BDC-FR: Faster R-CNN with Balanced Domain Classifier for Cross-Domain Object Detection

Shouhong Wan, Rui Wang, Peiquan Jin, Xuebin Yang
School of Computer Science and Technology
University of Science and Technology of China
Hefei, China
wansh@ustc.edu.cn, wr666@mail.ustc.edu.cn, jpq@ustc.edu.cn, yangxb@mail.ustc.edu.cn

***Abstract*—Object detectors trained with massive labeled data often suffer performance degradation in some particular scenarios with data distribution gap. Domain classifier is a commonly used method in the existing domain adaptation algorithms to alleviate the domain discrepancy. However, it has problems of instability in training, difficulty in obtaining optimal solutions and converging to an equilibrium point. To tackle this issue, we propose a novel balanced domain classifier (BDC), which not only eliminates domain discrepancy but also makes the domain classifier and the feature extractor maintain equilibrium during the adversarial learning. Furthermore, we propose an appropriate learning rate adjustment strategy, which makes the detection model converge to an equilibrium point more stably and more rapidly. Based on the domain-invariant region proposal network, we propose a cross-domain object detection model called Faster R-CNN with Balanced Domain Classifier (BDC-FR). The experimental results show that BDC-FR can effectively improve the performance of the cross-domain object detection model.**

***Index Terms*—object detection, domain adaptation**

I. Introduction

Applying the object detection models such as Faster R-CNN [1] trained on one image dataset directly to another image dataset will lead to significant performance drop, because the style, resolution, illumination, *etc.* of images are different. Conventionally, there are two fundamental data sets in cross-domain object detection problem, source domain dataset (with annotation information) and target domain dataset (without annotation information). There is always a distribution change between two domains, and it is crucial to develop approaches that enable better generalization of object detectors.

Recently, various domain adaptation approaches [2]–[6] have been proposed to solve this problem. To address this issue, these approaches attempt to build invariant feature representation by employing domain classifiers in adversarial learning. However, adversarial learning has been known to be unstable to train due to training instability and sensitivity to hyper-parameters. The relationship between the feature extractor and the domain classifier can easily become unbalanced during this process. When the feature extractor can easily deceive the domain classifier after a few iterations of training, the prediction result of the domain classifier is similar to a random value and cannot provide an effective optimization gradient for the feature extractor. On the other hand, if the domain classifier has a strong learning ability, it can accurately predict the domain label of the image every time, which will cause the phenomenon of gradient disappearance.

To overcome such unbalanced relationship between the domain classifier and the feature extractor, we design a novel balanced domain classifier network, which can effectively make the feature extractor and the domain classifier maintain a balanced state during training. Furthermore, considering that learning rate also has an impact on the convergence of the model, we propose a simple yet effective learning rate adjustment strategy to make the update of network weight more reasonable. Finally, we build a cross-domain detection model called Faster R-CNN with Balanced Domain Classifier (BDC-FR). We conduct several experiments to evaluate BDC-FR in multiple datasets, and the results demonstrate the effectiveness of our model.

The contribution of this work can be summarized as follows: (i) we design balanced domain classifiers to solve the unstable training problem in cross-domain object detection with domain classifier; (ii) we propose a learning rate adjustment strategy to make the detection model converge to an equilibrium point more stably and more rapidly; (iii) we propose a novel cross-domain detection model called BDC-FR and conduct extensive experiments to validate the effectiveness of proposed BDC-FR.

II. Related Work

Unsupervised domain adaptation (UDA) aims to transfer the information learned from a large number of labeled samples in the source domain to the target domain to solve the same problem, while the available samples in the target domain are unlabeled. Chen et al. [2] initially build a method based on Faster R-CNN, which minimizes the domain discrepancy by utilizing domain classifier at image- & instance-level. MeGA-CDA [5] employs category-wise domain classifiers to ensure category-aware feature alignment for learning domain-invariant discriminative features. Zhao et al. [6] strengthen both the classification and localization capabilities of the cross-domain detector by developing fine-grained feature alignment in separate task spaces. Domain classifiers have limited classification ability due to the unstable adversarial training process. In this paper, we propose a balanced domain classifier network to solve the imbalance between the feature extractor and the domain classifier.

DOI reference number: 10.18293/SEKE2023-114

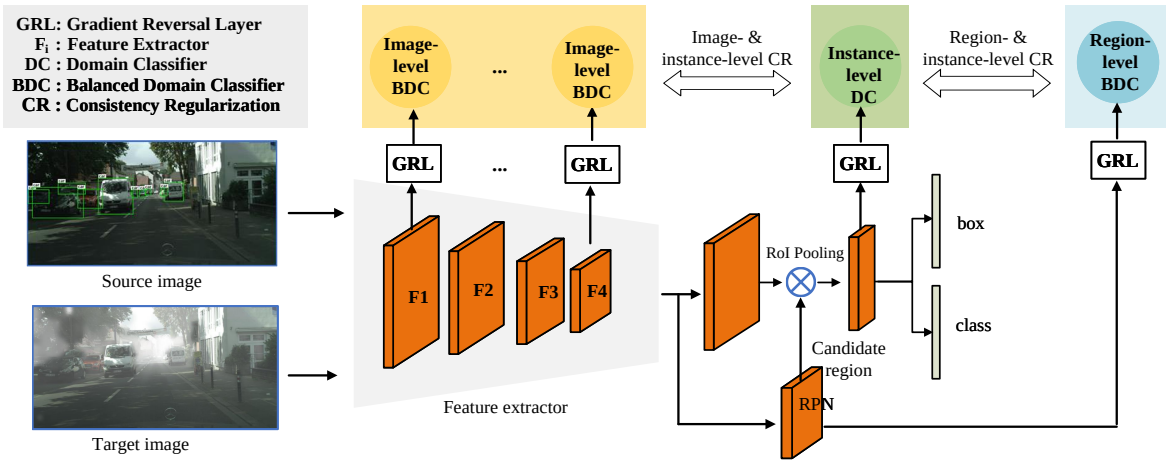

Fig. 1. Overview of the proposed Faster R-CNN with Balanced Domain Classifier (BDC-FR). By applying the proposed balanced domain classifier in image-level and region-level adaptation, the domain classifier and feature extractor can maintain balance in the training process.

III. The Proposed Model

A. BDC-FR Model

In this subsection, we overview the architecture of Faster R-CNN with Balanced Domain Classifier (BDC-FR). Fig. 1 illustrates the framework of our proposed model. Our model contains three major components, including the basic feature alignment network, the domain-invariant region proposal network and the balanced domain classifier.

B. Domain-Invariant Region Proposal Network

Region proposal plays an important role in object detectors. To get better region proposals, we use an RPN domain classifier to minimize domain shift between domains. Specifically, we extend RPN by embedding an RPN domain classifier, and train the classifier in an adversarial learning manner by using a Gradient Reversal Layer (GRL) [7]. The optimization objective of the RPN domain classifier is defined by (1):

$$\mathcal{L}_{rpn} = -\sum_{i,u,v}[D_i \log R_i^{(u,v)} + (1-D_i)\log(1-\log R_i^{(u,v)})], \tag{1}$$

where D_i denotes the domain label of the i^{th} image, and $R_i^{(u,v)}$ is the output of the region-level domain classifier at (u,v) of the RPN feature map.

The basic feature alignment network consists of image-level adaptation and instance-level adaptation. The image-level adaptation and the instance-level adaptation aim to reduce the domain discrepancy in image-level and instance-level, they are defined by (2) and (3):

$$\mathcal{L}_{img} = -\sum_{i,u,v}[D_i \log F_i^{(u,v)} + (1-D_i)\log(1-F_i^{(u,v)})], \tag{2}$$

$$\mathcal{L}_{ins} = -\sum_{i,j}[D_i \log N_{i,j} + (1-D_i)\log(1-\log N_{i,j})], \tag{3}$$

where $F_i^{(u,v)}$ denotes the output of the image-level domain classifier at (u,v) of the base feature extractor, $N_{i,j}$ represents the output of the instance-level domain classifier at the j^{th} instance of the i^{th} image,

In order to ensure that all the domain classifiers are consistent, we design the Double-Consistency Regularization (DCR), which includes two kinds of regulation, namely image- & instance-level consistency regularization and region- & instance-level consistency regularization. The loss of DCR is defined as follows:

$$\mathcal{L}_{dou\_cs} = \sum_{i,j}\|\frac{1}{|I|}\sum_{u,v}F_i^{(u,v)} - N_{i,j}\|_2 + \sum_{i,j}\|\frac{1}{|I|}\sum_{u,v}R_i^{(u,v)} - N_{i,j}\|_2, \tag{4}$$

where $|I|$ denotes the total number of pixels in the i^{th} image and $\|\cdot\|_2$ denotes the Euclidean norm.

C. Balanced Domain Classifier Network

Similar to GANs [8], the domain classifier attempts to accurately distinguish the domain labels of images, while the feature extractor deceives the domain classifier by aligning the features of images. The most important problem is to ensure that the extractor and the classifier are on par to each other. Generally, GANs uses alternating iterative training method to keep the generator and the discriminantor balanced. Different from GANs, the detection model with domain classifier is end-to-end and can't use alternating iterative training method to keep balanced.

In addition to training rounds, network parameters are also a way to control network capability. The parameters of domain classifier network, including the number of convolutional layers, the number and size of convolution kernels and the step size, jointly control the receptive field and feature extraction ability of domain classifier. So the learning ability of the domain classifier can be balanced by controlling these parameters. To obtain the optimal parameters effectively, we propose an iterative control variable method (As shown in Algorithm

Algorithm 1: Iterative control variable method

Input:
(1) Number of convolutional layer parameters n;
(2) Convolutional layer parameters p_i (i from 1 to n);
(3) Threshold t.
Output: Convolutional layer parameters p_i.

Initialize p_i, $i = 0$ and current accuracy $acc_{now} = 1$;
do
  Fix parameter values other than p_i and optimize p_i;
  $acc_{pre} = acc_{now}$;
  Calculate the current accuracy of the cross-domain detection model acc_{now};
  $i = (i + 1)$ mod n;
while $|acc_{pre} - acc_{now}| > t$;

1) to search the optimal parameters of the domain classifier network. During each search, only one parameter value is controlled as a variable, and the rest are fixed values. Search the optimal value of the variable parameters, and repeat the above steps iteratively until the model reaches a better solution. By using Algorithm 1, we can find the optimal parameters of domain classifier, then build the balanced domain classifier.

D. Learning Rate Adjustment Strategy

The learning rate is mainly used to control the strength of adjusting parameters of the detection model. When the prediction error is large, the model has a large learning space. Its parameters then can be adjusted with a large learning rate to speed up the convergence. When the error is small, the model has converged closely to the equilibrium point. At this time, it only needs a small learning rate to fine tune the parameters. Therefore, we design a learning rate adjustment strategy that is suitable for training based on adversarial learning methods. The specific adjustment is is defined by (5):

$$\alpha = \begin{cases} \alpha_{up} & ;if \quad loss \geq t_{up} \\ \alpha_{low} + \dfrac{loss - t_{low}}{t_{up} - t_{low}} (\alpha_{up} - \alpha_{low}) & ;if \quad t_{up} > loss > t_{low} \\ \alpha_{low} & ;if \quad t_{low} \geq loss \end{cases}, \quad (5)$$

where α is the learning rate of the current training round of the model, and $loss$ is the prediction error of the current iteration. α_{up} and α_{low} are the maximum and minimum learning rates, respectively. t_{up} and t_{low} are the upper and lower thresholds of prediction loss.

E. Overall Loss

The overall loss function for training our BDC-FR network can be summarized as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{det} + \lambda_{domain}(\mathcal{L}_{img} + \mathcal{L}_{ins} + \mathcal{L}_{rpn} + \mathcal{L}_{dou\_cs}), \quad (6)$$

where $\mathcal{L}_{det}$ is the detection loss generated by the Faster R-CNN. λ_{domain} is the trade-off to balance the loss of object detection and adaptive module. In our experiments, we set the value of λ_{domain} to 0.1.

IV. Experiments

A. Datasets and Settings

To validate the effectiveness of our proposed BDC-FR, we perform our model on popular image data sets: Cityscapes [9], Foggy Cityscapes [10], SIM 10k [11], and KITTI [12]. We design three different scenario experiments: Adverse Weather Adaptation, Synthetic Data Adaptation, and Cross Camera Adaptation. We implement BDC-FR with Pytorch and use the VGG16 network as the backbone of our model. Besides, in our learning rate adjustment strategy, we set the parameters α_{up} = 2e-3 and α_{low} = 2e-5, the parameters t_{up} and t_{low} are set to 12 and 1, respectively. We report mAP with an IoU threshold of 0.5 for evaluation.

B. Experiment Results

TABLE I
Results of the Adverse Weather Adaptation experiment.

Method	person	rider	car	truck	bus	train	cycle	bcycle	mAP
Base FR [1]	24.5	32.7	35.4	12.7	26.7	9.2	9.9	30.0	22.6
DAF [2]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
SWDA [3]	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
MDA [13]	33.2	44.2	44.8	28.2	41.8	28.7	30.5	36.5	36.0
DIR-FR [4]	36.9	45.8	49.4	28.2	44.6	34.9	35.1	38.9	39.2
HTCN [14]	47.4	37.1	47.9	32.3	33.2	47.5	40.9	31.6	39.8
UMT [15]	56.5	37.3	48.6	30.4	33.0	46.7	46.8	34.1	41.7
MeGA-CDA [5]	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8
TIA [6]	52.1	38.1	49.7	37.7	34.8	46.3	48.6	31.1	42.3
BDC-FR	38.2	48.4	52.9	29.8	51.0	43.3	37.1	41.9	**42.9**

Weather is a common factor causing domain shift. It is important for the detection model to perform faithfully in different weather conditions. In adverse weather adaptation experiment, we use Cityscapes as the source domain while Foggy Cityscapes as the target domain. The Foggy Cityscapes dataset is rendered from the original clear-weather images by simulating fog on real scenes. Table I shows that BDC-FR outperforms all other work and improves up to 42.9%. Specifically, the detection performance of our BDC-FR exceeds DAF [2], SWDA [3], DIR-FR [4] HTCN [14], HTCN [14], UMT [15], MeGA-CDA [5], TIA [6] by 15.3%, 8.6%, 6.9%, 3.7%, 3.1%, 1.2%, 1.1%, 0.6%.

TABLE II
Results of Synthetic Data Adaptation Experiment.

Method	car AP
Base FR [1]	34.2
DAF [2]	39.0
SWDA [3]	42.3
MDA FR [13]	42.0
DIR-FR [4]	**45.5**
HTCN [14]	42.5
UMT [15]	43.1
MeGA-CDA [5]	44.8
BDC-FR	45.3

A large amount of labeled synthetic data is easy to obtain by computer graphics technique. In synthetic data adaptation experiment, our source domain dataset is SIM 10k, which is

rendered by the game Grand Theft Auto (GTA-V). The target domain dataset is Cityscapes, which is an urban scene dataset from the real world. The results are summarized in Table II. Note that our method reduces the training time. During the model training, our BDC-FR model only trained 9 iterative rounds, which is one less than DIR-FR, which shows that the balanced domain classifier network makes the model converge more rapidly. We argue that the performance degradation of BDC-FR in comparison to DIR-FR is mainly caused by the decrease in the difficulty of the detection task.

TABLE III
RESULTS OF CROSS CAMERA ADAPTATION EXPERIMENT.

Method	person	rider	car	truck	train	mAP
Base FR [1]	43.3	28.6	73.9	13.6	14.0	34.7
DAF [2]	40.9	16.1	70.3	23.6	21.2	34.4
MDA FR [13]	53.0	24.5	72.2	28.7	25.3	40.7
C2F [16]	50.4	29.7	73.6	29.7	21.6	41.0
DIR-FR [4]	58.5	37.2	75.4	30.6	18.5	44.0
BDC-FR	54.4	37.5	73.1	39.0	15.0	**44.1**

Cameras with different parameters can also cause domain discrepancy even in the same scene. In cross camera adaptation experiment, we evaluate our model on Cityscapes and KITTI. We take Cityscapes as the source domain and KITTI as the target domain. As table III shows, our proposed BDC-FR achieves the best score in most categories. It again takes 9 epochs for BDC-FR to converge, which is faster than DIR-FR (10 epochs).

C. Ablation Experiment

TABLE IV
RESULTS OF ABLATION EXPERIMENT IN ADVERSE WEATHER ADAPTATION.

Method	$\mathcal{L}_{img}$&$\mathcal{L}_{ins}$	$\mathcal{L}_{rpn}$	BDC	LR	**mAP**
Ours (w/o all)	✓				36.0
Ours	✓	✓			39.2
	✓	✓	✓		42.6
	✓	✓	✓	✓	42.9

We conduct a ablation study of our proposed method on Adverse Weather Adaptation. Table IV shows the results of ablation study. BDC denotes the balance domain classifier, and LR refers to the learning rate adjustment strategy. The mAP of BDC-FR has improved by 3.4 (from 39.2% to 42.6%) with the balance domain classifier. And the learning rate adjustment strategy allows for a shorter training time, as well as improves the detection accuracy to 42.9%.

V. CONCLUSION

In this paper, we propose a cross-domain object detector called Faster R-CNN with Balanced Domain Classifier (BDC-FR). Our key contribution is the balanced domain classifier, which can help the cross-domain object detection model converge steadily by making feature extractor and domain classifier achieve better equilibrium state in training. Furthermore, we propose a learning rate adjustment strategy to improve the convergence of the cross-domain object detection model. In order to verify the validity of the BDC-FR model, we conduct extensive experiments on multiple cross-domain scenarios. Extensive experimental results, as well as ablation studies, demonstrate the effectiveness of the proposed model.

ACKNOWLEDGMENT

This work is supported by Natural Science Foundation of Anhui Province (Grant No. 2208085MF157).

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[2] 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.

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

[4] X. Yang, S. Wan, and P. Jin, "Domain-invariant region proposal network for cross-domain detection," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6.

[5] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4514–4524.

[6] L. Zhao and L. Wang, "Task-specific inconsistency alignment for domain adaptive object detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 197–14 206.

[7] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 1180–1189.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Neural Information Processing Systems*, 2014.

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[10] 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.

[11] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, "Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?" *arXiv preprint arXiv:1610.01983*, 2016.

[12] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[13] R. Xie, F. Yu, J. Wang, Y. Wang, and L. Zhang, "Multi-level domain adaptive learning for cross-domain detection," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[14] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, "Harmonizing transferability and discriminability for adapting object detectors," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8866–8875.

[15] J. Deng, W. Li, Y. Chen, and L. Duan, "Unbiased mean teacher for cross-domain object detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4089–4099.

[16] Y. Zheng, D. Huang, S. Liu, and Y. Wang, "Cross-domain object detection through coarse-to-fine feature adaptation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.