfrak{D}_1(\mathbf{F}_{\text{ins}}) \in \mathbb{R}^{N\times 1}$. Combining Eqs. (4) and (5) yields final weighted instance-level adversarial loss:

$$\mathcal{L}_{adv}^{ins} = \|\tilde{\mathbf{w}}_{\text{ins}} \odot \mathbf{f}_{\text{ins}}\|_1, \tag{6}$$

where $\|\cdot\|_1$ represents the ℓ_1 norm.

Uncertain-based Foreground-Oriented Alignment

As previously discussed, another functionality is required to explicitly guide the model by prioritizing the alignment of regions containing foreground objects. This part details our foreground-oriented alignment scheme. It comprises a mask generator for (approximately) indicating foreground and background areas of an image, and splitting the feature maps into two parts accordingly for subsequent discrimination and reweighting operations.

Mask generator. It is not difficult to divide images into foreground and background areas, if with the help of ground-truth annotations. However, for target-domain data, the ground truths are absent. Alternatively, we resort to the pseudo labels generated by the teacher model. Figure 4 exhibits an example marked with the generated pseudo labels. We can observe that, although these pseudo bounding boxes may be inaccurate, their union can largely zone foreground areas. Hence, we form the foreground mask $\mathbf{M}$ by utilizing the union of detection boxes: for the labeled source domain, ground-truth bounding boxes are used to construct the mask, *i.e.* the elements within the regions enclosed by these bounding boxes are set to 1 (and 0 otherwise), while for the unlabeled target domain, detected bounding-boxes by the teacher model (pseudo labels) are employed to accomplish the mask. As a consequence, the image feature $\mathbf{F}_{\text{img}}$ can be split simply via:

$$\mathbf{F}_{\text{img}}^{fg} = \mathbf{M} \odot \mathbf{F}_{\text{img}}, \quad \mathbf{F}_{\text{img}}^{bg} = \quad \bar{\mathbf{M}} \odot \mathbf{F}_{\text{img}}, \tag{7}$$

where the feature F_img refers to the P2 layer within the FPN, which is the highest-resolution feature map in the FPN, capturing fine-grained spatial details. $\mathbf{F}_{\text{img}}^{fg}$ and $\mathbf{F}_{\text{img}}^{bg}$ correspond to foreground and background feature parts, respectively. In addition, $\bar{\mathbf{M}}$ represents the complementary of $\mathbf{M}$.

Uncertainty-based image-level alignment. It is worth noting that, different from O^{2}net (Gong et al. 2022), our UFOA

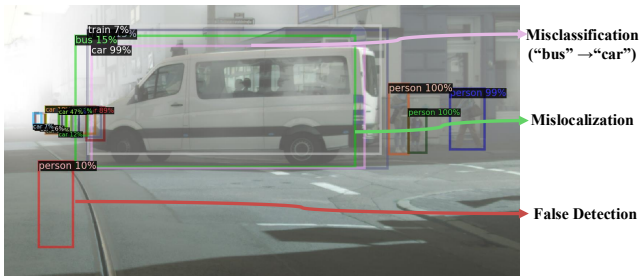

Figure 4: Visualization of pseudo labels generated by the teacher model. Despite misclassification, mislocalization and false detection errors exist, the union of these inaccurate bounding boxes can still largely indicate foreground areas.

module retains background regions along with regions of interest, which are both fed into the discriminator. This operation ensures that while foreground information remains the primary focus during alignment, background information is also considered, preventing it from being entirely discarded. Subsequently, an uncertainty factor is introduced to balance the relative importance of these two types of information during image-level alignment. The respective adversarial losses on the separated regions of interest and background regions are obtained through:

$$\begin{aligned}\mathcal{L}_{adv}^{fg} &= -d\log(\mathfrak{D}_2(\mathbf{F}_{\text{img}}^{fg})) - \bar{d}\log(1-\mathfrak{D}_2(\mathbf{F}_{\text{img}}^{fg})),\\ \mathcal{L}_{adv}^{bg} &= -d\log(\mathfrak{D}_2(\mathbf{F}_{\text{img}}^{bg})) - \bar{d}\log(1-\mathfrak{D}_2(\mathbf{F}_{\text{img}}^{bg})).\end{aligned} \tag{8}$$

Note that $\mathfrak{D}_2(\cdot)$ is the image-level discriminator, which takes the feature $\mathbf{F}_{\text{img}}$ as input and outputs the domain discriminating result $\mathfrak{D}_2(\mathbf{F}_{\text{img}}^{f}) \in [0,1]$ for the entire image-level feature. Here, $d \in \{0,1\}$ is a domain flag. Finally, the adversarial loss at the image level is a weighted sum of the above two terms as:

$$\mathcal{L}_{adv}^{img} = \gamma\mathcal{L}_{adv}^{fg} + (1-\gamma)\mathcal{L}_{adv}^{bg}. \tag{9}$$

By tuning the hyper-parameter uncertainty factor γ, we can modulate the relative emphasis on aligning these two components at the image level, thereby facilitating differential alignment. Notably, when setting $\gamma = 1$ in Eq. (9), only foreground regions are taken into account in the image-level alignment, as the alignment pattern in (Gong et al. 2022). By contrast, our alignment pattern is a balanced foreground-background alignment, the superiority of which will be validated in the ablation study.

Overall Objective Function

Combining all the presented parts, the overall objective function turns out to be:

$$\max_{\mathfrak{D}_1,\mathfrak{D}_2}\min_{\mathfrak{G}} \mathcal{L}_{sup} + \mathcal{L}_{unsup} + \lambda(\mathcal{L}_{adv}^{ins} + \mathcal{L}_{adv}^{img}), \tag{10}$$

where $\mathfrak{G}$ designates the feature extractor in the network. It includes all learnable components except for the two discriminators. In our experiments, we set λ as 0.01 by default. To be clear, the supervised detection loss $\mathcal{L}_{sup}$ is calculated with respect to ground truth labels, while $\mathcal{L}_{unsup}$ receives supervision from pseudo labels as:

$$\begin{aligned}\mathcal{L}_{sup} &= \mathcal{L}_{cls}^{s}(\tilde{c}_s, c_s) + \mathcal{L}_{reg}^{s}(\tilde{b}_s, b_s),\\ \mathcal{L}_{unsup} &= \mathcal{L}_{cls}^{t}(\tilde{c}_t, \hat{c}_t) + \mathcal{L}_{reg}^{t}(\tilde{b}_t, \hat{b}_t),\end{aligned} \tag{11}$$

where c, $\tilde{c}$, and $\hat{c}$ stand for the classification ground truth of source domain, the classification result and pseudo-classification label, respectively. The same principle applies to the bounding-box symbols b, $\tilde{b}$, and $\hat{b}$. $\mathcal{L}_{cls}$ adopts the cross-entropy loss for classification, and $\mathcal{L}_{reg}$ is the ℓ_1 loss used for regression.

Experimental Validation

Datasets

Following recent DAOD approaches (Deng et al. 2023; Zhao et al. 2023; Li et al. 2024), we evaluate our method on three widely-used benchmarks. Specifically, we perform adaptation experiments on three common scenarios: (1) weather adaptation with Cityscapes → Foggy Cityscapes, (2) synthetic to real adaptation with Sim10k → Cityscapes, and (3) small to large-scale dataset adaptation with Cityscapes → BDD100K-daytime.

Cityscapes (Cordts et al. 2016) comprises 2,975 training images and 500 validation images, covering various urban environments and traffic conditions. For our experiments, the semantic segmentation labels provided by Cityscapes are converted into bounding box annotations, allowing the dataset to be repurposed for object detection tasks.

Foggy Cityscapes (Sakaridis et al. 2018) is a synthetic dataset rendered from Cityscapes with three levels of foggy density (0.005, 0.01, 0.02). We use the highest density (0.02) as the target domain for a fair comparison following existing methods (Li et al. 2024; Zhang et al. 2024; Zhao et al. 2023).

Sim10k (Johnson-Roberson et al. 2016) contains 10,000 images rendered from GTA engine. In one of our adaptation experiments, Sim10k is used as the source domain, while the Cityscapes dataset, representing real-world scenes, serves as the target domain. The experiment focuses specially on the detection of 'car' category, aiming to evaluate the model's ability to adapt from synthetic to real-world scenarios.

BDD100k-daytime (Yu et al. 2020) is a subset of the larger BDD100k dataset, specially designed for daytime scenarios. It contains 36,728 training images and 5,258 validation images, which provides a diverse environment that mirrors real-world daytime driving, making it crucial for evaluating and improving detection models across domains.

Implementation Details

Network architecture. We use Faster R-CNN (Girshick 2015) as our base detector, with ResNet-50 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009) and Feature Pyramid Network (Lin et al. 2017) as the backbone. The teacher model is updated only by EMA from the student model, and we freeze the first two layers of the ResNet backbone network. Our implementation is based on the PyTorch framework and the model is trained on 4 NVIDIA RTX3090 GPUs with 24 GB of memory each.

Method	Detector	person	rider	car	truck	bus	train	mcycle	bicycle	AP_{50}^{val}
FCOS (Tian et al. 2020)	FCOS	29.6	26.3	37.1	7.9	14.1	6.3	12.9	28.1	20.3
EPM (Hsu et al. 2020)	FCOS	41.9	38.7	56.7	22.6	41.5	26.8	24.6	35.5	36.0
SIGMA (Li et al. 2022a)	FCOS	44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2
CSDA (Gao et al. 2023)	FCOS	46.6	46.3	63.1	28.1	56.3	53.7	33.1	39.1	45.8
HT (Deng et al. 2023)	FCOS	52.1	55.8	67.5	32.7	55.9	49.1	40.1	50.3	50.4
Def DETR (Zhu et al. 2020)	Def DETR	37.7	39.1	44.2	17.2	26.8	5.8	21.6	35.5	28.5
SFA (Wang et al. 2021)	Def DETR	46.5	48.6	62.6	25.1	46.2	29.4	28.3	44.0	41.3
O^2net (Gong et al. 2022)	Def DETR	48.7	51.5	63.6	31.1	47.6	47.8	38.0	45.9	46.8
MRT (Zhao et al. 2023)	Def DETR	<u>52.8</u>	51.7	<u>68.7</u>	<u>35.9</u>	<u>58.1</u>	<u>54.5</u>	41.0	47.1	51.2
Faster R-CNN (Girshick 2015)	FRCNN	26.9	38.2	35.6	18.3	32.4	9.6	25.8	28.6	26.9
DA-Faster (Chen et al. 2018)	FRCNN	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0
SADA (Chen et al. 2021)	FRCNN	48.5	52.6	62.1	29.5	50.3	31.5	32.4	45.4	44.0
PT (Chen et al. 2022)	FRCNN	40.2	48.8	63.4	30.7	51.8	30.6	35.4	44.5	42.7
AT* (Li et al. 2022b)	FRCNN	43.7	54.1	62.3	31.9	54.4	49.3	35.2	<u>47.9</u>	47.4
MIC (Hoyer et al. 2023)	FRCNN	50.9	55.3	67.0	33.9	52.4	33.7	40.6	47.5	47.6
MGA (Zhang et al. 2024)	FRCNN	47.0	54.6	64.8	28.5	52.1	41.5	40.9	49.5	47.4
REACT (Li et al. 2024)	FRCNN	52.1	<u>57.1</u>	66.3	35.0	56.7	52.8	<u>42.9</u>	<u>53.8</u>	<u>52.1</u>
Ours	FRCNN	**59.8**	**62.8**	**73.7**	**40.3**	**59.4**	**56.1**	**47.8**	**58.3**	**57.3**

Table 1: Results on adaptation from Cityscapes to Foggy Cityscapes (0.02). AT* denotes that the results of AT on *Foggy (0.02)* are taken from (Zhao et al. 2023). The best results are in **bold**, while the second-best results are <u>underlined</u>.

Data augmentations. We employ common weak augmentation techniques in DAOD, including RandomContrast, RandomBrightness, RandomSaturation, RandomGrayscale, and RandomBlur. For the source images, strong augmentation consists of the weak augmentations combined with the RandomErasing (DeVries and Taylor 2017) method, while for target domain images, it includes the weak augmentation along with the MIC (Hoyer et al. 2023) method.

Optimization. We optimize the network using the SGD optimizer with a momentum of 0.9. The initial learning rate is set to 0.01 and decreases in the final iteration. We use a batch size of 32, consisting of 16 labeled source images and 16 unlabeled target ones, and train the network for 25k iterations in total, including 10,000 iterations for burn-in and 15,000 iterations for teacher-student mutual learning.

Evaluation metric. We assess adaptation performance by reporting mean Average Precision (mAP) with an IoU threshold of 0.5, following standard protocols on the three aforementioned benchmarks. The results for prior methods are based on the values reported in their original papers. [1]

Comparisons with Other Methods

Adaptation from normal to foggy weather. The adaptation results of Cityscapes to Foggy Cityscapes are shown in Table 1. Ours achieves the best AP_{50} of 57.3% and surpasses the second best REACT (Li et al. 2024) with 52.1% by a margin of 5.2%. Notably, our method demonstrates a marked improvement in the first three categories: person (+7.7%), rider (+5.7%), and car (+7.4%). The significant performance gain in the long-tail category 'rider' further underscores the efficacy of our differential alignment strategy, which enables the model to better align regions that have a greater impact on detection results.

Adaptation from small to large-scale dataset. Table 2 presents the adaptation performance from Cityscapes to BDD100K. Our method achieves 45.8% AP_{50}, outperforming all baseline models by a considerable margin. Specifically, it outperforms the best performing one-stage adaptive detector, HT, by 5.6% AP_{50}, and exceeds the two-stage DAOD approach REACT (Li et al. 2024), by 10.0% AP_{50}. Although our method does not surpass the current state-of-the-art in the 'bicycle' category, it achieves the second-best performance, demonstrating competitive effectiveness. Despite this, our method shows superior performance in most categories, underscoring its robustness and generalizability.

Adaptation from synthetic to real images. We investigate the synthetic-to-real domain adaptation scenario, specifically evaluating the adaptation from Sim10K to Cityscapes, with results presented in Tab. 3. Our method achieves 69.7% AP_{50}, surpassing previous state-of-the-art method by a margin of 4.2%. Sim-to-Real adaptation task requires transferring the semantics of a synthetic scene from the GTA engine to a real scene, and our performance improvement on this task also demonstrates robustness in the face of synthetic scenarios.

Ablation Studies

Ablation on network components. We conduct ablation studies to evaluate the significance of key network components, as shown in Tab. 4. We replace PDFA and UFOA with standard instance-level and image-level alignment modules, which employ an equal alignment strategy. The replacements lead to 2.1% and 1.1% decreases in AP_{50}, respec-

[1]Due to discrepancies in experimental settings between ALDI (Kay et al. 2024) and other DAOD works, such as the use of COCO (Lin et al. 2014) pre-trained weights in ALDI, we did not perform a direct comparison here.

Method	Detector	person	rider	car	truck	bus	mcycle	bicycle	AP_{50}^{val}
EPM (Hsu et al. 2020)	FCOS	39.6	26.8	55.8	18.8	19.1	14.5	20.1	27.8
SIGMA (Li et al. 2022a)	FCOS	46.9	29.6	<u>64.1</u>	20.2	23.6	17.9	26.3	32.7
HT (Deng et al. 2023)	FCOS	<u>53.4</u>	<u>40.4</u>	63.5	<u>27.4</u>	<u>30.6</u>	<u>28.2</u>	**38.0**	<u>40.2</u>
Def DETR (Zhu et al. 2020)	Def DETR	38.9	26.7	55.2	15.7	19.7	10.8	16.2	26.2
SFA (Wang et al. 2021)	Def DETR	40.2	27.6	57.5	19.1	23.4	15.4	19.2	28.9
AQT (Huang et al. 2022)	Def DETR	38.2	33.0	58.4	17.3	18.4	16.9	23.5	29.4
O^2net (Gong et al. 2022)	Def DETR	40.4	31.2	58.6	20.4	25.0	14.9	22.7	30.5
MTTrans (Yu et al. 2022)	Def DETR	44.1	30.1	61.5	25.1	26.9	17.7	23.0	32.6
MRT (Zhao et al. 2023)	Def DETR	48.4	30.9	63.7	24.7	25.5	20.2	22.6	33.7
Faster R-CNN (Girshick 2015)	FRCNN	28.8	25.4	44.1	17.9	16.1	13.9	22.4	24.1
DA-Faster (Chen et al. 2018)	FRCNN	28.9	27.4	44.2	19.1	18.0	14.2	22.4	24.9
ICR-CCR-SW (Xu et al. 2020)	FRCNN	32.8	29.3	45.8	22.7	20.6	14.9	25.5	27.4
REACT (Li et al. 2024)	FRCNN	-	-	-	-	-	-	-	35.8
Ours	FRCNN	**61.4**	**45.4**	**75.4**	**33.0**	**36.2**	**29.5**	<u>36.7</u>	**45.8**

Table 2: Results on adaptation from Cityscapes to BDD100k-daytime. The best results are in **bold**, while the second-best results are <u>underlined</u>.

Method	Detector	carAP_{50}^{val}
FCOS (Tian et al. 2020)	FCOS	39.8
EPM (Hsu et al. 2020)	FCOS	49.0
SIGMA (Li et al. 2022a)	FCOS	53.7
HT (Deng et al. 2023)	FCOS	<u>65.5</u>
Def DETR (Zhu et al. 2020)	Def DETR	47.4
SFA (Wang et al. 2021)	Def DETR	52.6
O^2net (Gong et al. 2022)	Def DETR	54.1
MTTrans (Yu et al. 2022)	Def DETR	57.9
MRT (Zhao et al. 2023)	Def DETR	62.0
Faster R-CNN (Ren et al. 2017)	FRCNN	39.4
DA-Faster (Chen et al. 2018)	FRCNN	41.9
MeGA-CDA (Vs et al. 2021)	FRCNN	44.8
D-adapt (Jiang et al. 2021)	FRCNN	51.9
PT (Chen et al. 2022)	FRCNN	55.1
REACT (Li et al. 2024)	FRCNN	58.6
Ours	FRCNN	**69.7**

Table 3: Results on adaptation from Sim10k to Cityscapes (category 'car').

Baseline	Strong Aug	PDFA	UFOA	AP_{50}^{val}
✓				50.0
✓	✓			53.9
✓	✓	✓		56.2
✓	✓		✓	55.2
✓	✓	✓	✓	57.3

Table 4: Ablation of proposed modules on adaptation from Cityscapes to Foggy Cityscapes.

γ	0	0.5	0.8	1.0
AP_{50}^{val}	55.1	55.9	57.3	55.6

Table 5: Effect of the hyper-paramter γ in UFOA on adaptation from Cityscapes to Foggy Cityscapes.

tively, underscoring the effectiveness of our proposed differential alignment strategy. Additionally, we observe that MIC strong augmentation significantly improve the model's perception of target domain distributions. Specifically, the baseline consists of a teacher-student framework, image-level and instance-level alignment modules, without the differential attention mechanisms introduced by PDFA and UFOA.

Ablation on uncertainty factor γ. The results of different γ in Eq. (9) is shown in Tab. 5, which highlights the effectiveness of our foreground-oriented alignment module. When the value of γ is greater than 0.5, i.e. more attention is given to the foreground regions, the model obtains a greater performance improvement. But when the value of γ is equal to 1.0 (*i.e.* foreground alignment pattern in O^2net), the performance decreases due to the lack of alignment of the background information, validating the necessity of our balanced foreground-background alignment pattern.

Conclusion

In this paper, we tackled the issue of ineffective alignment in domain adaptive object detection by introducing two innovative modules: the adaptive Prediction-Discrepancy Feedback instance Alignment (dubbed PDFA) and the Uncertainty-based Foreground-Oriented image Alignment (UFOA). The PDFA module prioritizes instances with higher teacher-student prediction discrepancies, ensuring more accurate alignment of critical domain-specific features. Furthermore, the UFOA module guides the model's attention toward foreground regions, effectively mitigating the limitations of previous equal alignment strategies. Comprehensive evaluations on widely used DAOD datasets, along with ablation studies, have validated the effectiveness of our proposed method and demonstrated its significant advantages over other state-of-the-art approaches.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant numbers 62372251 and 62072327.

References

Arpit, D.; Wang, H.; Zhou, Y.; and Xiong, C. 2022. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *Advances in Neural Information Processing Systems*, 35: 8265–8277.

Cai, Q.; Pan, Y.; Ngo, C.-W.; Tian, X.; Duan, L.; and Yao, T. 2019. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11457–11466.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, M.; Chen, W.; Yang, S.; Song, J.; Wang, X.; Zhang, L.; Yan, Y.; Qi, D.; Zhuang, Y.; Xie, D.; et al. 2022. Learning domain adaptive object detection with probabilistic teacher. *arXiv preprint arXiv:2206.06293*.

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 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.

Chen, Y.; Wang, H.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2021. Scale-aware domain adaptive faster r-cnn. *International Journal of Computer Vision*, 129(7): 2223–2243.

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

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, J.; Xu, D.; Li, W.; and Duan, L. 2023. Harmonious teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23829–23838.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Gao, C.; Liu, C.; Dun, Y.; and Qian, X. 2023. Csda: Learning category-scale joint feature for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11421–11430.

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

Gong, K.; Li, S.; Li, S.; Zhang, R.; Liu, C. H.; and Chen, Q. 2022. Improving transferability for domain adaptive detection transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1543–1551.

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

Hoyer, L.; Dai, D.; Wang, H.; and Van Gool, L. 2023. MIC: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11721–11732.

Hsu, C.-C.; Tsai, Y.-H.; Lin, Y.-Y.; and Yang, M.-H. 2020. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, 733–748. Springer.

Huang, W.-J.; Lu, Y.-L.; Lin, S.-Y.; Xie, Y.; and Lin, Y.-Y. 2022. AQT: Adversarial Query Transformers for Domain Adaptive Object Detection. In *IJCAI*, 972–979.

Jiang, J.; Chen, B.; Wang, J.; and Long, M. 2021. Decoupled adaptation for cross-domain object detection. *arXiv preprint arXiv:2110.02578*.

Johnson-Roberson, M.; Barto, C.; Mehta, R.; Sridhar, S.; Rosaen, K.; and Vasudevan, R. 2016. Driving in the Matrix: Can Virtual Worlds Replace Human-Generated Annotations for Real World Tasks? *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*.

Kay, J.; Haucke, T.; Stathatos, S.; Deng, S.; Young, E.; Perona, P.; Beery, S.; and Van Horn, G. 2024. Align and Distill: Unifying and Improving Domain Adaptive Object Detection. *arXiv preprint arXiv:2403.12029*.

Li, H.; Zhang, R.; Yao, H.; Zhang, X.; Hao, Y.; Song, X.; and Li, L. 2024. REACT: Remainder Adaptive Compensation for Domain Adaptive Object Detection. *IEEE Transactions on Image Processing*.

Li, P.; Chen, X.; and Shen, S. 2019. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7644–7652.

Li, W.; Liu, X.; Yuan, Y.; and Bob. 2022a. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5291–5300.

Li, Y.-J.; Dai, X.; Ma, C.-Y.; Liu, Y.-C.; Chen, K.; Wu, B.; He, Z.; Kitani, K.; and Vajda, P. 2022b. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7581–7590.

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

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft

coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Nascimento, J. C.; and Marques, J. S. 2006. Performance evaluation of object detection algorithms for video surveillance. *IEEE Transactions on Multimedia*, 8(4): 761–774.

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

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(6): 1137–1149.

Sakaridis, C.; Dai, D.; Van Gool, H.; Katte, B.; and Luccy, L. 2018. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126: 973–992.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2020. FCOS: A simple and strong anchor-free object detector. *IEEE transactions on pattern analysis and machine intelligence*, 44(4): 1922–1933.

Vs, V.; Gupta, V.; Oza, P.; Sindagi, V. A.; and Patel, V. M. 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.

Wang, W.; Cao, Y.; Zhang, J.; He, F.; Zha, Z.-J.; Wen, Y.; and Tao, D. 2021. Exploring sequence feature alignment for domain adaptive detection transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1730–1738.

Xu, C.-D.; Zhao, X.-R.; Jin, X.; and Wei, X.-S. 2020. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11724–11733.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 2872–2893.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu, J.; Liu, J.; Wei, X.; Zhou, H.; Nakata, Y.; Gudovskiy, D.; Okuno, T.; Li, J.; Keutzer, K.; and Zhang, S. 2022. MTTrans: Cross-domain object detection with mean teacher transformer. In *European Conference on Computer Vision*, 629–645. Springer.

Zhang, L.; Zhou, W.; Fan, H.; Luo, T.; and Ling, H. 2024. Robust domain adaptive object detection with unified multi-granularity alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; and Chen, J. 2024. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16965–16974.

Zhao, Z.; Wei, S.; Chen, Q.; Li, D.; Yang, Y.; Peng, Y.; and Liu, Y. 2023. Masked retraining teacher-student framework for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19039–19049.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.