Dual Prototype-driven Objectness Decoupling for Cross-Domain Object Detection in Urban Scene

Taehoon Kim[1], Jaemin Na[2], Joong-won Hwang[3], Hyung Jin Chang[4], and Wonjun Hwang[1]

[1]Ajou University, Korea [2]Tech. Innovation Group, KT
[3]Electronics and Telecommunications Research Institute
[4]University of Birmingham, UK
th951113@ajou.ac.kr, jaemin.na@kt.com, jwhwang@etri.re.kr,
h.j.chang@bham.ac.uk, wjhwang@ajou.ac.kr

Abstract. Unsupervised domain adaptation aims to mitigate the domain gap between the source and the target domains. Despite domain shifts, we have observed intrinsic knowledge that spans across domains for object detection in urban driving scenes. First, it includes consistent characteristics of objects within the same category of extracted ROIs. Second, it encompasses the similarity of patterns within the extracted ROIs, relating to the positions of the foreground and background during object detection. To utilize these, we present DuPDA, a method that effectively adapts object detectors to target domains by leveraging domain-invariant knowledge to separable objectness for training. Specifically, we construct categorical and regional prototypes, each of which operates through their specialized moving alignments. These prototypes serve as valuable references for training unlabeled target objects using similarity. Leveraging these prototypes, we determine and utilize a boundary that trains separately the foreground and background regions within the target ROIs, thereby transferring the knowledge to focus on each respective region. Our DuPDA surpasses previous state-of-the-art methods in various evaluation protocols on six benchmarks.

Keywords: Unsupervised Domain Adaptation · Object Detection · Urban Scene · Domain-invariance · Objectness Decoupling

1 Introduction

Object detection is essential for autonomous driving, involving the detection of various objects such as cars, pedestrians, and traffic signs. However, the domain shift problem remains a significant challenge due to variations in lighting and weather conditions. These variations cause a degradation of the detection performance. To overcome this issue without the necessity of labeling datasets for each new domain, unsupervised domain adaptation (UDA) has emerged as an effective solution [9,11]. The UDA aims to minimize the discrepancy between the labeled source and the unlabeled target domain, enhancing the performance of the latter. A prime example of utilizing UDA can be seen in object detection within

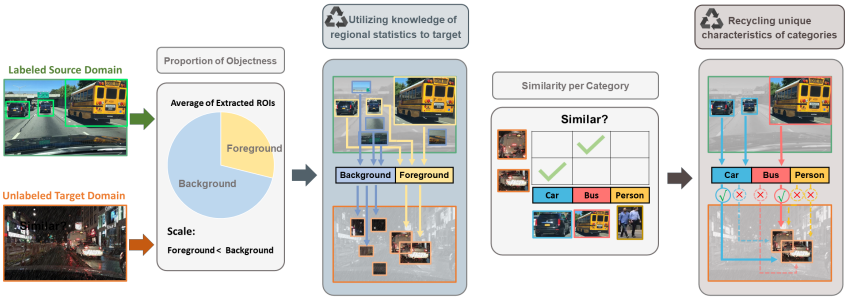

Fig. 1: DuPDA leverages domain-invariant regional and categorical knowledges to effectively train the target domain. We recycle each object's category-specific characteristics and utilize each region's proportion within ROI's position.

urban driving scenes. When trained with daytime driving scenes, often fail to perform under nighttime or adverse weather conditions. However, it is observed that the intrinsic characteristics of each object in driving scenes maintain their consistency, regardless of the domain changes; i.e., vehicles are typically found on roads and people on sidewalks, exhibiting predictable patterns. Moreover, the aspect ratio of objects, such as the horizontally elongated form of vehicles and the vertical orientation of people, remains unchanged whether day or night time.

For these characteristics, we introduce the dual prototype-driven objectness decoupling (DuPDA), which is applicable to two-stage object detectors in driving scenes. As in Fig. 1, our DuPDA is based on two key concepts. First, objects that need to be detected in urban scenes include consistent and unique characteristics in various domains. This necessitates the detector consistently extraction and recognizing objects of the same category, regardless of the domain. Therefore, effective training of the invariant characteristics of objects can facilitate adaptation to detection tasks across various domains without labels. Second, a two-stage detector is designed to extract a large number of Regions of Interest (ROI) from images, accommodating the variable number of objects to be detected. For this reason, most of these ROIs correspond to background regions rather than to the objects themselves. Although the source domain can easily identify these regions using its own labels, the absence of labels in the target domain makes it challenging for the detector to recognize them. Therefore, if the model can successfully distinguish both regions in the target domain, it can focus more on the target features corresponding to the actual objects.

In each observation, we construct two prototypes - categorical and regional - by capturing domain-invariant knowledge of objects to guide the training of the unlabeled target domain. Typically, the prototype encapsulates the common representative characteristics of each object. However, previous works [19, 45] only use category-wise prototype, which is solely accumulated by labeled source domain, to determine the category of each object in the target domain. In contrast, DuPDA uses two different prototypes to train the unlabeled target domain more effectively. Our categorical prototype adaptively differentiates the foreground re-

gions of the target ROIs. It helps assign each ROI target to its respective category by computing similarity. Furthermore, we introduce a regional prototype that plays a crucial role in controlling ROI regions. It regulates according to the statistical characteristics of the foreground and background regions within the extracted target ROIs, based on their positions and proportions. Note that we collectively named across both regions as 'objectness'.

However, generating prototypes solely from the source domain can interfere with bridging the gap with the target domain, causing performance degradation. To address this issue, we utilized the exponential moving average (EMA) to propose two moving alignment methods. This methods gradually transitions from the source to the target domain in both prototypes, making invariant characteristics within the domain which serve as effective guidance for the target domain. Using our prototypes, we generate the boundary for distinguishing the target ROIs into foreground and background regions, training each region through dedicated loss functions. This process enables the detector to focus more on each region, thus easily adapting and training effectively for the unlabeled target domains. We evaluated the effectiveness of our DuPDA in UDA-OD scenarios in the driving scene, using established evaluation metrics from previous research [36]. Our results demonstrate improved performance over existing state-of-the-art methods in different weather conditions, synthetic-to-real, and scene adaptation. In summary, the key contributions of our work are as follows:

- Our DuPDA introduces categorical and regional prototypes, each of which operates through their specialized moving alignments. This method generates domain-invariant prototypes that serve as a reference for target domain.
- Based on our prototypes, we propose objectness decoupling, which separates target ROIs into foreground and background regions, thus focusing on each region for effectively training the unlabeled target domain.
- DuPDA achieves competitive performance with previous works in six UDA-OD scenarios, with its effectiveness corroborated by a related ablation study.

2 Related Work

Urban Scene Analysis. Urban scene analysis has been actively studied for autonomous vehicles, and previous studies [6, 23] have mainly focused on the semantic segmentation of urban scenes. [23] suggested a perspective estimation network to learn the global perspective geometry of urban scenes. [6] proposed a reality-oriented adaptation approach to urban scene semantic segmentation by learning from synthetic data. Among the representative datasets used in urban scene analysis, the well-known datasets include Cityscapes [7], KITTI [13], Sim10k [20], and BDD100K [41]. In this paper, we measure object detection performance in UDA scenarios consisting of these urban scene datasets.

Unsupervised Domain Adaptation. In recent years, unsupervised domain adaptation (UDA) has garnered attention in diverse vision domains [9, 11, 17, 28, 30, 35]. Within object detection, various UDA methods have emerged. [5]

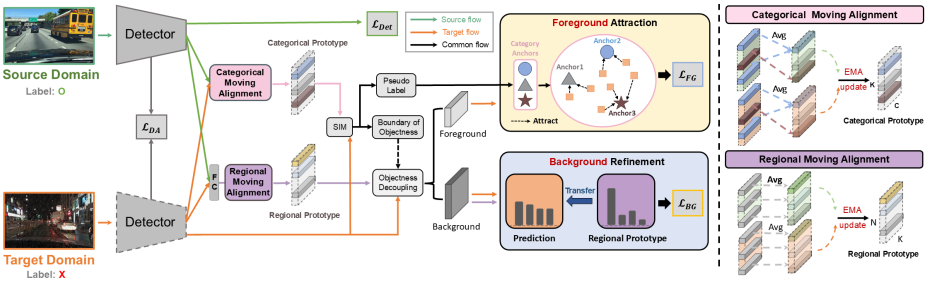

Fig. 2: DuPDA separates the target ROIs using prototypes, which are generated through the proposed alignments. By comparing target ROI features with categorical prototypes using similarity (SIM), we generate pseudo-labels and the separation boundary between foreground and background (objectness). Using this boundary, we split the target ROI scores and the regional prototypes and then calculate our proposed losses for each distinct region. **Categorical Moving Alignment:** Use EMA between accumulated ROI features of two domains to generate a categorical prototype for each object's category. **Regional Moving Alignment:** Use EMA between accumulated ROI scores of two domains to generate a regional prototype for each ROI position.

used adversarial approaches to ensure the consistency of the feature at the image and instance level. [30] focused on global dissimilarity through strong local and weak global alignments. [47] incorporates auxiliary predictors for classification and localization, leveraging their inconsistencies as indicators of domain specificity. In the context of distillation-based methods, such as [1,15], have incorporated similarity measurements and self-distillation techniques. Recent studies have shown performance improvements in UDA-OD with mean teacher framework [2, 4, 8, 10, 24, 50] and transformer [42, 44, 48]. Contrary to previous works, DupDA leverages the inherent domain-invariant knowledge. We introduce two distinct prototypes, which are used to decouple the foreground and background regions, focusing on each region separately to train the unlabeled target domain.

Prototype-based alignment. The prototype is a representative feature vector that captures the essential characteristics of a particular object category. In UDA-OD, prototypes are usually accumulated class-wise features from the labeled source features and aligned with target features to adapt the detector in the unlabeled target domain. [49] minimize the distance of the same category between both domains using global prototypes. [45] proposed RPN feature alignment using prototypes to generate pseudo-labels for proposals in target domain. [19] generate class-wise prototypes to aid contrast adaptation with unlabeled target features, providing pseudo-classes for the semantic segmentation. In previous works [19, 38, 39, 45, 49], they commonly used a fixed rate when adding a current feature to the prototype and only using the labeled source domain to solely construct the class-wise prototype. However, our method uses not only categorical but also regional prototype that accumulates statistical features of the foreground and background regions. We also propose specialized moving alignments to update each prototype for gradual transition from the source to the target domain by adaptively adjust the weight to grant domain-invariance.

3 Proposed Methods

3.1 Overview

As illustrates in Fig. 2, our DuPDA framework aims to recycle domain-invariant knowledge for gradual adaptation from the source domain (denoted as S) to the target domain (denoted as T). We reuse in two aspects: categorical and regional knowledge. For categorical knowledge, we observe that the unique characteristics of the object enable differentiation among various categories across different domains in the urban scene. To achieve this, we propose a categorical prototype, recycling ROIs that encompass all the necessary features to define objects. In addition to utilizing regional knowledge, we observe that the proportion of foreground and background regions within the extracted ROIs is consistently maintained across domains in urban scenes. Furthermore, there is a consistent tendency for the background region to occupy a larger proportion than the foreground. To accomplish this, we introduce a regional prototype which exploits the characteristics of each position-specific ROI region. Moreover, we employ EMA to adjust both prototypes, thereby improving their domain-invariance. Using these refined prototypes, we perform an objectness decoupling procedure to focus on each specific region within the target domain during training.

3.2 Categorical Moving Alignment

Utilizing our observation discussed in the previous sections, we generate categorical prototypes as a criterion to assign a category for ROIs in the unlabeled target domain. The prototype is obtained by ROI features based on their corresponding object categories. These ROI features are extracted from the detector's ROI-Align [14] output on the detector. Since the labeling information is available only for the source images, we can create a categorical prototype accumulating ROI features from the source domain that exceed a specified threshold τ_c, similar to the methods in [38, 49]. This categorical prototype of the source domain includes the distinct characteristics of the objects in each category. Therefore, each unlabeled target ROI can be classified into the most similar specific category, as also mentioned in previous prototype methods [19, 39, 49]. For this reason, we first simply generate this categorical prototype that uses only the source domain $P_{C_S} \in \mathbb{R}^{K \times C}$. Therefore, our DuPDA uses this prototype compared to the ROI features of the target domain $\Phi_T \in \mathbb{R}^{N \times C}$ by calculating the similarity map. K and N denote the number of categories and the number of ROIs, respectively, and C denote the channel size. The similarity map $S \in \mathbb{R}^{N \times K}$ is as follows:

$$S^{nk} = \frac{P_{C_S}^{k} \cdot \Phi_T^{n\top}}{||P_{C_S}^{k}||\ ||\Phi_T^{n}||}, \tag{1}$$

where n denote n-th ROI feature ($n \in N$), k denote k-th category ($k \in K$). Using this similarity to distinguish each unlabeled target ROI, because ROI features within the same category typically exhibit greater similarity as compared to ROI

features from disparate categories. Then, we assign a pseudo-label for the n-th target ROI feature using the similarity map, which is determined by identifying the index of the maximum similarity value in $S^n \in \mathbb{R}^{1\times K}$.

However, generating the categorical prototype using only source ROI features does not fully bridge the domain discrepancy with the target domain, causing performance degradation. To address this, we introduce categorical moving alignment (CMA). This method gradually transitions the categorical prototype from the source domain to the target domain by updating the confident target domain knowledge into the categorical prototype. To implement this, we construct the categorical prototype of the target domain, denoted P_{C_T}. This is achieved by accumulating only the target ROI features that have corresponding pseudo-labels with high confidence above the threshold τ_c. Subsequently, we apply the EMA to incrementally update the categorical prototype, thereby ensuring the domain-invariance of the prototype. The final P_{C_S} is as follows:

$$P_{C_S} \leftarrow \alpha \cdot P_{C_T} + (1-\alpha) \cdot P_{C_S}, \tag{2}$$

where α represents a hyperparameter for which we gradually increase the decay rate as the training iterations proceed. Since the target domain does not have labels in training process, it is relatively more challenging for the detector to extract ROI features compared to the source domain. Consequently, during the early iterations, there are fewer target ROI features accumulated above the threshold to the categorical prototype of the target domain than source domain, making it unstable to represent each target category. Therefore, a smaller value of α is required to more proportion to labeled source domain's prototype when updating the existing P_{C_S} with P_{C_T}. As the training progresses, α is gradually increased to give more proportion to unlabled target domain's prototype due to the greater quantity and accuracy of target ROI features accumulated in P_{C_T}. This progressive process generates a more domain-invariant prototype. Therefore, the prototype effectively minimizes the domain gap when calculating the similarity between the target ROI features and the categorical prototype.

3.3 Regional Moving Alignment

In the two-stage detector, the output ROIs exhibit a consistent pattern across different positions, regardless of datasets; in particular, the output ROIs often contain a higher proportion of background compared to foreground ROIs. To verify this, as shown in Fig. 3, we visualized the averaged background prediction scores, calculated at each index position for ROIs, from the model trained in a supervised manner on the source domain. As observed in the green line (i.e., source domains), the regions with high average background scores occupy a larger ROI range compared to the regions with low scores (i.e., foreground) and also seem clearly distinct. This statistical pattern remains consistent across different source domain datasets on various UDA-OD scenarios in terms of foreground and background regions. Note that the number of ROIs extracted by the detector remains constant [29]. In contrast, the orange line (i.e., target domains) does

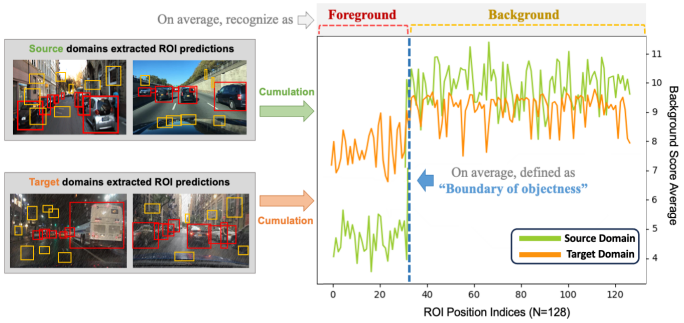

Fig. 3: To verify the consistent pattern related to label usage, we computed the average background classification scores (y-axis) for each position's ROI index (x-axis), extracted from a detector that was trained only on the source domain. It reveals distinct trend (index ≈ 37) - labeled source domains (green line) exhibit a clear division at a particular position, which is not observed in unlabeled target domains (orange line).

not relatively exhibit a clear statistical pattern. The source domain supervision signals enable the detector to precisely recognize the distinction between background and foreground regions. However, the lack of labeled data in the target domain makes it challenging to accurately determine whether an extracted ROI corresponds to a foreground or background region. To address this, we introduce a regional prototype designed to proficiently guide the detector in associating specific target ROI with their corresponding regions. By aggregating the ROIs that surpass the confidence threshold, we accumulate the ROIs retrieved for each specific position. This process yields the statistical correlation between each position and a particular region. When generating the regional prototype, We use the output category prediction scores from the ROI head for cumulative averaging, considering both the spatial positions within ROIs and their sequential orders. This ROI head, comprised of fully-connected layers, utilizes ROI features Φ as input, and produces outputs referred to as ROI scores ($\Psi \in \mathbb{R}^{N \times K}$). First, we generate a regional prototype for the source domain denoted as $P_{R_S} \in \mathbb{R}^{N \times K}$ and for the target domain denoted as $P_{R_T} \in \mathbb{R}^{N \times K}$. Next, we separately accumulate the ROI scores that surpass the threshold τ_c for each domain. However, updating the regional prototype exclusively by accumulating the source ROI scores may limit the performance in the target domains because of the domain shift. To address this, we propose regional moving alignment (RMA), akin to CMA, which gradually increase the proportion of each domain's regional prototype from the source to the target domain using confident target ROI scores. Therefore, using RMA, we can statistically determine if each ROI position is predominantly foreground or background. By using RMA, the final P_{R_S} is as follows:

$$P_{R_S} \leftarrow \beta \cdot P_{R_T} + (1 - \beta) \cdot P_{R_S}, \tag{3}$$

where β is a decay rate for RMA. Similar to CMA in Eq. 2, using RMA initial the small proportion of the target domain during the early train iterations, due to

Algorithm 1 Categorical and Regional Prototypes

Require: categorical regional prototype P_C; regional prototype P_R; ROI features Φ; ROI scores Ψ; threshold τ_c; domains d;
  for d **in** $\{source(S), target(T)\}$ **do**
    for $j = 1$ **to** $number\,of\,ROIs$ **do**
      $k = argmax(\Psi_d^j)$
      if $softmax(\Psi_d^j) > \tau_c$ **then** ▷ $P_{C_d}^k$ *is not empty*
        $P_{C_d}^k \leftarrow \frac{1}{2}(P_{C_d}^k + \Phi_d^j)$
      end if
      if $softmax(\Psi_d^j) > \tau_c$ **then** ▷ $P_{R_d}^j$ *is not empty*
        $P_{R_d}^j \leftarrow \frac{1}{2}(P_{R_d}^j + \Psi_d^j)$
      end if
    end for
  end for
  $P_{C_S} \leftarrow \alpha \cdot P_{C_T} + (1-\alpha) \cdot P_{C_S}$ ▷ CMA
  $P_{R_S} \leftarrow \beta \cdot P_{R_T} + (1-\beta) \cdot P_{R_S}$ ▷ RMA

the limited reliability of the target ROIs; then gradually increase for easily adapt to the target domain. As a result, P_{R_S} effectively encompasses domain-invariant characteristics, representing the distribution of ROI regions by position. Hence, it serves as a criterion for the proposed background refinement of target ROIs. The generating process of proposed prototypes are summarized in Algorithm 1.

3.4 Objectness Decoupling

Using our prototypes, we propose the objectness decoupling method that distinguishes ROIs of the target domain between the foreground and background regions. As mentioned in the previous sections, we observe that the ratio of each region within ROI features remains consistent across domains. However, during the training in the target domain, the labels that help to differentiate each region are not utilized, unlike in the labeled source domain. This can lead to potentially ambiguous output for the detector. Therefore, we suggest that providing guidance on the regions in the unlabeled target domain will facilitate easier object recognition within the target ROIs. In Fig. 3, the green line shows a clear separation in the distribution of background scores around a specific point, which we defined as the boundary of objectness, unlike the orange line. This suggests that the labeled source domain is able to recognize the background well through their own labels, while the unlabeled target domain lacks this recognition, causing ambiguity. Hence, if we can train the detector by training from the positions where the background frequently appears and using the accurate pseudo-labels, we can potentially reduce the ambiguity in the unlabeled target domain. For

this purpose, we introduce similarity-based boundary of objectness and pseudo-labels, designed to help distinguish between foreground and background regions for target domain. We first calculate the similarity S between the categorical prototype and the target ROI features, as defined in Eq. 1. Then, pseudo-labels $\hat{y}_T$ are assigned according to the categories represented by the specific categorical prototype that exhibits the highest similarity to the target ROI features, as shown below:

$$\hat{y}_T^i = \begin{cases} argmax(S^i), & if\ max(S^i) \geq \tau_p \\ background, & if\ max(S^i) < \tau_p \end{cases}. \tag{4}$$

Here, the $argmax$ function returns the index of maximum value in S^i, indicating the similarity map for the i-th target ROI feature. If the specific similarity score exceeds the threshold, the ROI feature is assigned as the foreground; otherwise, it is categorized as the background. To set the threshold, inspired by [28], we adopt an adaptive threshold τ_p instead of using a fixed value. This threshold is dynamically determined by the sum of the mean and variance from the similarity map. Further studies on the threshold are shown in Sec. 4.3.

Using the generated $\hat{y}_T^i$, we set the boundary of objectness H_T, where $0 < H_T < N$, thereby allowing the segregation of target ROIs into foreground ($foreground \in \mathbb{R}^{H_T \times K}$) and background ($background \in \mathbb{R}^{(N-H_T)\times K}$) regions. In other words, H_T is generated based on the ROI index of the target that is consistently classified as background, using the generated $\hat{y}_T$ as a reference. This guarantees that H_T dynamically adjusts to the target ROIs. Hence, the adjusted boundary enables the detector to focus on each region per each image individually, thereby mitigating confusion using proposed losses for each region.

3.5 Objectness Decoupled Loss

To separately train in foreground and background regions using H_T, we introduce an objectness decoupled loss. This loss consists of the foreground attraction loss and the background refinement loss, allowing separate training of each region within the target domain. For the foreground region of the target ROIs, we compute the foreground attraction loss, $\mathcal{L}_{FG}$, using the following formula.

$$\mathcal{L}_{FG} = -\sum_{i=1}^{H_T} \hat{y}_T^i \log\left(softmax\left(\Psi_T^i\right)\right). \tag{5}$$

The $\Psi_T \in \mathbb{R}^{N\times K}$ denotes the output of the classifier using the target ROI feature as input. Therefore, we train by aligning the foreground regions of the target ROI, which are derived from confident similarity-based pseudo-labels. For the background region, we employ the regional prototype as reference to compute the background refinement loss, $\mathcal{L}_{BG}$. Prior to computing the loss, we first extract the background from the target ROI scores and the regional prototype using the boundaries H_T and H_{P_R}, respectively. Here, H_{P_R} is identified as the boundary in the regional prototype where the background category appears consecutively for each ROI position. This is possible because the regional prototype applies

Table 1: Comparison results (%) in three scenarios using VGG-16. **(a)** Weather adaptation on the Cityscapes, **(b)** Synthetic-to-real (S→C) and scene adaptation (K↔C).

Method	Detector	bus	bike	car	mtor	prsn	rider	train	truck	mAP
SIGMA [22]	FCOS	50.7	41.4	63.7	34.7	46.9	48.4	35.9	27.1	43.5
OADA [40]	FCOS	48.0	39.5	62.8	34.6	47.3	45.6	49.4	30.7	44.8
MTTrans [42]	Def DETR	45.9	46.5	65.2	32.6	47.7	49.9	33.8	25.8	43.4
DA-DETR [43]	Def DETR	45.8	46.3	63.1	31.6	49.9	50.0	37.5	24.0	43.5
MEga-CDA [33]	FRCNN	49.2	39.0	52.4	34.5	37.7	49.0	46.9	25.4	41.8
TIA [47]	FRCNN	52.1	38.1	49.7	37.7	34.8	46.3	48.6	31.1	42.3
CIGAR [26]	FRCNN	50.0	40.4	61.6	31.9	45.3	45.3	51.0	32.1	44.7
CSDA [12]	FRCNN	33.3	50.5	44.7	42.9	43.1	58.4	37.3	50.0	45.0
PT [4]	FRCNN	56.6	48.7	63.4	41.3	43.2	52.4	37.8	33.4	47.1
CMT [2]	FRCNN	63.2	53.1	64.5	40.3	47.0	55.7	51.9	39.4	51.9
SWDA [30]	FRCNN	36.2	35.3	43.5	30.0	29.9	42.3	32.6	24.5	34.3
Ours+SWDA	FRCNN	53.8	41.0	53.3	35.0	38.1	48.0	52.5	32.0	**44.2**
AT [24]	FRCNN	56.3	51.9	64.2	38.5	45.5	55.1	54.3	35.0	50.9
Ours+AT	FRCNN	62.1	52.8	64.7	42.6	46.3	54.8	52.1	41.1	**52.1**

(a) Foggy Cityscapes

Method	S→C	K→C	C→K
DAF [5]	39.0	38.5	64.1
MAF [16]	41.1	41.0	72.1
RPA [45]	45.7	-	75.1
MeGA-CDA [33]	44.8	43.0	75.5
TIA [47]	-	44.0	75.9
TDD [15]	53.4	47.4	-
SIGMA [22]	53.7	45.8	-
CSDA [12]	56.9	48.6	-
CIGAR [26]	58.5	48.5	-
NSA-UDA [50]	55.6	56.3	-
SWDA [30]	40.1	37.9	71.0
Ours+SWDA	**57.6**	**49.4**	**81.9**
PT [4]	55.1	60.2	-
Ours+PT	**59.7**	**61.7**	**83.1**

(b) Sim10K & KITTI

an accumulative average to the output ROI scores, which gathering average predictive statistics for each ROI position; it ensures representativeness at each position. Hence, H_{P_R} indicate a boundary of the region where the background is likely to occur in the regional prototype which is suitable for serves as a reference for the background region in the target ROI scores. The $\mathcal{L}_{BG}$ is defined as follows:

$$\mathcal{L}_{BG} = \begin{cases} \sum_{i=H_T}^{N} ||\Psi_T^i - P_R^i||_2^2, & H_T \geq H_{P_R} \\ \sum_{i=H_{P_R}}^{N} ||\Psi_T^i - P_R^i||_2^2, & H_T < H_{P_R} \end{cases}. \tag{6}$$

Here, the condition of Eq. 6 indicates that by considering only the overlapping regions between Ψ_T and P_R. Detailed studies about $\mathcal{L}_{BG}$ are described in supp Sec 4. Hence, $\mathcal{L}_{BG}$ contributes to reducing false positives and refining the demarcation of the estimated background regions. This enables the detector to recognize the proportion of regions within unlabeled target domain. The reason for calculating differently from the foreground attraction loss in Eq. 5 is that Since the target domain lacks labels, objectness decoupling often may lead to unexpected misclassifications due to inaccurate pseudo-labels; i.e., a specific ROI that is foreground may be incorrectly classified as background or vice versa. To alleviate this issue, we use Eq. 6 instead of Eq. 5 to avoid direct classification by pseudo-labels, while remaining the possibility for re-classification into other categories in future training, considering their size and proportion among categories. Note that, in $\mathcal{L}_{BG}$, we do not consider the bounding box, as precise localization of the background is not required. The total objective function is defined as:

$$\mathcal{L} = \mathcal{L}_{Det} + \mathcal{L}_{DA} + \lambda_{FG} \cdot \mathcal{L}_{FG} + \lambda_{BG} \cdot \mathcal{L}_{BG}, \tag{7}$$

where $\mathcal{L}_{Det}$ represents the detection loss in the source domain, $\mathcal{L}_{DA}$ denotes the domain adversarial loss in our baseline. We adjust the ratios for each $\mathcal{L}_{FG}$ and $\mathcal{L}_{BG}$ based on the proportions $\lambda_{FG} = \lambda \cdot (H_T/N)$ and $\lambda_{BG} = \lambda \cdot ((N - H_T)/N)$.

Table 2: The comparison results (%) of different scenarios in BDD100K. **(a)** Two adverse weather adaptations on the BDD100K using ResNet-101. **(b)** Scene adaptation from Cityscapes to BDD100K daytime subset using VGG-16.

	Daytime-sunny to Night-rainy								Daytime-sunny to Dusk-rainy							
Method	bus	bike	car	mtor	prsn	rider	truck	mAP	bus	bike	car	mtor	prsn	rider	truck	mAP
CoT [46]	22.4	9.7	27.4	0.6	9.3	9.3	13.4	13.1	35.5	20.3	50.9	7.9	21.6	16.1	34.4	26.7
HTCN [3]	22.8	9.4	30.7	0.7	11.9	4.8	22.0	14.6	35.9	21.1	51.1	13.7	24.0	16.6	44.2	31.5
SCL [31]	20.0	9.2	33.2	0.3	11.9	10.6	26.4	15.9	34.8	19.2	50.8	13.2	25.9	18.0	38.1	28.6
DAF [5]	23.8	12.0	37.7	0.6	13.5	10.4	29.1	17.4	43.6	27.5	52.3	16.1	28.5	21.7	44.8	33.5
ICCR [37]	32.5	12.1	36.2	1.3	16.1	17.0	29.3	20.6	43.8	28.5	52.4	22.7	29.2	21.9	45.6	36.9
VDD [36]	31.7	15.3	38.0	11.1	18.2	16.7	30.8	23.1	46.1	31.1	54.4	25.3	31.0	22.4	47.6	36.9
SWDA [30]	24.7	10.0	33.7	0.6	13.5	10.4	29.1	17.4	40.0	22.8	51.4	15.4	26.3	20.3	44.2	31.5
Ours+SWDA	44.7	18.4	43.3	14.3	26.1	20.9	46.8	**30.6**	44.6	32.1	60.9	23.4	34.8	27.7	50.8	**39.3**

(a) BDD100K Night-rainy & BDD100K Dusk-rainy

Method	Detector	mAP
EPM [18]	FCOS	27.8
SIGMA [22]	FCOS	32.7
SFA [34]	DefDETR	28.9
MTTrans [42]	DefDETR	32.6
MRT [48]	DefDETR	33.7
TDD [15]	FRCNN	33.6
NSA-UDA [50]	FRCNN	35.5
PT [4]	FRCNN	34.9
Ours+PT	FRCNN	**35.7**

(b) BDD100K Daytime

4 Experiments

We evaluate our method on various scenarios, that show its efficacy in prior works.

Weather Adaptation. We evaluate on the Cityscapes [7] and BDD100K [41] datasets. The Cityscapes and Foggy Cityscapes encompass 8 categories, with 2,975 training and 500 validation, which are used as source and target domains, respectively. The BDD100K includes more challenging scenes spanning various weather conditions, containing 7 categories, using subsets of 27,708 daytime-sunny (source domain), 2,494 night-rainy and 3,501 dusk-rainy (target domains).

Synthetic-to-Real Adaptation. We validate in a synthetic-to-real scenario using the Sim10k [20] dataset, which contains 10,000 synthetic images rendered using Grand Theft Auto. These dataset serve as the source domain, while the Cityscapes dataset acts as the target domain, considering only the car category.

Scene Adaptation. We experimented on the KITTI [13] and Cityscapes datasets, each gathered using distinct camera setups in real-world conditions. Both datasets served as source and target domains, with bidirectional adaptation. The KITTI dataset contains 7,481 images and we only verified the car category, following the protocol described in [5]. We also experimented on Cityscapes as source and BDD100K as target domain, which also categorized small- to large-scale adaptation. We used the daytime subset of the BDD100K, which includes the images, 36,728 for training, and 5,258 for validation in 8 categories.

4.1 Implementation Details.

We adopted the settings from [5] and utilized the Faster-RCNN [29] model with VGG-16 [32], pre-trained on ImageNet [21]. We also established the default UDA baseline using SWDA [30]. We set the default values to $\tau_c = 0.6$ and $\lambda = 0.7$. The decay rates of both moving alignments, α and β, gradually increased from 0.5 to 0.8. We conducted warm-up training for 10k iterations at a learning rate of 10^{-4} to enable the model to detect objects in the source domain. The model was trained with learning rate of 10^{-3} for 50k iterations, then reduced to 10^{-4}. Detail analysis related to hyperparameters is provided in supplementary materials.

Table 3: Ablation study(%) of different training settings with Table 1-(a). * denotes larger training and testing scales.

Method	model	bus	bike	car	mtor	prsn	rider	train	truck	mAP
EPM* [18]	V16	41.5	35.5	56.7	24.6	41.9	38.7	26.8	22.6	36.0
SSAL* [27]	V16	50.0	38.7	59.4	26.0	45.1	47.4	25.7	24.5	39.6
Ours	V16	53.8	41.0	53.3	35.0	38.1	48.0	52.5	32.0	44.2
Ours*	V16	55.6	43.3	54.5	38.1	39.2	51.3	53.7	32.8	**46.1**
NLTE [25]	R50	49.9	39.6	54.8	29.9	37.0	46.9	43.5	32.1	41.8
NLTE* [25]	R50	56.7	43.3	58.7	33.7	43.1	50.7	42.7	33.6	45.4
Ours	R50	52.5	42.4	57.7	39.0	40.7	51.4	43.0	29.2	44.5
Ours*	R50	60.4	44.7	58.3	40.1	40.8	53.2	47.6	33.6	**47.3**

Table 4: Ablation study(%) of the performance changes for weather adaptation on the Cityscapes.

Objectness Decoupling	Categorical Prototype	Regional Prototype	CMA	RMA	mAP
Baseline					34.3
✓					39.9
✓	✓				41.2
✓	✓	✓			42.2
✓	✓	✓	✓		43.1
✓	✓	✓	✓	✓	**44.2**

4.2 Comparison Results.

Weather Adaptation on the Cityscapes. In Table 1-(a), we compare DuPDA with various state-of-the-art (SOTA) methods using VGG-16 for weather adaptation on the Cityscapes. DuPDA outperforms the default baseline [30], achieving an improvement of +9.9%. Moreover, DuPDA shows slight performance improvement over previous SOTA methods, achieving 52.1% using the baseline of [24].

Weather adaptation on the BDD100K. Table 2-(a) shows the results on the BDD100K dataset, which poses more challenging weather conditions than Table 1-(a). Our DuPDA surpasses the SOTA method [36] in both scenarios, with improvements from 23.1% to 30.6% (+7.5%) in Night-rainy and from 36.9% to 39.3% (+2.4%) in Dusk-rainy. This result highlights the effectiveness of DuPDA, which trains the characteristics of objects within the target domain under various adverse weather conditions.

Synthetic-to-Real adaptation. We provide results of DuPDA in the S→C column of Table 1-(b). Using our DuPDA with [30] improved performance from 40.1% to 57.6% (+17.5%). We also achieved the highest mAP of 57.6% and 59.7%, when compared equitably with previous methods. This indicates that DuPDA assists the model in training domain-invariant knowledge of synthetic car characteristics, which is then successfully applied to the real-world car scene.

Scene adaptation. Table 1-(b) shows the results of the scene adaptation using KITTI dataset. Our DuPDA surpasses the baseline [30] by +11.5% and the baseline [4] by +4.6% in KITTI to Cityscapes (K→C). Also, achieving +10.9% using the baseline [30] in Cityscapes to KITTI (C→ K). Compared to the previous SOTA, DuPDA improves by +5.4% from K → C and +7.2% from C → K. We also conduct experiments on adapting the Cityscapes to BDD100K, as shown in Table 2-(b). Our DuPDA outperforms various methods in various detectors and slightly surpasses the previous SOTA [50]. These results show that our DuPDA can effectively address domain shift problems in various scene adaptations.

4.3 Ablation Studies.

We perform a comprehensive ablation analysis using the SWDA baseline [30] with VGG-16, within the weather adaptation in the Cityscapes scenario.

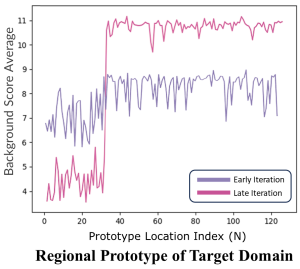

Fig. 4: Background classification score distribution in regional prototype of target domain across different iterations.

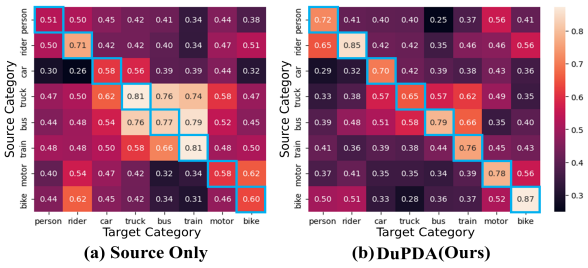

Fig. 5: Comparison of feature similarity between source and target ROI features for each category, using models trained: **(a)** Supervised on the source domain only, and **(b)** Using DuPDA.

Consistency of different training settings. In Table 3, we show the consistent effectiveness of DuPDA by experimenting with different backbone, depth and input size, unlike the settings in Table 1-(a). We replaced the VGG-16 with the deeper ResNet-50, leading to a slight performance increase. Following the settings in [25], the results marked with a * in the table show that a larger input size led to improved performance compared to the default setting in [30] (without the * mark). It also outperforms the previous methods under the same settings. The results show that DuPDA consistently maintains the effectiveness of adaptation, regarding changes in backbone and input size, without saturation.

Effect of individual components. Table 4 shows the impact of DuPDA's components on overall performance. First, to evaluate the basic objectness decoupling, we compared it with the baseline [30]. We used batch-specific ROI features extracted for each domain, without our prototypes and their corresponding moving alignment schemes, achieving +5.6%. This shows that our decoupling method effectively captures the characteristics of the target domain by focusing separately on foreground and background regions. However, it cannot be guaranteed that all categories are always present within a batch. Hence, the inclusion of proposed prototypes that accumulate information only from the source domain led to a +2.3% increase. Finally, using CMA and RMA to gradually transition prototypes from the source to the target domain leads to a +2.0% improvement.

Effect of objectness decoupling. Fig. 4 shows the distribution of the background category in the regional prototype of the target domain. The x-axis denotes the prototype indices (N = 128), the y-axis denotes the average of the classified background scores. In early iterations, the results reveal an ambiguous distinction between the foreground and background regions. This is because the unlabeled target domain can potentially lead to excessive background detection across all ROI positions, as it makes recognizing the position and proportion of background regions challenging in early iterations. As training progresses, clear distinctions emerge, intensifying the separation between the foreground (index $\leq$ 37) and the background (index $>$ 37) regions. These results indicate that DuPDA effectively mitigates misidentification by guiding the ratio of both regions using

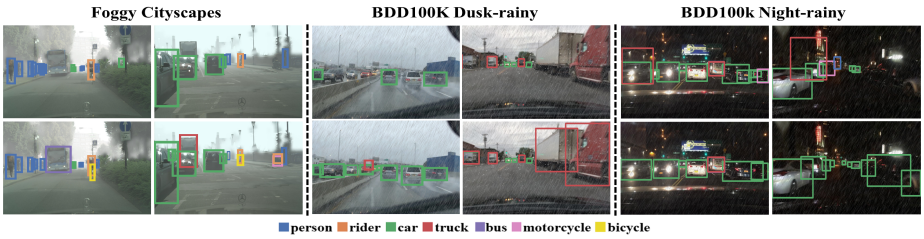

Fig. 6: Qualitative analysis on the target domains across three different UDA-OD scenarios. **Top row:** Results from baseline [30]. **Bottom row:** Results from our DuPDA.

our objectness decoupled loss to unlabeled target ROIs. Moreover, compared to source-only supervised results (orange line) in Fig. 3, DuPDA makes it clearer to separate both regions at specific ROI position in unlabeled target domain.

Fig. 5 shows a comparison of feature similarity for each category between ROIs in both domains. (a) shows the results of the model trained only in the source domain, and (b) presents the results after training with our DuPDA. The diagonal (blue box) denotes similarity scores between both domains within the same category, while the remaining areas represent scores across different categories. As a result, DuPDA enhances similarity at diagonal positions compared to the source-only model, while reducing similarity at other positions. These results show the effectiveness of our $\mathcal{L}_{FG}$, which augments the training of the foreground target ROIs by calculating similarity using our categorical prototype.

Qualitative results. Fig. 6 shows qualitative results for different scenarios of weather adaptation. The baseline [30] results shows satisfactorily in detecting easily distinguishable objects under adverse weather conditions; however, it struggles with slightly distorted or obscured objects. In contrast, DuPDA maintains consistent detection, enabling it to identify objects where the baseline fails.

5 Conclusions

We introduce DuPDA, a novel framework for UDA-OD, which involves categorical and regional moving alignments, as well as objectness decoupling to train by focusing on each region within an unlabeled target domain. Our alignments generate specialized domain-invariant prototypes that serve as a criterion to effectively training target domain. Additionally, objectness decoupling splits target ROIs into two regions, enhancing foreground attraction and background refinement through each loss; achieving successful results in six UDA-OD scenarios.

Acknowledgement. This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2014-3-00123, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis), (IITP-2024-No.RS-2023-00255968, AI Convergence Innovation Human Resources Development), and Korea NRF grant (NRF-2022R1A2C1091402). W. Hwang is the corresponding author.

References

1. Cai, Q., Pan, Y., Ngo, C.W., Tian, X., Duan, L., Yao, T.: Exploring object relation in mean teacher for cross-domain detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11457–11466 (2019)
2. Cao, S., Joshi, D., Gui, L.Y., Wang, Y.X.: Contrastive mean teacher for domain adaptive object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23839–23848 (2023)
3. Chen, C., Zheng, Z., Ding, X., Huang, Y., Dou, Q.: Harmonizing transferability and discriminability for adapting object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8869–8878 (2020)
4. Chen, M., Chen, W., Yang, S., Song, J., Wang, X., Zhang, L., Yan, Y., Qi, D., Zhuang, Y., Xie, D., et al.: Learning domain adaptive object detection with probabilistic teacher. arXiv preprint arXiv:2206.06293 (2022)
5. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3339–3348 (2018)
6. Chen, Y., Li, W., Van Gool, L.: ROAD: Reality oriented adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7892–7901 (2018)
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
8. Deng, J., Li, W., Chen, Y., Duan, L.: Unbiased mean teacher for cross-domain object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4091–4101 (2021)
9. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 994–1003 (2018)
10. Do, D.P., Kim, T., Na, J., Kim, J., Lee, K., Cho, K., Hwang, W.: D3t: Distinctive dual-domain teacher zigzagging across rgb-thermal gap for domain-adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23313–23322 (2024)
11. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)
12. Gao, C., Liu, C., Dun, Y., Qian, X.: Csda: Learning category-scale joint feature for domain adaptive object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11421–11430 (2023)
13. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
14. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
15. He, M., Wang, Y., Wu, J., Wang, Y., Li, H., Li, B., Gan, W., Wu, W., Qiao, Y.: Cross domain object detection by target-perceived dual branch distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9570–9580 (2022)

16. He, Z., Zhang, L.: Multi-adversarial faster-rcnn for unrestricted object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6668–6677 (2019)
17. Hong, W., Wang, Z., Yang, M., Yuan, J.: Conditional generative adversarial network for structured domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1335–1344 (2018)
18. Hsu, C.C., Tsai, Y.H., Lin, Y.Y., Yang, M.H.: Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In: European Conference on Computer Vision. pp. 733–748. Springer (2020)
19. Jiang, Z., Li, Y., Yang, C., Gao, P., Wang, Y., Tai, Y., Wang, C.: Prototypical contrast adaptation for domain adaptive semantic segmentation. In: European Conference on Computer Vision. pp. 36–54. Springer (2022)
20. Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., Vasudevan, R.: Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? Proceedings of International Conference on Robotics and Automation (2017)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
22. Li, W., Liu, X., Yuan, Y.: Sigma: Semantic-complete graph matching for domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5291–5300 (2022)
23. Li, X., Jie, Z., Wang, W., Liu, C., Yang, J., Shen, X., Lin, Z., Chen, Q., Yan, S., Feng, J.: FoveaNet: Perspective-aware urban scene parsing. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 784–792 (2017)
24. Li, Y.J., Dai, X., Ma, C.Y., Liu, Y.C., Chen, K., Wu, B., He, Z., Kitani, K., Vajda, P.: Cross-domain adaptive teacher for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7581–7590 (2022)
25. Liu, X., Li, W., Yang, Q., Li, B., Yuan, Y.: Towards robust adaptive object detection under noisy annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14207–14216 (2022)
26. Liu, Y., Wang, J., Huang, C., Wang, Y., Xu, Y.: Cigar: Cross-modality graph reasoning for domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23776–23786 (2023)
27. Munir, M.A., Khan, M.H., Sarfraz, M., Ali, M.: SSAL: Synergizing between self-training and adversarial learning for domain adaptive object detection. Advances in Neural Information Processing Systems **34**, 22770–22782 (2021)
28. Na, J., Jung, H., Chang, H.J., Hwang, W.: Fixbi: Bridging domain spaces for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1094–1103 (2021)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
30. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6956–6965 (2019)
31. Shen, Z., Maheshwari, H., Yao, W., Savvides, M.: Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. arXiv preprint arXiv:1911.02559 (2019)

32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
33. VS, V., Gupta, V., Oza, P., Sindagi, V.A., Patel, V.M.: 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. pp. 4516–4526 (2021)
34. Wang, W., Cao, Y., Zhang, J., He, F., Zha, Z.J., Wen, Y., Tao, D.: Exploring sequence feature alignment for domain adaptive detection transformers. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1730–1738 (2021)
35. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 79–88 (2018)
36. Wu, A., Liu, R., Han, Y., Zhu, L., Yang, Y.: Vector-decomposed disentanglement for domain-invariant object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9342–9351 (2021)
37. Xu, C.D., Zhao, X.R., Jin, X., Wei, X.S.: Exploring categorical regularization for domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11724–11733 (2020)
38. Xu, M., Wang, H., Ni, B., Tian, Q., Zhang, W.: Cross-domain detection via graph-induced prototype alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12355–12364 (2020)
39. Xu, Y., Sun, Y., Yang, Z., Miao, J., Yang, Y.: H2fa r-cnn: Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14329–14339 (2022)
40. Yoo, J., Chung, I., Kwak, N.: Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In: European Conference on Computer Vision. pp. 691–708. Springer (2022)
41. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv preprint arXiv:1805.04687 **2**(5), 6 (2018)
42. Yu, J., Liu, J., Wei, X., Zhou, H., Nakata, Y., Gudovskiy, D., Okuno, T., Li, J., Keutzer, K., Zhang, S.: Mttrans: Cross-domain object detection with mean teacher transformer. In: European Conference on Computer Vision. pp. 629–645. Springer (2022)
43. Zhang, J., Huang, J., Luo, Z., Zhang, G., Zhang, X., Lu, S.: Da-detr: Domain adaptive detection transformer with information fusion. arXiv preprint arXiv:2103.17084 (2021)
44. Zhang, J., Huang, J., Luo, Z., Zhang, G., Zhang, X., Lu, S.: Da-detr: Domain adaptive detection transformer with information fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23787–23798 (2023)
45. Zhang, Y., Wang, Z., Mao, Y.: RPN prototype alignment for domain adaptive object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12425–12434 (2021)
46. Zhao, G., Li, G., Xu, R., Lin, L.: Collaborative training between region proposal localization and classification for domain adaptive object detection. In: European Conference on Computer Vision. pp. 86–102. Springer (2020)

47. Zhao, L., Wang, L.: Task-specific inconsistency alignment for domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14217–14226 (2022)
48. Zhao, Z., Wei, S., Chen, Q., Li, D., Yang, Y., Peng, Y., Liu, Y.: Masked retraining teacher-student framework for domain adaptive object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19039–19049 (2023)
49. Zheng, Y., Huang, D., Liu, S., Wang, Y.: Cross-domain object detection through coarse-to-fine feature adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13766–13775 (2020)
50. Zhou, W., Fan, H., Luo, T., Zhang, L.: Unsupervised domain adaptive detection with network stability analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6986–6995 (2023)