IEEE Signal Processing Society
IEEE Transactions on Image Processing

Rethinking Cross-Domain Pedestrian Detection: A Background-Focused Distribution Alignment Framework for One-Stage Detectors

Journal:	*Transactions on Image Processing*
Manuscript ID	TIP-27931-2022
Manuscript Type:	Regular Paper
Date Submitted by the Author:	27-Jun-2022
Complete List of Authors:	Cai, Yancheng; Fudan University, School of Information Science and Technology Zhang, Bo; Shanghai Artificial Intelligence Laboratory Li, Baopu; Oracle Corp Chen, Tao; Fudan University, School of Information Science and Technology Yan, Hongliang; Shanghai Artificial Intelligence Laboratory Zhang, Jingdong; Fudan University, School of Information Science and Technology Xu, Jiahao; Fudan University, School of Information Science and Technology
Subject Category Please select at least one subject category that best reflects the scope of your manuscript:	Image & Video Processing Techniques
EDICS:	4. SMR-Rep Image and Video Representation < Image & Video Sensing, Modeling, and Representation, 36. ARS-SRV Image and Video Synthesis, Rendering, and Visualization < Image and Video Analysis, Synthesis and Retrieval

Rethinking Cross-Domain Pedestrian Detection: A Background-Focused Distribution Alignment Framework for One-Stage Detectors

Yancheng Cai, Bo Zhang, Baopu Li, *Member, IEEE*, Tao Chen, *Senior Member, IEEE*, Hongliang Yan, Jingdong Zhang, Jiahao Xu

Abstract—**Cross-domain pedestrian detection aims to generalize pedestrian detectors from one label-rich environment to another label-scarce environment, which is vital in enormous real-world applications. Recent works generally rely on domain alignment to train domain-adaptive detectors, either on image-level or instance-level. Due to the proposal-free rapid detection, we focus on the one-stage domain-adaptive detector design in this work. We find that the lack of instance-level proposals for the one-stage detector makes it only be able to do image-level feature alignment, causing the foreground-background misalignment issue, that is, the foreground features in the source domain image are falsely aligned with background features in the target domain image. To resolve the conflict between foreground and background in the alignment stage, we systematically analyze the importance of foreground and background in image-level cross-domain alignment, and learn that background plays a more important role in image-level cross-domain alignment. Therefore, we focus on background cross-domain feature alignment while minimizing the influence of foreground features on the cross-domain alignment stage. This paper proposes a novel Background-focused Distribution Alignment Framework (BFDA) to train domain adaptive one-stage pedestrian detectors. Specifically, BFDA first decouples the background features from the whole image feature maps and then aligns them via a novel long-short-range discriminator. Extensive experiments show that BFDA significantly enhances the cross-domain pedestrian detection performance, compared with the mainstream domain adaptation technologies for either one-stage or two-stage detectors. Moreover, by employing the efficient one-stage detector (YOLOv5) as the backbone, BFDA can reach 217.4 FPS (640×480 pixels) on NVIDIA Tesla V100 (7∼12 times FPS of the existing frameworks), which is very meaningful for practical applications. The code will be made publicly available.**

Index Terms—**Cross-domain pedestrian detection, one-stage object detectors, image-level feature alignment**

I. Introduction

PEDESTRIAN Detection (PD) has been a long-standing and essential task due to its key role in many fields such as autonomous driving [1], pedestrian re-identification [2], and video surveillance [3]. Thanks to the Convolutional Neural Networks (CNNs), PD frameworks have made significant progress recently. However, current PD methods highly rely on the consistency assumption between training and test data distribution, which is hard to be guaranteed in the real world. As a result, many well-trained PD models [4]–[8] in one environment (e.g., nature) fail to generalize to another environment (e.g., dense fog, heavy rain, illumination variation), resulting in significant performance drops.

To solve the above problem, researchers have suggested some cross-domain PD methods [9]–[15]. Most of them relieve the detectors from cross-domain performance degradation by aligning between source and target domains, either from image-level or instance-level. For example, DA-Faster-RCNN [16] uses image-level and instance-level adaptation to enhance the detector's domain generalization ability. Further, a Selective Alignment Network (SAN) [11] is designed to alleviate the inter-instance difference, suggesting that aligning each subtype of instances is more reasonable for cross-domain PD. However, these methods are all based on two-stage detectors (*e.g.*, Faster RCNN [17]), which lack sufficient inference speed for practical applications. One-stage detectors (*e.g.*, YOLOv5[*]) have sufficient speed, but due to the lack of the instance-level proposals, they cannot truly utilize the most mainstream instance-level feature alignment algorithms. In this case, one-stage cross-domain detectors mainly rely on image-level feature alignment.

However, in image-level feature alignment, the misalignment of background and foreground (red arrows in Fig. 1) has not been resolved. Therefore, one-stage pedestrian detectors suffer a considerable drop in cross-domain accuracy due to imperfect alignment algorithms. In this work, we reduce the risk of foreground-background misalignment[†] by focusing on the alignment of cross-domain background features and trying to avoid the participation of foreground features in cross-domain alignment. Specifically, we systematically study the respective importance of foreground and background in cross-domain tasks. We reveal an essential observation that background alignment plays a key role in the domain adaptive PD task. Our idea is based on the following two findings:

First, the image-level adaptation directly suffers from the foreground-background feature misalignment issue for dense

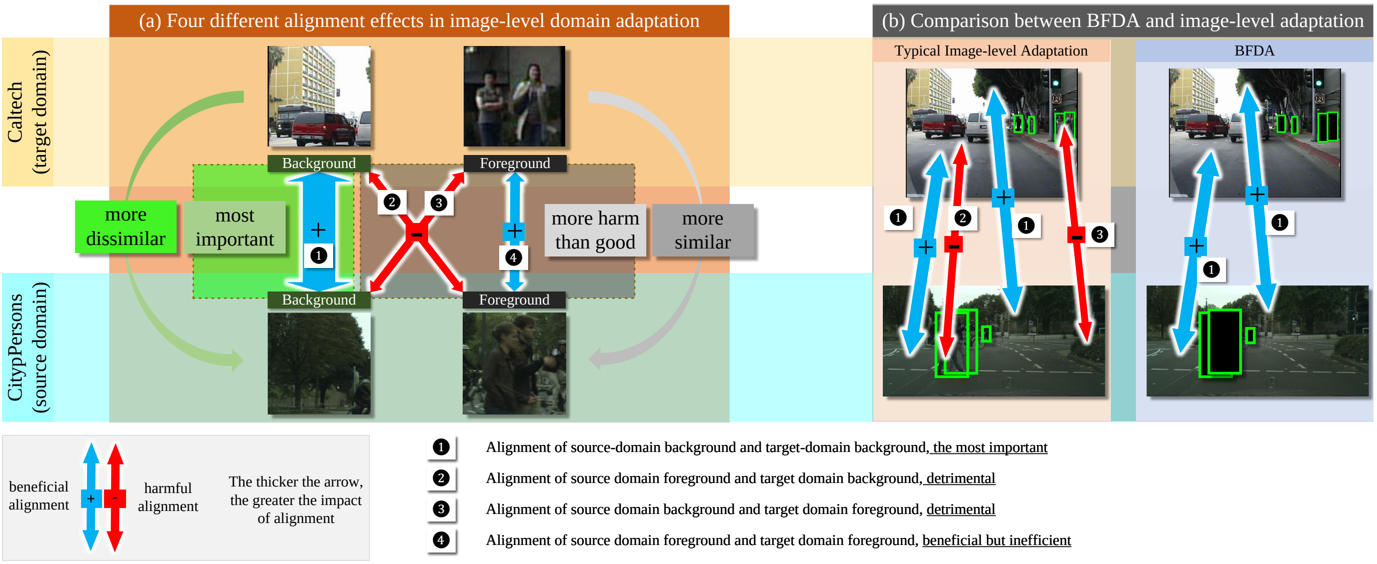
Yancheng Cai, Bo Zhang, Jingdong Zhang, Jiahao Xu, Tao Chen are with the School of Information Science and Technology, Fudan University, Shanghai 200433, China (Corresponding author: Tao Chen, e-mail: eetchen@fudan.edu.cn, tel: +86-2131242503).

Baopu Li is with Oracle, 100 Oracle Pkwy, Redwood City, CA 94065 (e-mail: Bpli.cuhk@Gmail.com).

Hongliang Yan is with Shanghai AI Laboratory, Shanghai, 200232, China (e-mail: yhldhit@gmail.com).

[*] https://github.com/ultralytics/YOLOv5

[†] Note that we believe that instance-level domain adaptation does not have the foreground-background feature misalignment issue, since instance-level proposals can naturally decouple foreground and background for alignment. However, the one-stage detector does not have the conditions for instance-level domain adaptation, and can only perform image-level domain adaptation.


(a) Four different alignment effects in image-level domain adaptation
Caltech (target domain)
CitypPersons (source domain)
Background
Foreground
Background
Foreground
more dissimilar
most important
more harm than good
more similar
❶
❷
❸
❹
+
(b) Comparison between BFDA and image-level adaptation
Typical Image-level Adaptation
BFDA
beneficial alignment
harmful alignment
The thicker the arrow, the greater the impact of alignment

❶ Alignment of source-domain background and target-domain background, the most important

❷ Alignment of source domain foreground and target domain background, detrimental

❸ Alignment of source domain background and target domain foreground, detrimental

❹ Alignment of source domain foreground and target domain foreground, beneficial but inefficient

Fig. 1. The foreground-background feature misalignment issue and the importance of background feature. In Image-level Adaptation, the presence of foreground results in three alignments (❷,❸,❹) simultaneously (because the foreground positions of different images are different, it will inevitably lead to the misalignment of the foreground of some images and the background of other images). We demonstrate that the background alignment **plays a major role**, and the three alignments (❷,❸,❹) in the grey box **yield more harmful effects than good ones**. Therefore, we focus on background feature alignment, while minimizing foreground interference in image-level cross-domain alignment.

prediction tasks due to the variable instance positions in different images. For example, Fig. 2 shows the profitless effects of such an alignment strategy. It can be seen that the foreground regions (pedestrians) have the highest feature response peaks. Due to the different locations of pedestrian foreground and background positions in different images, the pedestrian foreground of one source image and the background of another target image may occupy the same spatial position. In this case, the image-level adaptation process will erroneously align the foreground and background and vice versa (as shown in Fig.1). The foreground-background feature misalignment issue is the core issue in cross-domain PD that we want to solve.

Second, ensuring the background feature consistency between domains is indispensable for cross-domain PD. To better demonstrate this, we conduct preliminary studies on the mainstream one-stage detector YOLOv5 and two-stage detector Faster RCNN. We find that the accuracy drop caused by the feature change in background regions is much more significant than that in foreground regions, further hinting that the current pedestrian detectors are relatively sensitive to the background variations. Furthermore, when performing the background-focused distribution alignment, we observe that short-range background changes result in a more considerable decline in detection accuracy compared with long-range background (far away from pedestrian instances). It may be because nearer contextual background information around pedestrians is more critical for their position predictions than the further background. Based on this finding, we attempted to address the foreground-background feature misalignment issue by focusing on background alignment.

Inspired by the above two insights, we are motivated to rethink the cross-domain PD pipeline to alleviate the negative impact of the foreground-background feature misalignment issue on the existing one-stage detectors. Our method aims to focus on the background features' cross-domain alignment while mitigating the interference of foreground features on cross-domain alignment. Specifically, we first develop a Background Decoupling Module (fed with the feature maps from the detection head) and decouple the background feature with the help of a Feature Generation Module (inspired by CycleGAN's algorithm [18]) to solve the foreground-background feature misalignment issue as mentioned above. Second, we propose a Long-short-range Domain Discriminator using a Transformer-CNN-based parallel structure, assigning global and local attention to different background ranges according to their distance from pedestrian instances. Comprehensive tests show much better performance of the proposed novel scheme.

Our main contributions can be summarized as follows:

(1) Our research reveals the foreground-background feature misalignment issue that one-stage pedestrian detectors face when performing image-level feature alignment. At the same time, we uncover a critical but underappreciated aspect in achieving transferable PD between different domains: ensuring inter-domain consistency of background features. To the best of our knowledge, we are the first to propose focusing on background alignment in cross-domain detection.

(2) A novel background-focused domain adaptive PD framework is proposed, consisting of three key modules: the Background Decoupling Module (*BDM*), the Feature Generation Module (*FGM*), and the parallel Transformer-CNN-based Long-short-range Domain Discriminator (*LSD*). Such a framework can effectively decouple the background feature from original feature maps to mitigate the foreground-background feature misalign-

ment issue.

(3) Cross-domain PD experiments are performed on BFDA, and the results demonstrate that the proposed BFDA can produce SOTA results on one-stage detector YOLOv5.

II. Related Work

A. Pedestrian Detection.

The rise of deep learning technology has promoted the development of PD research [4]–[8], [19], [20]. They can be roughly divided into anchor-based and anchor-free methods. Anchor-based methods detect objects in a given image by classifying and regressing anchor boxes. They can be subdivided into one-stage and two-stage methods, in which the two-stage methods generate proposals and then calculate the confidence score for each proposal. For example, MGAN's [6] attention network emphasizes visible pedestrian areas while adjusting physical characteristics to suppress occluded areas. The one-stage methods directly process the detection classification and regression in one step. For example, ALFNet [4] stacks a series of predictors to gradually evolve the default anchor boxes of SSD [21]. The anchor-free methods like CSP [7] focus on other pedestrian features such as the center and corners. However, these methods face a big challenge for test images with large differences in feature distribution. We intend to overcome such a challenge in this work. Before deep learning, some traditional vision works have studied how to use the background to solve the tracking tasks [22]–[24], while our work differs from these works in both the background exploration way and the aiming task.

B. Cross-domain Object Detection.

Aiming at the problem that the detector cannot be generalized to datasets with significant domain gaps, cross-domain object detection technology is proposed. As the pioneer in this field, [16] proposed image-level and instance-level adaptation methods, then aligned these features simultaneously. [25] designed an adaptive method based on strong-local and weak-global alignment. [26] deployed an ancillary net parallel to the chief net and formulated an asymmetric tri-way architecture to avoid the model collapse in aligning procedure. [27] integrate the intermediate domain image generator and multi-scale adversarial feature alignments into a single framework to bridge the domain divergence progressively. [28] introduce a reinforcement learning-based method to gradually refine both source and target instances and alleviate the negative transfer. However, the above methods are all based on two-stage detection frameworks, which are highly dependent on region proposal and the region feature based on the ROI pooling. For one-stage detectors, [29] introduces a weak self-training method to suppress the effects of False Negatives and False Positives, and the adversarial background score regularization to extract discriminative features for target backgrounds to aid foreground alignment. [30] proposed a semantic enhance module to strengthen the foreground and multi-scale features for cross-domain adaptation. [31] proposed to reweight the image level align procedure and match foreground objects Pattern guided by the categorical information. [32] addressed the conflict among foreground classes. Unfortunately, these methods are not designed for pedestrians and face the foreground-background feature misalignment issue, so these cross-domain frameworks may be suboptimal when recognizing pedestrians with diverse appearances. We discard all foreground classes and only study the background (more complex than a class and can contain arbitrary objects) for alignment.

C. Transformer in Vision.

Transformer originates from natural language processing. ViT [33] demonstrates that an improved Transformer can achieve SOTA results in the image classification task with sufficient data (e.g., ImageNet-22k, JFT-300M). Convolution usually only has local attention and cannot focus on global features. However, Transformer is born with global attention, so it has been applied to various computer vision tasks. For example, DETR [34] successfully applies Transformer encoder-decoder architecture to the object detection task. On the other hand, some methods [35]–[38] have successfully reduced the Transformer architecture's parameters considering its enormous storage and computation requirements. For instance, CvT [35] takes convolutional token embedding and convolutional projection technique. In our work, we take advantage of the global attention capability of the Transformer architecture and mix the long-range attention module with a short-range convolution attention module, which is called the Long-short-range Domain Discriminator.

III. Rethinking Cross-domain PD

This section first reviews image-level cross-domain adaptation, the primary method in cross-domain PD research, and the most important algorithm that one-stage PD detectors can use. Here we present some important observations on the discriminator's theory. Then, we visualize the foreground-background feature misalignment issue and demonstrate the importance of background in PD tasks through experiments. Finally, we theoretically propose a new cross-domain PD paradigm.

A. Typical Image-level Adaptation

In fact, as early as 2010, before the emergence of deep learning, [39] first proved that the use of domain discriminator can effectively reduce the the difference $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ in feature distribution $F(I)$ between domains. [39] showed that this method can control the upper bound of the domain generalization error of a classifier model in the target domain.

DA-Faster-RCNN [16] first formulated the image-level adaptation. Object detection can be regarded as learning the posterior distribution $P(C, B|I)$, where I is the image representation, B are the bounding boxes of objects, and $C \in \{1, \cdots, K\}$ are different classes. The joint distribution of training samples can be expressed as $P(C, B, I)$, with $P_{\mathcal{S}}(C, B, I)$ and $P_{\mathcal{T}}(C, B, I)$ representing the joint distribution of source and target samples, respectively. When domain gaps exist: $P_{\mathcal{S}}(C, B, I) \neq P_{\mathcal{T}}(C, B, I)$. The joint distribution can be decomposed into:

$$P(C, B, I) = P(C, B|I)P(I). \tag{1}$$

According to the covariate shift assumption [40]:

$$P_{\mathcal{S}}(C, B|I) = P_{\mathcal{T}}(C, B|I). \tag{2}$$

Researchers are committed to using the adversarial feature learning methods to train a feature extractor F so that:

$$P_{\mathcal{S}}(F(I)) \approx P_{\mathcal{T}}(F(I)), \tag{3}$$

$$P_{\mathcal{S}}(C, B, F(I)) \approx P_{\mathcal{T}}(C, B, F(I)). \tag{4}$$

In fact, DA-Faster-RCNN [16] directly uses the theoretical proof result of [39] and moves it directly to the cross-domain object detection task. However, it is very worth noting that the theory of [39] is specially designed for the cross-domain image classification task. Let's review the key proof steps (Theorem 1 in [39]). [39] define a domain as a pair consisting of a distribution $\mathcal{D}$ on inputs $\mathcal{X}$ and a labeling function $f : \mathcal{X} \to [0, 1]$. They denote by $\langle \mathcal{D}_S, f_S \rangle$ the source domain and $\langle \mathcal{D}_T, f_T \rangle$ the target domain. A hypothesis is a function $h : \mathcal{X} \to \{0, 1\}$. The probability according to the distribution $\mathcal{D}_S$ that a hypothesis h disagrees with a labeling function f (which can also be a hypothesis) is defined as: $\epsilon_S(h, f) = \mathrm{E}_{\mathbf{x}\sim\mathcal{D}_S}[|h(\mathbf{x}) - f(\mathbf{x})|]$.

$$\begin{aligned}\epsilon_T(h) &= \epsilon_T(h) + \epsilon_S(h) - \epsilon_S(h) + \epsilon_S\left(h, f_T\right) - \epsilon_S\left(h,_T\ \right)\\ &\leq \epsilon_S(h) + \left|\epsilon_S\left(h, f_T\right) - \epsilon_S\left(h, f_S\right)\right| + \left|\epsilon_T\left(h, f_T\right) - \epsilon_S\left(h, f_T\right)\right|\\ &\leq \epsilon_S(h) + \mathrm{E}_{\mathcal{D}_S}\left[\left|f_S(\mathbf{x}) - f_T(\mathbf{x})\right|\right] + \left|\epsilon_T\left(h, f_T\right) - \epsilon_S\left(h, f_T\right)\right|\\ &\leq \epsilon_S(h) + \mathrm{E}_{\mathcal{D}_S}\left[\left|f_S(\mathbf{x}) - f_T(\mathbf{x})\right|\right] + \\ &\int \left|\phi_S(\mathbf{x}) - \phi_T(\mathbf{x})\right|\left|h(\mathbf{x}) - f_T(\mathbf{x})\right| d\mathbf{x}\\ &\leq \epsilon_S(h) + \underline{\mathrm{E}_{\mathcal{D}_S}\left[\left|f_S(\mathbf{x}) - f_T(\mathbf{x})\right|\right]} + d_1\left(\mathcal{D}_S, \mathcal{D}_T\right)\end{aligned} \tag{5}$$

We only need to observe the underlined item, which is considered small in the original text in [39] and discarded (page 155): "... and the third is the difference in labeling functions across the two domains, which we expect to be small." However, this assumption is only applicable for the image classification task, and for object detection tasks, this assumption does not hold. This is because this term can be regarded as a kind of intra-domain gap in feature maps. On image classification or semantic segmentation tasks, the intra-domain gap of feature maps is much smaller than the inter-domain gap of primary alignment. However, for the object detection task, since the feature maps have foreground peaks (Fig. 2), the peak positions may be different between different foreground maps, so this item cannot be ignored. Therefore, foreground feature alignment is not a very favorite part of image-level cross-domain.

B. The foreground-background feature misalignment issue

However, the above-mentioned image-level domain adaptation methods based on adversarial feature learning inevitably face the foreground-background feature misalignment issue. To illustrate this issue more clearly, we visualize the change of feature maps throughout the process of image-level cross-domain alignment (as shown in Fig. 2).

The alignments can bring non-negligible hazards in dense prediction tasks such as object detection, especially PD. That is the **foreground-background feature misalignment issue**. The feature map's most prominent parts (peaks) represent possible pedestrian instances, which are also the most concern during the feature alignment process. However, the location information of pedestrian instances in different images may be quite different. In this case, typical image-level adaptation results in many foreground regions of some images incorrectly aligning with the background regions of other images. As a result, the features in background regions are also inevitably accounted for in the alignment process, leading to fake peaks (wrong boxes in Fig. 2), which leads to false detections.

C. The importance of background in pedestrian detection

Considering the foreground-background feature misalignment issue, we turn to look into whether we can achieve cross-domain PD by aligning **only** the background, which has never been considered before. To be persuasive, we use the mainstream detector YOLOv5 to conduct experiments on the CityPersons and BDD10k datasets. At the same time, in order to explore whether the importance of the background is only a feature of the one-stage detector, we also use the mainstream two-stage detector Faster RCNN for experiments. The general evaluation metrics $MR^{-2}(\%)$ for pedestrian detection and $\mathrm{AP}_{50}(\%)$ for object detection are used for performance comparison. **The smaller (↓)** $MR^{-2}(\%)$ **and larger (↑)** $\mathrm{AP}_{50}(\%)$ **indicate that the method is better.** PD task uses MR^{-2} as the standard evaluation metric (Therefore, we only report this criterion in the section V), and we show $\mathrm{AP}_{50}(\%)$ here just to make our findings more convincing.

First, we study the degradation of PD accuracy caused by feature change in three regions (outer-bounding-box background, inner-bounding-box background, and foreground), as shown in Fig. 3 and Table I. These experimental results demonstrate that dominant person detectors are very sensitive to changes in background features. Therefore, in the image-level feature alignment process, the background feature alignment is even more important than the foreground feature alignment.

Second, we study the influence of different background regions (with different ranges to the foreground instances) on the detection results, as shown in Fig. 4 and Table II. The experimental results show that different background regions have different effects on the results. Mainstream detectors are more sensitive to feature change in background regions close to the instance (short-range).

D. New Paradigm: Background-focused Feature Alignment

Motivated by above studies, we propose a new paradigm named background-focused distribution alignment. Given an input image, $\boldsymbol{A}$ denotes the outer-bounding-box background (The **red part** in Fig. 3). Meanwhile, $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{w}_i\}, i = 1, \ldots, n$ denote the inner-bounding-box background (The **green part** in Fig. 3) and pixel-level foreground (The **blue part** in Fig. 3), respectively. $\boldsymbol{F}_S$ refers to the detector trained only on the source domain, and K_S, K_T respectively represent the detection performance on the source and target domain test datasets. We assume that $\{\boldsymbol{w}_k, \boldsymbol{w}_q\}, k \neq q$ are independent, and $\{\boldsymbol{x}_k, \boldsymbol{x}_q\}, k \neq q$ are independent,

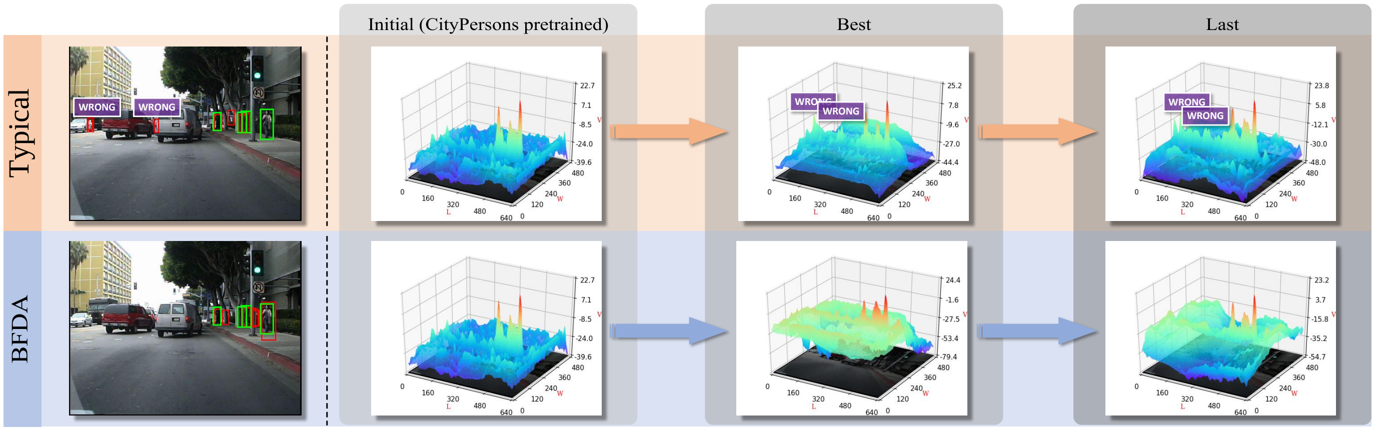


Fig. 2. Illustration of the feature map evolving process of the typical image-level adaptation (first row) and BFDA (second row) (CityPersons→Caltech, **test on Caltech**). The left part denotes the detection results (green and red boxes denote the ground truths and predictions, respectively). We study the feature map with the largest resolution after the first convolution layer in the detection head of YOLOv5. Initial, Best, and Last represent the beginning of cross-domain training (pretrained), the epoch with the best detection results, and the last epoch, respectively. In the 3D feature map visualization, two horizontal axes represent the spatial dimensions, while the vertical axis represents the channel dimension (sum along the channel). It can be seen that, due to the misalignment between foreground instances and background regions, **wrong peaks** gradually appear on the feature map during training, resulting in false detections (purple boxes) when typical image-level adaptation is used.). In fact, this is because the background features of some images are aligned with the foreground features of other images, so the detector's ability to distinguish the foreground and background of the image decreases. We call it **the foreground-background feature misalignment issue**. Our method alleviates this problem by focusing on the background feature alignment between domains and reducing the interference of the foreground in image-level cross-domain feature alignment.

TABLE I

Ablation study of feature change in different regions of images. We use YOLOv5 and Faster RCNN to conduct experiments on the CityPersons [41] and BDD10K [42] datasets. BDD10K does not provide occlusion labels, so partial and heavy (Section V-B) cannot be reported. Obviously, changes in background features have a greater impact on detection accuracy than the foreground feature change. ($*$ The result in the seventh-to-last row is correct.)

Method	Dataset	Foreground	Inner-bounding-box background	Outer-bounding-box background	MR^{-2}(%) ↓				AP$_{50}$(%) ↑
					reasonable	bare	partial	heavy	
YOLOv5	CityPersons	✓	✓	✓	10.45	7.38	8.63	40.28	83.5
YOLOv5	CityPersons	✓	✓		18.74	11.02	19.33	46.22	65.1(-18.4)
YOLOv5	CityPersons	✓		✓	34.56	26.53	37.76	72.99	56.0(-27.5)
YOLOv5	CityPersons		✓	✓	38.07	31.04	40.78	71.16	58.9(-24.6)
YOLOv5	CityPersons	✓			47.36	41.33	46.96	74.43	47.5(-36.0)
YOLOv5	BDD10K	✓	✓	✓	13.46	13.46	-	-	73.8
YOLOv5	BDD10K	✓	✓		50.20	50.20	-	-	25.5(-48.3)
YOLOv5	BDD10K	✓		✓	18.02	18.02	-	-	68.8(-5.0)
YOLOv5	BDD10K		✓	✓	11.28	11.28	-	-	82.5(+8.7)*
YOLOv5	BDD10K	✓			38.55	38.55	-	-	39.7(-34.1)
Faster RCNN	CityPersons	✓	✓	✓	19.65	10.84	21.33	85.10	68.3
Faster RCNN	CityPersons	✓	✓		69.13	63.83	67.00	93.25	20.5(-47.8)
Faster RCNN	CityPersons	✓		✓	45.53	34.98	51.33	87.96	35.4(-32.9)
Faster RCNN	CityPersons		✓	✓	56.29	46.43	62.49	94.90	35.5(-32.8)
Faster RCNN	CityPersons	✓			52.68	41.35	54.22	91.16	31.9(-36.4)

which means that (P stands for feature probability distribution): $P\left(\boldsymbol{w_S}\right)=\prod_{i=1}^{n_S} P\left(\boldsymbol{w}_{Si}\right)$, $P\left(\boldsymbol{w_T}\right)=\prod_{i=1}^{n_T} P\left(\boldsymbol{w}_{Ti}\right)$, $P\left(\boldsymbol{x_S}\right)=\prod_{i=1}^{n_S} P\left(\boldsymbol{x}_{Si}\right)$, $P\left(\boldsymbol{x_T}\right)=\prod_{i=1}^{n_T} P\left(\boldsymbol{x}_{Ti}\right)$.

Also, we have Equation 6 and Equation 7. $\boldsymbol{F}_S$ refers to the detector trained only on the source domain. In simple terms, the detection accuracy is related to the detector and its three inputs:

$$K_S \propto \boldsymbol{F}_S\left(P\left(\boldsymbol{w_S}\right), P\left(\boldsymbol{x_S}\right), P\left(\boldsymbol{A}_S\right)\right), \tag{6}$$

$$K_T \propto \boldsymbol{F}_S\left(P\left(\boldsymbol{w_T}\right), P\left(\boldsymbol{x_T}\right), P\left(\boldsymbol{A}_T\right)\right). \tag{7}$$

We derive the following:

$$K_S \propto \boldsymbol{F}_S\left(\prod_{i=1}^{n_S} P\left(\boldsymbol{w}_{Si}\right), \prod_{i=1}^{n_S} P\left(\boldsymbol{x}_{Si}\right), P\left(\boldsymbol{A}_S\right)\right), \tag{8}$$

$$K_T \propto \boldsymbol{F}_S\left(\prod_{j=1}^{n_T} P\left(\boldsymbol{w}_{Tj}\right), \prod_{j=1}^{n_T} P\left(\boldsymbol{x}_{Tj}\right), P\left(\boldsymbol{A}_T\right)\right). \tag{9}$$

Using the total differential equation ($w = w(x, y, z) \Longrightarrow$

TABLE II
ABLATION STUDY OF FEATURE CHANGE IN DIFFERENT RANGES OF BACKGROUNDS. CHANGES IN THE SHORT-RANGE BACKGROUND FEATURE SIGNIFICANTLY IMPACT DETECTION ACCURACY MORE THAN THE LONG-RANGE BACKGROUND FEATURE CHANGE.

Method	Dataset	Range	$MR^{-2}(\%)\downarrow$				$\mathrm{AP}_{50}(\%)\uparrow$
			reasonable	bare	partial	heavy	
YOLOv5	CityPersons	*original image*	10.45	7.38	8.63	40.28	83.5
YOLOv5	CityPersons	*no_1.0_2.0*	25.30	18.96	26.86	55.80	58.9(-24.6)
YOLOv5	CityPersons	*no_1.5_2.5*	12.50	7.77	11.54	41.96	75.6(-7.9)
YOLOv5	CityPersons	*no_2.0_3.0*	10.21	6.88	8.22	39.24	81.0(-2.5)
YOLOv5	BDD10K	*original image*	13.46	13.46	-	-	73.8
YOLOv5	BDD10K	*no_1.0_2.0*	46.92	46.92	-	-	27.2(-46.6)
YOLOv5	BDD10K	*no_1.5_2.5*	20.64	20.64	-	-	49.2(-24.6)
YOLOv5	BDD10K	*no_2.0_3.0*	16.31	16.31	-	-	58.1(-15.7)
Faster RCNN	CityPersons	*original image*	19.65	10.84	21.33	85.10	68.3
Faster RCNN	CityPersons	*no_1.0_2.0*	76.15	72.57	76.57	94.23	16.6(-51.7)
Faster RCNN	CityPersons	*no_1.5_2.5*	53.86	46.93	54.77	89.41	35.6(-32.7)
Faster RCNN	CityPersons	*no_2.0_3.0*	37.65	28.92	39.54	86.94	47.6(-20.7)

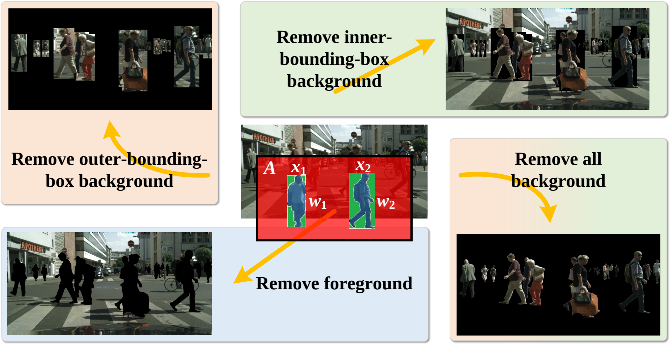

Remove inner-bounding-box background
Remove outer-bounding-box background
A
x1
x2
w1
w2
Remove all background
Remove foreground

Fig. 3. Feature change in different regions of background. We blacken different regions of the image (represents that the features in this region have changed). In these cases, we can study the dependence of mainstream detectors on each part of the feature.

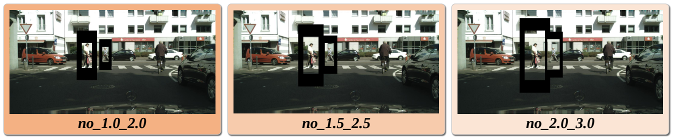


Fig. 4. Feature change in different ranges of background. *no_x_y* represents the images whose background features (from x times to y times the size of the bounding boxes) have changed.

$\Delta w = \frac{\partial w}{\partial x}\Delta x + \frac{\partial w}{\partial y}\Delta y + \frac{\partial w}{\partial z}\Delta z$):

$$\Delta K \propto C_1 \cdot \frac{\partial \boldsymbol{F}_S^e}{\partial(P(\boldsymbol{w}))} \cdot \Delta P(\boldsymbol{w}) + C_2 \cdot \frac{\partial \boldsymbol{F}_S^e}{\partial(P(\boldsymbol{x}))} \cdot \Delta P(\boldsymbol{x}) + C_3 \cdot \frac{\partial \boldsymbol{F}_S^e}{\partial(P(\boldsymbol{A}))} \cdot \Delta P(\boldsymbol{A}), \tag{10}$$

where C_1, C_2, C_3 are coefficients, which represent the coefficients due to the partial derivative of the cumulative product. $\boldsymbol{F}_S^e$ denotes $\boldsymbol{F}_S\left\{P\left(\boldsymbol{w}\right), P\left(\boldsymbol{x}\right), P\left(\boldsymbol{A}\right)\right\}$. Δ refers to domain gap, *e.g.*, $\Delta K = K_S - K_T$, $\Delta P(\boldsymbol{A})$ is the gap between $P\left(\boldsymbol{A}_S\right)$ and $P\left(\boldsymbol{A}_T\right)$.

The experimental results in Table 1 in the original text demonstrate that (Although the influence of the feature changes of each region on the detection accuracy is different, it is of the same order of magnitude.):

$$C_1 \cdot \frac{\partial \boldsymbol{F}_S^e}{\partial(P(\boldsymbol{w}))} \approx C_2 \cdot \frac{\partial \boldsymbol{F}_S^e}{\partial(P(\boldsymbol{x}))} \approx C_3 \cdot \frac{\partial \boldsymbol{F}_S^e}{\partial(P(\boldsymbol{A}))}. \tag{11}$$

At the same time, for the cross-domain pedestrian detection task, the inter-domain difference of the background is much larger than the inter-domain difference of the foreground. In fact, the reason why we feel that the two images have domain differences is mainly because of the domain differences in the background (such as fog or no fog background). The domain difference for pedestrian foreground is not large:

$$\Delta P(\boldsymbol{w}) < \Delta P(\boldsymbol{x}) \ll \Delta P(\boldsymbol{A}). \tag{12}$$

In this case, by Equation 10, Equation 11 and Equation 12, we can get:

$$\Delta K \propto O\left(\frac{\partial \boldsymbol{F}_S^e}{\partial(P(\boldsymbol{A}))} \cdot \Delta P(\boldsymbol{A})\right). \tag{13}$$

Therefore, we can make $\frac{\partial \boldsymbol{F}_S^e}{\partial(P(\boldsymbol{A}))} \rightarrow 0$ through the background-focused distribution alignment, greatly reducing $\Delta K = K_S - K_T(\Delta K \rightarrow 0)$.

Equation 13 means that the inconsistency of background features in cross-domain detection causes the greatest accuracy impact, so we can perform image-level cross-domain feature alignment by focusing only on the background. The background-focused feature alignment can not only play the same cross-domain adaptation role as the original image-level cross-domain feature alignment but also effectively avoid the foreground-background feature misalignment issue.

IV. METHODOLOGY

To solve the above issue, we propose three main modules: a Background Decoupling Module (*BDM*) with the aid of a Feature Generation Module (*FGM*), and a Long-short-range Domain Discriminator (*LSD*), as shown in Fig. 5. This section introduces these modules from two levels: background features decoupling and long-short-range attention discriminator.

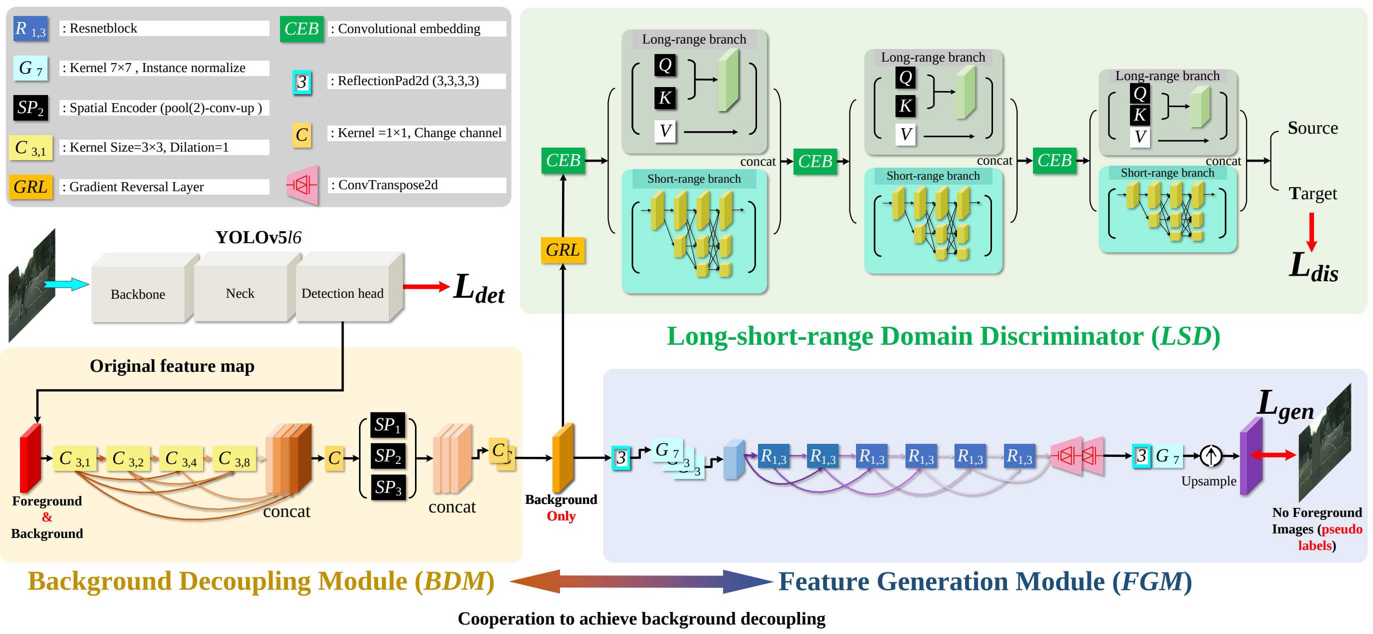

R 1,3 : Resnetblock
G 7 : Kernel 7×7 , Instance normalize
SP2 : Spatial Encoder (pool(2)-conv-up)
C 3,1 : Kernel Size=3×3, Dilation=1
GRL : Gradient Reversal Layer
CEB : Convolutional embedding
3 : ReflectionPad2d (3,3,3,3)
C : Kernel =1×1, Change channel
: ConvTranspose2d
YOLOv5l6
Backbone
Neck
Detection head
Ldet
Original feature map
C 3,1
C 3,2
C 3,4
C 3,8
concat
C
SP1
SP2
SP3
concat
C
Foreground & Background
Background Only
Background Decoupling Module (BDM)
Long-range branch
Q
K
V
Short-range branch
concat
CEB
GRL
Source
Target
Ldis
Long-short-range Domain Discriminator (LSD)
3
G 7
R1,3
Upsample
Lgen
No Foreground Images (pseudo labels)
Feature Generation Module (FGM)
Cooperation to achieve background decoupling

Fig. 5. The overall pipeline of the proposed framework consists of four main parts: a **YOLOv5***l6* Detector, a Background Decoupling Module extracting background information from the original feature map, a Feature Generation Module generating images containing only the background regions (using **pseudo labels** to remove the foreground regions of the original images) from the feature map generated by the *BDM*, and a Long-short-range Domain Discriminator which uses the feature map generated by the *BDM* for background-focused distribution alignment. *BDM*, *LSD*, *FGM* are only used during training and will not exist during testing.The function of *BDM* needs the help of *FGM* to complete.

A. Background Features Decoupling

Section III-B raises the foreground-background feature misalignment issue and Section III-C reveals the background importance in PD. Therefore, our proposed framework decouples the background features from the original feature maps (The original feature map we studied is extracted from the first layer of convolution in the detection head of YOLOv5) and only aligns the background features between domains. It can perfectly solve the foreground-background feature misalignment issue because no foreground features interfere with the alignment. Our framework mainly decouples background features by the Background Decoupling Module and the Feature Generation Module.

Visual analysis of the original feature maps (Sec. V-E and Fig. 6) shows that background and foreground features are fully mixed in the original feature maps. **We cannot completely decouple the features corresponding to the background part of the original image based only on the spatial position information.** Therefore, the background feature decoupling algorithm we designed is necessary.

First, the Background Decoupling Module only decouples the background features of the sub-feature map with the largest resolution in the original feature map (the original feature map of **YOLOv5***l6* (an excellent model in the YOLOv5 series) has four sub-feature maps with different resolutions). This sub-feature map has rich spatial information, which can well characterize background features, and meanwhile, it does not contain much pedestrian semantic information, which is convenient for decoupling background features. Although the semantic information of the feature map is insufficient, the multi-level spatial encoder of *BDM* can analyze the semantic information of the background nicely.

Second, the Feature Generation Module helps the Background Decoupling Module decoupling background features (Both are optimized by the same loss function L_{gen}). This module consists of a resnet-based [43] feature encoder and a transposed convolution module. Its goal is to restore images that only contain background from the feature map generated by the Background Decoupling Module. We use pseudo labels to remove all possible foreground regions from the input image as **ground truth**. These pseudo labels are the prediction boxes with confident scores greater than 0.01 in the pedestrian detection result of the previous epoch, and these regions are filled with the **average pixel value** of the original image. We use Manhattan distance to measure the loss:

$$L_{gen} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \left| I_{ij} - I'_{ij} \right|, \tag{14}$$

where the length and width of the image are H and W, the pixel value at point (i, j) of the ground truth (using pseudo labels) and restored image (by the Feature Generation Module) is expressed as I_{ij} and I'_{ij}, respectively.

B. Long-short-range: Dual-branch Discriminator

Fig. 4 and Table II demonstrate that a detector is more sensitive to the short-range background feature change than the long-range one. We design a dual-branch structure: the Transformer-CNN-based long-short-range discriminator to better analyze the local and global background features.

First, because the feature map fed into the discriminator represents a wide range of backgrounds, global spatial and semantic information is essential, requiring a long-range-attention module to encode long-distance background features. The self-attention-based Transformer is a natural long-range information encoder, driving us to use a convolutional-embedding-based Transformer (inspired by CvT [35]) as a long-range branch in the discriminator.

Second, only paying attention to long-range background is insufficient for the discriminator. It needs to encode short-range background features and analyze spatial and semantic information, which is more critical in our task. Convolution pays attention to local information, so we design a short-range-attention branch based on the multi-level structure of HRNet [44].

D_i is the domain label of the i-th image, $D_i = 0(1)$ means the i-th image comes from the source (target) domain. p_i is the probability of determining the i-th image belonging to the target domain (1) by the discriminator. The discriminator cross-entropy loss can be expressed as:

$$L_{dis} = -\sum_{i} \left[D_i \log\left(p_i\right) + \left(1 - D_i\right) \log\left(1 - p_i\right)\right]. \quad (15)$$

The adversarial feature alignment requires training the discriminator network to minimize L_{dis} while training the base detector to maximize L_{dis}. A Gradient Reversal Layer (GRL) module can help implement this algorithm (Fig. 5).

The total loss can be expressed as:

$$L = \alpha \cdot L_{det} + \beta \cdot L_{gen} + \gamma \cdot L_{dis}, \quad (16)$$

where L_{det} is the loss when YOLOv5 uses source domain images and annotations for training, and α, β, γ are trade-off parameters to balance these losses.

V. Experiments

In this section, we first evaluate our proposed BFDA on the cross-domain PD task under two scenarios, i.e., Scene Adaptation and Weather Adaptation. Moreover, empirical analyses are then provided from two aspects, i.e., ablation studies, to shed light on BFDA's sub-modules role and the generalization performance of BFDA on the common object detection task.

A. Datasets

Caltech [45]: The Caltech dataset has about 42,782 images for training and 4024 images for test with the resolution of 640×480 pixels. We use the new annotations provided by [46] for experiments. This is one of the most commonly used datasets for pedestrian detection.

CityPersons [41] : The CityPersons dataset is built from Cityscapes [47], which has about 2975, 500, 1525 images for training, validation, test (researchers often use Cityscapes validation dataset for test, so do we), with the resolution of 2048×1024 pixels.

Foggy Cityscapes [48]: The Foggy Cityscapes dataset is also built from Cityscapes, and there are three levels of fog images. We use the image with the thickest fog.

BDD10K [42]: The BDD10K dataset is an auxiliary dataset of the BDD100K dataset, which contains 7000, 2000, 1000 images for training, validation, test, with the resolution of 1280×720 pixels.

The BDD10K dataset is very similar to the Caltech dataset, but Caltech and BDD100K does not provide semantic segmentation-level annotations, so we can only use the BDD10K dataset to generate the dataset we need in Sec. III.

B. Experiment Settings

Evaluation Settings. We use two types of cross-domain settings: (i) *CityPersons →Caltech*: Scene Adaptation, where the source domain is CityPersons (Cityscapes), and the target domain is Caltech. (ii) *CityPersons→Foggy Cityscapes*: Weather Adaptation, where the source domain is CityPersons (Cityscapes), and the target domain is Foggy Cityscapes.

Metrics. (i) We utilize the standard log average Miss Rate over False Positive Per Image (FPPI) in the range of $\left[10^{-2}, 10^{0}\right]$, dubbed by MR^{-2} . Lower MR^{-2} indicates better performance. To gain a deeper understanding of the model's performance under different occlusion conditions, we further divide the test set into four parts according to the degree of occlusion (reasonable, bare, partial, and heavy), and report the results separately. (ii) We also use the general object detection evaluation metric AP_{50} in the section III. PD task uses MR^{-2} (not AP_{50}) as the standard evaluation metric (Therefore we only report MR^{-2} in the section V).

Implementation Details. We follow the standard protocols of UDA, where all samples in the source are labeled while those in the target are unlabeled. BFDA employs the excellent one-stage detector **YOLOv5*l6*** as base detector. Since many SOTA UDA methods designed for two-stage detectors are inapplicable on one-stage detector (*e.g.*, YOLOv5), Faster RCNN backbone is used in Table VI for a fair comparison. The input images maintain their original resolution, but their feature maps are resized to 224×224 before being fed into the discriminator. When performing cross-domain adaptation, we first initialize the model with pretrained weights. The initial learning rate of **YOLOv5*l6*** is 10^{-3}, which is reduced to 2×10^{-4} by cosine annealing, and the learning rates of the other three modules are 10^{-4}. Also, we do not use the mosaic trick. We use NVIDIA Tesla V100 to test the FPS of the frameworks.

C. Comparison Results

Scene Adaptation. Different scenes are often captured via different devices or setups, resulting in domain shifts among scenes. To study the effectiveness of our proposed framework for Scene Adaptation, we use CityPersons as the source domain and Caltech as the target domain.

Table IV compares our BFDA with current SOTA cross-domain PD models. Although the SOTA cross-domain PD model SAN [11] achieves impressive performance, our overall framework (BFDA) exceeds it by a large margin, almost reaching the accuracy of **Oracle** (train on the labeled target domain dataset). It shows that our framework is very effective in the Scene Adaptation task. One of the main reasons is that the domain gap of foreground (pedestrian) features in the Scene Adaptation task is very small, so the domain

TABLE III
A DESCRIPTION OF THE COMPOSITION OF DIFFERENT FRAMEWORKS FOR ABLATION EXPERIMENTS. THE MODULES HAVE AN ORDER IN WHICH THEY ARE ADDED. THIS IS BECAUSE THE BACKGROUND DECOUPLING MODULE REQUIRES THE FEATURE GENERATION MODULE TO CONSTRUCT THE LOSS FUNCTION. AFTER THE BACKGROUND FEATURES ARE DECOUPLED, WE CAN USE THE LONG-SHORT-RANGE DOMAIN DISCRIMINATOR TO GIVE DIFFERENT ATTENTION TO BACKGROUNDS IN DIFFERENT RANGES.

	Background Decoupling Module	Feature Generation Module	Short-range Discriminator	Long-range Discriminator
BFDA$_L$				✓
BFDA$_{LF}$		✓		✓
BFDA$_{LBF}$	✓	✓		✓
BFDA	✓	✓	✓	✓

TABLE IV
SCENE ADAPTATION: *CityPersons→Caltech* (FPS ON **V100**)

Method	$MR^{-2}(\%) \downarrow$				FPS
	reasonable	bare	partial	heavy	
Source-only	15.91	15.67	18.51	31.41	217.4
SCDA [51]	28.93	28.93	-	-	16.7
DAFR [16]	18.42	18.42	-	-	12.0
SAN [11]	14.27	14.27	-	-	17.1
BFDA$_L$	9.40	9.30	**9.29**	**24.77**	**217.4**
BFDA$_{LF}$	8.83	8.57	13.40	25.38	
BFDA$_{LBF}$	8.40	8.11	13.00	25.38	
BFDA	**7.71**	**7.26**	15.51	27.41	
Oracle(Train-on-target)	5.38	5.05	0.00	38.37	217.4

TABLE V
WEATHER ADAPTATION: *CityPersons→Foggy Cityscapes* (FPS ON **V100**)

Method	$MR^{-2}(\%) \downarrow$				FPS
	reasonable	bare	partial	heavy	
Source-only	26.64	19.75	27.13	54.35	42.7
DAFR [16]	54.71	54.71	-	-	5.7
SW-ICR-CCR [50]	49.54	37.95	55.13	89.69	6.3
BFDA$_L$	23.92	16.73	24.85	52.53	**42.7**
BFDA$_{LF}$	24.58	18.00	26.37	57.28	
BFDA$_{LBF}$	20.62	14.95	20.99	**50.95**	
BFDA	**18.57**	**12.84**	**19.35**	52.21	
Oracle(Train-on-target)	14.33	9.17	14.32	44.22	42.7

difference of background features plays a major role in the performance degradation. Therefore, our background-focused distribution alignment method can mitigate the misalignment issue. Moreover, by employing an efficient one-stage base detector, the FPS is more than ten times those based on Faster RCNN, which could appeal to a wide range of real-time applications, like automatic drive.

Weather Adaptation. Weather changing is another inevitable factor inducing domain gap in real-world applications and could degrade the model performance greatly, as suggested in [16], [49]. To study the effectiveness of BFDA, we use CityPersons as the source domain and Foggy Cityscapes as the target domain.

Results on Weather Adaptation are shown in Table V. The existing advanced framework SW-ICR-CCR [50] is not satisfactory. The experimental results of our complete framework (BFDA) drastically outperform the current advanced results and almost reach the accuracy of Oracle, which demonstrates the effectiveness of BFDA under Weather Adaptation scenarios. Further, our one-stage-detector-based BFDA is more than six times faster than those based on two-stage detectors, demonstrating our framework's efficiency once again.

D. Empirical Analysis

$MR^{-2}(reasonable)$ (represented as $MR^{-2}(r)$ for brevity) is the **most crucial metric** because there are far **more reasonable pedestrians** than other kinds (bare, partial, heavy) in Caltech and Foggy Cityscapes.

Ablation Study: To take a closer look at the proposed BFDA, ablation studies are conducted by additionally developing three variants of BFDA. The modules have an order in which they are added. This is because the Background Decoupling Module requires the Feature Generation Module to construct the loss function. After the background features are decoupled, we can use the Long-short-range Domain Discriminator to give different attention to backgrounds in different ranges .

Specifically, based on our base detector YOLOv5, **BFDA**$_L$ only takes the long-range discriminator (Transformer) and performs typical image-level domain adaptation. **BFDA**$_{LF}$ then integrates the Feature Generation Module (+*FGM*) based on BFDA$_L$. **BFDA**$_{LBF}$ further introduces a Background Decoupling Module (+*BDM*) based on BFDA$_{LF}$. Finally, **BFDA** (our full framework) introduces a long-short-range discriminator (+*LSD*) on top of **BFDA**$_{LBF}$ (as shown in Table III). Results are presented in both Table IV and Table V, and the following conclusions could be drawn:

(1) *Effects of the Long-range Domain Discriminator*: To demonstrate that we are not simply borrowing Transformer to improve accuracy, we take the long-range discriminator (Transformer only) as our **baseline** and use the typical image-level adaptation method to achieve domain adaptation. **BFDA**$_L$ generally outperforms source-only results. However, the improvement is limited, suggesting the commonly used image-level domain adaptation is far from sufficient. There are only L_{det} and L_{dis} here (no L_{gen}).

(2) *Effects of the Feature Generation Module*: **BFDA**$_{LF}$ employ the Feature Generation Module to force the feature map to contain only background features. However, the experimental results do not seem to be very satisfactory: in the Scene Adaptation task, the module successfully reduces $MR^{-2}(r)$

TABLE VI

Generalization evaluation: *Cityscapes→Foggy Cityscapes*. **We use framework-level comparison because of the non-portability of the various methods.** (†: YOLOv5-based; ‡: Faster RCNN (vgg16)-based). [**Note** that our BFDA is designed for one-stage detectors (like YOLOv5). We include experiments with Faster RCNN here just for a fair comparison, but most of the existing Faster RCNN-based SOTA methods cannot be used on one-stage detectors.]

Method	mAP(%)↑	person	rider	car	truck	bus	train	mcycle	bicycle
Source-only‡	25.8	33.7	35.2	13.0	28.2	9.1	18.7	31.4	24.4
Source-only†	46.0	55.0	58.3	63.9	30.1	37.9	28.1	44.8	49.8
MeGA-CDA [32]‡	41.8	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0
UMT [52]‡	41.7	33.0	46.7	48.6	34.1	56.5	46.8	30.4	37.3
HTCN [53]‡	39.8	33.2	47.5	47.9	31.6	47.4	40.9	32.3	37.1
CRDA [54]‡	37.4	32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6
SWDA [25]‡	34.3	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3
CADA [55]†	40.2	41.5	43.6	57.1	29.4	44.9	39.7	29.0	36.1
SSAL [56]†	39.6	45.1	47.4	59.4	24.5	50.5	25.7	26.0	38.7
S-DAYOLO [57]†	39.0	42.6	42.1	61.9	23.5	40.5	39.5	24.4	37.3
DA-YOLO [58]†	36.1	29.5	27.7	46.1	9.1	28.2	4.5	12.7	24.8
BFDA(Ours)‡	41.4	41.4	48.1	60.5	27.2	47.9	32.6	31.8	41.9
BFDA(Ours)†	**58.1**	**64.2**	**65.3**	**74.2**	**38.8**	**62.2**	**51.8**	**50.6**	**58.1**
Oracle‡	43.5	37.2	48.3	52.7	35.2	52.2	48.5	35.3	38.8
Oracle†	66.4	71.4	73.6	83.3	51.6	72.8	61.4	56.9	60.2

from 9.40% to 8.83%, while in the Weather Adaptation task, $MR^{-2}(r)$ increases from 23.92% to 24.58%. The main reason is that if the foreground is suppressed (by L_{gen}) directly on the original feature map (*i.e.*, the feature map used by YOLOv5 for detection), it will inevitably lead to the decline of detection accuracy since the foreground features are destroyed. In other words, the L_{det} promotes the feature map to contain foreground features, while the L_{gen} promotes the feature map to contain no foreground features, and the two losses cannot be optimized simultaneously.

(3) *Effects of the Background Decoupling Module*: **BFDA**$_{LBF}$ adopts a Background Decoupling Module, which takes the original feature map as input and extracts background features with L_{gen} 's help. As is observed in both Table IV and Table V, **BFDA**$_{LBF}$ further improves the cross-domain performance, especially on the Weather Adaptation scenario, obtaining a performance gain over 3%. The underlying reason could be that by adding the Background Decoupling Module, L_{gen} mainly trains the newly introduced module instead of YOLOv5, which could alleviate the contradiction between L_{det} and L_{gen}.

(4) *Effects of the Long-short-range Domain Discriminator*: The above modules have decoupled the background feature from the original feature map, and the next step is to use the domain discriminator to analyze the background features. Table II demonstrates that different ranges of backgrounds have different levels of importance, and short-range backgrounds are more important and should be focused on. Our complete framework **(BFDA)** has greatly improved over BFDA$_{L}$: in the Scene Adaptation task, $MR^{-2}(r)$ reduces from 9.40% to 7.71%, and in the Weather Adaptation task, it reduces from 23.92% to 18.57%. We contribute that the Long-short-range Domain Discriminator combines global and local attention capabilities, successfully analyzing complex backgrounds.

E. Visual Analysis

Background decoupling visualization: To demonstrate that our framework does decouple background, we visualize the *FGM* output images in Fig. 7. Hence, if the generated map only contains background, this input feature map should only contain background features. It can be clearly seen that the output of *FGM* contains almost all the background features. We can't clearly illustrate the effect of background decoupling by visualizing *BDM* generated feature maps (Fig. 6).

The coupling of foreground and background features in the feature map: The Background Decoupling Module (*BDM*) and the Feature Generation Module (*FGM*) are proposed to decouple background features from the original feature maps. This section discusses our third finding: background features and foreground features are completely coupled in the feature maps, implying the necessity of our two modules (*BDM* and *FGM*).

We conduct ablation experiments (YOLOv5 on Citypersons (Cityscapes [47])) to study the influence of three different regions (outer-bounding-box background, inner-bounding-box background, and foreground) on the YOLOv5 original feature maps. As illustrated in Fig. 6, the visualization results reveal an interesting phenomenon: **the peaks** on each feature map corresponding to the foreground region of the original image turned out to **be mainly generated by background information**. One reason is that the detection network often models the spatial contextual relations between the foreground and background regions. After removing the background information, these ”foreground” peaks have been weakened a lot; but after removing the foreground information, these ”foreground” peaks have not changed significantly.

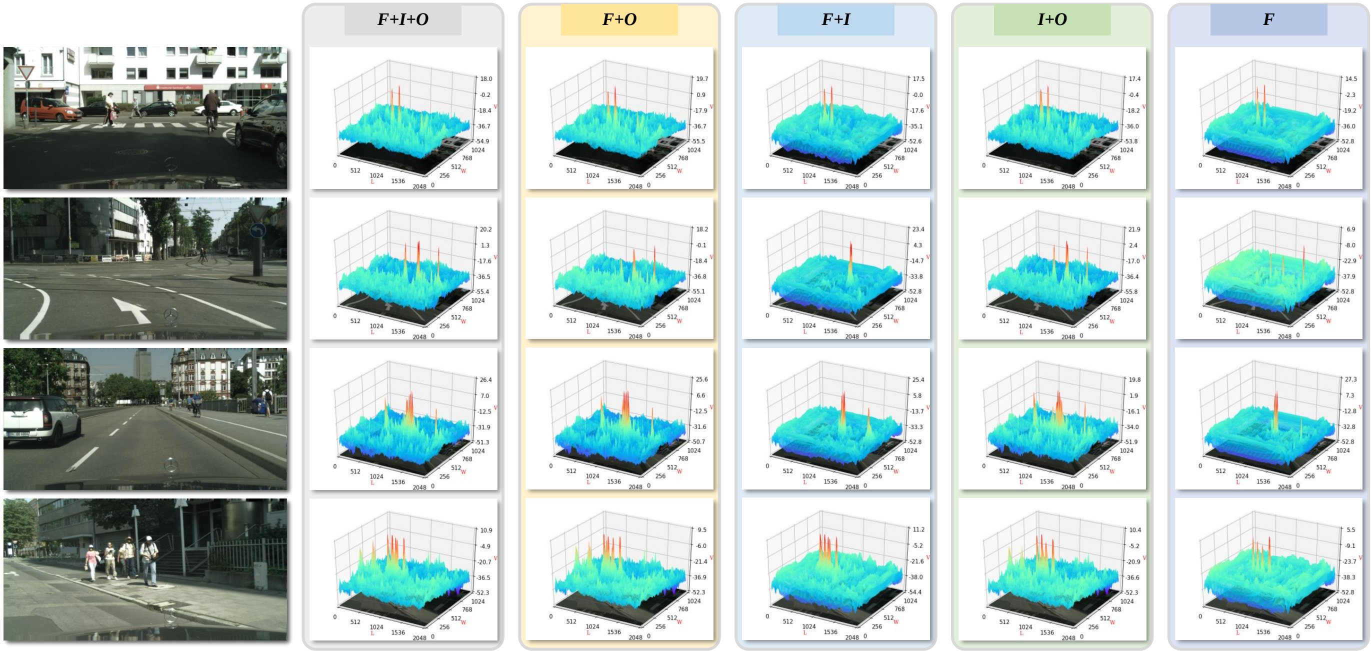
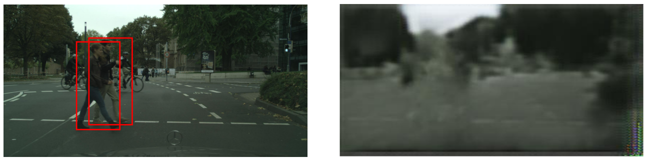


Fig. 6. Ablation study of the foreground-background feature coupling (YOLOv5-based). F, I, and O refer to Foreground, Inner-bounding-box background, and Outer-bounding-box background, respectively. Each column in the figure represents the feature maps generated by retaining only a specific part of the original image. For example, the second column ($F+O$) means that this column is the corresponding feature map after removing the Inner-bounding-box background (I) from the original images. At the same time, we **cannot** estimate the quality of our background decoupling by visualizing the feature maps generated by BDM (even if all are backgrounds, the feature maps still have some peaks).



(a) Original image (Cityscapes) (b) Output of FGM (Cityscapes)

Fig. 7. Background decoupling visualization.*CityPersons→Caltech*

Visualization of the detection results: We visualize the detection results (Fig. 8) on the scene adaptation task (*CityPersons→Caltech*) and the weather adaptation task (*CityPersons→Foggy Cityscapes*). The visualization results clearly demonstrate that the Background-focused distribution alignment has much fewer wrong predictions than the Typical image-level adaptation, and further verifies the foreground-background feature mismatch issue (the main reason for the wrong prediction boxes) proposed in the main text.

F. Generalization Evaluation.

So far, we have performed comprehensive analyses of BFDA on the cross-domain pedestrian detection task and achieved promising results. Our key finding is that the background inconsistency dominates the domain gap, which may happen in other detection tasks and could be general. Motivated by this, we extend BFDA to conduct experiments on the general object detection task by treating all classes as foreground at once. The results on BDD10k are presented in Table VI, where we compare BFDA with a series of SOTA domain adaptation methods based on Faster RCNN. As is shown, BFDA greatly outperforms almost all other SOTA frameworks [25], [32], [49], [52]–[54], [60], validating the proposed BFDA benefits adaptation of general object detector.

Note that our BFDA is designed for one-stage detectors (such as YOLOv5) because one-stage detectors can only use image-level cross-domain and must face this issue. (instance-level cross-domain adaptation can partially solve this problem, as the RPN can propose foregrounds individually for alignment). We include experiments with Faster RCNN here for a fair comparison since the existing SOTA records are mainly from Faster RCNN-based methods. Also, we cannot move existing two-stage detector-based SOTA methods to one-stage detectors because one-stage detectors lack all the conditions required for instance-level cross-domain adaptation.

VI. Conclusions

We uncover a problem with direct application of image-level domain adaptation on one-stage detectors and investigate cross-domain PD tasks from a new perspective. We also find that mainstream detectors are generally sensitive to the background variations, further inspiring us to develop a new background-focused distribution alignment framework BFDA. The BFDA comprises three essential parts: the Background Decoupling Module, the Feature Generation Module, and the Long-short-range Domain Discriminator. We conduct extensive experiments on multiple benchmark datasets, and their results clearly show that our BFDA surpasses the existing SOTA frameworks with great advantages in detection accuracy. Meanwhile, as our framework is based on advanced YOLOv5, the inference speed can reach 7∼12 times FPS of the existing SOTA frameworks.

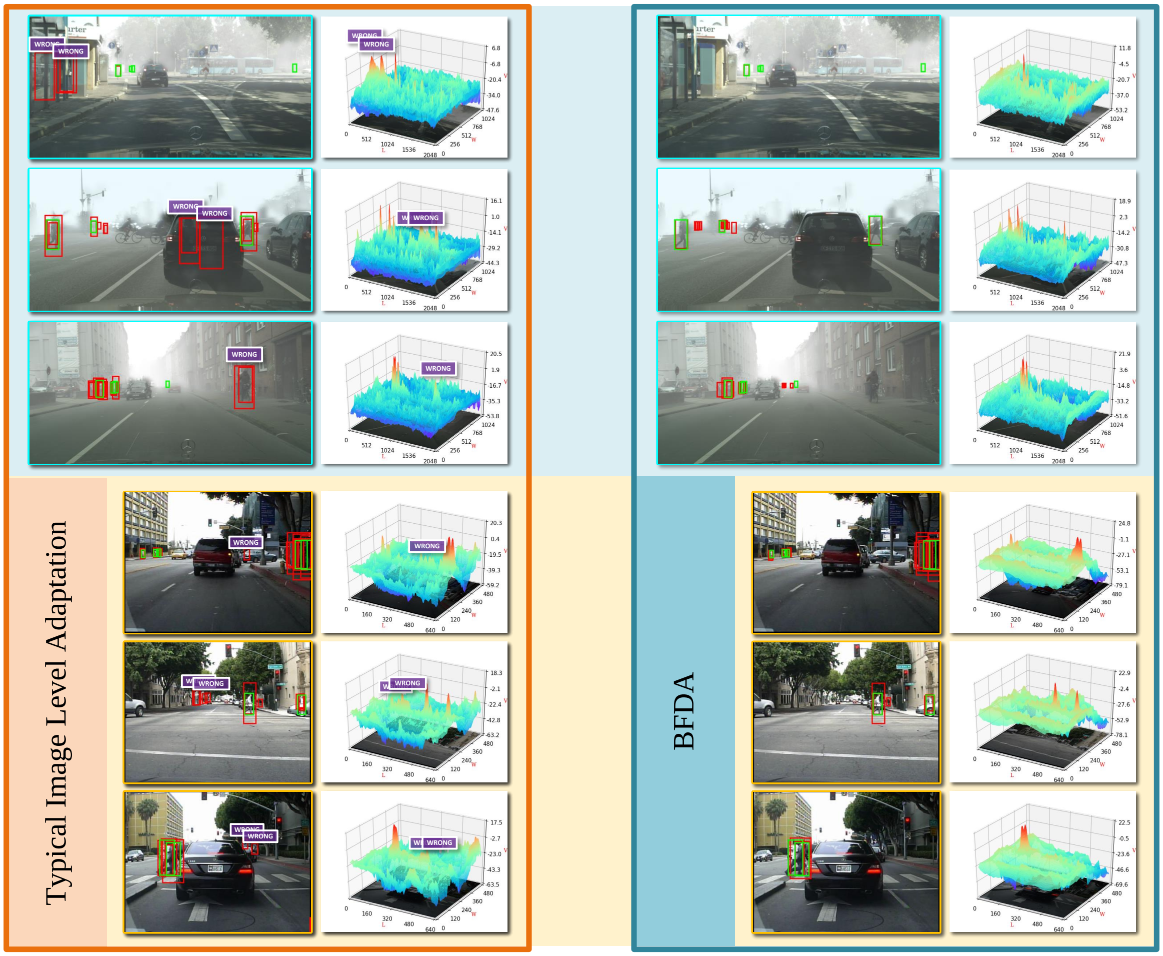


Fig. 8. Cross-domain pedestrian detection results comparison on *CityPersons→Foggy Cityscapes* (first three lines) and *CityPersons→Caltech*(last three lines). The green boxes are ground truths, and the red boxes are prediction boxes with confidence scores greater than 0.01 before Non-Maximum Suppression [59]. Purple boxes (marked with "WRONG") indicate wrong prediction bounding boxes.

References

[1] V. Campmany, S. Silva, A. Espinosa, J. C. Moure, D. Vázquez, and A. M. López, "Gpu-based pedestrian detection for autonomous driving," *Procedia Computer Science*, vol. 80, pp. 2377–2381, 2016.

[2] Y. Yan, J. Li, J. Qin, S. Bai, S. Liao, L. Liu, F. Zhu, and L. Shao, "Anchor-free person search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7690–7699.

[3] H. Hattori, V. Naresh Boddeti, K. M. Kitani, and T. Kanade, "Learning scene-specific pedestrian detectors without real data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3819–3827.

[4] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 618–634.

[5] W. Lan, J. Dang, Y. Wang, and S. Wang, "Pedestrian detection based on yolo network model," in *2018 IEEE international conference on mechatronics and automation (ICMA)*. IEEE, 2018, pp. 1547–1551.

[6] J. Xie, Y. Pang, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, "Mask-guided attention network and occlusion-sensitive hard example mining for occluded pedestrian detection," *IEEE transactions on image processing*, vol. 30, pp. 3872–3884, 2020.

[7] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5187–5196.

[8] W. Liu, S. Liao, and W. Hu, "Efficient single-stage pedestrian detector by asymptotic localization fitting and multi-scale context encoding," *IEEE transactions on image processing*, vol. 29, pp. 1413–1425, 2019.

[9] T. Guo, C. P. Huynh, and M. Solh, "Domain-adaptive pedestrian detection in thermal images," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1660–1664.

[10] L. Liu, W. Lin, L. Wu, Y. Yu, and M. Y. Yang, "Unsupervised deep domain adaptation for pedestrian detection," in *European Conference on Computer Vision*. Springer, 2016, pp. 676–691.

[11] Y. Jiao, H. Yao, and C. Xu, "San: selective alignment network for cross-domain pedestrian detection," *IEEE transactions on image processing*, vol. 30, pp. 2155–2167, 2021.

[12] M. Kieu, A. D. Bagdanov, M. Bertini, and A. Del Bimbo, "Task-conditioned domain adaptation for pedestrian detection in thermal imagery," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*. Springer, 2020, pp. 546–562.

[13] D. Guan, X. Luo, Y. Cao, J. Yang, Y. Cao, G. Vosselman, and M. Ying Yang, "Unsupervised domain adaptation for multispectral pedestrian detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[14] D. Vázquez, A. M. López, and D. Ponsa, "Unsupervised domain adaptation of virtual and real worlds for pedestrian detection," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 3492–3495.

[15] W. Chen, Y. Guo, S. Yang, Z. Li, Z. Ma, B. Chen, L. Zhao, D. Xie, S. Pu, and Y. Zhuang, "Box re-ranking: Unsupervised false positive suppression for domain adaptive pedestrian detection," *arXiv preprint arXiv:2102.00595*, 2021.

[16] 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.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time

object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[18] 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.

[19] J. Li, S. Liao, H. Jiang, and L. Shao, "Box guided convolution for pedestrian detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1615–1624.

[20] J. Zhang, L. Lin, J. Zhu, Y. Li, Y.-c. Chen, Y. Hu, and C. S. Hoi, "Attribute-aware pedestrian detection in a crowd," *IEEE Transactions on Multimedia*, 2020.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[22] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2411–2418.

[23] T. Zhang and D. Freedman, "Improving performance of distribution tracking through background mismatch," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 2, pp. 282–287, 2005.

[24] H. T. Nguyen and A. W. Smeulders, "Robust tracking using foreground-background texture discrimination," *International Journal of Computer Vision*, vol. 69, no. 3, pp. 277–293, 2006.

[25] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6956–6965.

[26] Z. He, L. Zhang, Y. Yang, and X. Gao, "Partial alignment for object detection in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[27] H. Wang, S. Liao, and L. Shao, "Afan: Augmented feature alignment network for cross-domain object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 4046–4056, 2021.

[28] J. Chen, X. Wu, L. Duan, and L. Chen, "Sequential instance refinement for cross-domain object detection in images," *IEEE Transactions on Image Processing*, vol. 30, pp. 3970–3984, 2021.

[29] S. Kim, J. Choi, T. Kim, and C. Kim, "Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6092–6101.

[30] B. Zhang, T. Chen, B. Wang, X. Wu, L. Zhang, and J. Fan, "Densely semantic enhancement for domain adaptive region-free detectors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1339–1352, 2021.

[31] C. Chen, Z. Zheng, Y. Huang, X. Ding, and Y. Yu, "I3net: Implicit instance-invariant network for adapting one-stage object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 576–12 585.

[32] V. VS, V. Gupta, P. Oza, V. A. Sindagi, and V. M. Patel, "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*, 2021, pp. 4516–4526.

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[34] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[35] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," *arXiv preprint arXiv:2103.15808*, 2021.

[36] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *arXiv preprint arXiv:2102.12122*, 2021.

[37] ——, "Pvtv2: Improved baselines with pyramid vision transformer," *arXiv preprint arXiv:2106.13797*, 2021.

[38] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 603–612.

[39] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1, pp. 151–175, 2010.

[40] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation for covariate shift adaptation," *Annals of the Institute of Statistical Mathematics*, vol. 60, no. 4, pp. 699–746, 2008.

[41] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *CVPR*, 2017.

[42] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[44] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[45] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2011.

[46] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Proceedings of the iEEE conference on computer vision and pattern recognition*, 2016, pp. 1259–1267.

[47] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[48] 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.

[49] 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.

[50] 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.

[51] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 687–696.

[52] 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.

[53] 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.

[54] 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. 11 724–11 733.

[55] 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 *European Conference on Computer Vision*. Springer, 2020, pp. 733–748.

[56] M. A. Munir, M. H. Khan, M. Sarfraz, and M. Ali, "Ssal: Synergizing between self-training and adversarial learning for domain adaptive object detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 770–22 782, 2021.

[57] G. Li, Z. Ji, X. Qu, R. Zhou, and D. Cao, "Cross-domain object detection for autonomous driving: A stepwise domain adaptative yolo approach," *IEEE Transactions on Intelligent Vehicles*, 2022.

[58] S. Zhang, H. Tuo, J. Hu, and Z. Jing, "Domain adaptive yolo for one-stage cross-domain detection," in *Asian Conference on Machine Learning*. PMLR, 2021, pp. 785–797.

[59] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[60] F. Rezaeianaran, R. Shetty, R. Aljundi, D. O. Reino, S. Zhang, and B. Schiele, "Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9204–9213.