DIVERGENCE-GUIDED FEATURE ALIGNMENT FOR CROSS-DOMAIN OBJECT DETECTION

Zongyao Li[†] *Ren Togo*[††] *Takahiro Ogawa*[†††] *Miki Haseyama*[†††]

[†] Graduate School of Information Science and Technology, Hokkaido University
[††] Education and Research Center for Mathematical and Data Science, Hokkaido University
[†††] Faculty of Information Science and Technology, Hokkaido University
E-mail:{li, togo, ogawa}@lmd.ist.hokudai.ac.jp, miki@ist.hokudai.ac.jp

ABSTRACT

Domain shift causes performance drop in cross-domain object detection. To alleviate the domain shift, a prevailing approach is global feature alignment with adversarial learning. However, such simple feature alignment has defects of unawareness of foreground/background regions and well-aligned/poorly-aligned regions. To remedy the defects, in this paper, we propose a novel divergence-guided feature alignment method for cross-domain object detection. Specifically, we generate source-like images of the target domain and seek cues of foreground regions and poorly-aligned regions from prediction divergence of the source-like and original images. The feature alignment is guided by the divergence maps and consequently results in adaptation performance superior to alignment unaware of the cues. Different from most previous studies focusing on two-stage object detection, this paper is devoted to adapting one-stage object detectors which have simpler and faster inference. We validated the effectiveness of our method by conducting experiments in cross-weather, cross-camera, and synthetic-to-real adaptation scenarios.

Index Terms— Cross-domain object detection, one-stage object detection, unsupervised domain adaptation.

1. INTRODUCTION

Object detection is a fundamental computer vision task which relies on labor-intensive annotations of bounding boxes. To mitigate the dependence on labeled data, cross-domain object detection borrows label information from a well-labeled source domain that shares the same label space with the target domain, which is also called an unsupervised domain adaptation problem. However, discrepancies between the domains may significantly hinder performance in the target domain. Furthermore, cross-domain object detection must transfer both object classification and localization knowledge across domains and is hence a challenging task.

Cross-domain object detection has been studied extensively, mainly on the basis of Faster R-CNN [1] which is a prevailing two-stage object detector. Most of previous studies conducted feature alignment on global features and/or instance features with adversarial learning [2, 3, 4, 5, 6, 7], and some also applied image-to-image translation for pixel-level adaptation [8, 9]. In addition, some researchers aligned the domains by using self-training [10, 11], introducing Mean Teacher architecture [12, 13], performing prototype alignment [14, 15], or exploring categorical information with an auxiliary multi-label classifier [16, 17]. Due to the prevalence of Faster R-CNN, only a few studies [18, 19, 20] adapted one-stage object detectors such as single shot multibox detector (SSD) [21] and fully convolutional one-stage object detector (FCOS) [22], which have simpler and faster inference and are this paper's focus. Following many previous methods proposed for Faster R-CNN, we try to adapt one-stage object detectors by aligning global features with adversarial learning.

Global feature alignment proposed in [2] performed domain adversarial learning between the object detector and a domain discriminator which predicts domain labels of feature maps extracted from the detector backbone. However, such simple feature alignment is unaware of some feature information that has a considerable effect on the alignment: foreground/background and well-aligned/poorly-aligned. First, since aligning background features which involve no categorical information is far less meaningful for knowledge transfer than aligning foreground features, awareness of foreground/background regions is intuitively valuable for the feature alignment. Moreover, awareness of well-aligned/poorly-aligned regions can be used to adaptively adjust the alignment process. Some methods improved the feature alignment in the above respects by weighting the backbone features or discriminator losses with attention maps [11, 14, 19, 23]. However, some of them relied on network predictions which may be unreliable in the target domain, and more importantly, none of them solved both the defects mentioned above.

In this paper, we propose a novel divergence-guided feature alignment method to solve the above-mentioned problems in cross-domain object detection. Specifically, we focus on adaptation for one-stage object detectors, which has yet to be widely studied, and innovatively introduce a divergence-based guidance mechanism for the feature alignment with adversarial learning. To seek cues of foreground regions and poorly-aligned regions in the target domain, we translate target domain images to the source domain by pixel-level adaptation and calculate the divergence of classification results for the source-like and original target domain images. The obtained divergence maps are then used as attention to guide the feature alignment by spatially weighting the discriminator losses, assuming that foreground regions are highlighted in the divergence maps and poorly-aligned features tend to result in larger prediction divergence than well-aligned features. In this manner, information of foreground regions and poorly-aligned regions is perceived by the feature alignment and contributes to improvements in adaptation performance. We conduct experiments in three representative

This study was partly supported by JSPS KAKENHI Grant Number JP17H01744 and JP21H03456. This study was conducted on the Data Science Computing System of Education and Research Center for Mathematical and Data Science, Hokkaido University.

adaptation scenarios including cross-weather, cross-camera, and synthetic-to-real adaptation, and our method shows superior performance to the previous methods.

2. PRELIMINARIES: CROSS-DOMAIN OBJECT DETECTION WITH GLOBAL FEATURE ALIGNMENT

2.1. Problem Definition

Let $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ be the source domain which consists of a total number of n_s images x_i^s and corresponding object annotations y_i^s. And let $\mathcal{D}_t = \{x_i^t\}_{i=1}^{n_t}$ be the target domain which contains only n_t unlabeled images x_i^t and shares the same object categories as $\mathcal{D}_s$. Given the two domains $\mathcal{D}_s$ and $\mathcal{D}_t$, we aim to train an object detector that can generalize well to the target domain by transferring label knowledge from the source domain to the target domain.

2.2. Base Object Detector

In this paper, we construct our method on the basis of FCOS [22], a well-performed framework for one-stage object detection. Note that we believe that our method can also adapt to other prevailing one-stage detectors such as SSD [21] and RetinaNet [24] since our method does not depend on any specific characteristic of FCOS. Here, we introduce briefly the architecture of FCOS.

FCOS consists of a backbone network B, for which feature pyramid network (FPN) [25] is adopted to produce feature maps of multiple levels, and a detection head H which predicts objects on the feature maps. Each object is assigned to one of the feature levels according to its bounding box size, and H is shared across all the levels. Unlike anchor-based detectors [1, 21, 24], FCOS directly regresses the target bounding box at each location instead of using pre-defined anchor boxes as references. The detection head consists of a classification branch for predicting categories of objects, a regression branch for regressing bounding boxes, and a center-ness branch for suppressing low-quality bounding boxes. The center-ness branch is trained to predict the distance between a location and the center of the object that the location is responsible for. Bounding boxes regressed at a location far from the object center are considered low-quality and thus down-weighted in the final non-maximum suppression (NMS) process.

Corresponding to the above branches of the detection head, FCOS is trained with a classification loss $\mathcal{L}_{\text{cls}}$, a bounding box regression loss $\mathcal{L}_{\text{loc}}$, and a center-ness loss $\mathcal{L}_{\text{ctr}}$. For details, please refer to [22]. The final loss function of FCOS can be written as:

$$\mathcal{L}_{\text{fcos}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{ctr}}. \tag{1}$$

2.3. Global Feature Alignment

Global feature alignment with adversarial learning has been widely used to reduce the domain shift in cross-domain object detection. By conducting domain adversarial learning between the backbone network and a discriminator, the backbone feature distributions in the two domains become close, which benefits detection on the target domain feature maps by the detection head. Here, we apply the general global feature alignment to FCOS and then propose our divergence-guided feature alignment in Section 3.

Given the multi-level feature maps $\{B_k(x)\}_{k=1}^{N_l}$ of image x from the backbone network B of FCOS where N_l is the number of feature levels, a discriminator D_k ($k = 1, 2, \ldots, N_l$) is trained to predict domain labels of the feature maps for each level. B and D_k are connected with a Gradient Reversal Layer (GRL) [26] which reverses gradients derived from the domain classification loss and back-propagates the reversed gradients towards B. Consequently, B and D_k are trained in an adversarial manner, and $B_k(x)$ becomes gradually domain-invariant. The domain classification loss for global feature alignment is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{glo}} = \sum_{k=1}^{N_l} [&\mathbb{E}_{x^s \sim \mathcal{D}_s} \log D_k(B_k(x^s)) + \\ &\mathbb{E}_{x^t \sim \mathcal{D}_t} \log(1 - D_k(B_k(x^t)))].\end{aligned} \tag{2}$$

Note that D_k produces spatial results for locations on the feature maps, and the spatial dimensions are omitted in this paper for simplicity. The training process is conducted by optimizing the following objective function:

$$\min_{B,H} \max_{D} \mathcal{L}(B, H, D) = \mathcal{L}_{\text{fcos}}(B, H) + \lambda \mathcal{L}_{\text{glo}}(B, D), \tag{3}$$

where λ is a trade-off parameter.

3. DIVERGENCE-GUIDED FEATURE ALIGNMENT FOR CROSS-DOMAIN OBJECT DETECTION

3.1. Motivation

The global feature alignment method described in Section 2.3 pays equal attention to all regions in an image and may thereby lead to inferior alignment. This can be explained by two factors: (1) foreground features are more informative for domain adaptation than background features, and (2) there may exist both well-aligned and poorly-aligned regions in an image. The first factor is intuitive considering that the aim is to transfer label knowledge of foreground objects. The second factor causes insufficient alignment for the poorly-aligned features since the objects that are relatively easy to align are usually in the majority among the target objects and thus dominate the training process. To tackle the above problems, we propose a divergence-based guidance mechanism which emphasizes alignment for foreground regions and poorly-aligned regions on the feature maps. An overview of our method is shown in Fig. 1.

3.2. Production of Divergence Maps with Pixel-level Adaptation

Pixel-level adaptation with image-to-image translation models such as CycleGAN [27] which reduces visual differences between the domains has shown effectiveness for domain adaptation. However, domain translation is typically challenging in real-world applications and can only provide limited improvements in detection performance. Therefore, rather than using the pixel-level adaptation to reduce the domain shift, we seek cues of foreground regions and poorly-aligned regions on the feature maps from divergence maps which measure the prediction divergence of images before and after translation. Our conception is that predictions for foreground objects are more prone to be affected by the translation than those for background contents since foreground features are richer with categorical information, and moreover, features of poorly-aligned regions are domain-specific and thus affected more by the translation than domain-invariant features of well-aligned regions. Consequently, foreground regions and poorly-aligned regions have higher prediction divergence and are highlighted on the divergence maps.

As shown in Fig. 1 (a), we only produce the divergence maps for target domain images, and for source domain images, we directly use the ground-truth label maps as indications of foreground

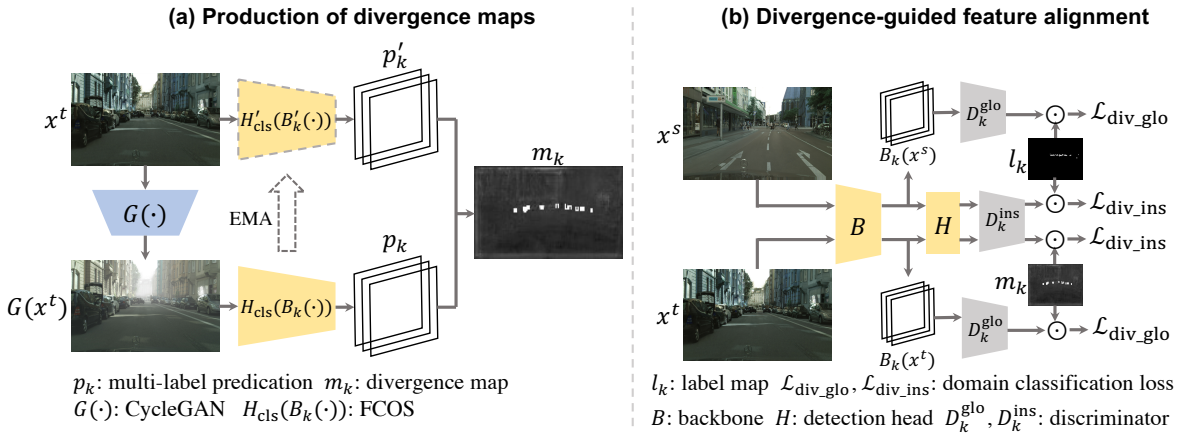

Fig. 1. An overview of the proposed method. Note that we conduct feature alignment for feature maps of multiple levels respectively (k=1, 2, . . ., N_l), while only one level is shown in the figure.

regions. Specifically, to produce the divergence maps, we first translate the target domain image x^t to the source domain with a pre-trained CycleGAN model G. Next, we obtain outputs of the classification branch for x^t and $G(x^t)$ as follows:

$$p'_k = H'_{\text{cls}}(B'_k(x^t)), \tag{4}$$

$$p_k = H_{\text{cls}}(B_k(G(x^t))), \tag{5}$$

where $H_{\text{cls}}(\cdot)$ is multi-label prediction produced by the classification branch of H, and $H'_{\text{cls}}(B'_k(\cdot))$ is a self-ensembling model of $H_{\text{cls}}(B_k(\cdot))$. We use the self-ensembling model, which is an exponential moving average (EMA) of the detector and updated after each iteration, for producing more stable and more reliable predictions. With the obtained classification outputs p_k and p'_k, the divergence map is calculated as the following equation:

$$m_k^{(h,w)} = \sum_{i=1}^{N_c} ||p_k^{(h,w,i)} - p_k'^{(h,w,i)}||_2^2, \tag{6}$$

where h, w, i are indexes of the spatial dimensions and category dimension respectively, and N_c is the number of categories of foreground objects. m_k is then normalized to (0, 1). For any source domain image x_s, we produce a label map l_k ($k = 1, 2, \ldots, N_l$) on which locations labeled as positive by any object category are assigned 1 and 0 otherwise. Both m_k and l_k are further clipped with a minimum value of 0.05 to avoid completely eliminating alignment for background regions.

Figure 2 shows an example of the produced divergence maps. We can see that objects in the target image are highlighted by the divergence maps of different feature levels. The left divergence map highlights small-size objects since it is derived from predictions on low-level feature maps. Medium-size objects are highlighted by the middle one, and the largest objects by the right one. The example shown in Fig. 2 implies the capability of our method to highlight foreground regions on the divergence maps, as assumed in our conception mentioned above. Since there are no definite indicators of well-aligned/poorly-aligned regions for qualitative analysis, we will validate the capability of our method to highlight poorly-aligned regions and its significance in quantitative experiments.

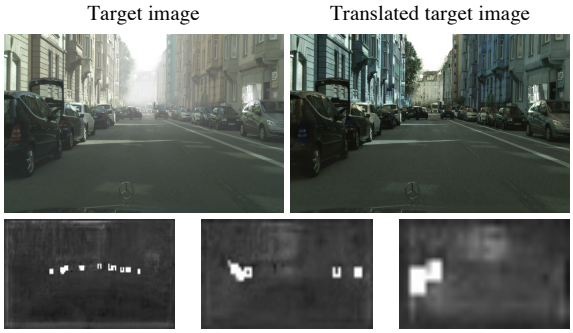

Fig. 2. An example of the produced divergence maps. The divergence maps were produced at the first level, the third level, and the fourth level (from left to right) of the total five feature levels.

3.3. Feature Alignment Guided with Divergence Maps at Two Levels

Using the divergence maps m_k and label maps l_k, we can guide the feature alignment process to emphasize the alignment for foreground regions and poorly-aligned regions, which is shown in Fig. 1 (b). Specifically, on the basis of the feature alignment method described in Section 2.3, we introduce m_k and l_k into the alignment as weights of the domain classification loss in Eq. (2). Our divergence-guided feature alignment loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{div\_glo}} = \sum_{k=1}^{N_l} [&\mathbb{E}_{x^s \sim \mathcal{D}_s} l_k \odot \log D_k(B_k(x^s)) + \\ &\mathbb{E}_{x^t \sim \mathcal{D}_t} m_k \odot \log(1 - D_k(B_k(x^t)))],\end{aligned} \tag{7}$$

where "$\odot$" indicates the element-wise product. $\mathcal{L}_{\text{glo}}$ in Eq. (3) is replaced by $\mathcal{L}_{\text{div\_glo}}$ in the objective function of our method.

Different from two-stage detectors that use RoIPool [28] to obtain instance-level features, FCOS uses a detection head composed of only convolutional layers to detect on the backbone feature maps. As a result, instance-level feature maps which have the same spatial

size as the backbone feature maps can be obtained from the detection head by concatenating outputs of the second last convolutional layers of the classification branch and the regression branch. Moreover, the divergence-guided alignment for the backbone features can also be performed for the instance-level features similarly. Since the feature alignment at the instance level is complementary with that at the backbone level, we jointly conduct feature alignment at the two levels by optimizing the final objective function defined as follows:

$$\min_{B,H} \max_{D^{\mathrm{glo}},D^{\mathrm{ins}}} \mathcal{L}(B,H,D^{\mathrm{glo}},D^{\mathrm{ins}}) = \mathcal{L}_{\mathrm{fcos}}(B,H) + \lambda(\mathcal{L}_{\mathrm{div\_glo}}(B,D^{\mathrm{glo}}) + \mathcal{L}_{\mathrm{div\_ins}}(B,D^{\mathrm{ins}})), \tag{8}$$

where D^{glo} and D^{ins} are the discriminators for aligning the backbone features and instance-level features respectively, and $\mathcal{L}_{\mathrm{div\_ins}}$ is similar to $\mathcal{L}_{\mathrm{div\_glo}}$ replacing $B_k(\cdot)$ with the instance-level features.

4. EXPERIMENTS

4.1. Implementation Details

We used ResNet-50 [29] as the backbone network of FCOS. The discriminators are composed of four 3×3 convolutional layers. In the discriminators for the largest feature maps and second largest feature maps, the outputs were down-scaled by two and one convolutional layers with stride 2 respectively, and the weight maps were down-sampled to the output size with max-pooling. The networks were trained with a stochastic gradient descent (SGD) optimizer for 24,000 iterations with an initial learning rate of 0.01 and a mini-batch of 8 images. The learning rate was decreased to 0.001 at iteration 18,000. The scale parameter of GRL and loss weight λ were set as 0.1 and 1.0 respectively in normal-to-foggy scenario, and 0.01 and 0.1 in the other scenarios.

4.2. Datasets and Adaptation Scenarios

Normal-to-Foggy. Cityscapes [30] and Foggy Cityscapes [31] are the source domain and target domain in this scenario. Cityscapes is a street scene dataset which consists of 2,975 images for training and 500 images for validation. All the images are collected in clear weather. Foggy Cityscapes derives from Cityscapes and consists of synthetic foggy images. The validation set of Foggy Cityscapes was used for evaluation. The categories include "person", "rider", "car", "truck", "bus", "train", "motorcycle", and "bicycle".

Cross-Camera. KITTI [32] and Cityscapes are the source domain and target domain in this scenario. KITTI is also a street scene dataset but collected with a different camera setup from Cityscapes. 7,481 images are provided for training in KITTI. Evaluation was performed on the validation set of Cityscapes for five common categories including "person", "rider", "car", "truck", and "train".

Synthetic-to-Real. Synscapes [33] and Cityscapes are the source domain and target domain in this scenario. Synscapes is a synthetic dataset which consists of 25,000 photo-realistic street scene images. Evaluation was performed on the validation set of Cityscapes for five common categories including "person", "car", "truck", "bus", and "train".

4.3. Methods for Comparison

We conducted experiments with three previous methods most related to ours for comparison. All of them align the features from the backbone network with adversarial learning. Global feature alignment [2] which has been described in Section 2.3 aligns the backbone features with no additional information. Center-aware alignment [19] introduces the output of the center-ness branch of FCOS into the feature alignment to highlight foreground features. Uncertainty-aware alignment [23] uses prediction uncertainty to assess the alignment and emphasizes the alignment for poorly-aligned features. Our method considers both foreground and poorly-aligned features with the proposed divergence-based guidance mechanism. Since the methods [2, 23] are proposed for Faster R-CNN, we retained only the alignment for backbone features to apply the methods to FCOS.

Table 1. Adaptation performance of the proposed method and previous methods in three scenarios. $C \rightarrow F$: Cityscapes to Foggy Cityscapes. $K \rightarrow C$: KITTI to Cityscapes. $S \rightarrow C$: Synscapes to Cityscapes.

Method	mAP		
	$C \rightarrow F$	$K \rightarrow C$	$S \rightarrow C$
Source only	41.1	18.2	33.3
Global feature alignment [2]	49.9	19.0	36.7
Center-aware alignment [19]	50.6	19.8	38.3
Uncertainty-aware alignment [23]	50.8	20.6	37.6
Ours ($\mathcal{L}_{\mathrm{div\_glo}}$ only)	51.9	21.2	38.7
Ours ($\mathcal{L}_{\mathrm{div\_glo}}$+$\mathcal{L}_{\mathrm{div\_ins}}$)	**52.2**	**21.4**	**38.8**

4.4. Results

Table 1 shows adaptation performance of the proposed method and the three previous methods. We also report the results of training with only source domain data as baselines. As the evaluation metric, we compute the mean average precision (mAP) of the common categories with an intersection over union (IoU) threshold of 0.5.

As shown in Table 1, all of the domain adaptation methods improved the detection performance compared to the baseline of training with only source data in the three adaptation scenarios, and the methods [19, 23] that introduce additional guidance information for adaptation outperformed the general alignment method [2]. In all of the scenarios, our method that conducts feature alignment at only the backbone level ("$\mathcal{L}_{\mathrm{div\_glo}}$ only" in Table 1) achieved superior performance to all of the previous methods which also conduct feature alignment at the backbone level, which proved the effectiveness of guiding the feature alignment with the proposed divergence maps. Moreover, the performance was slightly improved by jointly aligning the features at both the backbone level and the instance level ("$\mathcal{L}_{\mathrm{div\_glo}}$+$\mathcal{L}_{\mathrm{div\_ins}}$" in Table 1), which indicated the complementarity of feature alignment at the two levels.

5. CONCLUSION

In this paper, we have proposed a novel unsupervised domain adaptation method for one-stage cross-domain object detection. Our method produces divergence maps from predictions for target images to guide feature alignment. With the proposed divergence-based guidance mechanism, our method is aware of both foreground regions and poorly-aligned regions and leverages such information to emphasize the alignment for these regions, which has not been realized in previous works to our knowledge. Experiment results in three representative adaptation scenarios show that our method outperformed the previous methods that conduct simple global feature alignment or alignment with additional guidance information.

6. REFERENCES

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, vol. 28, pp. 91–99.

[2] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool, "Domain adaptive faster r-cnn for object detection in the wild," in *CVPR*, 2018, pp. 3339–3348.

[3] Zhenwei He and Lei Zhang, "Multi-adversarial faster-rcnn for unrestricted object detection," in *ICCV*, 2019, pp. 6668–6677.

[4] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko, "Strong-weak distribution alignment for adaptive object detection," in *CVPR*, 2019, pp. 6956–6965.

[5] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin, "Adapting object detectors via selective cross-domain alignment," in *CVPR*, 2019, pp. 687–696.

[6] Zhenwei He and Lei Zhang, "Domain adaptive object detection via asymmetric tri-way faster-rcnn," in *ECCV*. Springer, 2020, pp. 309–324.

[7] Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel, "Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection," in *CVPR*, 2021, pp. 4516–4526.

[8] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou, "Harmonizing transferability and discriminability for adapting object detectors," in *CVPR*, 2020, pp. 8869–8878.

[9] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang, "Progressive domain adaptation for object detection," in *WACV*, 2020, pp. 749–757.

[10] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready, "A robust learning approach to domain adaptive object detection," in *ICCV*, 2019, pp. 480–490.

[11] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin, "Collaborative training between region proposal localization and classification for domain adaptive object detection," in *ECCV*. Springer, 2020, pp. 86–102.

[12] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao, "Exploring object relation in mean teacher for cross-domain detection," in *CVPR*, 2019, pp. 11457–11466.

[13] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan, "Unbiased mean teacher for cross-domain object detection," in *CVPR*, 2021, pp. 4091–4101.

[14] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang, "Cross-domain object detection through coarse-to-fine feature adaptation," in *CVPR*, 2020, pp. 13766–13775.

[15] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang, "Cross-domain detection via graph-induced prototype alignment," in *CVPR*, 2020, pp. 12355–12364.

[16] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei, "Exploring categorical regularization for domain adaptive object detection," in *CVPR*, 2020, pp. 11724–11733.

[17] Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye, "Adaptive object detection with dual multi-label prediction," in *ECCV*. Springer, 2020, pp. 54–69.

[18] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim, "Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection," in *ICCV*, 2019, pp. 6092–6101.

[19] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang, "Every pixel matters: Center-aware feature alignment for domain adaptive object detector," in *ECCV*. Springer, 2020, pp. 733–748.

[20] Chaoqi Chen, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu, "I3net: Implicit instance-invariant network for adapting one-stage object detectors," in *CVPR*, 2021, pp. 12576–12585.

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multibox detector," in *ECCV*. Springer, 2016, pp. 21–37.

[22] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He, "Fcos: Fully convolutional one-stage object detection," in *ICCV*, 2019, pp. 9627–9636.

[23] Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, and Yanpeng Cao, "Uncertainty-aware unsupervised domain adaptation in object detection," *IEEE Transactions on Multimedia*, 2021.

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2980–2988.

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017, pp. 2117–2125.

[26] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*. PMLR, 2015, pp. 1180–1189.

[27] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232.

[28] Ross Girshick, "Fast r-cnn," in *ICCV*, 2015, pp. 1440–1448.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[30] 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," in *CVPR*, 2016, pp. 3213–3223.

[31] Christos Sakaridis, Dengxin Dai, and Luc Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018.

[32] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012, pp. 3354–3361.

[33] Magnus Wrenninge and Jonas Unger, "Synscapes: A photorealistic synthetic dataset for street scene parsing," *arXiv preprint arXiv:1810.08705*, 2018.