TARGET-AWARE AUTO-AUGMENTATION FOR UNSUPERVISED DOMAIN ADAPTIVE OBJECT DETECTION

Zhaoyang Li[1], Long Zhao[1], Weijie Chen[1,2], Shicai Yang[1], Di Xie[1], Shiliang Pu[1,†]

[1]Hikvision Research Institute, [2]Zhejiang University

ABSTRACT

Recent researches show that data auto-augmentation strategies can enhance the performance of object detection models. However, the existing works mainly focus on in-domain generalization. There is still a blank in out-of-domain generalization. In this paper, for the first time, we propose an auto-augmentation problem under unsupervised domain adaptation scenarios. To solve this problem, we propose a simple yet effective target-aware auto-augmentation technique to search for an optimal data augmentation strategy on labeled source data, so as to boost the detection ability on the given unlabeled target data. Our method can be easily plugged into the existing domain adaptation methods. Extensive experiments have been carried out to verify the effectiveness.

Index Terms— Data Auto-Augmentation, Unsupervised Domain Adaption, Object Detection

1. INTRODUCTION

Object detection aims to identify and localize certain categories of object instances in an image, which is a fundamental problem in computer vision. However, unless remarkably large amounts of labeled data are supported, object detection models with large capacity will usually suffer from overfitting. Data augmentation [1, 2] has been shown to be an effective regularization technique that can increase the quantity and the diversity of training data.

With the recent advancement of Automated Machine Learning (AutoML), some efforts exist to design an automated process of searching for augmentation strategies directly from a dataset. Although a series of auto augmentation methods [1, 2, 3, 4] have shown advanced performance in supervised learning, there are still some problems in how to apply auto-augmentation in unsupervised domain adaptive object detection (UDAOD) tasks [5, 6, 7, 8, 9, 10, 11]. When a labeled validation dataset is given, these methods usually utilize a controller to automatically search for data augmentation strategies. However, this limits their generalization ability when facing a new scenarios where the appearance of objects, backgrounds, and even weather condition are significantly different from the source domain training images. Meanwhile, due to the high cost of box annotations, it is not always feasible to obtain sufficient annotated validating images from the new scenarios. Namely, the new scenarios (target domain) cannot effectively provide labeled data for the detection model to verify and find the optimal data augmentation strategy. In this case, the existing search strategies are caught in a problem.

In this paper, we present a novel Target-Aware Auto-Augmentation (TA^3) approach for UDAOD. As illustrated in Figure 1, we divide UDAOD into a two-stage framework. The consideration behind our proposed solution is that we empirically found the bottleneck of UDAOD performance comes from the severe false negatives problem due to domain shift. To this end, we propose a target-aware auto-augmentation TA^3 method in the first stage to recall the true positives (or false negatives) as more as possible. Specifically, we propose a TArget Cumulative Scores (TACS) indicator on the unlabeled target data to guide data auto-augmentation search during source domain pre-training. However, TACS indicator is not an absolutely accurate performance evaluation metric compared with the ground-truth labels on the target data, which means it may introduce a slight part of false positives during this process. In this way, we still need to conduct UDAOD in the second stage for domain alignment and false positives suppression for further performance boost. We have to highlight that our TA^3 can be combined with the existing UDAOD methods in the second stage to obtain optimal object detector in a plug-and-play manner, which has been validated in the following extensive experiments.

2. METHODOLOGY

2.1. Problem Formulation

In the task of cross-domain object detection, we are given a labeled source domain $D_S = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, where x_i^s and $y_i^s = (b_i^s, c_i^s)$ denote the i-th image and its corresponding labels, i.e., the coordinates of the bounding box b and its associated category c respectively. In addition, we have access to an unlabeled target domain $D_T = \{x_i^t\}_{i=1}^{N_t}$. We assume that the source and target samples are drawn from different distributions (i.e., $D_S \neq D_T$) but the categories are exactly the same. The goal is to improve the detection performance

† Corresponding author.

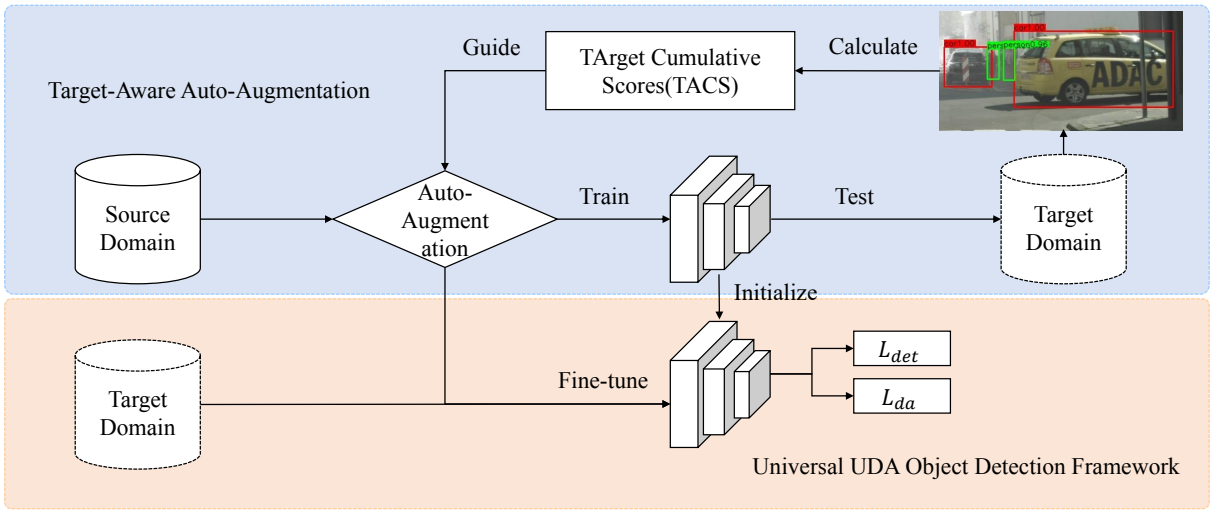

Fig. 1. A two-stage framework for UDAOD. In the first stage, we propose TArget Cumulative Scores (TACS) indicator on unlabeled target validation data to guide the data auto-augmentation search in the source data pre-training. In the second stage, the transformation strategy searched for source data and the corresponding pre-trained model are provided to combine with the existing UDAOD methods (*e.g.*, DA-Faster [5] or SW-Faster [6]) to obtain optimal object detector.

in D_T using the knowledge in D_S.

The existing auto-augmentation object detection methods are trained on the labeled source training data and tested on the labeled source validation data, which calculate the performance indicator to search the augmentation strategies:

$$\underset{augmentation}{\arg\max} \quad mAP(D_S) \tag{1}$$

where mAP denotes the mean average precision indicator. However, the domain shift happens in open conditions and the data distribution of validation set between source domain and target domain is inconsistent. Meanwhile, due to the high cost of box annotations, it is not always feasible to acquire sufficient annotated validating images from target domain. Therefore, we should hunt for a discriminative indicator F on the unlabeled target validation data to search for source data augmentation strategies so as to boost the detection ability on the given unlabeled target data:

$$\underset{augmentation}{\arg\max} \quad F(D_T) \tag{2}$$

2.2. Target-Aware Auto-Augmentation Approach

In this section, we propose TA^3 approach for UDAOD. Preliminarily, as shown in Table 1, we choose five transformations and the corresponding probability or magnitude interval (from the minimum values MIN to the maximum values MAX) to form the search space. Specifically, three to five transformations are randomly selected to form each sub-policy in the search space. We totally construct P different sub-policies. To reduce the search space further, following RandAugment [4], we postulate that a single global distortion

Transformations	Types	MIN	MAX
RandomGrayscale	probability	0.3	0.5
ColorJitter	magnitude	0.3	0.5
RandomHSV	probability	0.3	0.5
RandomNoise	magnitude	0.003	0.005
RandomScale	magnitude	0.3	0.5

Table 1. Data transformation types and strengths.

may suffice for parameterizing all transformations. Specifically, the parameter m controls the probability or magnitude values of each transformation in each sub-policy. The transformation strengths can be formulated as:

$$val = \frac{m}{M} \times (MAX - MIN) + MIN, m \in [0, M] \tag{3}$$

In our experiment, the value of m is an integer. So the transformation strength is contained to $M + 1$ discrete values. Therefore, the data augmentation search space is composed of $P * (M + 1)$ potential strategies, which greatly reduces the search space for data augmentation.

TA^3 can be applied to the training of any object detection model, and Faster-RCNN [12] is dropped here to illustrate. Faster-RCNN is a two-stage detector consisting of three major components: a feature extractor G, a region proposal network (RPN) and a region-of-interest based detection head (ROIHead). Generally, the training loss is composed of a RPN loss and a ROIHead loss:

$$L_{det} = L_{rpn} + L_{roi} \tag{4}$$

During UDAOD, the biggest problem with cross-domain detection is that lots of False Negatives (FN) exist with "Source only" model as shown in Figure 2. Our experiments

Methods	TA^3	truck	car	rider	person	train	motor	bicycle	bus	mAP
Source only	×	19.5	43.7	42.3	33.4	16.3	26.6	38.0	37.8	32.2
DA-Faster[5]	×	25.7	48.5	41.8	33.3	25.8	26.7	38.6	43.6	35.5
SW-Faster[6]	×	30.1	52.5	44.5	37.2	30.7	32.3	41.3	47.7	39.5
CR-DA-DET[8]	×	27.2	49.2	43.8	32.9	36.4	30.3	34.6	45.1	37.4
SFOD[10]	×	30.4	51.9	44.4	34.1	25.7	30.3	37.2	41.8	37.0
ART-PSA[9]	×	30.8	52.1	46.9	34.0	29.9	34.7	37.4	43.2	38.6
MeGA-CDA[11]	×	25.4	52.4	49.0	37.7	46.9	34.5	39.0	49.2	41.8
Oracle	×	34.9	58.9	48.8	39.2	44.3	37.8	43.4	55.8	45.4
Source only	✓	31.6	52.3	48.0	38.1	36.5	36.3	43.3	47.5	41.7
DA-Faster	✓	31.4	52.8	**50.2**	39.8	**42.5**	37.1	**44.6**	48.9	**43.4**
SW-Faster	✓	**32.5**	51.9	**50.2**	39.1	37.7	**38.0**	43.7	**51.5**	43.1

Table 2. Results of adaptation from Cityscapes to Foggy Cityscapes.

empirically found that a model with better detection performance can obtain more high-confidence detection boxes in the target domain validation set. Therefore, we use Faster-RCNN model trained on the augmented source data to obtain all detection boxes with confidence information in the target domain validation set, and calculate the discrete integral of confidence distribution on the number of detection boxes to obtain the TACS indicator. The calculation of TACS indicator is as follows:

$$TACS = \sum_{b}^{N} p_b \tag{5}$$

where N is the number of detection boxes obtained by the Faster-RCNN model on the target domain validation set (after post-processed by NMS), and p_b is the classification confidence of the detection box b.

As shown in Figure 1, TA^3 samples a data augmentation strategy in the search space, and then employs the source data equipped with the augmentation strategy to train Faster-RCNN. After training, the TACS indicator of the target validation dataset is used to select the augmentation strategy:

$$\underset{augmentation}{\arg\max} \quad TACS(D_T) \tag{6}$$

To alleviate the fluctuations in the training process, we apply an exponential moving average (EMA) to smooth the model parameters in training and exploit the final smoothed model to calculate the TACS indicator. The larger TACS score indicates recalling the more true positives. However, it is inevitable to induce a part of false positives. Under this consideration, we have to highlight that our TA^3 can be combined with the existing UDAOD methods to obtain optimal object detector via domain alignment and false positives suppression. For example, we summarize a universal form of the existing UDAOD methods (*e.g.*, DA-Faster [5] or SW-Faster [6]), in which the total loss function is formulated as:

$$L = L_{det} + \lambda L_{da} \tag{7}$$

L_{da} represents an unsupervised domain alignment loss, combined with a supervised object detection loss L_{det} on source data, where λ is a balance parameter between L_{det} and L_{da}. TA^3 is utilized to automatically search for the optimal data augmentation strategy for source data to train the detection model with L_{det} and adapt the target data with L_{det}.

Methods	TA^3	K2C	S2C
Source only	×	41.5	38.3
DA-Faster[5]	×	43.1	42.3
SW-Faster[6]	×	43.4	44.9
SFOD[10]	×	44.6	43.1
ART-PSA[9]	×	–	43.8
MeGA-CDA[11]	×	43.0	44.8
Oracle	×	61.0	61.0
Source only	✓	48.2	49.6
DA-Faster	✓	**49.1**	49.9
SW-Faster	✓	48.9	**50.1**

Table 3. K2C and S2C denote the results of adaptation from KITTI to Cityscapes and Sim10k to Cityscapes, respectively.

3. EXPERIMENTS

3.1. Empirical Setup

Datasets. Five public datasets are utilized in our experiments, including Cityscapes[13], Foggy Cityscapes[14], BDD100k[15], Sim10k[16], and KITTI[17].

Baselines and Comparison Methods. We consider the Source only trained Faster-RCNN (without adaptation) as our baseline methods. In addition, we introduce DA-Faster[5], SW-Faster[6], CR-DA-DET[8], ART-PSA[9], SFOD[10] and MeGA-CDA[11] for the comparison.

Implementation Details. Following the default settings in[5, 6], all training and test images are resized such that the shorter side has a length of 600 pixels. We use the pre-trained weights of VGG-16[18] on ImageNet[19] as the backbone of the Faster-RCNN framework. The detector is trained using Stochastic Gradient Descent (SGD) with the momentum of 0.9 and the weight decay of 5×10^{-4}. We fine-tune the network with a learning rate of 10^{-3} for 50k iterations and then

Methods	TA^3	truck	car	rider	person	motor	bicycle	bus	mAP
Source only	×	22.5	50.1	28.5	36.2	23.1	23.2	22.2	29.4
DA-Faster[5]	×	21.8	50.6	31.9	35.9	22.9	24.5	23.0	30.3
SW-Faster[6]	×	20.6	51.0	32.5	37.6	19.8	28.4	25.3	30.8
CR-DA-DET[8]	×	19.5	46.3	31.3	31.4	17.3	23.8	18.9	26.9
SFOD[10]	×	20.6	50.4	32.6	32.4	18.9	25.0	23.4	29.0
Oracle	×	58.5	62.1	47.7	51.3	41.1	42.6	58.5	51.6
Source only	✓	**30.7**	55.0	40.3	41.9	28.8	32.1	22.8	35.9
DA-Faster	✓	30.2	**57.0**	40.2	**43.6**	**31.2**	31.5	23.9	36.8
SW-Faster	✓	29.5	55.6	**40.5**	43.0	28.7	**33.5**	**27.5**	**36.9**

Table 4. Results of adaptation from Cityscapes to BDD100k.

a) Ground Truth b) Source only c) SW-Faster d) Source only+TA^3

Fig. 2. Top: Visualization from Cityscapes to Foggy Cityscapes. Bottom: Visualization from Sim10k to Cityscapes.

reduce the learning rate to 10^{-4} for another 20k iterations.

3.2. Results Comparison

Adaptation from KITTI to Cityscapes. In this part, the KITTI and Cityscapes datasets are used as source and target domains, respectively. In Table 3, the proposed method TA^3 reaches 48.2% mAP with a gain of +6.7% over the Source only model.

Adaptation from Sim10k to Cityscapes. Synthetic images offer an alternative to alleviate the data annotation problem. However, there is a distribution gap between synthetic data and real data. To adapt the synthetic scenes to the real one, we utilize the entire Sim10k dataset as the source domain and the training set of Cityscapes as the target domain. The results are displayed in Table 3. We obtain 49.6% mAP. This shows that the TACS indicator can guide the auto-augmentation process. Moreover, compared to DA-Faster and SW-Faster, DA-Faster+TA^3 and SW-Faster+TA^3 have also achieved improvements. The data augmentation strategy and pre-training model of our approach are instrumental to these adaptation methods.

Adaptation from Cityscapes to Foggy Cityscapes. To study the changing environment adaptation from normal weather to a foggy condition, Cityscapes and Foggy Cityscapes are used as the source domain and the target domain, respectively. In Table 2, our method TA^3 improves baseline (source-only) by 9.5% and performs the best compared to other methods in mAP.

Adaptation from Cityscapes to BDD100k. Finally, we use Cityscapes and BDD100k as source and target domain dataset, respectively. Similarly, in Table 4, our method significantly improves the detection results compared to other adaptation methods.

Visualization. In Figure 2, we show some detection examples from two target datasets, *i.e.*, Foggy Cityscapes [14] and Cityscapes [13]. Compared to Source Only and SW-Faster, our proposed method produces more accurate detection results under complex environments and large domain shifts. From (b), (c) to (d), we can observe that the proposed method increases true positives, which is consistent with the previous analysis.

4. CONCLUSION

In this work, we present a Target-Aware Auto-Augmentation framework for Unsupervised Domain Adaptive Object Detection to improve the detection performance. Specifically, we propose TArget Cumulative Scores (TACS) indicator on unlabeled target validation data to guide the data auto-augmentation search in source data training. Experiments and visualization can validate the effectiveness of our method.

5. REFERENCES

[1] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 113–123. 1

[2] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim, "Fast autoaugment," *arXiv preprint arXiv:1905.00397*, 2019. 1

[3] Daniel Ho, Eric Liang, Xi Chen, Ion Stoica, and Pieter Abbeel, "Population based augmentation: Efficient learning of augmentation policy schedules," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2731–2741. 1

[4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 702–703. 1, 2

[5] 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*, 2018, pp. 3339–3348. 1, 2, 3, 4

[6] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6956–6965. 1, 2, 3, 4

[7] W. Chen, Y. Guo, S. Yang, Z. Li, and Y. Zhuang, "Box re-ranking: Unsupervised false positive suppression for domain adaptive pedestrian detection," 2021. 1

[8] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei, "Exploring categorical regularization for domain adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11724–11733. 1, 3, 4

[9] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang, "Cross-domain object detection through coarse-to-fine feature adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13766–13775. 1, 3

[10] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang, "A free lunch for unsupervised domain adaptive object detection without source data," *arXiv preprint arXiv:2012.05400*, 2020. 1, 3, 4

[11] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4516–4526. 1, 3

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015. 2

[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," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223. 3, 4

[14] 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. 3, 4

[15] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell, "Bdd100k: A diverse driving video database with scalable annotation tooling," *arXiv preprint arXiv:1805.04687*, vol. 2, no. 5, pp. 6, 2018. 3

[16] 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?," *arXiv preprint arXiv:1610.01983*, 2016. 3

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361. 3

[18] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 3

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. 3