

Decompose to Adapt: Cross-domain Object Detection via Feature Disentanglement

Dongnan Liu, *Member, IEEE*, Chaoyi Zhang, Yang Song, *Member, IEEE*, Heng Huang, Chenyu Wang, Michael Barnett and Weidong Cai, *Member, IEEE*

arXiv:2201.01929v1 [cs.CV] 6 Jan 2022

Abstract—Recent advances in unsupervised domain adaptation (UDA) techniques have witnessed great success in cross-domain computer vision tasks, enhancing the generalization ability of data-driven deep learning architectures by bridging the domain distribution gaps. For the UDA-based cross-domain object detection methods, the majority of them alleviate the domain bias by inducing the domain-invariant feature generation via adversarial learning strategy. However, their domain discriminators have limited classification ability due to the unstable adversarial training process. Therefore, the extracted features induced by them cannot be perfectly domain-invariant and still contain domain-private factors, bringing obstacles to further alleviate the cross-domain discrepancy. To tackle this issue, we design a Domain Disentanglement Faster-RCNN (DDF) to eliminate the source-specific information in the features for detection task learning. Our DDF method facilitates the feature disentanglement at the global and local stages, with a Global Triplet Disentanglement (GTD) module and an Instance Similarity Disentanglement (ISD) module, respectively. By outperforming state-of-the-art methods on four benchmark UDA object detection tasks, our DDF method is demonstrated to be effective with wide applicability.

Index Terms—domain adaption, object detection, feature disentanglement, automatic drive

I. INTRODUCTION

OBJECT detection, which aims to assign a bounding box and category prediction for each foreground instance, is essential for modern computer vision. Taking advantages from the deep learning techniques, previous object detection methods based on convolutional neural networks (CNN) have achieved appealing performance on various benchmarks [1]–[5]. However, these fully-supervised models have been criticized for the lack of generalization ability and suffer from severe performance drop when validated on other unseen datasets, since they tend to bias towards the data distribution of the training domain [6], [7]. On the other hand, collecting sufficient annotations for each new domain is impractical in real applications, due to time-consuming and expensive annotation procedure.

To address this dilemma, unsupervised domain adaptation (UDA) methods have been proposed to transfer the domain-invariant information from the labeled source domain to an

D. Liu, C. Zhang, and W. Cai are with the School of Computer Science, University of Sydney, NSW, 2008 Australia. Corresponding author: Dongnan Liu (dongnan.liu@sydney.edu.au)

Y. Song is with the School of Computer Science and Engineering, University of New South Wales, NSW, 2052 Australia.

H. Huang is with the Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15261 USA.

C. Wang and M. Barnett are with the Brain and Mind Centre, University of Sydney, NSW, 2050 Australia.

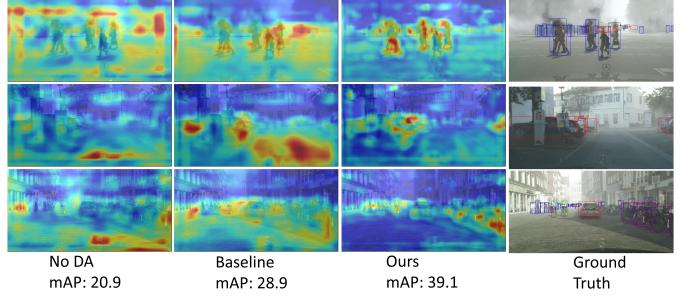


Fig. 1. The visualization results on the image-level features. ‘No DA’ is the Faster R-CNN without adaptation, and ‘Baseline’ is the Faster R-CNN with basic adversarial domain discriminator. For the No DA and Baseline, the features are collected and visualized from the backbone network. For our method, we simultaneously visualize the global-level domain-shared and domain-private features. All the features and labels are overlapped on the original images. (Best viewed in color and zoom in)

unlabeled target domain [8]–[10]. For the UDA object detection methods, the majority of them incorporate the adversarial domain discriminators with the object detection architectures (e.g Faster-RCNN), to alleviate the cross-domain discrepancy by inducing the domain-invariant features generation [11]–[14].

Despite the state-of-the-art performance of the adversarial learning-based UDA object detection methods, they overlook the entanglement between the *domain-shared* and *domain-private* features in the latent manifold space [15]–[17]. Due to the unstable learning process, the decision boundary for the adversarial domain classifiers in these methods is inaccurate [18]. As such, the extracted features in the typical adversarial UDA methods cannot be perfectly domain-invariant and inevitably contain domain-private factors. Optimizing the UDA models with these entangled features induces the model bias towards the source domain, and further degrades the performance on the target domain. As indicated in Figure 1, the Baseline method with the adversarial domain discriminator tends to focus on the source-specific information in the background, which leads to suboptimal adaptation performance.

Previously, several methods have been proposed to solve the feature entanglement issue in UDA classification [15]–[17], [19], [20]. They are established based on auto-encoder architectures, incorporated with latent code independence induction mechanism, and classifier entropy regularization to decompose the domain-private factors from the features for classifier learning. However, directly extending the feature disentanglement-based UDA classification methods to cross-

domain object detection suffers from challenges. First, [15]–[17] decompose the shared and private features in each domain by inducing their distributions to be independent, which always requires the batch size of the features to be sufficiently large to describe the characteristics of the distribution. However, the batch size for UDA object detection is very limited (sometimes even equals to 1 [11], [13]), due to the mandatory larger model sizes than those for classification. Second, these methods ignore the feature entanglement at the local level for each foreground object since their objective is to assign a category label for the whole image. For cross-domain object detection which typically contains multiple object instances in an image, the adaptation ability of these UDA disentanglement-based classification model is limited, resulting from the instance-level feature entanglement. Although PD [21] was previously proposed for UDA object detection via feature decomposition based on mutual information minimization [17], the limited batch size for the object detection framework causes the suboptimal adaptation performance. Subsequently, there currently still lack feature disentanglement UDA methods particularly for cross-domain object detection.

Motivated by the aforementioned observations, we propose a Domain Disentanglement Faster-RCNN (DDF) method in this work, to improve the typical adversarial learning-based UDA object detection via feature disentanglement. Specifically, our DDF achieves feature disentanglement at the global and local levels. Throughout this paper, the global-level features represent the output of the backbone network, which contain the information of the object structure, and spatial distribution for the whole images. The local-level features represent the region of interests (ROIs) for object localization and classification. Domain-invariant features represent the features containing shared factors between the source and target domains, and the instance-invariant features represent the domain-invariant features at the instance-level for the ROIs. First, we design a Global Triplet Disentanglement (GTD) module jointly optimized with a domain discriminator, which improves the feature adaptation ability at the global level via a triplet feature disentanglement strategy. To further facilitate the feature disentanglement at the local stage, an Instance Similarity Disentanglement (ISD) module is proposed, based on the similarity regularization between the shared and private features for instance objects. Our DDF method is validated on four benchmark UDA object detection tasks and outperforms the state-of-the-art methods.

II. RELATED WORK

A. Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) aims at bridging the gap between an annotated source domain and an unannotated target domain. Typically, UDA methods transfer the knowledge in four ways: 1) directly minimizing the statistical distribution distance between two domains [22], [23]; 2) inducing the domain-invariant feature generation [9], [10], [24]; 3) learning from the synthesized images [25]–[29], and 4) self-training via pseudo labels [30], [31]. By alleviating the cross-domain discrepancy at the feature and appearance levels, UDA

methods have achieved outstanding performance in cross-domain classification [9], [10], [27], segmentation [32]–[34], and detection [11], [13], [14], [30], [35], [36]. Since our proposed method aims at tackling the feature entanglement issue for cross-domain object detection, only the literature of UDA object detection and feature disentanglement are reviewed in detail.

B. UDA for Cross-domain Object Detection

Recent UDA object detection methods are designed based on the two-stage detector Faster-RCNN and induce the domain-invariant feature learning by domain adaptive modules at the feature and appearance levels. Among the feature-level adaptation methods, Chen *et al.* [11] first proposed to learn domain-invariant features for the global image and local instance features. To align the image-level features with different scales, strong-weak feature alignment was then proposed [13]. [35] designed a selective adaptation architecture to focus on the local regions. Later, hierarchical feature alignment with multi-stage domain discriminators further improves the baseline [12], [36], [37]. Additionally, category-aware feature alignment is effective in alleviating the class-imbalance issues during adaptation [38]–[41]. In [42], a domain randomization strategy is proposed based on the spectrum learning at the frequency space, which has achieved appealing performance on the cross-domain object detection under the domain generalization setting. For the appearance-level adaptation, target-like synthetic images and pseudo labels are commonly employed [30], [43]. Even though the previous methods alleviated the domain-bias by various feature-level or appearance-level adaptations, few of them have discussed and attempted to explicitly address the entanglement of the domain-shared and domain-private factors, limiting the model’s adaptation abilities.

C. UDA with Feature Disentanglement

Previous feature disentanglement methods decomposed the features by encouraging the independence between them, based on the variation autoencoders (VAE) [44], [45] and generative adversarial networks (GAN) [46], [47]. By separating the appearance-level (e.g., texture, brightness) and the global-level factors (e.g., object structure, spatial distribution) in the latent feature space, [26], [48] achieved compelling performance on multi-modal image synthesis. For UDA classification, [15]–[17], [19] have been proposed to decompose the domain-private and domain-shared information using autoencoder-based structures. As essential steps for these methods, the independence induction for the latent feature distribution requires large batch sizes, which is impractical for UDA object detection. In addition, [49] proposed to decompose the features via the attention mechanism based on uncertainty learning. In [50], a variational information bottleneck was designed to filter out the redundant domain-private information and achieve the disentanglement. However, the feature entanglement issue for each object instance has not been considered among these methods, which is essential for UDA object detection. Therefore, it is not intuitive to

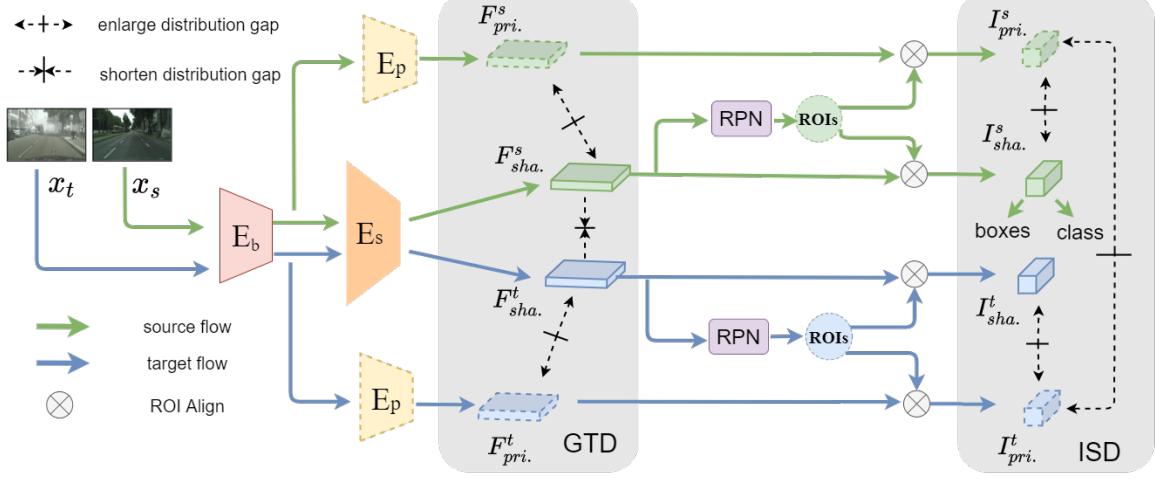


Fig. 2. The overall architecture of our proposed DDF method. Please refer to Section III-A for detailed definitions on all the symbols. Note that the MLP module before $I_{sha.}^s$, $I_{sha.}^t$, $I_{pri.}^s$, and $I_{pri.}^t$ is omitted. **(Best viewed in color)**

directly extend current UDA feature disentanglement methods for cross-domain object detection.

In [51], a Conditional Domain Normalization (CDN) mechanism was proposed for UDA object detection via feature decomposition, by incorporating the source-specific factors with the target features. However, there lacks feature decomposition for the target features, which would cause the model to suffer from domain bias during inference. Although Wu *et al.* [21] designed a PD method with progressive feature disentanglement mechanism for UDA object detection, their feature entanglement modules are based on Mutual Information minimization [17], which requires batch-wise feature shuffling [52]. Based on the implementation of [17], the features have a batch size of 128, sufficiently large to contain the distribution characteristics of their corresponding domains. When adapting the method to UDA object detection, the disentanglement module in [21] has to work with small batch sizes and hence could not fully utilize the Mutual Information minimization strategy. To tackle this issue, we propose a DDF method for cross-domain object detection via feature disentanglement at the global and local levels, without requiring a large image batch size for training. Particularly, our DDF method has achieved better performance than PD [21] (detection mAP: Ours: 39.1; PD: 36.6). This further demonstrates the superiority of our designed feature decomposition modules compared to the typical ones, when the batch sizes for the framework are limited for cross-domain object detection.

III. DOMAIN DISENTANGLEMENT FASTER-RCNN

In this section, we present the detailed framework of our Domain Disentanglement Faster-RCNN (DDF) model. Denote a labeled source domain with N_s i.i.d images as $D_s = \{(x_i^s, c_i^s, b_i^s)\}_{i=1,2,\dots,N_s}$, where c_i^s and b_i^s represent the category labels and bounding box coordinates for all the foreground objects in each image x_i^s , respectively. Then, an unlabeled target domain with N_t i.i.d images is defined as $D_t = \{(x_j^t)\}_{j=1,2,\dots,N_t}$, with a different data distribution from

the D_s . Our DDF method aims at transferring the knowledge from the labeled D_s to the unlabeled D_t and achieving competitive detection performance on the target domain.

A. Framework Overview

The overall diagram of our proposed DDF model is shown in Figure 2. First, we establish a baseline UDA object detection model, by incorporating a global-level adversarial domain discriminator with a Faster-RCNN detection model. During each training iteration, the source and target images, defined as x_s and x_t , are fed into the network. First, a backbone is employed to extract the global-level features of the input images, constructed by a basic feature encoder E_b with fixed weights, and a domain-shared feature extractor E_s with dynamically updated weights.

To facilitate the feature disentanglement, we design a domain-private feature encoder E_p following the E_b , to obtain the feature maps containing specific factors of each domain. As shown in Figure 2 and Equation 1, $F_{sha.}^s$ and $F_{sha.}^t$ represent the domain-shared features across the source and target domains, and $F_{pri.}^s$ and $F_{pri.}^t$ indicate the domain-private features specific to the source and target domains:

$$\begin{aligned} F_{sha.}^s &= E_s(E_b(x_s)), & F_{sha.}^t &= E_s(E_b(x_t)), \\ F_{pri.}^s &= E_p(E_b(x_s)), & F_{pri.}^t &= E_p(E_b(x_t)). \end{aligned} \quad (1)$$

For the global-level feature disentanglement, $F_{sha.}^s$, $F_{sha.}^t$, $F_{pri.}^s$, and $F_{pri.}^t$ are optimized with the Global Triplet Disentanglement (GTD) module, which aligns the distributions of the domain-shared features between the source and target domains, as well as enlarges the discrepancy between the domain-shared and domain-private features within each domain. Leveraging the GTD module, the global-level domain-specific factors are thus decomposed from the domain-shared features for the detection task learning.

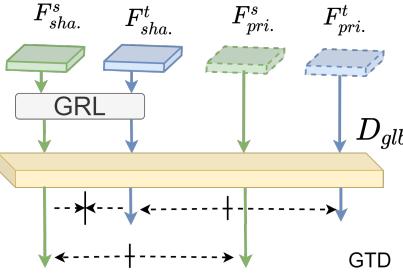


Fig. 3. The detailed structure of our proposed Global Triplet Disentanglement (GTD) module in Section III-B. The definitions for all the symbols are the same as Figure 2. **(Best viewed in color, and in conjunction with Figure 2)**

At the local level, we employ a Region Proposal Network (RPN) and a ROIAlign layer to extract the local instance features in $F^s_{sha.}$, $F^t_{sha.}$, $F^s_{pri.}$, and $F^t_{pri.}$:

$$\begin{aligned} I^s_{sha.} &= \text{MLP}\left(\text{ROIAlign}(F^s_{sha.}, RPN(F^s_{sha.}))\right), \\ I^t_{sha.} &= \text{MLP}\left(\text{ROIAlign}(F^t_{sha.}, RPN(F^t_{sha.}))\right), \\ I^s_{pri.} &= \text{MLP}\left(\text{ROIAlign}(F^s_{pri.}, RPN(F^s_{pri.}))\right), \\ I^t_{pri.} &= \text{MLP}\left(\text{ROIAlign}(F^t_{pri.}, RPN(F^t_{pri.}))\right), \end{aligned} \quad (2)$$

where MLP contains a flatten layer followed by three fully connected layers. In Equation 2, $I^s_{sha.}$ and $I^t_{sha.}$ denote the instance-wise domain-shared features captured from each domain, while $I^s_{pri.}$ and $I^t_{pri.}$ represent the domain-private ones. Then, the instance-level $I^s_{sha.}$, $I^t_{sha.}$, $I^s_{pri.}$, and $I^t_{pri.}$ are fed into the Instance Similarity Disentanglement (ISD) module to facilitate the feature disentanglement at the local stage, based on feature similarity optimization. Eventually, the domain-shared features at the instance-level ($I^s_{sha.}$) are employed for the object classification and location regression learning. The detailed structure of the private components of the domain-specific feature extractor is shown in Table I.

TABLE I

DETAILED MODEL STRUCTURE OF THE OVERALL ARCHITECTURE. *in* AND *out* REPRESENT THE INPUT AND OUTPUT CHANNEL NUMBERS OF EACH LAYER, RESPECTIVELY. *k*, *s*, AND *p* ARE THE KERNEL SIZE, STRIDE, AND PADDING OF EACH LAYER.

Layer	E_p	Hyperparameters
1	Conv, ReLU	$k = (3, 3), s = 1, p = 1, in = 256, out = 256$
2	Maxpooling	$k = (2, 2), s = 2, p = 0$
3	Conv, ReLU	$k = (3, 3), s = 1, p = 1, in = 256, out = 512$
4	Maxpooling	$k = (2, 2), s = 2, p = 0$
5	Conv, ReLU	$k = (3, 3), s = 1, p = 1, in = 512, out = 512$

Layer	D_{glb}	Hyperparameters
1	Conv, ReLU	$k = (3, 3), s = 1, p = 1, in = 512, out = 512$
2	Conv, ReLU	$k = (3, 3), s = 1, p = 1, in = 512, out = 256$
3	Conv, ReLU	$k = (3, 3), s = 1, p = 1, in = 256, out = 128$
4	Conv, ReLU	$k = (3, 3), s = 1, p = 1, in = 128, out = 64$
5	Conv	$k = (3, 3), s = 1, p = 1, in = 64, out = 2$

B. Global Triplet Disentanglement

In previous UDA object detection methods, the invariance between the domain-shared $F^s_{sha.}$ and $F^t_{sha.}$ is induced by

optimizing an adversarial domain discriminator at the global level [11], [13], [14] through:

$$L_{di} = \max_{\theta_{E_s}} \min_{\theta_D} L_{ce}\left(D_{glb}(F^s_{sha.}), 0\right) + L_{ce}\left(D_{glb}(F^t_{sha.}), 1\right), \quad (3)$$

where $L_{ce}(\cdot, \cdot)$ measures the cross entropy between the feature maps and labels, while θ_{E_s} and θ_D indicate the parameters for the domain-shared feature extractor E_s and domain discriminator D_{glb} , respectively. We denote the domain label as 0 for the source domain and 1 for the target domain. However, as suggested in [16], [18], obtaining the precise classification decision boundaries for the adversarial discriminators in typical UDA methods is always challenging due to the unstable learning procedure. As such, the features induced by D_{glb} in Equation 3 cannot be perfectly domain-invariant, and some lurking domain-private factors would inevitably bring domain bias into the detection task learning. Although recent autoencoder-based feature disentanglement methods attempt to alleviate this issue for UDA classification tasks [15]–[17], they are suboptimal for UDA object detection architectures. Due to the large model size, the batch size for the typical UDA object detection framework is always set to 1 [11], [13], [21]. However, the feature independence learning paradigms for feature disentanglement in [15]–[17] involve the calculations across a batch with average and shuffling, which would become redundant when the batch size is 1.

To this end, we propose a Global Triplet Disentanglement (GTD) module, which aims to improve the adversarial learning-based feature adaptation at the global level via feature disentanglement. Our GTD module focuses on the triplet optimization among the domain-private and domain-shared features across two domains without batch-wise calculations as [15]–[17], hence mitigating the influence of feature batch size. Detailed illustrations of the GTD module are shown in Figure 3. We consider that under the ideal situation, the domain-private $F^s_{pri.}$ and $F^t_{pri.}$ contain the specific factors for the source and target domains, respectively. Subsequently, the domain discriminator D_{glb} should be able to distinguish them. Based on this assumption, we first introduce a domain-specific classification loss L_{ds} to enhance the classification ability of D_{glb} :

$$L_{ds} = \min_{\theta_D, \theta_{E_p}} L_{ce}\left(D_{glb}(F^s_{pri.}), 0\right) + L_{ce}\left(D_{glb}(F^t_{pri.}), 1\right), \quad (4)$$

where θ_{E_p} indicates the parameters of the domain-private encoder E_p .

To further enlarge the discrepancy between the domain-shared and private features within each domain, and meanwhile to align the domain-shared features across the domains, we propose a triplet optimization strategy for $F^s_{sha.}$, $F^t_{sha.}$, $F^s_{pri.}$, and $F^t_{pri.}$. Specifically, the triplet feature decomposition loss L_{tri} is defined as:

$$\begin{aligned} L_{tri} = & \frac{1}{2} \left(\max \left(d(F^s_{sha.}, F^t_{sha.}) - d(F^s_{sha.}, F^s_{pri.}) + m, 0 \right) \right. \\ & \left. + \max \left(d(F^s_{sha.}, F^t_{sha.}) - d(F^t_{sha.}, F^t_{pri.}) + m, 0 \right) \right), \end{aligned} \quad (5)$$

where $d(f_1, f_2) = ||D_{glb}(f_1) - D_{glb}(f_2)||^2$, measuring the square Euclidean distance between the *SoftMax* domain predictions of f_1 and f_2 . We have set the margin value m as 0.25 to ensure the domain probability predictions of the F_{sha}^s and F_{sha}^t close to 0.5. In Equation 3, D_{glb} for distinguishing F_{sha}^s from F_{sha}^t is trained in an adversarial manner with the feature extractor E_s . Under the ideal optimization scenario, the features from E_s should be able to confuse D_{glb} for being domain-invariant. To this end, the domain category predictions for both F_{sha}^s and F_{sha}^t are expected to be 0.5, which is consistent with the objective of Equation 5.

Overall, the GTD module improves the adaptation ability of the adversarial domain discriminator by decomposing the global-level source-specific factors. Meanwhile, the GTD module could be intuitively implemented without any batch-wise calculations, which is more practical-friendly to UDA object detection tasks by loosening the restrictions on large batch size. The loss function for the GTD module is defined as:

$$L_{GTD} = L_{ds} + L_{tri}. \quad (6)$$

C. Instance Similarity Disentanglement

The GTD module promotes the global-level feature adaptation by removing the domain-specific factors. However, there still exist domain-bias issues in the local-level features of each object. In previous work, adversarial domain discriminators are employed to alleviate the discrepancy for the cross-domain instance features at the local levels [11], [12]. However, similar to the global-level issues, the local-level adversarial learning-based feature adaptation also suffers from the feature entanglement, due to its inaccurate decision boundary for domain classification.

To this end, we propose an Instance Similarity Disentanglement (ISD) to facilitate the instance-level feature alignment based on feature similarity optimization. First, we obtain the domain-shared and domain-private features at the local level, following Equation 2. Since the batch sizes of the local-level I_{sha}^s , I_{sha}^t , I_{pri}^s , and I_{pri}^t features equal to the number of ROIs, they are sufficiently large for dependency-aware optimization. Based on the assumption that the domain-shared factors in each domain should be independent from the domain-private ones [15]–[17], we next propose to enlarge the distribution distance between the domain-shared and domain-private instance features within each domain. This step could be achieved by minimizing their cosine similarity as:

$$L_{ISD-intra} = sim(I_{sha}^s, I_{pri}^s) + sim(I_{sha}^t, I_{pri}^t), \quad (7)$$

where $sim(\cdot, \cdot)$ denotes the cosine similarity between two vectors: $sim(a, b) = \frac{a^T b}{\|a\| \|b\|}$. Since I_{sha}^s , I_{sha}^t , I_{pri}^s , and I_{pri}^t are acquired after the *ReLU* activation layers, the cosine similarity between each pair of them is in range [0, 1]. In addition to the intra-domain feature similarity minimization in Equation 7, we then propose to minimize the feature similarity between the domain-private features across two domains. This is motivated by [53]: under the ideal disentanglement scenario, each pair of domain-private factors in the manifold space from

two different domains should not intersect. Therefore, enlarging the distribution distance between the domain-private I_{pri}^s and I_{pri}^t is beneficial to facilitate the instance-level feature disentanglement, which could be mathematically described as minimizing:

$$L_{ISD-inter} = sim(I_{pri}^s, I_{pri}^t). \quad (8)$$

The loss function for the ISD module is defined by integrating the intra- and inter-domain feature disentanglement at the local stage:

$$L_{ISD} = L_{ISD-intra} + L_{ISD-inter}. \quad (9)$$

D. Training and Inference

Our DDF method is build on Faster-RCNN with a VGG16 and ResNet50 backbone. For model initialization, the backbones are pretrained on the ImageNet classification task [54] and the weights of the remaining layers are initialized with normal distribution initialization. Following the implementation of [13], [36], [40], the fixed-weight E_b is the model before the *conv3_3* layer of the VGG16 backbone, and E_s is the rest. The batch size is 1 and we do not use batch normalization due to the small mini-batch size. Each batch contains two images, one from the source domain and the other from the target domain. For all the experiments, the shorter side of the training image is resized to 600, while maintaining the aspect ratio. During training, we employ the SGD optimizer with a momentum of 0.9, and the weight decay is set to 0.0005. The initial learning rate is 0.001 for the first 50K training iterations and decreased to 0.0001 for the last 20K iterations.

The overall loss function of our proposed DDF is defined as:

$$L_{ddf} = L_{det} + L_{di} + L_{GTD} + L_{ISD}, \quad (10)$$

where L_{det} represents the detection loss function in the standard Faster-RCNN, including the classification and bounding boxes regression loss for the RPN and the final task. Note that we do not include any trade-off weight for each loss in L_{ddf} , which avoids the time-consuming process of hyperparameter fine-tuning. In addition, the overall DDF framework is optimized in an end-to-end manner. The GRL layer between D_{glb} and E_s was employed when optimizing Equations 3 and 5.

During inference, the target images directly pass through the vanilla Faster-RCNN with off-the-shelf weights for object detection. For evaluation, we report the mean average precision (mAP) with a threshold of 0.5. Our experiment is implemented with PyTorch [55] on one NVIDIA GeForce 1080Ti GPU.

IV. EXPERIMENTAL DETAILS

A. Dataset Description

Our extensive experiments are conducted on four public datasets including Cityscapes [63], Foggy Cityscapes [64], SIM10K [65], and KITTI [66].

CityScapes [63] dataset contains 5000 images of the real urban scenes collected from 27 cities in different seasons. In this work, the ground truth for UDA object detection is the

TABLE II

QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE CITYSCAPES → FOGGY CITYSCAPES EXPERIMENT. NO DA REPRESENTS THE FASTER-RCNN TRAINED WITH THE SOURCE IMAGES AND DIRECTLY TESTED ON THE TARGET IMAGES, WITHOUT ANY DOMAIN ADAPTATION. FOR EACH COMPARISON METHOD, MAP \star AND GAIN REPRESENT ITS RESULT UNDER “NO DA” SETTING SETTING AND THE CORRESPONDING IMPROVEMENT ACHIEVED VIA ADAPTATION, RESPECTIVELY. ORACLE DENOTES THE FULLY SUPERVISED FASTER-RCNN TRAINED ON THE TARGET DOMAIN.

Methods	person	rider	car	trunk	bus	train	moto	bicycle	mAP	mAP \star	gain
Backbone: VGG-16											
No DA	23.2	27.2	32.8	13.0	23.5	9.3	12.9	25.3	20.9	20.9	—
DAF [11]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6	18.8	8.8
MAF [12]	28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0	18.8	15.2
SWDA [13]	36.2	35.3	43.5	30.0	29.9	42.3	32.6	24.5	34.3	27.8	6.5
CFA [56]	41.9	38.7	56.7	22.6	41.5	26.8	24.6	35.5	36.0	18.8	17.2
CDN [51]	35.8	45.7	50.9	30.1	42.5	29.8	30.8	36.5	36.6	26.1	10.5
PD [21]	33.1	43.4	49.6	22.0	45.8	32.0	29.6	37.1	36.6	22.8	13.8
CDTD [37]	31.6	44.0	44.8	30.4	41.8	40.7	33.6	36.2	37.9	20.3	17.6
C2F [41]	43.2	37.4	52.1	34.7	34.0	46.9	29.9	30.8	38.6	20.8	17.8
FAA-SW [57]	39.5	41.3	47.0	34.5	39.3	44.0	31.9	28.4	38.3	22.0	16.3
MCAR [58]	32.0	42.1	43.9	31.3	44.1	43.4	37.4	36.6	38.8	23.4	15.4
DDF (ours)	37.2	46.3	51.9	24.7	43.9	34.2	33.5	40.8	39.1	20.9	18.2
Prior-DA [59]	36.4	47.3	51.7	22.8	47.6	34.1	36.0	38.7	39.3	24.4	14.9
Oracle	37.0	47.2	55.7	31.1	54.4	29.6	38.5	40.2	41.7	20.9	20.8
Backbone: ResNet-50											
No DA	28.7	33.7	36.7	18.3	30.1	16.8	21.7	27.3	26.7	—	—
MTOR [60]	30.6	41.4	44.0	21.9	38.6	40.6	28.3	35.6	35.1	26.9	8.2
GPA [40]	32.9	46.7	54.1	24.7	45.7	41.1	32.4	38.7	39.5	26.9	12.6
DIDN [61]	38.3	44.4	51.8	28.7	53.3	34.7	32.4	40.4	40.5	26.9	13.6
UaDAN [62]	36.5	46.1	53.6	28.9	49.4	42.7	32.3	38.9	41.1	26.9	14.2
DDF (ours)	37.6	45.5	56.1	30.7	50.4	47.0	31.1	39.8	42.3	26.7	15.6

tightest bounding boxes for the instance mask annotations, following the previous practices in [11]–[13], [36]. The 5000 total images in the original dataset are officially split into training, validation, and testing sets with 2975, 500, and 1525 images, respectively.

Foggy CityScapes [64] dataset simulates foggy scenes in different intensity levels on the CityScapes dataset. Following the previous methods, we employed the images with the highest foggy intensity level. The data split and annotations of the Foggy Cityscapes dataset are the same as the Cityscapes dataset.

SIM10K [65] dataset contains 10000 simulated images synthesized by the Grand Theft Auto (GTA) engine, with around 60K bounding box annotations of car instances.

KITTI [66] dataset contains 14999 real autonomous driving images obtained from a mid-sized city, with around 80K bounding boxes annotations for object detection study. Following previous works [11], [12], [14], we employed the training set with 7481 images for all experiments.

B. Comparison Experiments

1) *Adapting from the Normal to Foggy Weather:* Enhancing the model generalization ability under different weather conditions is crucial for automatic driving. In this section, we evaluate the model adaptation capability from the urban scene images captured in normal weather to the foggy ones, under the Cityscapes → Foggy Cityscapes setting. The training set with 2975 images from the Cityscapes and Foggy Cityscapes datasets are employed as the source and target domains, respectively, to train the models. For testing, the models are validated on the validation set of the Foggy Cityscapes dataset with 500 images. Under the same data split, our method is directly compared with the state-of-the-art methods, as shown

in Table II. In addition, we notice that the performance of the vanilla Faster-RCNN without adaptation is different in each comparison method. To this end, we first report the detection performance without adaptation for each method, which is denoted by the $mAP\star$ and directly copied from the original paper. Next, we report the performance gain of each method ($gain = mAP - mAP\star$) for a fair comparison, as suggested in [41]. Note that the performance gain can also be regarded as the adaptation ability for each cross-domain object detection method.

As illustrated in Table II, our DDF method has achieved the highest performance gain to the Faster RCNN without UDA. Compared with the Progressive Disentanglement (PD) [21] and Conditional Domain Normalization (CDN) [51] methods, which tackle the UDA object detection via feature disentanglement, our method has a better detection performance among most categories (7 out of 8). For the global-level disentanglement in PD [21], the calculation for the Mutual Information is suboptimal for decomposing the features in the UDA object detection architectures. In addition, its instance-level feature disentanglement requires reconstruction task learning, which introduces a decoder with auxiliary parameters. On the other hand, our GTD can avoid the influence from the batch size issue, and the ISD module requires no extra parameters, which further demonstrates the superiority and efficiency of our DDF method. During the inference process of CDN [51], the target features for detection predictions contain the domain-specific information from both the source and target domains. Therefore, the detection model trained on the source data still suffers from domain bias, which would lead to suboptimal adaptation performance for the target data. By feature disentanglement for both the source and target domains, our DDF method is able to alleviate the influence of the target-specific factors during testing and achieve better adaptation performance.

Besides, we notice that the Coarse-to-Fine Feature Adaptation (C2F) method [41] has achieved the best performance under a few categories. Leveraging prototype alignment for the instance-level features, the C2F has a better object classification ability. However, our method can still achieve better overall performance by decomposing the irrelevant factors that cause domain shift. Although the mAP for Prior-DA [59] has outperformed our DDF method, their mAP \star is still inferior to ours. MAP \star indicates the improvement of the proposed UDA method over the vanilla Faster R-CNN without adaption, which can alleviate the influence of the various implementation conditions for the comparison methods and therefore brings fairer comparison. Moreover, we notice that our DDF even outperforms the oracle fully supervised Faster-RCNN over two categories, which further demonstrates the feature adaptation ability of our DDF method.

To further evaluate the effectiveness of our proposed method with various backbone models, we have also incorporated our DDF with Faster R-CNN using the ResNet50 backbone and compared with the state-of-the-art UDA object detection methods using the same backbone [40], [60]–[62]. As indicated in Table II, our DDF method has achieved state-of-the-art detection and adaptation performance, which further demonstrates the robustness of the DDF method under different backbone models. Compared to DIDN [61] on the disentanglement-based cross-domain object detection via feature reconstruction, our DDF method has shown superior adaptation performance by decomposing the features based on global-level triplet optimization and local-level similarity learning.

TABLE III
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE SIM10K → CITYSCAPES EXPERIMENT.

Methods	mAP	mAP \star	gain
No DA	34.0	34.0	—
Baseline	39.5	34.0	5.5
DAF [11]	39.0	30.1	8.9
SWDA [13]	40.1	34.6	7.7
MAF [12]	41.1	30.1	11.0
SCDA [35]	43.0	34.0	9.0
ATF [14]	42.8	34.6	8.2
HTCN [36]	42.5	34.6	7.9
CDTD [37]	42.6	34.6	8.0
UMT [67]	43.1	34.3	8.8
C2F [41]	43.8	35.0	8.8
DDF (ours)	44.3	34.0	10.3

2) *Adaptation from the Synthetic to Real Scene:* Learning from synthetic images is essential for object detection since it alleviates the extensive costs in acquiring the bounding and category annotations for each instance. In this section, we study the UDA object detection for the car object from the synthesized dataset to the real one, i.e., the adaptation from the SIM10K to Cityscapes dataset. During training, the whole SIM10K dataset is employed as the source domain, and the Cityscapes training set with 2975 images as the target domain. The detailed results are shown in Table III. By outperforming the state-of-the-art methods such as HTCN [36] and UMT [67], the superiority of our proposed DDF method is further demonstrated. Among all the comparison methods, we notice that MAF [12] has a slightly better performance

gain than ours. By outperforming MAF under other settings (Table II, and IV), our DDF method is still more effective.

TABLE IV

QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE UDA OBJECT DETECTION TASKS BETWEEN THE KITTI AND CITYSCAPES DATASETS. $K \rightarrow C$ REPRESENTS ADAPTATION FROM THE KITTI TO THE CITYSCAPES DATASET, AND VICE VERSA.

Methods	$K \rightarrow C$			$C \rightarrow K$
	map	map \star	gain	map
No DA	32.6	32.6	—	53.5
Baseline	40.8	32.6	8.2	68.6
DAF [11]	38.5	30.2	8.3	64.1
SCDA [35]	42.5	37.4	5.1	—
MAF [12]	41.0	30.2	10.8	72.1
CDN [51]	44.9	37.1	7.8	—
DSS [68]	42.7	34.6	8.1	—
MeGA-CDA [69]	43.0	30.2	12.8	75.5
DDF (ours)	46.0	32.6	13.4	75.0

3) *Adaptation between Different Cameras:* Due to a large diversity of the hardware devices, there exists a domain shift between two datasets captured through various cameras. In this section, we conduct a UDA car object detection experiment between the KITTI and Cityscapes datasets. For the adaptation from the KITTI to Cityscapes, the whole KITTI dataset is employed as a source domain, and the training set for the Cityscapes dataset is adopted as the target domain. When adapting from the Cityscapes to KITTI, the training set of the Cityscapes is employed as the source domain, and the whole KITTI dataset is used as the target domain. The results are presented in Table IV. Under the KITTI → Cityscape setting, our method achieves the best performance, in terms of both the overall mAP and the performance gain. When adapting from Cityscapes to KITTI, we observe that the results for the “No DA” reported in all the other comparison methods [11], [12], [14], [69] are the same. To reach a fair comparison, we directly report the comparisons on the overall mAP for each method, where our DDF also achieves competitive performance.

We have also presented visual comparisons in Figure 4, where the Baseline is the basic adaptation method without disentanglement, and DAF is a state-of-the-art UDA object detection method [11]. It can be seen that our disentanglement-based method generated more true-positive predictions as well as fewer false predictions, further indicating the advantage of our method.

C. Ablation Studies

In this section, we conduct an ablation analysis of the effectiveness of the proposed modules under the Cityscapes → Foggy Cityscapes setting. The results are shown in Table V, where “No DA” represents the Faster-RCNN without domain adaptation and Baseline is the Faster-RCNN with global-level feature alignment (L_{di} in Equation 3). The w/o ISD method is implemented by removing L_{ISD} in Equation 10, which can also be treated as Baseline + GTD. Similarly, the w/o GTD method is equivalent to Baseline + ISD that neglects L_{GTD} in Equation 10. Apart from the ablated modules, all the methods have the same implementation details as in Section III-D.

Aligning the features at the global level, the Baseline outperforms the “No DA” method among all categories. To

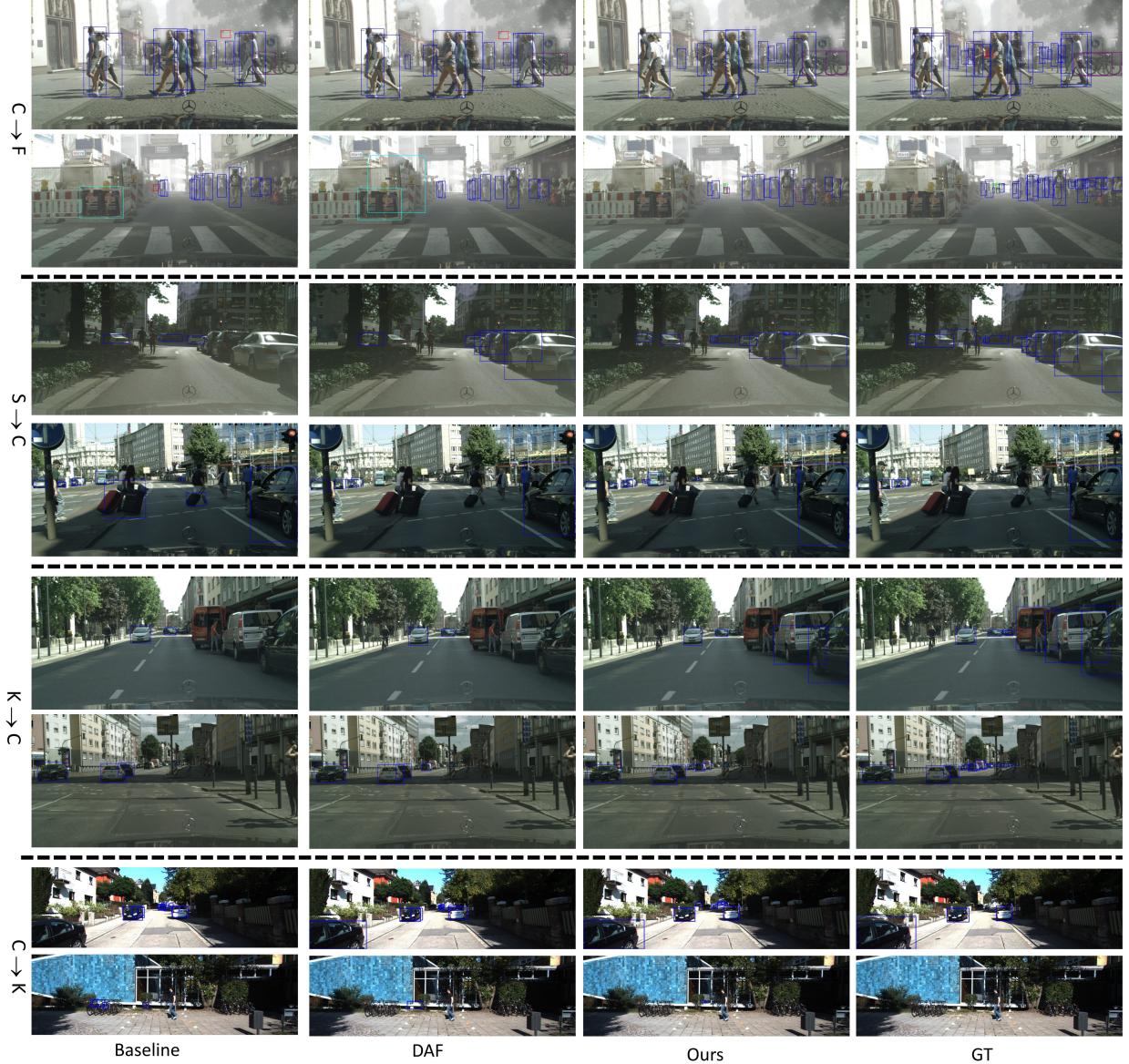


Fig. 4. Visual comparison of results achieved under various UDA settings: C → F : CityScapes → Foggy CityScapes, S → C : Sim10K → CityScapes, K → C : KITTI → CityScapes, and C → K : CityScapes → KITTI. The bounding box predictions are overlapped on the original images. (**Best viewed in color and zoom in**)

TABLE V
ABLATION STUDIES FOR THE GTD AND ISD MODULES ON THE CITYSCAPES → FOGGY CITYSCAPES EXPERIMENT.

Method	person	rider	car	trunk	bus	train	moto	bicycle	mAP
No DA	23.2	27.2	32.8	13.0	23.5	9.3	12.9	25.3	20.9
Baseline	25.8	32.8	46.1	21.4	36.9	24.7	17.9	25.7	28.9
w / o GTD	33.2	41.0	51.0	30.0	43.4	24.0	25.8	35.1	35.4
w / o L_{ds}	36.6	46.9	55.2	28.4	46.3	26.9	28.4	38.7	38.4
w / o L_{tri}	36.8	44.7	55.4	26.5	45.1	27.1	28.8	38.7	37.9
w / o ISD	35.0	44.4	51.0	23.4	42.0	22.3	30.0	38.2	35.8
w / o $L_{ISD-intra}$	36.0	42.2	54.0	27.2	42.5	25.1	32.2	37.4	37.1
w / o $L_{ISD-inter}$	35.7	44.0	51.5	25.6	40.5	36.0	28.6	38.5	37.5
DDF	37.2	46.3	51.9	24.7	43.9	34.2	33.5	40.8	39.1

further strengthen the model adaptation ability at the global and local levels, the GTD and ISD modules are proposed to bring consistent performance gains under most categories except for “trunk” and “train”. As two types of large-scale

vehicles, parts of the trains and trunks are more likely to be covered by the fog and thus become obscure and vague for the perception system, compared to the objects from the rest categories. To this end, detecting these two categories

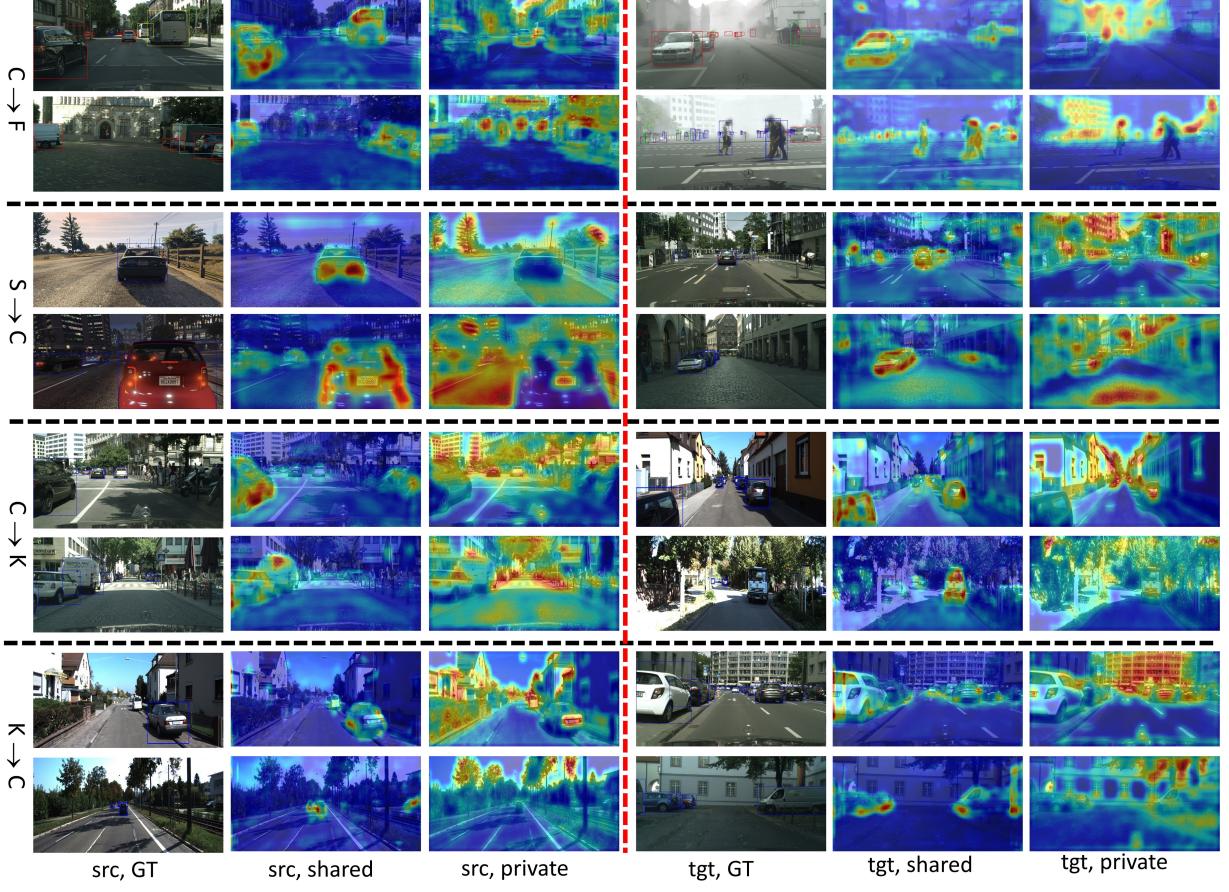


Fig. 5. Qualitative analysis on the disentanglement effectiveness of our DDF method under all UDA settings: $C \rightarrow F$: CityScapes \rightarrow Foggy CityScapes, $S \rightarrow C$: Sim10K \rightarrow CityScapes, $K \rightarrow C$: KITTI \rightarrow CityScapes, and $C \rightarrow K$: CityScapes \rightarrow KITTI. The left three columns are the results on the source domain (src), and the right three are on the target domain (tgt). All the features and labels are overlapped on the original images. **(Best viewed in color and zoom in)**

under the foggy situation is challenging, which is also proved by the inferior performance achieved by the fully supervised benchmark shown in Table II.

Although solely utilizing the GTD or ISD models incurs a slight performance drop on the “train” class, integrating them can still improve the performance significantly (about 10%). This indicates the cross-domain detection for the “train” class is challenging and requires feature disentanglement at both global and local stages. In addition, we notice that the DDF incorporating ISD and GTD degrades the performance than only using ISD. However, by achieving the best overall detection performance, incorporating GTG with ISD in DDF method is still demonstrated to be superior.

Moreover, we also include ablation studies on the inter-domain ISD ($L_{ISD-inter}$ in Equation 8), intra-domain ISD ($L_{ISD-intra}$ in Equation 7), global-level triplet optimization mechanism (L_{tri} in Equation 5), and the global-level domain-specific classification learning (L_{ds} in Equation 4). The results show performance drop by removing each module, hence demonstrating that all modules are useful. For the global-level GTD module, we notice the triplet optimization module has a stronger decomposition ability than the domain-specific classifier, as indicated by the larger performance drop when ablated. At the local level, the inter-domain ISD works slightly

better than the intra-domain one.

TABLE VI
QUANTITATIVE EVALUATION ON THE DOMAIN DISTANCE UNDER THE CITYSCAPES \rightarrow FOGGY CITYSCAPES SETTING. G AND I REPRESENT THE FEATURES AT THE GLOBAL AND LOCAL LEVELS, RESPECTIVELY.

	PAD \downarrow		EMD \downarrow	
	G	I	G	I
No DA	1.98	1.69	6.37	10.46
Baseline	1.51	0.53	1.83	10.34
Ours	0.75	0.39	1.38	9.64

D. Disentanglement Evidence Analysis

In this section, we first perform a qualitative analysis on the effectiveness of our disentanglement-based DDF method. For the No DA and Baseline, the visualized features are the outputs of the backbone network. For our DDF method, we visualize the domain-shared features as well as the domain-private features at the global level. All the features are firstly averaged along their channel dimensions and then normalized to range [0, 255]. Finally, the features are resized and overlapped with the original images for visualization. Note that we only visualize the image-level features. The instance-level features I_{sha}^s , I_{sha}^t , I_{pri}^s , and I_{pri}^t , on the other hand, are

vectors instead of tensors. Therefore, they do not have height and width and cannot be resized and overlaid on the original images for visualization.

As shown in Figures 5 and 1, the domain-shared features from our DDF at the global stages particularly focus on the instance objects which are crucial for detection task learning. However, the domain-shared features for the Baseline still focus on auxiliary background components, instead of particularly on the foreground instances. Therefore, in the Baseline method, the global-level domain-specific factors are introduced into the detection task learning, which results in the feature entanglement issue and further degrades the model adaptation performance. In addition, we observe the domain-private features for our DDF particularly concentrate on the information specific to its belonging domain, such as the components that reflect the weather conditions. This further demonstrates our E_p and E_s successfully extract the private and shared features for the given domains as expected.

To quantitatively evaluate the disentanglement ability of our DDF method, we calculate the feature distribution discrepancy of the domains under the Cityscape → Foggy Cityscape setting at both the global and local stages. At the global level, the feature distributions are acquired by average pooling applied on the outputs of the network backbone. For the local-level features, we randomly select 100 foreground object features from each category in each domain. To measure the domain distance, Proxy \mathcal{A} -distance (PAD) [70] and Earth Movers Distance (EMD) [71] are employed. As shown in Table VI, our method achieves a shorter feature distribution distance than the Baseline at both the global and local levels. Given a better representation of the domain-invariant features and a lower cross-domain feature discrepancy, our DDF method is proved to be effective in improving the Baseline via feature disentanglement.

E. Model Design Selections

In this section, we discuss the effectiveness of our DDF methods under different model design choices, including directly considering the similarity between the local-level shared cross-domain features, and introducing the adversarial learning strategies at the local level. The experiments are conducted under the Cityscape → Foggy Cityscape setting and the results are shown in Table VII.

TABLE VII

THE DDF METHOD WITH DIFFERENT MODEL DESIGNS ON THE CITYSCAPES → FOGGY CITYSCAPES EXPERIMENT. ‘INS-SIMMAX’ REPRESENTS THE DDF WITH THE DIRECT SIMILARITY MAXIMIZATION BETWEEN THE CROSS-DOMAIN SHARED FEATURES AT THE LOCAL LEVEL. ‘INS-TD’ INDICATES THE DDF EMPLOYING TRIPLET DISENTANGLEMENT MODULE AT THE LOCAL LEVEL.

Method	per	rid	car	tru	bus	tra	moto	bic	mAP
ins-simmax	32.4	36.6	43.5	24.1	40.7	16.5	22.4	32.2	31.1
ins-td	37.6	46.2	51.9	23.4	43.8	23.7	34.8	39.0	37.6
OG DDF	37.2	46.3	51.9	24.7	43.9	34.2	33.5	40.8	39.1

First, we include a similarity maximization mechanism for the local-level domain-shared features $I_{sha.}^s$ and $I_{sha.}^t$, by inducing the cosine similarity between $I_{sha.}^s$ and $I_{sha.}^t$ to be 1. This selection is referred to as ‘ins-simmax’ in Table VII. By

directly maximizing the similarity between the cross-domain shared features, the model performance drops significantly under many categories. One of the major reasons is that the cross-domain local-level features are ROIs under several categories, and directly narrowing the distribution gap between them incurs misalignment across different classes.

Next, we conduct an experiment replacing the similarity-based ISD module of the DDF method at the local level with the same disentanglement strategies as the global level, i.e., using the adversarial learning strategy and triplet optimization. Specifically, the optimization strategies in Equations 3, 4, and 5 are adopted into $I_{sha.}^s$, $I_{sha.}^t$, $I_{pri.}^s$ and $I_{pri.}^t$ at the local level. The overall method is referred to as ‘ins-td’ in Table VII, which achieves less competitive performance than the DDF method with similarity-based feature disentanglement at the local level. Due to the lack of annotations for the target images during training, the instance-level target features might not be accurate without the detection task learning, especially at the early training stage. Therefore, we think the inferior performance of the ‘ins-td’ method results from the influence of the unstable training process of the adversarial learning strategy. In addition, a similar phenomenon was also observed in [13] previously: including an adversarial domain discriminator for the local features in a UDA object detection model degrades the overall performance.

F. Computational Complexity Analysis

In this section, we conduct a computational complexity analysis for each principle module (GTD and ISD) in our proposed DDF method. The computational cost for the ablation studies under the $C \rightarrow F$ settings in Table V is as follows (number of parameters, and training time per iteration): 1) Baseline: $140M$, $0.34s/iter$; 2) Baseline + GTD: $145M$, $0.39s/iter$; 3) Baseline + ISD: $145M$, $0.51s/iter$; and 4) DDF: $145M$, $0.54s/iter$. For the model size, the auxiliary cost for both ISD and GTD modules is negligible. Although our ISD module brings extra training time, we think it is acceptable given the performance gain.

V. CONCLUSION

In this work, we present a novel Domain Disentanglement Faster-RCNN (DDF) method for cross-domain object detection via feature disentanglement. Specifically, we propose a GTD module to decompose the shared and private features within each domain at the global stage, as well as fitting the model batch size restrictions. At the instance level, an ISD module based on inter- and intra-domain similarity optimization is proposed to facilitate the feature disentanglement. Extensive experiments demonstrate the superiority of our DDF method by outperforming state-of-the-art methods on several benchmark UDA object detection tasks. Additionally, the qualitative and quantitative analysis on the disentanglement evidence further indicates the effectiveness of our method on decomposing the domain-specific factors and eliminating the domain-bias from them. In a larger perspective, the feature entanglement issues and the restrictions on model batch size are not limited to cross-domain object detection. With the

promising performance on cross-domain object detection, the DDF method can also be extended to other cross-domain object analysis tasks (e.g., segmentation, and tracing) in future work.

REFERENCES

- [1] R. Girshick, “Fast R-CNN,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.
- [2] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.
- [3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.
- [4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.
- [5] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.
- [6] A. Torralba and A. A. Efros, “Unbiased look at dataset bias,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1521–1528.
- [7] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” in *Advances in Neural Information Processing Systems*, 2014, pp. 3320–3328.
- [8] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
- [9] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 1180–1189.
- [10] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7167–7176.
- [11] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, “Domain adaptive faster R-CNN for object detection in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3339–3348.
- [12] Z. He and L. Zhang, “Multi-adversarial faster-rcnn for unrestricted object detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6668–6677.
- [13] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, “Strong-weak distribution alignment for adaptive object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6956–6965.
- [14] Z. He and L. Zhang, “Domain adaptive object detection via asymmetric tri-way faster-rcnn,” in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 309–324.
- [15] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, “Domain separation networks,” in *Advances in Neural Information Processing Systems*, 2016, pp. 343–351.
- [16] R. Cai, Z. Li, P. Wei, J. Qiao, K. Zhang, and Z. Hao, “Learning disentangled semantic representation for domain adaptation,” in *IJCAI: proceedings of the conference*, vol. 2019. NIH Public Access, 2019, p. 2060.
- [17] X. Peng, Z. Huang, X. Sun, and K. Saenko, “Domain agnostic learning with disentangled representations,” in *Proceedings of the International Conference on Machine Learning*, 2019, pp. 5102–5112.
- [18] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *International conference on machine learning*. PMLR, 2017, pp. 214–223.
- [19] X. Peng, Y. Li, and K. Saenko, “Domain2Vec: Domain embedding for unsupervised domain adaptation,” in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 756–774.
- [20] B. Gholami, P. Sahu, O. Rudovic, K. Bousmalis, and V. Pavlovic, “Unsupervised multi-target domain adaptation: An information theoretic approach,” *IEEE Transactions on Image Processing*, vol. 29, pp. 3993–4002, 2020.
- [21] A. Wu, Y. Han, L. Zhu, and Y. Yang, “Instance-invariant adaptive object detection via progressive disentanglement,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [22] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, “A kernel method for the two-sample-problem,” in *Advances in Neural Information Processing Systems*, 2007, pp. 513–520.
- [23] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, “Unsupervised visual domain adaptation using subspace alignment,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2960–2967.
- [24] T. Shermin, G. Lu, S. W. Teng, M. Murshed, and F. Sohel, “Adversarial network with multiple classifiers for open set domain adaptation,” *IEEE Transactions on Multimedia*, 2020.
- [25] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.
- [26] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, “Multimodal unsupervised image-to-image translation,” in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 172–189.
- [27] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, “CyCADA: Cycle-consistent adversarial domain adaptation,” in *Proceedings of the International Conference on Machine Learning*, 2018, pp. 1989–1998.
- [28] H. Emami, M. M. Aliabadi, M. Dong, and R. B. Chinnam, “Spa-gan: Spatial attention gan for image-to-image translation,” *IEEE Transactions on Multimedia*, vol. 23, pp. 391–401, 2020.
- [29] D. Liu, D. Zhang, Y. Song, F. Zhang, L. O’Donnell, H. Huang, M. Chen, and W. Cai, “PDAM: A panoptic-level feature alignment framework for unsupervised domain adaptive instance segmentation in microscopy images,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 154–165, 2020.
- [30] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, “Cross-domain weakly-supervised object detection through progressive domain adaptation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5001–5009.
- [31] J. Huang, D. Guan, A. Xiao, and S. Lu, “Cross-view regularization for domain adaptive panoptic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10133–10144.
- [32] Y. Li, L. Yuan, and N. Vasconcelos, “Bidirectional learning for domain adaptation of semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6936–6945.
- [33] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, “ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.
- [34] D. Liu, D. Zhang, Y. Song, F. Zhang, L. O’Donnell, H. Huang, M. Chen, and W. Cai, “Unsupervised instance segmentation in microscopy images via panoptic domain adaptation and task re-weighting,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4243–4252.
- [35] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, “Adapting object detectors via selective cross-domain alignment,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 687–696.
- [36] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, “Harmonizing transferability and discriminability for adapting object detectors,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8869–8878.
- [37] Z. Shen, M. Huang, J. Shi, Z. Liu, H. Maheshwari, Y. Zheng, X. Xue, M. Savvides, and T. S. Huang, “CDTD: A large-scale cross-domain benchmark for instance-level image-to-image translation and domain adaptive object detection,” *International Journal of Computer Vision*, pp. 1–20, 2020.
- [38] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, “Exploring categorical regularization for domain adaptive object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11724–11733.
- [39] C. Zhuang, X. Han, W. Huang, and M. Scott, “IFAN: Image-instance full alignment networks for adaptive object detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 13122–13129.
- [40] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang, “Cross-domain detection via graph-induced prototype alignment,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12355–12364.
- [41] Y. Zheng, D. Huang, S. Liu, and Y. Wang, “Cross-domain object detection through coarse-to-fine feature adaptation,” in *Proceedings of*

- the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 766–13 775.
- [42] J. Huang, D. Guan, A. Xiao, and S. Lu, “FSDR: Frequency space domain randomization for domain generalization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6891–6902.
- [43] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim, “Diversify and match: A domain adaptive representation learning paradigm for object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 456–12 465.
- [44] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [45] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, “beta-VAE: Learning basic visual concepts with a constrained variational framework.” in *International Conference on Learning Representations*, 2017.
- [46] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, “Disentangling factors of variation in deep representation using adversarial training,” in *Advances in Neural Information Processing Systems*, 2016, pp. 5040–5048.
- [47] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, “Infogan: Interpretable representation learning by information maximizing generative adversarial nets,” in *Advances in Neural Information Processing Systems*, 2016, pp. 2172–2180.
- [48] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, “Diverse image-to-image translation via disentangled representations,” in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 35–51.
- [49] V. K. Kurmi, S. Kumar, and V. P. Namboodiri, “Attending to discriminative certainty for domain adaptation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 491–500.
- [50] W. Deng, L. Zhao, Q. Liao, D. Guo, G. Kuang, D. Hu, M. Pietikainen, and L. Liu, “Informative feature disentanglement for unsupervised domain adaptation,” *IEEE Transactions on Multimedia*, 2021.
- [51] P. Su, K. Wang, X. Zeng, S. Tang, D. Chen, D. Qiu, and X. Wang, “Adapting object detectors with conditional domain normalization,” in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 403–419.
- [52] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, “Mutual information neural estimation,” in *Proceedings of the International Conference on Machine Learning*, 2018, pp. 531–540.
- [53] P. Chattopadhyay, Y. Balaji, and J. Hoffman, “Learning to balance specificity and invariance for in and out of domain generalization,” in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 301–318.
- [54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [55] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” *NeurIPS 2017 Autodiff Workshop*, 2017.
- [56] C.-C. Hsu, Y.-H. Tsai, Y.-Y. Lin, and M.-H. Yang, “Every pixel matters: Center-aware feature alignment for domain adaptive object detector,” in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 733–748.
- [57] J. Huang, D. Guan, A. Xiao, and S. Lu, “RDA: Robust domain adaptation via fourier adversarial attacking,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
- [58] Z. Zhao, Y. Guo, H. Shen, and J. Ye, “Adaptive object detection with dual multi-label prediction,” in *European Conference on Computer Vision*. Springer, 2020, pp. 54–69.
- [59] V. A. Sindagi, P. Oza, R. Yasarla, and V. M. Patel, “Prior-based domain adaptive object detection for hazy and rainy conditions,” in *European Conference on Computer Vision*. Springer, 2020, pp. 763–780.
- [60] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao, “Exploring object relation in mean teacher for cross-domain detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 457–11 466.
- [61] C. Lin, Z. Yuan, S. Zhao, P. Sun, C. Wang, and J. Cai, “Domain-invariant disentangled network for generalizable object detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8771–8780.
- [62] D. Guan, J. Huang, A. Xiao, S. Lu, and Y. Cao, “Uncertainty-aware unsupervised domain adaptation in object detection,” *IEEE Transactions on Multimedia*, 2021.
- [63] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.
- [64] C. Sakaridis, D. Dai, and L. Van Gool, “Semantic foggy scene understanding with synthetic data,” *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018.
- [65] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, “Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?” in *Proceedings of International Conference on Robotics and Automation*, 2017.
- [66] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.
- [67] J. Deng, W. Li, Y. Chen, and L. Duan, “Unbiased mean teacher for cross-domain object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4091–4101.
- [68] Y. Wang, R. Zhang, S. Zhang, M. Li, Y. Xia, X. Zhang, and S. Liu, “Domain-specific suppression for adaptive object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9603–9612.
- [69] V. VS, V. Gupta, P. Oza, V. A. Sindagi, and V. M. Patel, “MeGACDA: Memory guided attention for category-aware unsupervised domain adaptive object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4516–4526.
- [70] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira *et al.*, “Analysis of representations for domain adaptation,” *Advances in neural information processing systems*, vol. 19, p. 137, 2007.
- [71] Y. Rubner, C. Tomasi, and L. J. Guibas, “The earth mover’s distance as a metric for image retrieval,” *International journal of computer vision*, vol. 40, no. 2, pp. 99–121, 2000.