Received: 25 January 2021 | Revised: 7 March 2021 | Accepted: 5 April 2021

DOI: 10.1002/int.22434

RESEARCH ARTICLE

WILEY

Source data-free domain adaptation of object detector through domain-specific perturbation

Lin Xiong[1] | **Mao Ye**[1] | **Dan Zhang**[1] | **Yan Gan**[2] | **Xue Li**[3] | **Yingying Zhu**[4]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, People's Republic of China

[2]College of Computer Science, Chongqing University, Chongqing, People's Republic of China

[3]School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Queensland, Australia

[4]College of Engineering, University of Texas Arlington, Arlington, Texas, USA

Correspondence
Mao Ye, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Western High-Tech Industrial Zone, Chengdu 611731, P.R. China.
Email: cvlab.uestc@gmail.com

Funding information
National Key R&D Program of China, Grant/Award Number: 2018YFE0203900; National Natural Science Foundation of China, Grant/Award Number: 61773093; Important Science and Technology Innovation Projects in Chengdu, Grant/Award Number: 2018-YF08-00039-GX; Key R&D Programs of Sichuan Science and Technology Department, Grant/Award Number: 2020YFG0476; Postdoctoral Research Foundation of China, Grant/Award Number: 2020M683243

Abstract

The current unsupervised cross-domain detection methods need source domain data to retrain the detection model in target domain. However, the source domain data may be unavailable due to privacy, decentralization, or computation resource restrictions. A natural idea is to optimize the parameters of the source domain model by self-supervised learning based on pseudo labels. We propose another approach from the viewpoint of noise perturbation without pseudo-labeling. It can be assumed that the source and target domains are actually derived from a domain invariant space through domain-specific perturbations, respectively. A super target domain can be constructed by augmenting more target domain perturbations to the target domain images. The optimal direction of the target domain to the domain invariant space can be approximated as the alignment direction from the super target domain to the target domain. Based on this

idea, we propose a novel method called SOAP (SOurce data-free domain Adaptation through domain Perturbation) which can remove domain perturbation from the target domain. The image-level, instance-level, and category consistency regularizations based on Mean Teacher structure are proposed to learn the correct alignment direction. Specifically, the category consistency can also further improve the classification accuracy. Extensive experiments on multiple domain adaptation scenarios demonstrate that SOAP achieves better performance surpassing the baseline (Faster R-CNN) and multiple state-of-the-art domain adaptation methods which need to access source domain data.

KEYWORDS
domain adaptation, object detection, perturbation, transfer learning

1 | INTRODUCTION

Object detection problem aims to locate objects in a given image and gives the labels of these objects simultaneously. There are many solutions to object detection problem,[1–10] for example, the representative one-stage method YOLO[11] and the two-stage method Faster R-CNN.[12] Although these methods obtain good performance, they assume that the features from the source and target domains obey the same distribution. But domain shift often exists between the source and target domains. When the source domain model is applied to the target domain, the performance always drops significantly. To solve the domain shift problem, many domain adaptation methods are proposed. According to whether the target domain has some labels, the domain adaptation problem is divided into supervised,[13] weakly-supervised,[14–17] unsupervised,[18–22] and one-shot.[23,24] Since it is impractical to access source domain data, for example, due to privacy preservation, decentralization, or computing resource limitations, we focus on source data-free domain adaption with only source model and unlabeled target domain data available. Our problem scenario is that there are only unlabeled target domain and no source domain data.

Only few recent works studied source data-free domain adaptation in the following three categories: The first one adjusts the source model by self-supervised learning based on pseudo labels[25–27]; the second one is to generate some target domain style samples which can be correctly classified by the source model[28]; and the final is to transform the target domain samples to the source domain style samples.[29] First of all, without source domain data, it is not easy to do satisfied image translation between the target and source domains. Second, since object detection problem is a highly unbalanced classification problem, that is, lots of negative samples and inaccurate segmented objects, the high quality of pseudo labels is the key to success. These problems prompt us to find another way for source data-free domain adaptive object detection without pseudo-labeling.

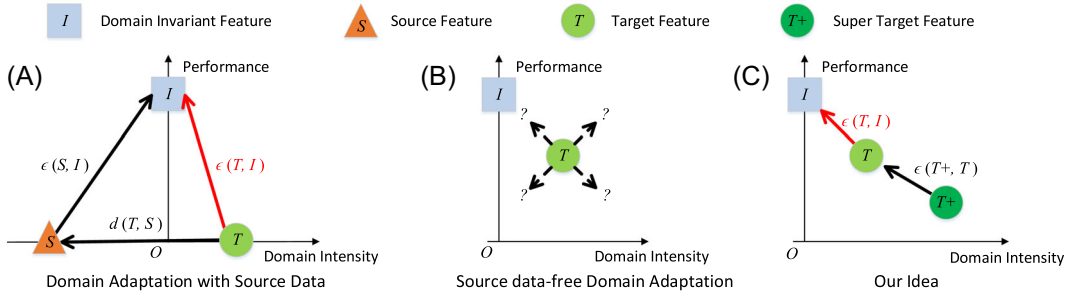

FIGURE 1 Analysis of domain adaptation problem. The horizontal and vertical axis represent domain intensity and detection performance respectively. At the coordinate origin, the domain intensity is zero which means the feature has not any domain information. The model trained on this kind feature has strong generalization ability and better performance. (A) Domain adaptation with source data. The way from target domain feature (T) to domain invariant feature (I) is formed by combing the directions from target domain feature (T) to source domain feature (S) and from source domain feature to domain invariant feature. (B) Source data-free domain adaptation. The difficulty is how to find an optimization direction. (C) Our idea. The super target domain is constructed to estimate the correct direction from target domain feature to domain invariant feature [Color figure can be viewed at wileyonlinelibrary.com]

We found that assuming source and target domains are obtained by perturbing the domain invariant space with domain-specific perturbation make it feasible in following perspectives. As Hoffman et al.[30] stated, the lower layer of CNN can extract domain invariant features, because the lower layer is domain-specific and can be directly used for domain agnostic higher layer networks trained in different domains. So the feature extraction network of the pretrained source domain model is actually trying to extract domain invariant features from source domain. And its classification network is also expected to be trained on domain invariant features. As shown in Figure 1, if we can map the target domain feature to the domain invariant feature and adjust the pretrained classification network correspondingly to deal with the slight source bias in actual training, then the source domain model can be applied to the target domain. Motivated by this hypothesis, a new domain called super target domain is constructed by perturbing the target domain with the target domain-specific perturbation. The alignment direction along the super target domain to the target domain is actually the direction from the target domain to the domain invariant space.

If the domain shift is large, it is a big challenge to get a satisfactory super target domain which can help to deeply adjust the source feature extraction network such that it can map the target domain feature to the domain invariant feature. While if the domain shift is not large, it is possible.[29] For this situation, domain-specific perturbations for the source and target domains are similar. The source feature extraction network only needs to be slightly modified to fit the target domain. Under this assumption, we first obtain the domain perturbation by averaging multiple target domain images, and then further augment the domain perturbation to the target domain images to obtain the super target domain. Although the construction is simple, we will show in the experiment that such super target domain also works when the domain shift is large. Different from the general data augment method, we generate super target domain images along more target domain perturbation direction to find the optimization direction from the target domain to the domain invariant space, as shown in Figure 1C.

Based on the above theory, we propose a new SOurce data-free domain Adaptation method through domain Perturbation (SOAP). The Mean Teacher[31] structure is used. The same backbone

of teacher and student models is from Faster R-CNN.[12] The super target domain and the corresponding target domain are input to the student and teacher models respectively. There are three consistency regularizations: image-level alignment, instance-level alignment, and category consistency. By these regularizations, the super target domain is aligned to the target domain. As the analysis in the previous paragraph, we actually line up the target domain feature to the domain invariant feature by aligning the super target domain to the target domain. At the same time, the category consistency also improves classification accuracy. Our method solves the domain adaptation problem by eliminating domain perturbation, which is basically irrelevant to the category. Compared with previous source data-free approaches which heavily depend on the quality of pseudo labels, our method treats each object category equally and does not use the pseudo labels. So the unbalanced problem of object detection is avoided in some sense implicitly.

Our contributions can be summarized as follows:

(1) We proposed a new approach for source data-free domain adaptive object detection by eliminating domain perturbation. The super target domain is constructed by adding more target domain perturbation, and then return to the target domain to get the direction to the domain invariant space. Of course, our approach is not contradictory to the traditional pseudo-labeling method, and they can be combined to achieve better results. As it is beyond the scope of this paper, we will not discuss it further.
(2) We proposed three consistency regularizations on aligning the super target domain to the target domain and optimized them using Mean Teacher model. The learned alignments from the super target domain to the target domain is further used to align the target domain to the domain invariant space.
(3) We designed an approximation method based on the Law of Large Numbers to obtain the domain perturbation, thereby constructing the super target domain. Although the construction is simple, experiment results on four domain adaption scenarios prove such super target domain assisted domain alignment idea works.

2 | RELATED WORKS

2.1 | Domain adaptation of object detection

Unsupervised domain adaptation of object detection can be divided into three categories: (1) feature alignment; (2) data augmentation, and (3) semi-supervised learning.

Feature alignment solves domain adaptation problem by eliminating domain shift between the source and target domains, which is currently the most popular method. For example, the earlier work DAF[32] eliminates domain shift by image-level and instance-level alignments based on H-divergence.[33] Many recent works[34-42] proposed different alignment methods to make the two domains more indistinguishable.

Data augmentation is also proposed from the perspective of eliminating domain shift between the source and target domains. The general approach is to transform the source domain image into the target domain style or use third-party domain to augment data. Such methods[16,43,44] usually combine the idea of feature alignment.

The final approach regards domain adaptation as a semi-supervised problem. A typical method is based on Mean Teacher model.[31] MTOR[45] proposes multiple consistency regulations to solve cross-domain detection problem. Then, a robust learning-based method[46] is proposed

by generating better target domain pseudo-labels. Recently, UMT[47] pays attention to the domain shift, and combines the ideas of feature alignment and data augmentation. These methods have achieved good results, but all of these methods need to access source domain data.

2.2 | Source data-free domain adaptation

The current research on source data-free domain adaptation focus on object classification problem. There is not much work on this problem yet. SHOT[25] proposes a clustering method to generate pseudo-labels of the target domain for adjusting the source domain model. At the same time, information maximization principle is used such that the correct classification probability in the target domain should be close to one-hot encodings. The solution to the source data-free problem of USF[26] is also minimizing the entropy similar to SHOT. Further, SFDA uses prototypes and filtering mechanism to improve the reliability of the pseudo labels.[27] The high quality of pseudo-labeling is the key for this kind of approach.

Different from the above methods, 3C-GAN[28] first generates images of each category as the style of target domain to adjust the source network and then also minimizes entropy. DAAS[29] trains a transformation network from the target domain to the source domain to make the entropy of the transformed target domain input to the classifier as small as possible than before. Because source domain data are not accessible, it is not easy to find a satisfactory mapping between the target domain and the source domain.

3 | PRELIMINARIES

3.1 | Problem definition

Suppose $D_s = \{(X_S, Y_S)\}$ where Y_S denotes object bounding box annotations for source domain sample X_S and $D_T = \{X_T\}$ for target domain sample without any annotations. For source data-free cross detection problem, when the pretrained source detector $f_S: X_S \rightarrow Y_S$ is applied to the target domain, the source data set D_S is not accessible. Given the source detector f_S and D_T, the task is to find a mapping $f_T: X_T \rightarrow Y_T$ where Y_T denotes object bounding box annotations for target domain sample X_T that can work well in the target domain.

3.2 | Mean teacher

Mean teacher[31] is first proposed for semi-supervised learning. It includes a student model M_{st} and a teacher model M_{te}. The labeled data X_L is input to the student model for standard supervised training. Thus, the cross-entropy loss L_{std} can be calculated. Then, two perturbed samples X_U^{st} and X_U^{te} from the same unlabeled data X_U are input into the student and teacher models respectively. Since X_U^{st} and X_U^{te} contain the same category, their predictions should be consistent. The consistency loss is defined as follows:

$$L_{cons} = \left\| M_{st}\left(X_U^{st}\right) - M_{te}\left(X_U^{te}\right) \right\|_2^2.$$

The student model is trained by minimizing the total loss $L_{MT} = L_{std} + L_{cons}$. The parameters of teacher model are updated from student model. Specifically, for the i-th iteration, $\theta_{te}^{(i)} = \alpha\theta_{te}^{(i-1)} + (1-\alpha)\theta_{st}^{(i)}$, where θ_{te} and θ_{st} are the parameters of teacher model and student model respectively. The parameter α is used to control the update of teacher model. Through multiple iterations, the parameters of teacher model are the exponential moving average of student model. Finally, the trained teacher model is used for test.

In the semi-supervised problem, the training of the model is not always optimized in a good direction, so the teacher model is proposed to save the historical parameters of student model training process. By ensuring the consistency of the unlabeled images in the two models, the student model can correct the previous learning direction. Recently, Mean Teacher model is applied to the problem of cross-domain detection.[45,47] The available reason is that it can be regarded as a special semi-supervised problem.

4 | PROBLEM ANALYSIS

4.1 | Domain alignment

We denote the source domain and target domain features as S and T respectively. If there exist ideal domain invariant features I that can be accurately classified, then the object detection problem is to train a network that can project the input image feature to I as much as possible. We abstractly define the distance between the input image feature and I as $\overrightarrow{\epsilon}$. The ultimate goal of solving domain adaptation problem is to make the target domain features close to I. Since the target domain has no label, we cannot find a way from T to I directly. So for the cross-domain detection problem with source domain data, many methods convert $\overrightarrow{\epsilon}(T, I)$ to $\overrightarrow{\epsilon}(S, I)$ and divergence loss $\overrightarrow{d}(T, S)$, that is, $\overrightarrow{\epsilon}(T, I) = \overrightarrow{\epsilon}(S, I) + \overrightarrow{d}(T, S)$. For the labeled source domain, $\overrightarrow{\epsilon}(S, I)$ can be minimized through supervised training. And the divergence loss between source and target domains can be minimized by traditional feature alignment method. This way can be simplified shown in Figure 1A.

Source data-free domain adaptation problem is shown in Figure 1B. If the source domain data and the labels of target domain do not exist, the technical route in Figure 1A does not work. Fortunately, we can estimate the optimization direction of $\overrightarrow{\epsilon}(T, I)$ by constructing a new domain called super target domain by augmenting domain-specific perturbation to the target domain. We denote the super target domain as $T+$, which corresponds to T one by one. From the perspective of domain perturbation, we can assume both S and T are the perturbations of I. The super target domain is transformed from the corresponding target domain, but has a stronger target domain style than the target domain (i.e., if the target domain is foggy, the super target domain has more fog) due to more domain-specific perturbation. As shown in Figure 1C, the optimization direction from T to I can be approximated by the alignment direction from $T+$ to T.

Remark : The above analysis shows that if we can construct a satisfactory super target domain $T+$ such that the optimization directions of $\overrightarrow{\epsilon}(T, I)$ and $\overrightarrow{\epsilon}(T + , T)$ are the same, by aligning the super target domain to the target domain, the target domain feature can be aligned to domain invariant feature even without source domain data. However, as we mentioned in Section 1, if the domain shift is very large, it is still an open question to construct a correct super target domain.

4.2 | Super target domain generation

We assume that the images of the source and target domains are obtained by their domain-specific perturbations on the domain invariant space, we have the following equation:

$$X_T^i = X_I^i + w_T^i N_T, \tag{1}$$

for ith sample, where N_T is domain-specific perturbation. The corresponding ith pure sample in the domain invariant space and the weight on domain-specific perturbation are denoted as X_I^i and w_T^i, respectively.

Since the domain perturbation is the same for all images in the same domain, the estimation of domain-specific perturbation can be obtained by averaging target domain images. $\hat{N_T} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_T^i$. By Equation (1), we can get the following equation for target domain,

$$\hat{N_T} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_I^i + \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} w_T^i N_T,$$

where X_I^i is a list of independent and identically distributed variables. Then according to Wiener–khinchin Law of Large Numbers, we have

$$\forall\, \varepsilon > 0,\ \lim_{n\to\infty} P\left\{ \left| \frac{1}{n} \sum_{i}^{n} X_I^i - E\left(X_I^i\right) \right| < \varepsilon \right\} = 1.$$

For the normalized input images (removing mean value), the assumption of $E(X_I^i) = 0$ is suitable. Therefore the domain-specific perturbation can be estimated by the average of the target samples,

$$\hat{N_T} = \lim_{n\to\infty} \frac{1}{n} \sum_{i}^{n} w_T^i N_T.$$

We add domain-specific perturbation to target domain as follows to obtain super target domain,

$$X_{T+}^i = \beta X_T^i + (1-\beta)\hat{N_T},$$

where β is used to control the portion of pure image. Although some super target domain looks very similar to the corresponding target domain, the distributions are different. It helps us find the direction of optimization.

Remark : The linear addition of domain-specific perturbation in Equation (1) implicitly assume that the shift between the source and target domains is not large. For the transfer scenario satisfying this condition, the experimental results are very good. Even so, it is not a big deal for the transfer scenario which does not meet this condition. Experiments show that our method SOAP also exceeds the benchmark method (Faster R-CNN).

In the practical implementation, the target domain images are input into the network in batches, so the domain perturbation is updated iteratively. Specifically, we first construct the local domain perturbation *LN* of the current batch, which is the average of the

images in the current batch. Then the jth global domain perturbation is updated as $N^{(j)} = (N^{(j-1)} + LN^{(j)})/2$, where $N^{(j-1)}$ is the previous global domain perturbation and $LN^{(j)}$ is the jth local domain perturbation. Due to the limitation of the number of images, the estimated domain perturbation still has random noise, which is not ideally completely clean. We do not intend to remove this random noise, because the adopted Mean Teacher structure in the next section can effectively use the random noise to increase the difficulty of training the student model, thereby enhancing the model stability. To show the effectiveness of our approach, we do not use a very complex method to construct the super target domain, such as GANs; and we also do not combine the self-supervised learning method based on pseudo-labeling, because the super target domain constructed by a simple linear method can achieve good results.

5 | THE PROPOSED METHOD

Based on the above analysis, it is natural to use Mean Teacher model to increase the stability in finishing our task, because the optimization process of Mean Teacher fits our idea very well. As shown in Figure 2, the backbone of the teacher and student models is the source Faster R-CNN model. The teacher and student models are in charge of the alignments from target domain to domain invariant feature and super target domain to target domain, respectively. Our method consists of four parts: (1) image-level alignment; (2) instance-level alignment; (3) category consistency; and (4) optimization process. The pretrained source model (Faster R-CNN) is first loaded for both student and teacher model, and super target domain is constructed. Then we

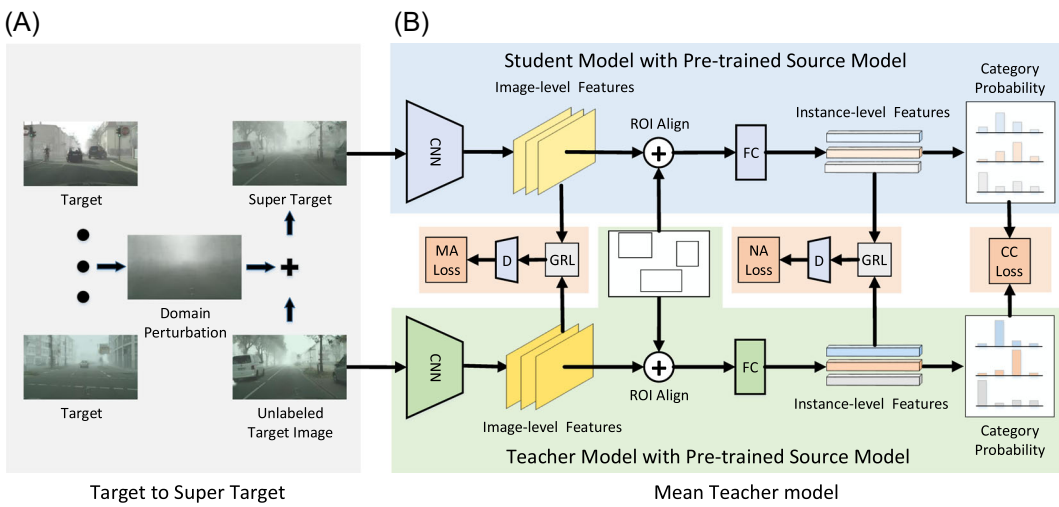

FIGURE 2 An overview of proposed method SOAP. (A) Target to Super Target. The super target domain images are constructed one by one from the corresponding target domain images. (B) Mean Teacher model. The backbone is the same Faster R-CNN model. The student model uses the region proposals of teacher model. Three consistencies are added, that is, iMage-level Alignment (MA), iNstance-level Alignment (NA), and Category Consistency (CC). The image and instance alignments are achieved by Gradient Reversal Layer (GRL).[48] The category consistency means the predictions from a target image and its corresponding super target image are the same [Color figure can be viewed at wileyonlinelibrary.com]

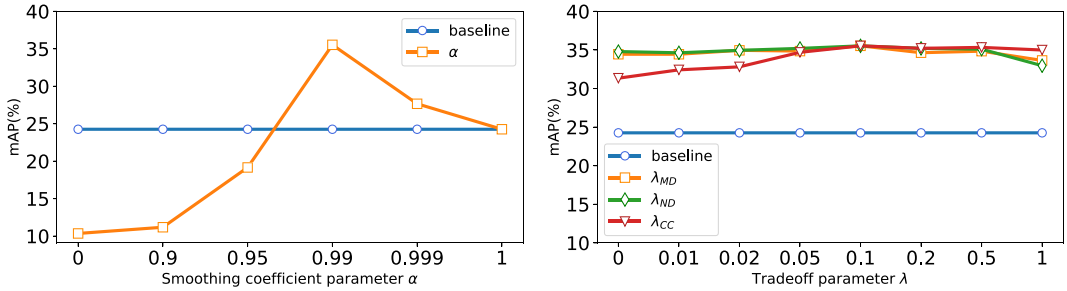

FIGURE 3 Parameter analysis of α, λ_{MA}, λ_{NA}, λ_{CC}. Our method is greatly affected by the smoothing coefficient parameter α but is not sensitive to the changes of the trade-off parameters λ_{MA}, λ_{NA}, λ_{CC} [Color figure can be viewed at wileyonlinelibrary.com]

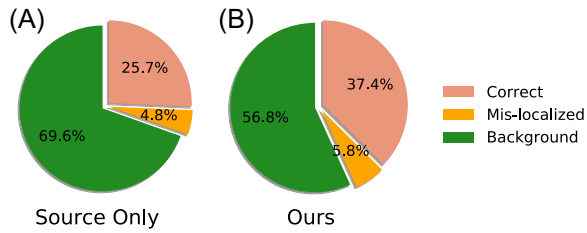

FIGURE 4 Error analysis of highest confident detections. Compared with baseline, our method can reduce background detections, thus detecting more correct objects. (A) Source Only and (B) Ours [Color figure can be viewed at wileyonlinelibrary.com]

feed the corresponding target and super target domain images to the teacher and student models, respectively. The student model shares region proposals from the teacher model. The student model is trained by three consistencies: iMage-level Alignment (MA), iNstance-level Alignment (NA), and Category Consistency (CC). The Gradient Reversal Layer (GRL)[48] is used for image-level and instance-level alignment. Category consistency minimizes the prediction loss between the student and teacher models. Finally, the student model is updated through backpropagation. Then the teacher model is updated according to the student model.

5.1 | Image-level alignment

To align the image-level features of the super target and target domains, the idea of GANs[49] is used. We construct an image-level domain discriminator to make the features from two domains indistinguishable. To reverse the influence of the discriminator on the feature, a GRL in front of the discriminator is used to reverse the gradient. The image-level loss of the ith image is the following,

$$L_{MA} = -\sum_{i,h,w}\left[D^i \log p^i_{h,w} + (1 - D^i)\log\left(1 - p^i_{h,w}\right)\right],$$

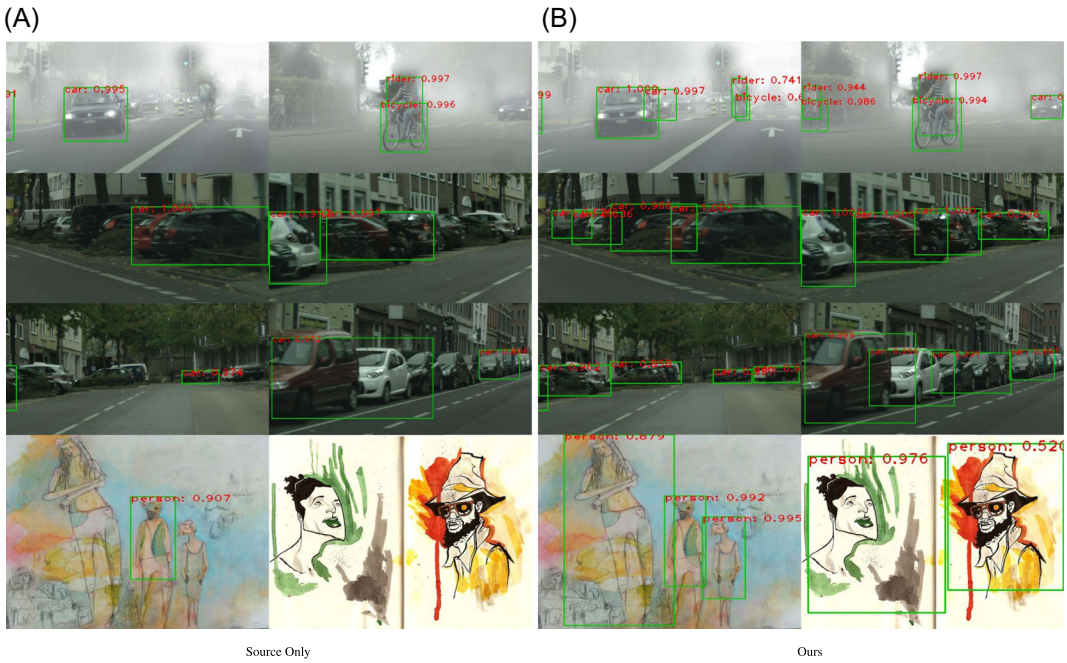

FIGURE 5 Detection visualization. The left two columns show the detection results of the source domain model, and the right two columns are the detection results of our transferred model. The four rows of images from top to bottom are taken from the four domain adaptation scenarios of *Cityscapes to Foggy Cityscapes*, *KITTI to Cityscapes*, *SIM10k to Cityscapes*, and *Pascal VOC to Watercolor2k*. (A) Source Only and (B) Ours [Color figure can be viewed at wileyonlinelibrary.com]

where h and w represent the position in the feature map, and p is the probability of belonging to the target domain predicted by the discriminator. D is the domain label, for the super target domain $D = 0$, and for the target domain $D = 1$.

5.2 | Instance-level alignment

Similar to the principle of image-level alignment, we construct an instance-level discriminator to align the instance-level features of the super target and target domains. The instance-level loss of the ith image is the following:

$$L_{NA} = -\sum_{i,r}\Big[D^i \log p_r^i + (1 - D^i)\log\Big(1 - p_r^i\Big)\Big],$$

where r represents the rth region proposal. Because super target domain always has more domain perturbation, the training of student model tries to filter out the extra domain perturbation. Since the teacher model will be updated according to the student model later, it also gradually filters the domain perturbation. Thus, the training process will force the features

from the target and super target domains moving to the domain invariant feature space together.

5.3 | Category consistency

The target and the corresponding super target images contain exactly the same objects at the same locations, but it is more difficult to detect objects in super target images. So we use the predictions in target images to supervise the predictions from the worse super target images. The category consistency loss of each pair of target domain and super target domain is defined as a cross-entropy loss as follows:

$$L_{CC} = -\frac{1}{N_r}\sum_{r,c} p_{r,c}\log\left(p^{+}_{r,c}\right),$$

where r and c represent the rth region proposal and the cth category respectively, and N_r is the number of region proposals. p and p^{+} are the prediction results of the target domain and the super target domain, respectively.

The student model learns from teacher model for better classification in the more "worse" super target domain; the classification model will be better for generation. Then, the teacher model is updated according to this student model, and the classification model of teacher model is also improved. Teaching benefits teachers as well as students.

5.4 | Optimization process

The parameters of student model are updated by minimizing the loss L, which is

$$L = \lambda_{MA}L_{MA} + \lambda_{NA}L_{NA} + \lambda_{CC}L_{CC},$$

where $\lambda_{MA}, \lambda_{NA}$ and λ_{CC} are trade-off parameters.

The teacher model is learned along the optimization direction of student model. The parameters of teacher model at current iteration are weighted sum of the values of previous iteration and the values of current iteration of the student model,

$$\theta^{(i)}_{te} = \alpha\theta^{(i-1)}_{te} + (1-\alpha)\theta^{(i)}_{st},$$

where θ_{te} are the parameters of the teacher model, θ_{st} are the parameters of the student model, and α is the coefficient used to control the update of the teacher model.

6 | EXPERIMENTS

We conduct evaluations on four domain adaptation scenarios including weather, cross-camera, synthetic-to-real, and real-to-artistic transfers to verify the effectiveness of the proposed method. Our method SOAP is also compared with multiple domain adaptation methods which use source domain data.

6.1 | Experimental settings

6.1.1 | Data sets

For the data sets used as target domain, we randomly divide the data set according to the proportion of training set to test set. The specific proportions are introduced in the following data set descriptions respectively. *Cityscapes*[50] is a data set of city scenery under normal weather with eight categories, including 2975 training images and 500 test images. *Foggy Cityscapes*[51] is a data set of city scenery in foggy weather synthesized from Cityscapes. It contains the same categories as the Cityscapes data set. The *Watercolor2k*[16] data set contains six categories of watercolor-style artistic images. It contains 2000 images, in which 1500 samples are choose as training set and the remaining 500 samples are used as test set. For the data sets used only as source domain, we use all images to train the source domain model. *KITTI*[52] contains 7481 images of city scenery with different camera setup from Cityscapes. To be consistent to the Cityscapes data set, only the car category is used in the experiment. *SIM10k*[53] contains 10,000 virtual city scene images obtained from GTAV games. *Pascal VOC*[54] is a data set containing 20 categories of real images. We use 5,011 training images from Pascal VOC 2007 in the experiment. Six categories shared with the data set Watercolor2k are used in the experiments.

6.1.2 | Scenarios of domain adaptation

Based on the above six data sets, we conduct evaluations on four different transfer scenarios.

(1) *Cityscapes to Foggy Cityscapes*. In this scenario, the source domain is the Cityscapes data set under normal weather condition, and the target domain is the Foggy Cityscapes data set under foggy weather condition. This scenario can effectively verify the adaptability of our method in different weather.
(2) *KITTI to Cityscapes*. The images contained in two data sets are city scenes under different camera settings, which can effectively verify the adaptation performance of the proposed method under cross-camera settings. The common category car is used for evaluation.
(3) *SIM10k to Cityscapes*. This case is to verify the adaptability of the proposed method from virtual to reality. The common category car is used for evaluation.
(4) *Pascal VOC to Watercolor2k*. This case can effectively verify the adaptability of our method from reality to artistic images. Domain shift is large. We use six common categories for evaluation.

6.1.3 | Implementation details

Following the setting of most domain adaptation methods which use source data, we take Faster R-CNN with VGG16 in the first three experiments and ResNet-101 in the fourth experiment as the baseline and report mean average precision (mAP) with an IoU threshold of 0.5 as an evaluation metric in all experiments. If not mentioned, all experiment settings in our model follow the setup in Faster R-CNN.[12] We train the source domain model (Faster R-CNN) with a batch size of 4, a learning rate of 4e-3 and a decay rate of 0.1 for the last epoch. The student and teacher models are first initialized by the pre-trained source domain model. Then

we train our model SOAP with the learning rate of 4e-4 and batch size of 4. One traversal of each unlabeled target domain image is enough to obtain good performance.

In all transfer experiments, the learning rates of the image-level and instance-level discriminators are 0.001 and the smoothing coefficient α is set to 0.99. The trade-off parameters β, λ_{MA}, λ_{NA}, and λ_{CC} are set to 0.8, 0.1, 0.1, and 0.1 for all experiments, respectively. The teacher model is the final model used for testing.

6.2 | Performance comparison

6.2.1 | Cityscapes to Foggy Cityscapes (C-F)

Table 1 shows the comparison results between our SOAP and the baseline (Source Only) and some state-of-the-art cross-domain detection methods which use source data. They are DAF,[32] DM,[43] WD,[34] FAF,[23] MAF,[37] SCDA,[36] MTOR,[45] and SWDA.[35] We also show the results of ablation analysis. The models SOAP(MA), SOAP(NA), SOAP(MA&CC), SOAP(NA&CC), and SOAP(All) mean the proposed SOAP has only MA, NA, MA&CC, NA&CC, and all consistency regularizations, respectively.

Because the source domain and target domain are similar under different weather conditions, the optimization performance by augmenting domain perturbation is particularly good. The proposed SOAP method has an 11.2% improvement over baseline, and surpasses multiple domain adaptation methods that use source domain data to transfer source model. The

TABLE 1 The mean average precision (mAP) of different models for transferring from *Cityscapes to Foggy Cityscapes* (C-F)

Method	Person	Rider	Car	Truck	Bus	Train	Mcycle	Bicycle	mAP
Source Only	25.8	33.3	35.2	13.0	26.4	9.1	19.0	32.3	24.3
DAF[32]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
DM[43]	30.8	40.5	44.3	**27.2**	38.4	34.5	28.4	32.2	34.6
WD[34]	30.2	42.0	44.2	22.2	**39.9**	36.5	25.4	34.4	33.1
FAF[23]	29.1	39.7	42.9	20.8	37.4	24.1	26.5	29.9	31.3
MAF[37]	28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
SCDA[36]	33.5	38.0	**48.5**	26.5	39.0	23.3	28.0	33.6	33.8
MTOR[45]	30.6	41.4	44.0	21.9	38.6	**40.6**	28.3	35.6	35.1
SWDA[35]	29.9	42.3	43.5	24.5	36.2	32.6	**35.3**	30.0	34.3
SOAP(MA)	32.5	40.5	43.1	21.5	35.4	22.9	26.2	37.4	32.4
SOAP(NA)	30.8	36.6	39.7	19.9	32.5	11.0	24.9	35.7	28.9
SOAP(MA&CC)	35.7	44.5	48.3	24.8	35.2	20.7	29.6	37.6	34.5
SOAP(NA&CC)	35.3	44.1	47.9	23.9	35.3	18.9	30.1	37.6	34.1
SOAP(All)	**35.9**	**45.0**	48.4	23.9	37.2	24.3	31.8	**37.9**	**35.5**

Note: We report the mAP results for eight categories and the results of baseline (Source Only) and multiple domain adaptation methods using source domain data. The best result is emphasized in bold.

performances of the ablation modules also exceed many domain adaptation methods, which shows that SOAP is effective and stable.

6.2.2 | KITTI to Cityscapes (K-C) and SIM10k to Cityscapes (S-C)

As shown in Table 2, SOAP improves the performances up to 5.2% and 5.8% compared to the baseline respect to the transfer scenarios K-C and S-C. SOAP also exceeds some domain adaptation methods using source domain data such as DAF,[32] WD,[34] and SWDA,[35] respectively. The ablation modules also show very good results. This experiment verifies the adaptability of SOAP under cross-camera and virtual-to-reality transfer cases.

6.2.3 | Pascal VOC to Watercolor2k (P-W)

The performance of SOAP is shown in Table 3. SOAP has a 3.2% improvement over the baseline. In this scenario, there are different image sizes and large domain shift. The performance of SOAP cannot surpass domain adaptation method using source domain data. The main reason is that it is not easy to obtain correct domain perturbation. The aligned target domain features by SOAP does not reach the domain invariant space in which the source model works well. The domain perturbation image for *Watercolor2k* is rather different from the above mentioned scenarios which is shown in Figure 6. For the domain adaptation methods which use source data set, by the help of source domain data, domain invariant feature can be more correctly obtained and reached. However, as shown in the experiments, even with not good domain perturbation, SOAP and its ablation modules also have higher performance than

TABLE 2 The mean average precision (mAP) of different models for transferring from *KITTI to Cityscapes* (K-C) and *SIM10k to Cityscapes* (S-C), respectively

Method	K-C	S-C
Source Only	36.7	35.0
DAF[32]	38.5	38.97
WD[34]	–	40.6
FAF[23]	–	41.2
MAF[37]	41.0	41.1
SWDA[35]	–	40.7
SCDA[36]	42.5	–
SOAP(MA)	40.2	41.4
SOAP(NA)	39.6	39.0
SOAP(MA&CC)	42.2	41.2
SOAP(NA&CC)	**42.7**	**41.6**
SOAP(All)	41.9	40.8

Note: We report mAP of the car category and the results of baseline (Source Only) and multiple domain adaptation methods using source domain data. The best result is emphasized in bold.

TABLE 3 The mean average precision (mAP) of different models for transferring from *Pascal VOC to Watercolor2k*

Method	Bike	Bird	Car	Cat	Dog	Person	mAP
Source Only	**83.7**	44.8	35.6	37.0	29.5	53.7	47.4
SOAP(MA)	82.0	42.3	36.8	40.5	28.1	53.8	47.3
SOAP(NA)	81.7	44.7	36.2	41.1	37.4	55.1	49.4
SOAP(MA&CC)	80.5	**45.4**	39.5	46.6	37.4	55.3	50.8
SOAP(NA&CC)	79.3	44.3	**41.4**	45.7	**39.3**	**55.9**	**51.0**
SOAP(All)	77.7	43.2	40.1	**48.2**	38.8	55.4	50.6

Note: We report mAP of six public categories and the baseline results. The best result is emphasized in bold.

that of baseline. The method to generate good domain perturbation is worth studying in the future.

6.3 | Parameter analysis

To analyze the influence of parameters on the proposed method SOAP, we conduct parameter analysis experiment on the C-F adaptation scenario. While ensuring that all settings are exactly the same, only one parameter to be analyzed is adjusted.

Figure 3 shows a line graph of the result changing with the parameter α. When α is set to 0, the teacher model parameters change completely with the student model. The parameters of two model are exactly the same. The Mean Teacher model is equivalent to a single model. The performance drops significantly, which shows that adopting Mean Teacher structure is indispensable. When α gradually increases to 0.99, the result is gradually improved to achieve optimal performance. However, when the value of α is continued to be increased, performance will decrease. Because when α is close to 1, the teacher model is hardly be trained. The performance will be close to that of the source domain model.

Figure 3 also shows the influences of three parameters $\lambda_{MA}, \lambda_{NA}$, and λ_{CC} on the experimental results. The changes of these parameters within a relatively large range have little influence on the experimental results. The performances are all better than that of Source Only. For the C-F adaptation scenario, when the three parameters all take the value of 0.1, the performance is best. For other transfer scenarios, to show the robustness of SOAP, we use the same parameter settings as the C–F adaptation scenario. Actually, we can get about 1% more improvement on average if we use different parameter settings.

6.4 | Error analysis

We follow Cai et al.[45] to conduct error analysis of the highest confidence detection experiment on the *Cityscapes to Foggy Cityscapes* scenario. We first select top-K detection results with the highest confidence in each category, where K is the number of ground-truth in each category. According to IoU with ground-truth, these detection results are divided into three types:

Correct: $IoU > 0.5$, Mis-localized: $0.3 < IoU \leq 0.5$, and Background: $IoU \leq 0.3$ or others. We calculated average percentages of three types of detection results in each category. As shown in Figure 4, our method SOAP effectively reduces background detection and detects more objects than Source Only (Faster R-CNN).

6.5 | Detection visualization

To show the visualization effect of the proposed method SOAP, we respectively show the visual detection results of the baseline (Source Only) and SOAP in the target domain for four domain adaptation scenarios. As shown in Figure 5, the left two columns shows visual detection results of baseline, and the right two columns are the visual detection results of SOAP. The four row images from top to down are taken from *Cityscapes to Foggy Cityscapes*, *KITTI to Cityscapes*, *SIM10k to Cityscapes,* and *Pascal VOC to Watercolor2k*. In each scenario, SOAP can detect objects that are not detected by baseline, and has a higher score on the jointly detected objects. Our method SOAP really works like a soap which removes the domain perturbations.

6.6 | Domain perturbation analysis

We conduct ablation analysis for domain perturbation on the *Cityscapes to Foggy Cityscapes* scenario to verify the effectiveness of domain perturbation theory. In the experiment, we try some other type of noises while keeping the same setting. Specifically, the global domain perturbation (Global) is replaced by all-zero tensor (Zero), Gaussian noise perturbation (Gaussian), and our local domain perturbation (Local).

As shown in Table 4, the accuracies of Zero and Gaussian are only 1% higher than Source Only. A tiny improvement is got since the noise can also help the teacher model learns something. However, their performance are far lower than ours (Local and Global). Because the local domain perturbation is not complete enough and there are more random noises, the performance is lower than that of the global perturbation. The analysis of the aforementioned domain perturbation proves that the proposed domain perturbation is effective.

TABLE 4 The mean average precision (mAP) with different type of perturbations

Method	Person	Rider	Car	Truck	Bus	Train	Mcycle	Bicycle	mAP
Source Only	25.83	33.30	35.16	12.98	26.38	9.11	18.98	32.31	24.26
Zero	25.69	34.59	35.53	19.06	25.13	10.50	21.41	30.08	25.25
Gaussian	25.69	34.73	35.55	19.03	24.58	10.10	21.15	30.00	25.10
Local	35.57	43.21	44.44	23.32	36.54	22.92	29.05	37.07	34.02
Global	**35.85**	**44.95**	**48.40**	**23.90**	**37.15**	**24.33**	**31.75**	**37.91**	**35.53**

Note: We replace the global domain perturbation (Global) with all-zero tensor (Zero), Gaussian noise perturbation (Gaussian), and local domain perturbation (Local) under the same settings. The best result is emphasized in bold.

6.7 | Domain perturbation visualization

For a more intuitive understanding of the domain perturbation and super target domain, we visualize the domain perturbation of three target domains including *Foggy Cityscapes*, *Cityscapes,* and *Watercolor2k*. As shown in Figure 6, the left, middle and right columns are the target domain image, the corresponding domain perturbation, and super target domain image. In *Foggy Cityscapes* and *Cityscapes*, we only use 100 images to generate domain perturbation. Due to the different sizes of the images in *Watercolor2k*, global domain perturbation cannot be

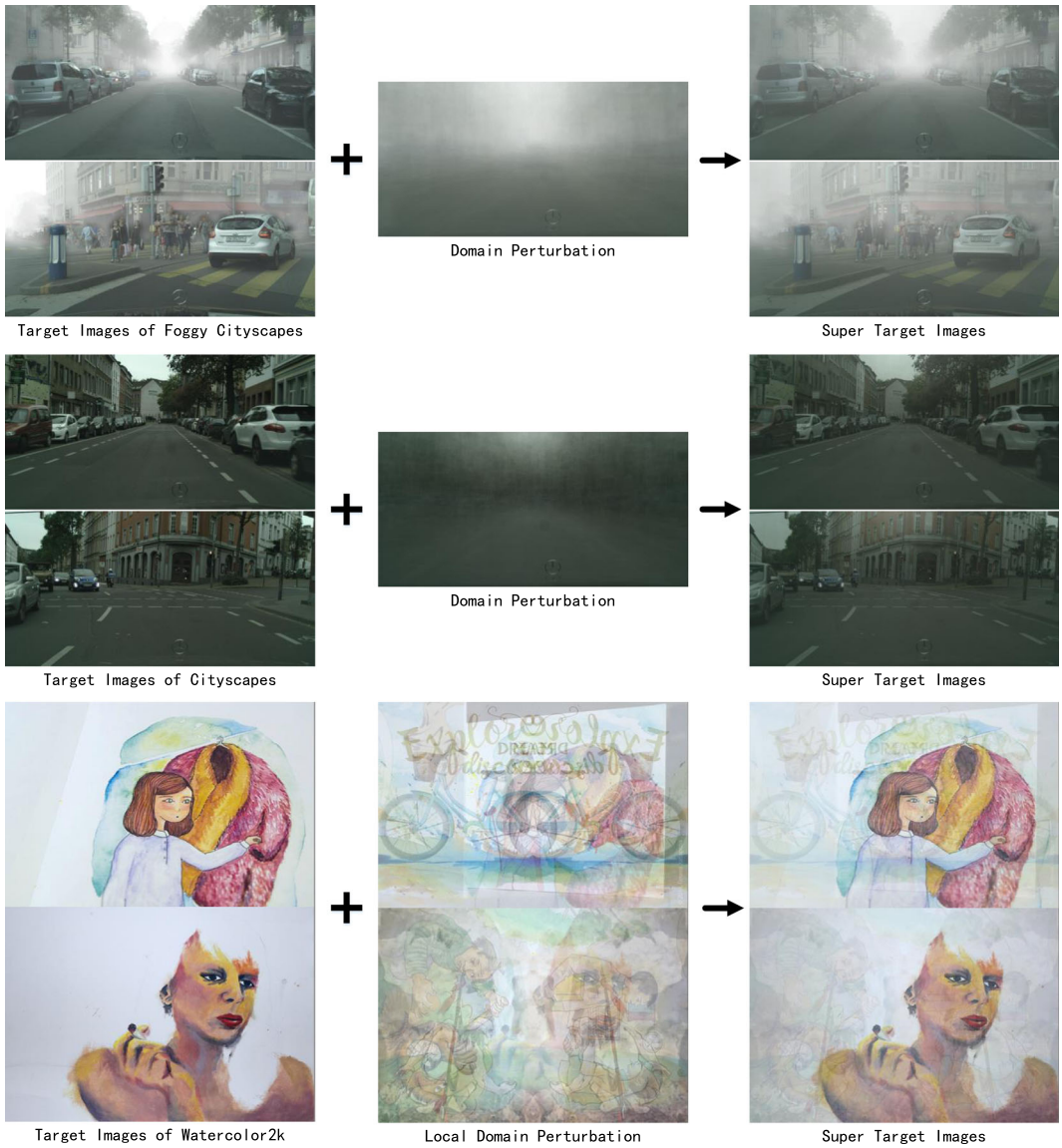

FIGURE 6 Domain perturbation visualization. The left column shows three target domain images. The middle and right columns are the corresponding domain perturbation and super target domain images [Color figure can be viewed at wileyonlinelibrary.com]

shown. So the local domain perturbation in a batch is shown. In the implementation of SOAP, we also use local domain perturbation to synthesize the super target domain. In some transfer scenarios, there is little difference between the target domain and the super target domain in human eyes, but there indeed exist differences which can be effectively captured by the computer. While in some transfer scenarios, the super target domain has some large different images from the target domain. No matter what their appearances are, these super target domain images can help us better adapt the object detector.

From Figure 6, it can be observed that if the source domain and the target domain have similar scene, then SOAP works very well such as these domain adaptation scenarios *Cityscapes to Foggy Cityscapes*, *KITTI to Cityscapes*, *SIM10k to Cityscapes*. While for the case that the source domain and the target domain have large distribution differences, without the help of source domain data, how to find a good optimization direction is still left for future study.

7 | CONCLUSION

We proposed a novel source data-free domain adaptation method (SOAP) to tackle the domain shift problem in cross-domain object detection. We assume that the target domain is derived from the domain invariant space by domain perturbation, then the super target domain is obtained by further domain perturbation. The alignment direction of target domain to the domain invariant feature can be approximated by the direction of super target domain to the target domain. Three consistency regularizations guarantee Mean Teacher structure achieves the learning goal. Experiments on four domain adaptation scenarios showed that our method can efficiently improve the accuracy of source model and even surpass some domain adaptation methods using source domain data. Our method SOAP really works like a soap which removes domain perturbations. SOAP provides a new way worthy of further study to solve the problem of domain adaptation for different tasks such as image segmentation, image classification, image enhancement, and so forth.

ACKNOWLEDGMENTS

This study was supported in part by the National Key R&D Program of China (2018YFE0203900), National Natural Science Foundation of China (61773093), Important Science and Technology Innovation Projects in Chengdu (2018-YF08-00039-GX), Key R&D Programs of Sichuan Science and Technology Department (2020YFG0476), and Postdoctoral Research Foundation of China (2020M683243).

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

ORCID

Lin Xiong https://orcid.org/0000-0002-8303-3216
Mao Ye https://orcid.org/0000-0003-4760-8702
Dan Zhang https://orcid.org/0000-0001-7429-3609
Yan Gan https://orcid.org/0000-0002-6716-0039
Xue Li https://orcid.org/0000-0002-4515-6792
Yingying Zhu https://orcid.org/0000-0003-3920-5890

REFERENCES

1. Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: Leibe B, Matas J, Sebe N, Welling M (eds), *Computer Vision—ECCV 2016. ECCV 2016. Lecture Notes in Computer Science*. 9905. Cham: Springer; 2016. https://doi.org/10.1007/978-3-319-46448-0_2
2. Lin T, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017:2999-3007. https://doi.org/10.1109/ICCV.2017.324
3. Law H, Deng J. CornerNet: detecting objects as paired keypoints. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds), *Computer Vision—ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*. 11218. Cham: Springer; 2018. https://doi.org/10.1007/978-3-030-01264-9_45
4. Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q. CenterNet: Keypoint triplets for object detection. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, 2019:6568-6577. https://doi.org/10.1109/ICCV.2019.00667
5. Tian Z, Shen C, Chen H, He T. FCOS: Fully convolutional one-stage object detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, 2019:9626-9635. https://doi.org/10.1109/ICCV.2019.00972
6. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014:580-587. https://doi.org/10.1109/CVPR.2014.81
7. Girshick R. Fast R-CNN. In: *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, 2015:1440-1448. https://doi.org/10.1109/ICCV
8. Dai J, Li Y, He K, Sun J. R-FCN: object detection via region-based fully convolutional networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16), Curran Associates Inc.*, Red Hook, NY, 2016:379-387.
9. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017:2980-2988. https://doi.org/10.1109/ICCV.2017.322
10. Cai Z, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake, 2018:6154-6162. https://doi.org/10.1109/CVPR.2018.00644
11. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, 2016:779-788. https://doi.org/10.1109/CVPR.2016.91
12. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In: Dickinson S, ed. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017;39(6), 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031.
13. Devaguptapu C, Akolekar N, Sharma MM, Balasubramanian VN. Borrow from anywhere: pseudo multi-modal object detection in thermal imagery. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, 2019:1029-1038. https://doi.org/10.1109/CVPRW.2019.00135
14. Chen H, Wang Y, Wang G, Qiao Y. LSTD: A low-shot transfer detector for object detection. arXiv preprint arXiv:1803.01529; 2018.
15. Zhu F, Shao L. Weakly-supervised cross-domain dictionary learning for visual recognition. *Int J Comput Vis*. 2014;109(1-2):42-59.
16. Inoue N, Furuta R, Yamasaki T, Aizawa K. Cross-domain weakly-supervised object detection through progressive domain adaptation. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake, 2018:5001-5009. https://doi.org/10.1109/CVPR.2018.00525
17. Wang Q, Gao J, Li X. Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes. *IEEE Trans Image Process*. 2019;28(9):4376-4386.
18. Tzeng E, Hoffman J, Saenko K, Darrell T. Adversarial discriminative domain adaptation. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Hawaii, 2017:2962-2971. https://doi.org/10.1109/CVPR.2017.316

19. Li X, Ye M, Liu Y, Zhu C. Adaptive deep convolutional neural networks for scene-specific object detection. In: Wu F, (ed). *IEEE Transactions on Circuits and Systems for Video Technology*, 2019;29(9), 2538-2551. https://doi.org/10.1109/TCSVT.2017.2749620
20. Kang G, Jiang L, Yang Y, Hauptmann AG. Contrastive adaptation network for unsupervised domain adaptation. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019:4888-4897. https://doi.org/10.1109/CVPR.2019.00503
21. Pan Y, Yao T, Li Y, Wang Y, Ngo C, Mei T. Transferrable prototypical networks for unsupervised domain adaptation. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019: 2234-2242. https://doi.org/10.1109/CVPR.2019.00234
22. Hsu H, Yao C-H, Tsai Y-H, et al. Progressive domain adaptation for object detection. In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass Village, CO, 2020:738-746. https://doi.org/10.1109/WACV45572.2020.9093358
23. Wang T, Zhang X, Yuan L, Feng J. Few-shot adaptive faster R-CNN. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019:7166-7175. https://doi.org/10.1109/CVPR.2019.00734
24. Motiian S, Jones Q, Iranmanesh SM, Doretto. G. Few-shot adversarial domain adaptation. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, Curran Associates Inc., Red Hook, NY, 2017:6673-6683.
25. Liang J, Hu D, Feng J. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. arXiv preprint arXiv:2002.08546; 2020.
26. Kundu JN, Venkat N, Rahul MV, Babu RV. Universal source-free domain adaptation. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, 2020:4543-4552. https://doi.org/10.1109/CVPR42600.2020.00460
27. Kim Y, Hong S, Cho D, Park H, Panda P. Domain adaptation without source data. arXiv preprint arXiv:2007.01524; 2020.
28. Li R, Jiao Q, Cao W, Wong H-S, Wu S. Model adaptation: unsupervised domain adaptation without source data. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, 2020:9638-9647. https://doi.org/10.1109/CVPR42600.2020.00966
29. Sahoo R, Shanmugam D, Guttag J Unsupervised domain adaptation in the absence of source data. In: ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning; 2020.
30. Hoffman J, Gupta S, Leong J, Guadarrama S, Darrell T. Cross-modal adaptation for RGB-D detection. In: Kragic D, Bicchi A, De Luca A, (eds) *2016 IEEE International Conference on Robotics and Automation (ICRA)*. Stockholm, Sweden; 2016:5032-5039. https://doi.org/10.1109/ICRA.2016.7487708
31. Tarvainen A, Valpola H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, Curran Associates Inc., Red Hook, NY, 2017:1195-1204.
32. Chen Y, Li W, Sakaridis C, Dai D, Van Gool L. Domain adaptive faster R-CNN for object detection in the wild. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake, 2018:3339-3348. https://doi.org/10.1109/CVPR.2018.00352
33. Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW. A theory of learning from different domains. *Mach Learn*. 2010;79(1-2):151-175.
34. Xu P, Gurram P, Whipps G, Chellappa R. Wasserstein distance based domain adaptation for object detection. arXiv preprint arXiv:1909.08675; 2019.
35. Saito K, Ushiku Y, Harada T, Saenko K. Strong-weak distribution alignment for adaptive object detection. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019: 6949-6958. https://doi.org/10.1109/CVPR.2019.00712
36. Zhu X, Pang J, Yang C, Shi J, Lin D. Adapting object detectors via selective cross-domain alignment. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019: 687-696. https://doi.org/10.1109/CVPR.2019.00078
37. He Z, Zhang L. Multi-adversarial faster-RCNN for unrestricted object detection. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, 2019:6667-6676. https://doi.org/10.1109/ICCV.2019.00677

38. Xie R, Yu F, Wang J, Wang Y, Zhang L. Multi-level domain adaptive learning for cross-domain detection. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, 2019: 3213-3219. https://doi.org/10.1109/ICCVW.2019.00401
39. Zheng Y, Huang D, Liu S, Wang Y. Cross-domain object detection through coarse-to-fine feature adaptation. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, 2020:13763-13772. https://doi.org/10.1109/CVPR42600.2020.01378
40. Xu M, Wang H, Ni B, Tian Q, Zhang W. Cross-domain detection via graph-induced prototype alignment. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, 2020: 12352-12361. https://doi.org/10.1109/CVPR42600.2020.01237
41. Xu C-D, Zhao X-R, Jin X, Wei X-S. Exploring categorical regularization for domain adaptive object detection. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, 2020:11721-11730. https://doi.org/10.1109/CVPR42600.2020.01174
42. Chen C, Zheng Z, Ding X, Huang Y, Dou Q. Harmonizing transferability and discriminability for adapting object detectors. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, 2020:8866-8875. https://doi.org/10.1109/CVPR42600.2020.00889
43. Kim T, Jeong M, Kim S, Choi S, Kim C. Diversify and match: a domain adaptive representation learning paradigm for object detection. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019:12448-12457. https://doi.org/10.1109/CVPR.2019.01274
44. Rodriguez AL, Mikolajczyk K. Domain adaptation for object detection via style consistency. arXiv preprint arXiv:1911.10033; 2019.
45. Cai Q, Pan Y, Ngo C, Tian X, Duan L, Yao T. Exploring object relation in mean teacher for cross-domain detection. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019:11449-11458. https://doi.org/10.1109/CVPR.2019.01172
46. Khodabandeh M, Vahdat A, Ranjbar M, Macready W. A robust learning approach to domain adaptive object detection. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, 2019:480-490. https://doi.org/10.1109/ICCV.2019.00057
47. Deng J, Li W, Chen Y, Duan L. Unbiased Mean Teacher for cross domain object detection. arXiv preprint arXiv:2003.00707; 2020.
48. Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. In: Bach F, Blei D, (eds) *International Conference on Machine Learning*. Lille: PMLR; 2015:1180-1189.
49. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 2 (NIPS'14)*. MIT Press, Cambridge, MA, 2014:2672-2680.
50. Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, 2016:3213-3223. https://doi.org/10.1109/CVPR.2016.350
51. Sakaridis C, Dai D, Van Gool L. Semantic foggy scene understanding with synthetic data. *Int J Comput Vis*. 2018;126(9):973-992.
52. Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: the Kitti dataset. *Int J Robotics Res*. 2013; 32(11):1231-1237.
53. Johnson-Roberson M, Barto C, Mehta R, Sridhar SN, Rosaen K, Vasudevan R. Driving in the matrix: can virtual worlds replace human-generated annotations for real world tasks? In: Okamura A, ed. *2017 IEEE International Conference on Robotics and Automation (ICRA)*. Singapore; 2017: 746-753. https://doi.org/10.1109/ICRA.2017.7989092
54. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. *Int J Comput Vis*. 2010;88(2):303-338.

How to cite this article: Xiong L, Ye M, Zhang D, Gan Y, Li X, Zhu Y. Source data-free domain adaptation of object detector through domain-specific perturbation. *Int J Intell Syst*. 2021;36:3746-3766. https://doi.org/10.1002/int.22434