BLENDA: DOMAIN ADAPTIVE OBJECT DETECTION THROUGH DIFFUSION-BASED BLENDING

Tzuhsuan Huang[1⋆]   *Chen-Che Huang*[2⋆]   *Chung-Hao Ku*[1]   *Jun-Cheng Chen*[1]

[1] Academia Sinica   [2] National Taiwan University
jason890425@citi.sinica.edu.tw, {hang6318179, mku666}@gmail.com, pullpull@citi.sinica.edu.tw

ABSTRACT

Unsupervised domain adaptation (UDA) aims to transfer a model learned using labeled data from the source domain to unlabeled data in the target domain. To address the large domain gap issue between the source and target domains, we propose a novel regularization method for domain adaptive object detection, BlenDA, by generating the pseudo samples of the intermediate domains and their corresponding soft domain labels for adaptation training. The intermediate samples are generated by dynamically blending the source images with their corresponding translated images using an off-the-shelf pre-trained text-to-image diffusion model which takes the text label of the target domain as input and has demonstrated superior image-to-image translation quality. Based on experimental results from two adaptation benchmarks, our proposed approach can significantly enhance the performance of the state-of-the-art domain adaptive object detector, Adversarial Query Transformer (AQT). Particularly, in the Cityscapes to Foggy Cityscapes adaptation, we achieve an impressive 53.4% mAP on the Foggy Cityscapes dataset, surpassing the previous state-of-the-art by 1.5%. It is worth noting that our proposed method is also applicable to various paradigms of domain adaptive object detection. The code is available at https://github.com/aiiu-lab/BlenDA

Index Terms— object detection, domain adaptation, unsupervised domain adaptation, blended images, target domain

1. INTRODUCTION

Unsupervised domain adaptation (UDA) has recently gained significant attention in the field of object detection, with its primary goal being the reduction of cross-domain discrepancies. It enables detectors trained on sufficient labeled source data to effectively generalize to unlabeled data in the target domain.

To bridge the domain gap, recent advancements in unsupervised domain adaptive object detection have prominently featured mean teacher (MT)-based approaches [1]. Deng et al. [2] introduce unbiased mean teacher (UMT), incorporating CycleGAN [3] to synthesize images and mitigate domain discrepancies. Another innovative approach is presented by Li et al. [4], who propose an adaptive teacher (AT) utilizing feature-level adversarial learning. This method ensures that features extracted from both source and target domains display similar distributions. Instead of applying adversarial learning, He et al. [5] utilize target-like images along with the proposed target-perceived dual-branch distillation (TDD) framework to enhance the student model by recognizing objects in the
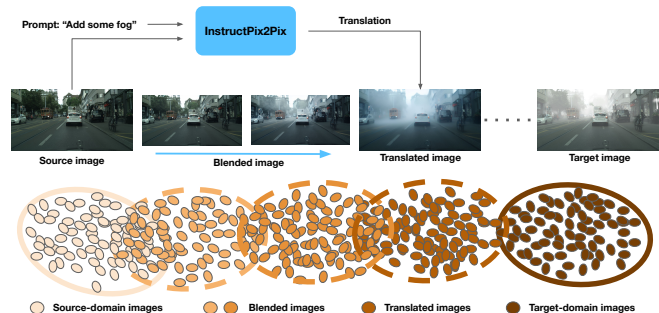
Fig. 1. Overview of BlenDA. We utilize a diffusion-based generative model [8] to generate a target-like translated image with an appropriate prompt and mix it with the source image into a series of blended images for the intermediate domains, which play the role of a bridge between two domains and can be used to train a detector progressively. The direction of the arrow (Blue one) represents an increasing proportion of the translated image used to blend with the source image.

target domain through iterative cross-attention. While these methods successfully narrow the domain discrepancy, the quality of the pseudo-labels predicted by the teacher model remains insufficient for domain adaptation.

To address this issue, Cao et al. [6] introduce Contrastive Mean Teacher (CMT), utilizing pseudo-labels to extract object-level features and refine them through contrastive learning. Deng et al. [7] propose a harmonious teacher (HT) that assesses the quality of pseudo-labels by computing the correlation between object confidence and bounding box location.

Unlike traditional approaches, Mattolin et al. [9] propose a method that incorporates a portion of the target domain image with the highest region-level detection confidence into the source image to facilitate domain adaptation. Inspired by data augmentation techniques in image classification tasks [10], Vu et al. [11] train the model by integrating the interpolated image and loss, computed using two distinct labels corresponding to the images used for interpolation.

The methods mentioned above are primarily based on convolution neural network (CNN). Different from CNN-based detection architectures, transformer-based detectors greatly simplifies the two-stage detection pipeline while achieving superior detection performance. Carion et al. [16] propose DETR, which concentrates on token-wise dependencies and eliminates hand-crafted components such as anchor generation and non-maximum suppression, replaced by the Hungarian algorithm. To address slow convergence in DETR, Zhu et al. [17] introduce Deformable DETR, which re-
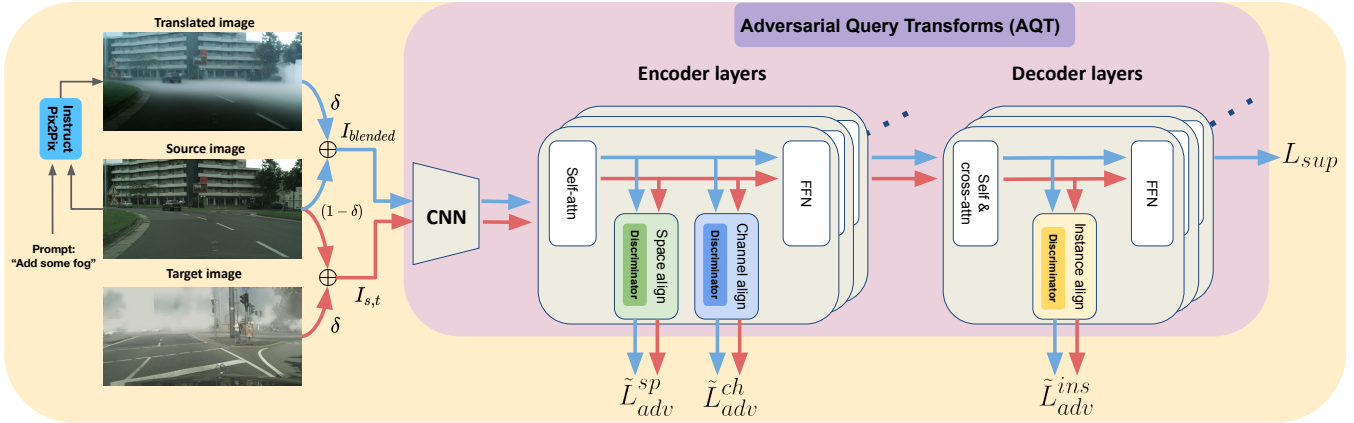
Fig. 2. Comprehensive overview of **BlenDA** on AQT [12]. We first generate translated images by InstructPix2Pix [8] and compute the δ using Eq (3). Subsequently, we feed the images $I_{blended}$ and $I_{s,t}$, which are mixed according to Eq (1), into the model. $I_{blended}$ replaces the original source image during the fine-tuning process, using the supervised loss L_{sup} to reduce the domain gap. Moreover, we adjust the original adversarial loss specified in Eq (5) to Eq (6). This new adversarial loss enables the discriminators to distinguish different domains in greater detail. Note that the model weights are initialized using the pre-trained weights released by Huang et al. [12] before fine-tuning.
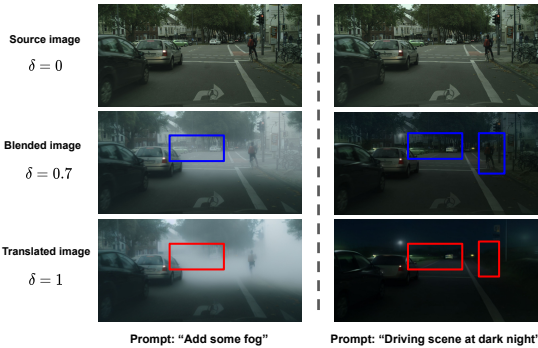

Fig. 3. Cityscapes [13] images are fed into InstructPix2Pix [8] with the prompts to create translated images for both Foggy Cityscapes [14] and BDD100K daytime [15]. Foreground objects are notably absent, as indicated by the red boxes in the translated images. After mixing the source and translated images, foreground objects become visible as outlined by blue boxes in the blended images. δ represents the proportion used to blend the translated image with the source one.

places the vanilla attention mechanism in DETR with deformable attention. Recent studies have started to investigate how to leverage transformer-based detectors in domain adaptive object detection, such as Sequence Feature Alignment (SFA) [18]. Based on Deformable DETR, Huang et al. [12] propose an Adversarial Query Transformer (AQT) for domain adaptive object detection, which performs adversarial feature alignment on transformers, with adversarial tokens attending to the feature tokens.

To improve the domain alignment between the source and target domains, we introduce a novel regularization strategy called BlenDA. The overview is shown in Fig. 1. During adaptation training, we synthesize high-quality pseudo samples of the intermediate domains by blending source images with their corresponding target-like translated images. These translated images are generated by the off-the-shelf diffusion-based generative model InstructPix2Pix [8], which utilizes the text label of the target domain as an instruction for image-to-image translation. Subsequently, the blended images act as a bridge between the source and target domains, serving as labeled data with the soft domain labels and ground-truth annotations

for fine-tuning a pre-trained model, which has been trained on data from both the original source and target domains. To further improve the adaptation performance, instead of a fixed mixing weight, we propose dynamically adjusting the mixing weight as the training iteration progresses. Meanwhile, we employ the corresponding soft domain labels to compute adversarial losses, which enhance the capability of discriminators to distinguish between different domains in greater detail. With extensive experiments, the results demonstrate that our approach significantly enhances the model's performance over the previous state-of-the-art method.

2. THE PROPOSED METHOD

To leverage the superior strength of a pre-trained foundation model for zero-shot text-to-image manipulation, we employ Instruct-Pix2Pix [8] to produce translated data, establishing a bridge between the source and target domains. In Section 2.1, we commence by pointing out the benefit of blending translated and source images and identifying the problem that may arise when directly using translated images as source data. In Section 2.2, we delve into the details of the mixed-domain adversarial loss, enhancing the discriminator's capacity to discern different domains with greater precision. In Section 2.3, we show the overall training objective.

2.1. Sample Generation for Intermediate Domains

Directly using the translated images as source data for training can lead to transferring inaccurate domain information due to the large domain gap. As indicated by the red boxes in the translated images in Fig. 3, it is evident that numerous foreground objects are either obscured by fog or engulfed in darkness after translation. Therefore, we first blend the paired source and translated data using a fixed mixing weight δ, as demonstrated in the following equation:

$$I_{blended} = \delta \cdot I_{translated} + (1 - \delta) \cdot I_{source}, \quad (1)$$

where δ is a scalar value, while $I_{blended}$, $I_{translated}$, and I_{source} correspond to the blended image, translated image, and source image, respectively. We use $\delta = 0.7$ as an example to demonstrate that combining source and translated images can make foreground objects visible, as indicated by the blue boxes in the blended images shown in Fig. 3. Subsequently, we utilize the blended image

$I_{blended}$ as source data to fine-tune the model, which has been pre-trained on the original source and target data. Intuitively, the presence of the blended image is expected to guide the model towards better alignment with the target domain over time. Nevertheless, as the supervised loss is computed from the blended images and δ stays constant in the whole training stage, the model may overfit the blended images and cannot generalize well to the target domain. To address this limitation, we propose a method that enables the adaptive adjustment of δ, allowing the model to gradually transfer the acquired domain knowledge to the target domain. Inspired by progressive pseudo-labeling [9], which dynamically adjusts the confidence threshold during non-maximum suppression, we instead use it as a method to blend the source and translated images, in which δ increases based on the current number of the training iterations:

$$\gamma = \frac{\text{current iteration}}{\text{total iterations}}, \qquad (2)$$

$$\delta = \left(\frac{2}{1 + exp(-\alpha \cdot \gamma)} - 1 \right) \cdot \beta, \qquad (3)$$

where γ denotes the training progress, taking on values ranging from 0 to 1. α serves as an adjustment factor; a higher value for α accelerates the removal of source data information. β is the upper bound of the mixing weight, ensuring the foreground objects remain visible and are not obscured. See Fig. 2, we generate the translated image and combine it with the paired source image using the dynamic mixing weight δ, as defined in Eq (3), and obtain the blended image by Eq (1). We also blend the target and source images without considering their structural relationship. The way to blend the target and source images is the same as Eq (1). As illustrated below:

$$I_{s,t} = \delta \cdot I_{target} + (1 - \delta) \cdot I_{source}. \qquad (4)$$

Subsequently, we use $I_{blended}$ and $I_{s,t}$ as input data to fine-tune the model. Using $I_{blended}$ to fine-tune the model with L_{sup}, which is the original detection loss in AQT [12]. Using $I_{blended}$ and $I_{s,t}$ to fine-tune the model with mixed-domain adversarial loss, which will be introduced in Section 2.2.

2.2. Mixed-domain Adversarial Loss

In this section, we explain how to incorporate the dynamic mixing weight δ into the adversarial losses of AQT [12], one of the state-of-the-art domain adaptive object detectors. To perform adversarial learning, gradient reverse layers (GRL) [19] are used to reverse the gradients that pass through the adversarial query tokens during gradient backpropagation. The adversarial losses L_{adv}^{sp}, L_{adv}^{ch}, and L_{adv}^{ins} in AQT are shown as below:

$$L_{adv}^{l} = d \cdot \log D_l(q_l) + (1 - d) \cdot \log(1 - D_l(q_l)), \qquad (5)$$

where $l \in \{sp, ch, ins\}$ indicates different levels of feature alignment, with sp, ch, and ins signifying space-level, channel-level, and instance-level alignment, respectively. q_l is the corresponding adversarial query token, and D_l is the discriminator. Note that d is a domain label, taking value 0 for the source domain and 1 for the target domain. To enable the discriminator to distinguish domains in greater detail, instead of the hard domain label, we can use a soft label $\tilde{d}$ and let $\tilde{d} = \delta$, as follows:

$$\tilde{L}_{adv}^{l} = \tilde{d} \cdot \log D_l(q_l) + (1 - \tilde{d}) \cdot \log(1 - D_l(q_l)). \qquad (6)$$

Note that $\tilde{d} = \delta$ signifies the mixing weight, indicating the proportion of non-source domain information in the mixing data. For example, it represents the proportion of I_{target} used to blend with the source image for $I_{s,t}$.

2.3. Overall Training Objective

The total loss is defined as:

$$L_{total} = L_{sup} + \lambda_{sp}\tilde{L}_{adv}^{sp} + \lambda_{ch}\tilde{L}_{adv}^{ch} + \lambda_{ins}\tilde{L}_{adv}^{ins}, \qquad (7)$$

where λ_{sp}, λ_{ch}, and λ_{ins} are the trade-off parameters. The optimization objective for detection transformer F is

$$F^* = \underset{F}{\operatorname{argmin}} \max_{\substack{D_{sp}, D_{ch}, \\ D_{ins}}} L_{total}. \qquad (8)$$

3. EXPERIMENTS

3.1. Datasets and Experimental Settings

Our experimentation encompasses three standard datasets, including Cityscapes [13], Foggy Cityscapes [14] and BDD100K [24]. We evaluate two adaptation settings: (1) **Source : Cityscapes → Target : Foggy Cityscapes** and (2) **Source : Cityscapes → Target : BDD100K daytime**. The details of the datasets are described below.
Cityscapes: A real urban scenes data containing 2,975 images for training and 500 images for validation, including 8 classes (person, rider, car, train, bicycle, motorbike, truck, and bus).
Foggy Cityscapes: A synthetic dataset generated from Cityscapes. We take the highest fog density images as the validation set to test the performance of our model.
BDD100K daytime: BDD100K is a large-scale driving dataset and the daytime subset is selected as the target domain. The training and validation sets have 70,000 and 10,000 images, respectively. We consider the common 7 categories (person, rider, car, bicycle, motorbike, truck, and bus) following [15].

3.2. Implementation Details

In our experiments, we retain the original architecture of AQT [12], while making modifications to the input data and adversarial losses. To create the translated data, in **Cityscapes to Foggy Cityscapes**, we use InstructPix2Pix [8] with the prompt "Add some fog" and set the text classifier free guidance (cfg) scale = 7.5, image cfg scale = 1.5 and seed = 58,912. In **Cityscapes to BDD100K daytime**, we use the prompt "Driving scene at dark night" and set text cfg scale = 9.0, image cfg scale = 1.5 and seed = 981. This translation is applied to all training images in Cityscapes, resulting in a total of 2,975 translated images. For the hyper-parameters, in **Cityscapes to Foggy Cityscapes**, we set $\alpha = 20$, $\beta = 1.0$ to control the mixing rate and the mixing upper bound, and λ_{sp}, λ_{ch}, and λ_{ins} are set to 10^{-1}. In **Cityscapes to BDD100K daytime**, we set $\alpha = 20$, $\beta = 0.5$, $\lambda_{sp} = 0.1$, $\lambda_{ch} = 10^{-5}$, and $\lambda_{ins} = 10^{-4}$. We use AdamW [25] as the optimizer with the weight decay rate of 10^{-4} and the constant learning rate of 2×10^{-5} for the entire training process. All experiments are trained for 150 epochs with a batch size of 3 (a source image, a translated image, and a target image), using two NVIDIA RTX A6000 GPUs.

3.3. Experiment Results Comparison

In this section, we present a performance comparison of AQT using BlenDA alongside other methods on the Cityscapes to Foggy Cityscapes dataset, as shown in Table 1. Additionally, we provide a performance analysis of AQT, both with and without the use of BlenDA in Table 1 and Table 2.
Cityscapes to Foggy Cityscapes: In Table 1, we compare AQT with BlenDA against existing methods built upon Faster RCNN [26], FCOS [27], or Deformable DETR [17]. "Source Only" refers to the baselines that are only trained on source data. "V16" and "R50" indicate the backbone architecture used, which are VGG-16 [28] and

Table 1. We present the outcomes and comparative analysis of cross-domain object detection on the validation set of Foggy Cityscapes for the adaptation from **Cityscapes to Foggy Cityscapes**. The average precision (AP, %) for all classes is reported.

Method	Detector	Backbone	**prsn**	**rider**	**car**	**truck**	**bus**	**train**	**motor**	**bike**	**mAP**
Source Only	Faster RCNN	R50	26.9	38.2	35.6	18.3	32.4	9.6	25.8	28.6	26.9
TDD [5]	Faster RCNN	R50	50.7	53.7	68.2	35.1	53.0	45.1	38.9	49.1	49.2
CMT [6]	Faster RCNN	V16	47.0	55.7	64.5	**39.4**	**63.2**	**51.9**	40.3	**53.1**	51.9
AT [4]	Faster RCNN	V16	45.5	55.1	64.2	35.0	56.3	54.3	38.5	51.9	50.1
PT [20]	Faster RCNN	V16	40.2	48.8	59.7	30.7	51.8	30.6	35.4	44.5	42.7
Source Only	FCOS	R50	36.9	36.3	44.1	18.6	29.3	8.4	20.3	31.9	28.2
SIGMA [21]	FCOS	R50	44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2
SCAN [22]	FCOS	V16	41.7	43.9	57.3	28.7	48.6	48.7	31.0	37.3	42.1
HT [7]	FCOS	V16	52.1	55.8	67.5	32.7	55.9	49.1	40.1	50.3	50.4
OADM [23]	FCOS	V16	47.8	46.5	62.9	32.1	48.5	50.9	34.3	39.8	45.4
Source Only	Deformable DETR	R50	37.7	39.1	44.2	17.2	26.8	5.8	21.6	35.5	28.5
SFA [18]	Deforamble DETR	R50	47.1	46.4	62.2	30.0	50.3	35.5	27.9	41.2	42.6
AQT [12]	Deformable DETR	R50	49.3	52.3	64.4	27.7	53.7	46.5	36.0	46.4	47.1
AQT(BlenDA)	Deformable DETR	R50	**55.0**	**57.5**	**73.6**	37.5	61.6	48.0	**43.2**	50.9	**53.4** (+6.3)

Table 2. We present a performance comparison of AQT, both with and without the integration of BlenDA, on the BDD100K validation set, focusing on the adaptation from **Cityscapes to BDD100K daytime.**

		prsn	**rider**	**car**	**truck**	**bus**	**motor**	**bike**	**mAP**
Cityscapes to BDD100K	AQT	38.2	33.0	58.4	17.3	**18.4**	16.9	23.5	29.4
	AQT(BlenDA)	**44.5**	**36.3**	**64.1**	**20.0**	18.1	**24.6**	**26.9**	**33.5** (+4.1)

Table 3. Ablation studies of (static/dynamic) δ and the experiments of different loss settings on **Cityscapes to Foggy Cityscapes**.

Different δ Settings				
		static		dynamic
	$\delta = 0.7$	$\delta = 0.9$	$\delta = 1.0$	δ in Eq (3)
mAP	49.1	48.0	32.7	**52.8**

W/ or w/o Mixed-domain Adversarial Loss				
	$I_{blended}$	I_{target}	$I_{s,t}$	mAP
L_{adv}	✓	✓		52.8
$\tilde{L}_{adv}$	✓		✓	53.4

ResNet-50 [29], respectively. After training with BlenDA, AQT improves by 6.3% in mAP, outperforming all other methods and surpassing the previous state-of-the-art by 1.5%. Notably, CMT [6] enhances rare class samples by utilizing target images and their corresponding pseudo labels, whereas BlenDA does not employ such labels. This accounts for the superior performance of rare classes, such as truck, bus, train, and bike, in CMT.

Cityscapes to BDD100K daytime: Our method demonstrates a exceptional performance compared to the original AQT, achieving a mAP higher by 4.1%. Notably, by employing our method, we observe enhancements across nearly all categories, leading to an overall performance surpassing the original AQT baseline. **Bold text** is employed to signify superior performance.

3.4. Ablation Studies

In this section, we analyze the proposed fine-tuning strategy in BlenDA by experiments on **Cityscapes to Foggy Cityscapes**.

Different δ Settings: As indicated in Table 3, it is observed that when we set δ as a hard mixing weight, which disregards the mixing between the source and translated images, the model performs poorly (shown in red). This reflects that even though image trans-

lation helps simulate the domain shifts in the image space, some discriminative details may be lost. Therefore, it is essential to account for mixing between I_{source} and $I_{translated}$ during fine-tuning to achieve better alignment. To further analyze the importance of dynamic mixing weight δ, we use the blended images as source data to fine-tune the model. We set δ to 0.7 and 0.9 as a static mixing weight to generate $I_{blended}$. Although these blended images with a static mixing weight improve the model's performance, utilizing dynamic δ can further improve the model as it can progressively control the mixing of source and domain knowledge. See Table 3, using dynamic mixing weight can achieve performance improvement by 3.7% compared to the setting of δ as 0.7, which gives the highest performance for static blending.

W/ or w/o Mixed-domain Adversarial Loss: To demonstrate that applying the mixed-domain adversarial loss can further enhance the model's performance, we compare the results in two different settings: one uses $I_{blended}$ and I_{target} with the hard domain labels 0 and 1, respectively, and the adversarial loss L_{adv}; the other utilizes $I_{blended}$ and $I_{s,t}$ with a soft label to compute the mixed-domain adversarial loss $\tilde{L}_{adv}$. As shown in Table 3, using the mixed-domain adversarial loss $\tilde{L}_{adv}$ for fine-tuning outperforms that without mixed-domain labels by 0.6% in terms of mAP.

4. CONCLUSION

In this paper, we introduce BlenDA, a regularization strategy for the training of domain adaptive object detectors that incorporate the mixing of the source and the translated images and their corresponding soft domain labels. In general, the proposed approach is also applicable to various domain adaptive object detection paradigms. The proposed BlenDA can effectively reduce domain shifts between the source and target domain by generating blended images for the intermediate domains. We exemplify the application of our training strategy on AQT and evaluate it on two standard benchmarks. The superior results confirm the effectiveness of our methods. The ablation studies further demonstrate that dynamically mixing both images and domain labels significantly improves the model's performance, leading to substantial progress.

5. REFERENCES

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

[2] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan, "Unbiased mean teacher for cross-domain object detection," *CVPR*, 2021.

[3] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *ICCV*, 2017.

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

[5] Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao, "Cross domain object detection by target-perceived dual branch distillation," *CVPR*, 2022.

[6] Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, and Yu-Xiong Wang, "Contrastive mean teacher for domain adaptive object detectors," *CVPR*, 2023.

[7] Jinhong Deng, Dongli Xu, Wen Li, and Lixin Duan, "Harmonious teacher for cross-domain object detection," *CVPR*, 2023.

[8] Tim Brooks, Aleksander Holynski, and Alexei A Efros, "Instructpix2pix: Learning to follow image editing instructions," *arXiv preprint arXiv:2211.09800*, 2022.

[9] Giulio Mattolin, Luca Zanella, Elisa Ricci, and Yiming Wang, "Confmix: Unsupervised domain adaptation for object detection," *WACV*, 2023.

[10] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz, "Mixup: Beyond empirical risk minimization," *ICLR*, 2018.

[11] Thanh Vu, Baochen Sun, Bodi Yuan, Alex Ngai, Yueqi Li, and Jan-Michael Frahm, "Lossmix: Simplify and generalize mixup for object detection and beyond," *arXiv preprint arXiv:2303.10343*, 2023.

[12] Wei-Jie Huang, Yu-Lin Lu, Shih-Yao Lin, Yusheng Xie, and Yen-Yu Lin, "Aqt: Adversarial query transformers for domain adaptive object detection," *IJCAI-ECAI*, 2022.

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," *CVPR*, 2016.

[14] Christos Sakaridis, Dengxin Dai, and Luc Van Gool, "Semantic foggy scene understanding with synthetic data," *IJCV*, 2018.

[15] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang, "Cross-domain detection via graph-induced prototype alignment," *CVPR*, 2020.

[16] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," *ECCV*, 2020.

[17] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *ICLR*, 2021.

[18] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao, "Exploring sequence feature alignment for domain adaptive detection transformers," *ACM MM*, 2021.

[19] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," *ICML*, 2015.

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

[21] Wuyang Li, Xinyu Liu, and Yixuan Yuan, "Sigma: Semantic-complete graph matching for domain adaptive object detection," *CVPR*, 2022.

[22] Wuyang Li, Xinyu Liu, Xiwen Yao, and Yixuan Yuan, "Scan: Cross domain object detection with semantic conditioned adaptation," *AAAI*, 2022.

[23] Jayeon Yoo, Inseop Chung, and Nojun Kwak, "Unsupervised domain adaptation for one-stage object detector using offsets to bounding box," *ECCV*, 2022.

[24] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multi-task learning," *CVPR*, 2020.

[25] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," *ICLR*, 2019.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NIPS*, 2015.

[27] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He, "Fcos: Fully convolutional one-stage object detection," *ICCV*, 2019.

[28] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CVPR*, 2016.