HIERARCHICAL DOMAIN-CONSISTENT NETWORK FOR CROSS-DOMAIN OBJECT DETECTION

Yuanyuan Liu[⋆†] *Ziyang Liu*[⋆] *Fang Fang*[⋆] *Zhanghua Fu*[†] *Zhanlong Chen*[⋆]

[⋆]School of Geography and Information Engineering, China University of Geosciences, Wuhan
[†]Institute of Robotics and Intelligent Manufacturing, The Chinese University of Hong Kong, Shenzhen

ABSTRACT

Cross-domain object detection is a very challenging task due to multi-level domain shift in an unseen domain. To address the problem, this paper proposes a hierarchical domain-consistent network (HDCN) for cross-domain object detection, which effectively suppresses pixel-level, image-level, as well as instance-level domain shift via jointly aligning three-level features. Firstly, at the pixel-level feature alignment stage, a pixel-level subnet with foreground-aware attention learning and pixel-level adversarial learning is proposed to focus on local foreground transferable information. Then, at the image-level feature alignment stage, global domain-invariant features are learned from the whole image through image-level adversarial learning. Finally, at the instance-level alignment stage, a prototype graph convolution network is conducted to guarantee distribution alignment of instances by minimizing the distance of prototypes with the same category but from different domains. Moreover, to avoid the non-convergence problem during multi-level feature alignment, a domain-consistent loss is proposed to harmonize the adaptation training process. Comprehensive results on various cross-domain detection tasks demonstrate the broad applicability and effectiveness of the proposed approach.

Index Terms— Cross-domain object detection, hierarchical feature alignment, domain-consistent loss, foreground-aware attention, adversarial learning

1. INTRODUCTION

Recent years have witnessed great progress in deep learning based object detection. However, due to the domain shift problem, applying off-the-shelf detectors to an unseen domain leads to significant performance drop [8]. Therefore, it is a very challenging task for a detection model to adapt the domain shift from the source domain to an unseen target domain [14].

To address the domain shift problem, existing cross-domain detection methods can be generally divided into two categories. The first category is generation and fine-tuning based methods. Inoue *et.al.* [12] and A. RoyChowdhury *et.al.* [18] improved adaptability of the detection model in the target domain through fine-tuning utilizing pseudo and soft labels. Wang *et.al.* [21] and V. F. Arruda *et.al.* [1] proposed a generator network to further learn the feature difference between the source and target domains. These methods have achieved promising results in some specific scenes. Nevertheless, it still remains difficulty to guarantee the quality of generated images and labels, especially in some extreme cases such as complex traffic scenes, which may undermine the adapted results.

Methods in the second category focus on domain feature alignment in different levels, such as the pixel, the image, and the instance level [10]. Most of the previous methods learn feature adaption in one or two-level feature alignment. Chen *et.al.* [4] used adversarial domain adaption at image-level and instance-level alignment; Xu *et.al.* [23] generated graph prototypes to guide instance-level domain alignment, and achieve good results on common tasks. Besides, Chen *et.al.* [3] performed adversarial learning and regularization for alignment at three levels. Although promising results have been reported, further improvement suffers from the following limitations. Firstly, due to the complexity of cross-domain object detection, one or two-level feature alignment methods are difficult to align the whole process of domain shift, which limits the ability of domain adaption learning. Secondly, the training of three-level feature-aligned method is easy to be non-convergent due to complex construction and lots of parameters.

To address the above limitations, this paper proposes a hierarchical domain-consistent network (HDCN) for cross-domain object detection. It consists of pixel-level, image-level, and instance-level feature alignment subnets for effectively learning domain-invariant features at three levels. The architecture of the proposed method is shown in Fig.1. Moreover, a joint multi-loss with three-level losses and a domain-consistent regularization loss is proposed to optimize the whole network in an end-to-end manner.

This study makes the following research contributions:

1. An effective three-level domain adaptation object detection method, called HDCN, is proposed for solving the multi-level domain shift problem. Extensive experiments show that the HDCN outperforms existing state-of-the-art methods, with the highest accuracies of 51.6% and 45.9% on two cross-domain tasks respectively.
2. At pixel-level feature alignment stage, two adaptive modules, i.e., the foreground-aware attention (FAM) and pixel-level adversarial learning, are adopted to focus on local foreground transferable information.
3. To avoid the non-convergence problem during multi-level feature alignment, a domain-consistent regularization loss is proposed to harmonize the adaptation training process.

2. PROPOSED METHOD

The structure of our proposed method is shown in Fig.1. To address domain shift in multiple levels, the HDCN consists of pixel-level, image-level, and instance-level feature alignment subnets and is op-

This work was supported by a NSFC grant (62076227) and Wuhan Applied Fundamental Frontier Project (2020010601012166). Corresponding author: *Fang Fang

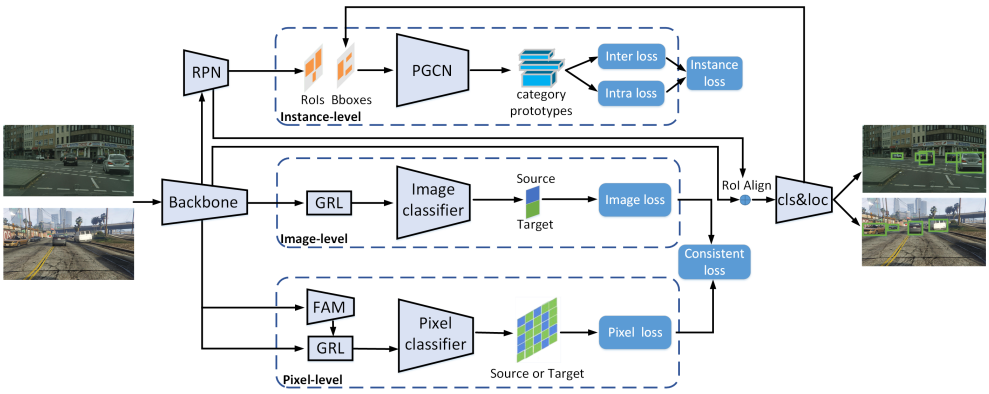


Fig. 1. HDCN architecture. Our method performs three-level feature alignment (i.e., pixel-level, image-level, and instance-level) for domain adaption in a mutually-reinforced manner. Note: GRL denotes the gradient reversal layer, FAM is a foreground-aware attention module, and PGCN is a prototype graph convolutional network.

timized by a joint multi-loss with a domain-consistent regularization loss. Details are given in the following.

2.1. Pixel-level feature alignment

At the pixel-level feature alignment, two adaption modules, i.e., foreground-aware attention learning and pixel-level adversarial learning, are designed for exploring and aligning foreground transferable information.

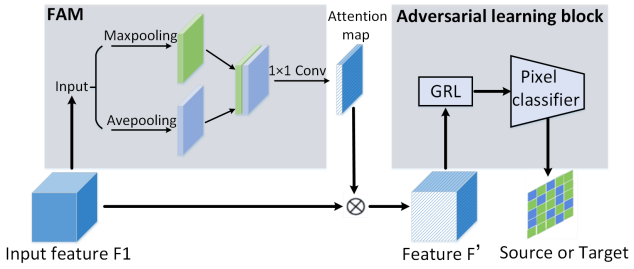


Fig. 2. Pixel-level domain alignment subnet.

2.1.1. *Foreground-aware attention for pixel transfer*

Pixel-level local information in object detection is not all transferable, such as the background. Forcefully aligning the untransferable information leads to negative transfer [22]. Therefore, we introduce the FAM to focus on foreground transferable information and weaken the background untransferable information.

As shown in Fig.2, given an input feature map F_1 extracted from the backbone, we first apply average-pooling and max-pooling operations for down-sampling and then concatenate them to generate an efficient feature descriptor. A 1×1 convolution and element-wise multiplication is used to achieve a foreground transferable map via learning a spatial attention. The foreground transferable attention map F' is given by:

$$F' = A(F_1) \otimes F_1 \tag{1}$$

where $A(\bullet)$ denotes the FAM and $\otimes$ denotes element-wise multiplication. The F' has more significant difference between the foreground and background. Spatial attention mechanism is capable of re-weighting each pixel value according to its contribution. As a result, foreground pixels are re-weighted by higher weights. Therefore, negative transfer caused by background could be addressed, resulting in more effective pixel-level feature alignment.

2.1.2. *Pixel-level adversarial learning for feature alignment*

After achieving foreground transferable information, pixel-level adversarial learning is used to further align local domain-invariant features through the gradient reversal layer (GRL) [6] and a pixel-level domain classifier. Specifically, a pixel domain classifier C_{pix} tries to distinguish which domain the foreground transferable attention feature F' comes from, while B', the shallow layers of the backbone, aim to confuse the classifier. In fact, C_{pix} and B' are connected by the GRL, which reverses the gradients that flow through B'. This module is trained in an adversarial learning manner. Formally, the loss function of pixel-level adversarial learning L_{pix} can be written as:

$$L_{pix_s} = \min_{\theta_{C_{pix}}} \max_{\theta_{B'}} \frac{1}{n_s HW} \sum_{i=1}^{n_s} \sum_{w=1}^{W} \sum_{H=1}^{H} C_{pix}(F'_{si})^2_{wh}, \tag{2}$$

$$L_{pix_t} = \min_{\theta_{C_{pix}}} \max_{\theta_{B'}} \frac{1}{n_t HW} \sum_{i=1}^{n_t} \sum_{w=1}^{W} \sum_{H=1}^{H} (1 - C_{pix}(F'_{ti})^2_{wh}), \tag{3}$$

$$L_{pix} = \frac{1}{2}(L_{pix_s} + L_{pix_t}), \tag{4}$$

where s and t respectively denote source and target domains and n represents the number of input images. F'_{ti} and F'_{si} are the i^{th} foreground transferable attention feature maps with the size of $H \times W$ from the target and source domain respectively. w and h are the coordinates on the above feature maps. During training, to obtain domain-invariant features, the network seeks the parameters $\theta_{B'}$ of the shallow layers of the backbone via maximizing the loss, while simultaneously seeking the parameters $\theta_{C_{pix}}$ of the domain classifier via minimizing the loss.

Combining adversarial learning with the spatial attention mechanism, the pixel-level subnet aligns the domain-invariant feature distributions of foreground regions that are more transferable for the detection task.

2.2. Image-level feature alignment

Due to large discrepancy at image-level features of different domains, we design an image-level feature alignment subnet composed of a GRL and image-level domain classifier C_{img}, after the backbone B. Unlike pixel-level domain classifier, the image-level domain classifier determines the domain of the whole feature map. Similar to the adversarial training in Sec.2.1, the whole backbone B tends to maximize the image-level loss and confuse image-level domain classifier. By this way, the ability to extract image-level domain-invariant features is obtained. To accomplish that, a jointly image-level loss function L_{img} with a focal loss [15] is defined as follows,

$$L_{img_s} = -\min_{\theta_{C_{img}}} \max_{\theta_B} \frac{1}{n_s} \sum_{i=1}^{n_s} (1 - C_{img}(F_{si})^{\gamma}) log(C_{img}(F_{si})), \tag{5}$$

$$L_{img_t} = -\min_{\theta_{C_{img}}} \max_{\theta_B} \frac{1}{n_t} \sum_{i=1}^{n_t} (C_{img}(F_{ti})^{\gamma}) log(1 - C_{img}(F_{ti})), \tag{6}$$

$$L_{img} = \frac{1}{2}(L_{img_s} + L_{img_t}), \tag{7}$$

where γ is the weight parameter focusing on hard samples. θ_B denotes the parameters of the whole backbone. F_{ti} and F_{si} are the i^{th} output feature maps of the backbone B from target and source domain respectively.

2.3. Instance-level feature alignment

Since supervisory signal is lacked on target domain, foreground object instances are normally represented by a bunch of inaccurate region proposals. To align source and target domain at instance-level, a prototype graph convolution network (PGCN) is conducted to guarantee the invariation of instance localization and classification in different domains. As shown in Fig.1., the PGCN is used to extract prototypes of each class from the embedding features learned from the RoIs by RPN [17] and bounding boxes by the classification and regression network. Following [23], an intra loss is introduced to minimize the distance between the prototypes of the same class in the source domain and the target domain, meanwhile, an inter loss is to increase the distance between different classes in two domains. The jointly instance-level loss L_{ins} with an intra loss and three inter losses are defined as

$$L_{ins} = \frac{1}{3}(L_{inter(s,s)} + L_{inter(t,t)} + L_{inter(s,t)}) + L_{intra(s,t)}, \tag{8}$$

where $L_{inter(s,s)}$ and $L_{inter(t,t)}$ are the inter loss from different classes in the same domain, while $L_{inter(s,t)}$ represents the inter loss from different classes in different domains. $L_{intra(s,t)}$ is the intra loss from the same class in different domains.

Through category prototype alignment, localization and classification of object instances tend to be domain-invariant.

2.4. Domain-consistent loss in three-level domain alignment

To optimize the HDCN in an end-to-end way, we propose a consistent regularization loss to constrain the domain adaptation in different levels. The domain-consistent loss is given by:

$$L_{cst} = \beta \left\| \frac{1}{uv} \sum_{u,v} p_{pix_{uv}} - p_{img} \right\|_2, \tag{9}$$

where Euclidean distance is used to measure the divergence between the prediction results in two levels. $p_{pix_{uv}}$ is the pixel-level classification probability at the pixel (u, v) of the feature map, and p_{img} is the classification probability of the whole feature map. β is an empirical parameter and set as 5 in this study. By minimizing L_{cst}, the prediction results of image-level and pixel-level domain tend to be the same, so as to make domain adaptation directions consistent.

Furthermore, to further constrain the instance-level training of domain adaptation, the total loss L_{Tol} with jointly three-level losses, the domain-consistent loss, and the detecting loss L_{det} of Faster RCNN is written as:

$$L_{Tol} = L_{pix} + L_{img} + L_{ins} + L_{cst} + L_{det}. \tag{10}$$

3. EXPERIMENTS

In this section, we provide comprehensive experimental results on two cross-domain detection tasks with distinct domain shift, including *Synthetic to Real* (SIM 10k [13] → Cityscapes [5]) and *Cross Camera Adaptation* (KITTI [7] → Cityscapes [5]).

3.1. Experimental setup

The experiment environment is a single Geforce GTX2080Ti of 64 bit Ubuntu operating system, which is implemented based on Pytorch framework. VGG16 [20] and ResNet50 [9] were used as the backbone in the synthetic-to-real task and the cross camera task, respectively. The SGD optimizer is selected for training, the initial learning rate is set to 0.001, and the learning rate decay rate and decay step are set to 0.1 and 5 respectively. Besides, we utilize the learning rate warm-up strategy during the first 200 steps when training. During training, bounding box labels only exist in source domain, while image domain labels exist in both two domains, guiding the domain alignment. For better comparison, we performed source-only evaluation with training only on the source datasets and testing on the target datasets, by using the Faster RCNN [17].

3.2. Synthetic to real task

In this experiment, SIM 10k [13] is employed as the source domain and Cityscapes [5] serves as the target domain. During training, we utilize the common car category with 10000 source samples and 2,975 target samples; for testing, we use the validation split of Cityscapes with 500 samples.

Table 1 compared our method with state-of-the-art methods, including Source-only [17], DA [4], SW-DA [19], MTOR [2], and GPA [23]. The AP of our method reached 51.6%, which outperformed all results of other methods. It demonstrates the advantage of the proposed three-level HDCN method.

Table 1. Experimental results of SIM 10k → Cityscapes

Methods	car AP(%)
Source-only [17]	34.6
DA[4]	41.9
SW-DA[19]	44.6
MTOR[2]	46.6
GPA[23]	47.6
HDCN(ours)	**51.6**

3.3. Cross camera adaptation

In this part, we study the adaptation between different camera settings. KITTI [7] dataset serves as source domain, and it contains 7,481 training images. Cityscapes [5] dataset is utilized as target domain, and its validation set is used for evaluation.

The shape and resolution of the images, as well as the weather, light, and other information in KITTI are markedly different from Cityscapes. The results of various methods on the common category car of the two datasets are presented in Table 2. The proposed method achieved the highest accuracy of 45.9%.

Table 2. Experimental results (%) of KITTI → Cityscapes

Methods	car AP
Source-only [17]	37.6
DA [4]	41.8
SW-DA [19]	43.2
SC-DA[24]	43.6
P-DA[11]	43.9
HDCN(ours)	**45.9**

3.4. Ablation analysis

Table 3 presents the ablation experiments on the task SIM 10k → Cityscapes. Here we used the GPA detector with ResNet50 as our experimental baseline, which employed the instance-level domain aligning method. The detection AP rises with the gradual addition of each domain alignment subnet (i.e., Ins, Img, and Pix in Table 3). Then, the AP improves by 1.1% after adding the consistency regularization loss. It is evident that the consistency loss tackles the training disorder caused by multi-level aligning method. Finally, thanks to adding the FAM module at pixel-level alignment, the detection result in target domain is significantly improved by 1.7%.

Table 3. Ablation study of the proposed HDCN. Note: Ins represents instance-level domain aligning, Img represents image-level domain alignment, Pix represents pixel-level domain aligning, Con represents consistency regularization loss, and FAM represents whether using pixel-based foreground attention module.

Methods	Ins	Img	Pix	Con	FAM	car AP
GPA	✓					47.6
ours	✓					45.5
	✓	✓				46.4
	✓	✓	✓			48.8
	✓	✓	✓	✓		49.9
	✓	✓	✓	✓	✓	51.6

To analyze the influence of γ, we evaluated our method with γ change on the task KITTI → Cityscapes. As shown in Fig.3, the model achieves the highest accuracy 45.9% when γ is 5.

3.5. Visualization

Fig.4 displays typical detection results on the task SIM 10k → Cityscapes, where green boxes represent correct positives while red boxes represent false positives. Obviously, GPA detects more wrong instances than ours. Furthermore, the localization of our model is more precise even when severe occlusion occurs.

Moreover, in Fig.5, we use t-SNE[16] to visualize the feature distributions of source and target domain on the task SIM 10k → Cityscapes, in which GPA and our method are employed for feature extraction. Due to domain-invariant feature learning, our method better confuses the feature distributions of different domains.

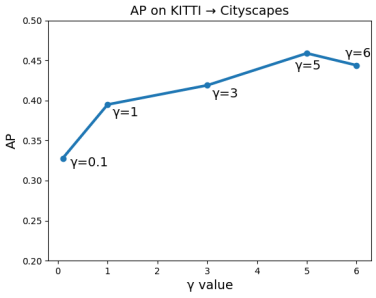
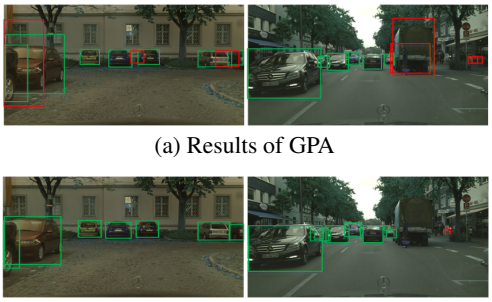


Fig. 3. Ablation study on the parameter γ.



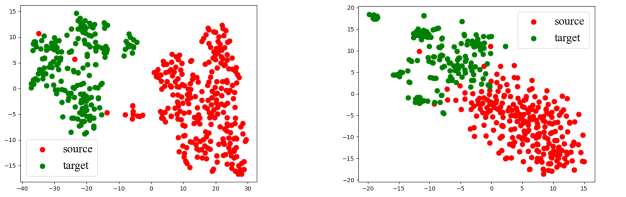
(a) Results of GPA

(b) Results of HDCN

Fig. 4. Detection results by different methods. (a) GPA, (b) HDCN

4. CONCLUSION

This paper proposes a hierarchical domain-consistent network for domain adaptation object detection, with three-level feature alignment subnets and a domain-consistent optimization machenism. Comprehensive experiments demonstrate the efficiency of the hierarchical adaptation method. Furthermore, pixel-based foreground attention and domain-consistent optimization have been proved beneficial for multi-level domain alignment in this study. In the future, a more effective and efficient multi-head self-attention method will be introduced into the method for better performance. To reduce the large computation we will consider to verify our method with lightweight detecting methods, for a better speed-accuracy trade-off.



(a) Features obtained by GPA (b) Features obtained by HDCN

Fig. 5. The comparison of different representations in 2D space by t-SNE feature visualization. (a) GPA, (b) our HDCN

5. REFERENCES

[1] V. F. Arruda, T. M. Paixão, R. F. Berriel, A. F. De Souza, C. Badue, N. Sebe, and T. Oliveira-Santos. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019.

[2] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 627–636, 2019.

[3] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8869–8878, 2020.

[4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc 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*, pages 3339–3348, 2018.

[5] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[6] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[7] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[8] Tiantong Guo, Cong Phuoc Huynh, and Mashhour Solh. Domain-adaptive pedestrian detection in thermal images. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1660–1664. IEEE, 2019.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6668–6677, 2019.

[11] 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 *The IEEE Winter Conference on Applications of Computer Vision*, pages 749–757, 2020.

[12] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018.

[13] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 746–753. IEEE, 2017.

[14] Wanyi Li, Fuyu Li, Yongkang Luo, and Peng Wang. Deep domain adaptive object detection: a survey. *arXiv preprint arXiv:2002.06797*, 2020.

[15] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[16] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[18] A. RoyChowdhury, P. Chakrabarty, A. Singh, S. Jin, H. Jiang, L. Cao, and E. Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 780–790, 2019.

[19] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[21] Tao Wang, Xiaopeng Zhang, Li Yuan, and Jiashi Feng. Few-shot adaptive faster r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7173–7182, 2019.

[22] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:5345–5352, 2019.

[23] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12355–12364, 2020.

[24] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin. Adapting object detectors via selective cross-domain alignment. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 687–696, 2019.