Source-Free Object Detection by Learning to Overlook Domain Style

Shuaifeng Li[1] Mao Ye[1] Xiatian Zhu[2] Lihua Zhou[1] Lin Xiong[1]
[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China
[2]Centre for Vision, Speech and Signal Processing, University of Surrey
hotwindlsf@gmail.com,maoye@uestc.edu.cn,xiatian.zhu@surrey.ac.uk

Abstract

Source-free object detection (SFOD) needs to adapt a detector pre-trained on a labeled source domain to a target domain, with only unlabeled training data from the target domain. Existing SFOD methods typically adopt the pseudo labeling paradigm with model adaption alternating between predicting pseudo labels and fine-tuning the model. This approach suffers from both unsatisfactory accuracy of pseudo labels due to the presence of domain shift and limited use of target domain training data. In this work, we present a novel Learning to Overlook Domain Style (LODS) method with such limitations solved in a principled manner. Our idea is to reduce the domain shift effect by enforcing the model to overlook the target domain style, such that model adaptation is simplified and becomes easier to carry on. To that end, we enhance the style of each target domain image and leverage the style degree difference between the original image and the enhanced image as a self-supervised signal for model adaptation. By treating the enhanced image as an auxiliary view, we exploit a student-teacher architecture for learning to overlook the style degree difference against the original image, also characterized with a novel style enhancement algorithm and graph alignment constraint. Extensive experiments demonstrate that our LODS yields new state-of-the-art performance on four benchmarks.

1. Introduction

The resurgence of deep convolutional neural networks has greatly promoted the development of object detection, for example, the one-stage YOLO [1] and two-stage Faster R-CNN [27] have made a big splash. However, when applied to a new scenario, a pre-trained detector often suffers a performance drop, due to the domain shift [5]. Moreover, considering to data privacy, distributed data storage, and inconvenient data transmission, *Source-Free Object Detection* (SFOD) [23] which assumes only the pre-trained model on the source domain is available and source data itself is unavailable, emerged recently as a promising topic.

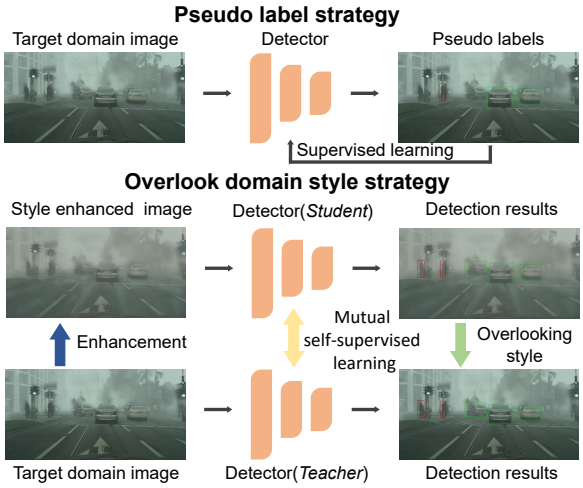

Figure 1. The comparison of pseudo label strategy (above) and our proposed strategy (below).

At present, there do not exist much researches on SFOD problem. The community pays more attention to *Source-Free Domain Adaptation* (SFDA). The methods for SFDA can be roughly divided into two categories. The first category is based on the idea of *sample generation* [11, 19, 21, 30]. Since the source data is not accessible, traditional domain adaptation techniques are not applicable. The labeled images with source domain style or target domain style, or the labeled features obeying the source distribution, are generated. The key to success is the satisfied samples generation which itself is sufficiently challenging and not well solved. Another category utilizes the self-training based *pseudo labeling* [18, 24, 25]. But obtaining reliable labels is not easy especially in the situation with large domain gap and always only high-confidence labeled samples are taken in the self-training process.

The domain adaptation methods for SFDA can not be applied to SFOD directly, owing to the complexity of background, diversity of objects and numerous negative samples (background). Recently, several SFOD methods [13,23,36] based on *pseudo-labeling* or *sample generation* strategies

are proposed. They use better predicted pseudo labels or domain noise perturbed images as the self-supervised signals. The state-of-the-art performance has been achieved. Similar as the SFDA methods, the unreliable pseudo labels and bad quality of generated samples limit their performance.

It is obvious that target domain style (e.g. imaging characteristics) contributes to a significant part of the domain shift against the source domain. Hence, minimizing the impact of target domain style on the model behaviour would be immediately effective in reducing the domain shift. On the basis of the above, as shown in Fig. 1, we propose a new domain adaption method, dubbed as *Learning to Overlook Domain Style* (LODS). It first enhances the target domain style for each target image while maintaining the original style of target images. In this way, an auxiliary view based on the style enhanced images is constructed. With this auxiliary view, our method lets object detector learn to overlook target domain style. The student-teacher framework is employed to do this task.

Specifically, our method consists of style enhancement module and overlooking style module. For the style enhancement module, to increase the degree of target domain style, it merges the target domain style in a non-linear way. The overlooking style module is based on the Mean-Teacher architecture. The target sample is input to teacher model; while the corresponding style enhanced version is input to student model. Both models are initialized by the pre-trained source model. To help both of the teacher and student models have the ability of overlooking target domain style, we devise graph alignment constrains at instance and image levels. By requiring the consistency of the object instance and image patch feature relationships between the image and its corresponding style enhanced version, the extracted feature will overlook the target domain style.

Our contributions are summarized as follows. (1) We propose a novel learning to overlook domain style strategy. Different from traditional *pseudo label strategy*, it not only makes a full use of all target data but also reduces the domain sensitivity of the object detector. Compared with the *sample generation* strategy, style enhancement is easier. (2) A style enhancement method is proposed. Different from the existing style transfer methods, it retains the original target domain style, and further adds more target domain style to the target domain image. (3) We propose a new Mean-Teacher framework variant which achieves a two-way knowledge distillation. It overlooks domain style by two graph alignments without any help of source data.

2. Related Work

2.1. Unsupervised domain adaptive object detection

Recent advances promoted the development of *Unsupervised Domain Adaptive Object Detection* (UDAOD) which can access source data (the difference from SFOD). Methods are roughly classified into four strategies. The first one is based on *distribution alignment strategy*, such as DA-Faster [5], SWDA [28], HTCN [4], SSA-DA [43], ICR-CCR [37], VDD [34], SGA-S [41], CST-DA [42]and DBGL [3] etc. It aligns different types of features at different levels via domain classifiers or prototypes. The second one uses *pseudo label strategy* such as NL [16] and CDG [22], which exploits pseudo labels for target samples. The third one is *sample generation strategy*, such as DM [17], AFAN [32], UMT [7], etc. They tend to transfer the style of source and target images by CycleGAN [44]. The final one is using *auxiliary model strategy*, which learns an auxiliary detector, multi-class classifier, or multi-label classifier to assist transferring detector, such as NL [16], ICR-CCR [37], MTOR [2], UMT [7]. The Mean-Teacher framework is a typical representative. Despite the great performance achieved, all of these methods need to access source domain data.

2.2. Source-free domain adaptation

Due to the lack of source data, *Source-Free Domain Adaptation* (SFDA) only relying on the pre-trained source model is more difficult than traditional unsupervised domain adaptation. There are two main routes to address this problem. One line is based on *sample generation strategy*. For example, 3C-GAN [21] and SDDA [19] generate labeled samples with target domain style for training; VDM-DA [30] generates source domain style features then aligns the generated features with target features; SFIT [11] utilizes the batch-norm layers of the source model to generate images with source domain style and aligns the output predictions. Another line uses *pseudo label strategy*. SHOT [24] and SHOT++ [25] use the centroid of each class to generate pseudo labels, and information maximization to ensure the balance between classes; DASD [18] constructs an adaptive prototype memory to exploit pseudo labels.

There are not many methods for *Source-Free Object Detection* (SFOD). SED [23] searches a confident threshold for pseudo labels generation according to self-entropy descent policy. Except for pseudo labels, HCL [13] also proposes historical contrastive instance discrimination to pull the current representation to its positive key. Both achieve good performance, but unreliable pseudo labels and only (confident) instance-level samples are used. SOAP [36] proposes to perturb the target images with domain noise and uses the adversarial learning technique to transfer the detector. It does not work in the case of existing large domain gap.

3. Methodology

3.1. Problem statement

Suppose source domain $D_s = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$ is labeled, where $y_s^i = (b_s^i, c_s^i)$ denotes the boxes and classes of ob-

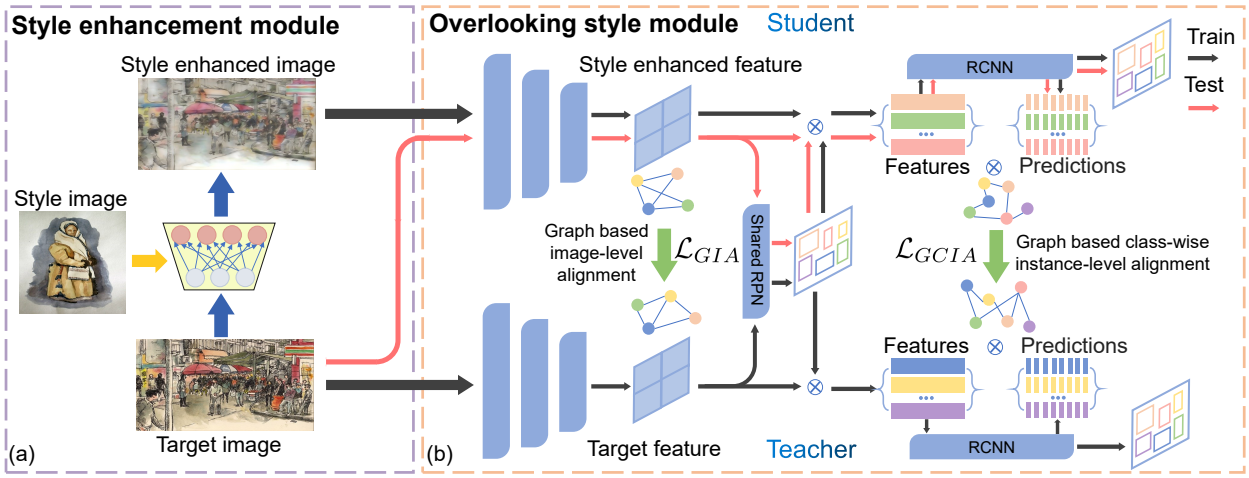

Figure 2. Overview of the proposed **Learning to Overlook Domain Style** (LODS) method. The black and red lines indicate the training and testing flows respectively. (a) The style enhancement module for increasing the style degree of a target domain image. (b) The overlooking style module formulated in a student-teacher architecture.

jects in the ith image of source domain, N_s denotes the total number of source images. Target domain $D_t = \{x_t^i\}_{i=1}^{N_t}$ is unlabeled, where N_t denotes the total number of target images and the target sample obeys the same distribution. Our goal is to transfer the source model to the target domain without performance dropping while the source data cannot be accessed in the domain adaptation process.

Overview. The proposed *Learning to Overlook Domain Style* (LODS) method consists of two parts. As shown in Fig. 2, one is the style enhancement module; another one is the overlooking style module. Style enhancement module (Fig .2(a)) first extracts the styles of each images, i.e., the channel-wise mean and variance. For an image, its enhanced target domain style is calculated as the non-linearly combination of styles of itself and any target image. Then, the style is enhanced by replacing with the enhanced style.

By considering the style enhanced images as another domain, the Mean-Teacher framework can be employed to leverage the style difference for model adaptation (Fig. 2(b)). The target image and style enhanced version are fed into the teacher and student models respectively. These two models are based on Faster-RCNN and initialized as the pre-trained source model. Class-wise instance-level alignment and image-level alignment based on graph matching are designed to help teacher and student learn from each other. Pseudo labels are also used to increase the discrimination of student model.

3.2. Style enhancement

Existing approach [14] already achieves arbitrary style transfer by simply replacing feature channel-wise mean and variance as the input style. We follow this technical route. But different from style transfer, we need to further manipulate the feature mean and variance.

Suppose we have an image x and any image y in the target domain; e_x and e_y are the corresponding features respectively. $\mu(e_x)$ and $\sigma(e_x)$ are the channel-wise mean and variance respect to e_x, and so do $\mu(e_y)$ and $\sigma(e_y)$. According to the work in [14], style transfer is done by replacing feature channel mean and variance. Suppose e_x^y is the style transferred feature by adding the image y style, the formula is denoted as

$$e_x^y = \sigma(e_y)\frac{e_x - \mu(e_x)}{\sigma(e_x)} + \mu(e_y). \tag{1}$$

Instead of transferring image style, our goal is to enhance similar style, so we come up with an idea to generate the enhanced style consisting of a new channel mean and variance. Since the images x and y are from the same distribution, it is natural to integrate their mean and variance to enhance the style. This process can be denoted as $\mu(\hat{e}_x) = \delta_1(\mu(e_x), \mu(e_y))$, $\sigma(\hat{e}_x) = \delta_2(\sigma(e_x), \sigma(e_y))$ where δ_1 and δ_2 are two non-linear functions. $\hat{e}_x$ is the style enhanced feature respect to x.

Based on the above inference, we design a style enhancement module as shown in Fig. 3. Two networks F_1 and F_2 are designed to approximate δ_1 and δ_2, respectively. Each of them is composed of two fully connected layers and a ReLU layer to be nonlinear with minimal parameters. The feature encoder E is derived from a pre-trained VGG-16 model and fixed during training and testing. The decoder D is the inverse of the encoder. Because style consistency is constrained on low layer feature [14], the encoder $E = E_2 \circ E_1$ is further divided into E_1 and E_2 parts, where $\circ$ is the function nesting operator. So does the decoder $D = D_2 \circ D_1$ as D_1 and D_2. Specifically, the first ReLU layer after the first down-sampling is the divided line to separate E. D is symmetrically divided as E.

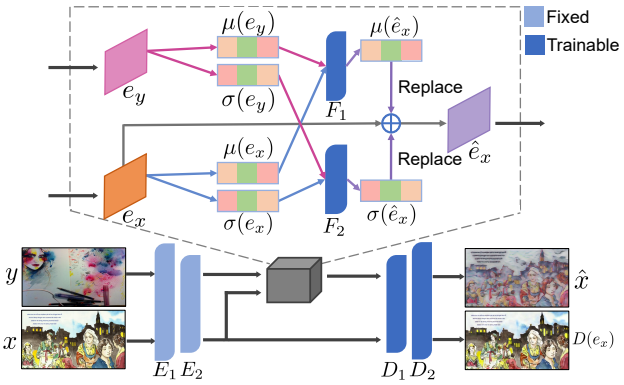

Figure 3. Architecture of style enhancement module. The style is enhanced by two networks F_1 and F_2.

For *training* the style enhancement module, the networks F_1 and F_2 are trained based on style consistency, which is denoted as the following,

$$\mathcal{L}_{sty} = \alpha L_s(E_1(\hat{x}), E_1(y)) + L_s(E_1(\hat{x}), E_1(x)), \quad (2)$$

where $\hat{x}$ is style enhanced image and $L_s(\phi, \psi) = \|\mu(\phi) - \mu(\psi)\|^2 + \|\sigma(\phi) - \sigma(\psi)\|^2$, for any two features ϕ and ψ, is a function to measure consistency in terms of feature channel mean and variance. α is a hyperparameter, fixed as 50 in all our experiments, that controls the rate at which style is added. The first and second items ensure that the styles of x and y are both included in $\hat{x}$ respectively. The decoder D is trained by the following loss function,

$$\begin{aligned} \mathcal{L}_{con} \quad = \quad & \|x - D(e_x)\|^2 + \|E_1(x) - D_1(e_x)\|^2 \\ & + \|\hat{e}_x - E(\hat{x})\|^2, \end{aligned} \quad (3)$$

where the three items respectively represent the content consistency from the perspectives of image, low layer feature, and high layer feature. To avoid be disturbed by each other, the decoder D and the networks F_1 and F_2 are trained alternately.

Remark. Noted that for enhancing the style of image x, we choose any image y as style image. We do not choose the mean style because it will be Gaussian noise (not diverse styles) if the backgrounds are scattered. While if the backgrounds are very similar, the mean of all target images can be used as the style image (e.g. Foggy-Cityscapes). Compared with domain randomization [12, 40] and domain randomization [17], our method sticks to the target domain style to mitigate the negative effect of it instead of using auxiliary domains for generalized representation learning.

3.3. Overlooking target domain style

Suppose two Faster R-CNN detectors Θ_{tea} and Θ_{stu} are regarded as the teacher model and the student model respectively, which are initialized by the pre-trained source model Θ_s. The target image and the corresponding style enhanced version are fed into the teacher model and the student model respectively, and the student model has the same region proposals by the teacher model. Graph based alignments are used to boost both of teacher and student models overlooking the target domain style.

3.3.1 Graph based alignments

Given a target image x and its corresponding style enhanced version $\hat{x}$, the image-level features $\boldsymbol{g}_x \in \mathbb{R}^{H\times W\times C}$ and $\boldsymbol{g}_{\hat{x}} \in \mathbb{R}^{H\times W\times C}$ are extracted by using the base feature extractor of Faster-RCNN (ResNet-101 or VGG-16). H and W represent the height and width of the feature map respectively; C is the number of feature channel. Subsequently, the instance-level feature features $\boldsymbol{f}_x \in \mathbb{R}^{R\times C'}$ and $\boldsymbol{f}_{\hat{x}} \in \mathbb{R}^{R\times C'}$ can also been extracted by utilizing Region Proposal Network (RPN) and ROI Pooling layer of Faster-RCNN. R represents the number of region proposals in an image; C' is the feature dimension. Denoting $\boldsymbol{p}, \hat{\boldsymbol{p}} \in \mathbb{R}^{R\times N_c}$ are the class predictions of the instance-level features $\boldsymbol{f}_x \in \mathbb{R}^{R\times C'}$ and $\boldsymbol{f}_{\hat{x}} \in \mathbb{R}^{R\times C'}$ respectively, where N_c is number of object categories including background.

Graph based class-wise instance-level alignment. Since the image $\hat{x}$ is just the style enhanced version of x, the instance features and their relationships between these two images should be consistent regardless of the impact of different strength styles on them. For improving the discrimination ability, we use class-wise instance-level features as the following,

$$\tilde{\boldsymbol{f}}_x = \boldsymbol{f}_x \odot \boldsymbol{p}, \quad \tilde{\boldsymbol{f}}_{\hat{x}} = \boldsymbol{f}_{\hat{x}} \odot \hat{\boldsymbol{p}}, \quad (4)$$

where $\tilde{\boldsymbol{f}}_x \in \mathbb{R}^{R\times(C'*N_c)}$ and $\tilde{\boldsymbol{f}}_{\hat{x}} \in \mathbb{R}^{R\times(C'*N_c)}$ are obtained by multilinear transformation $\odot$ of the predictions and the instance-level features.

Base on the above class-wise instance-level features, for the target image x and the corresponding style enhanced version $\hat{x}$, we define two graphs $G(\mathcal{V}, C)$ and $\hat{G}(\hat{\mathcal{V}}, \hat{C})$, respectively. $\mathcal{V}$ and $\hat{\mathcal{V}}$ are the corresponding class-wise instance-level features; C and $\hat{C}$ are the edge matrices, i.e., the cosine similarity matrices between these features respectively. We define *Graph based Class-wise Instance-level Alignment* (GCIA) loss by utilizing *Gromov-Wasserstein* (GW) discrepancy [26] as follows,

$$\mathcal{L}_{GCIA} = \sum_{i,j,m,n} L(C^{i,j}, \hat{C}^{m,n}) T^{i,m} T^{j,n}, \quad (5)$$

where $L(\cdot,\cdot)$ is the Kullback-Leibler divergence to measure the distance of the edges across graph. $\mathcal{L}_{GCIA}$ uses graph matching matrix $T \in \mathbb{R}^{R\times R}$ as weights to measure feature differences. Because each edge has two point, graph matching matrix T is used twice.

Next, we explain how to construct a class-wise graph matching matrix T. Since $T^{i,m}$ represents the matching degree between the features $\tilde{\boldsymbol{f}}_x^i$ and $\tilde{\boldsymbol{f}}_{\hat{x}}^m$, if their relationship is strong, the matching degree should be greater. So we define a category similarity matrix Γ based on cosine similarity,

$$\Gamma^{i,m} = \frac{\boldsymbol{p}^i \cdot \boldsymbol{p}^m}{\|\boldsymbol{p}^i\|^2 \cdot \|\boldsymbol{p}^m\|^2}. \tag{6}$$

Here, the predictions from teacher model are adapted since it is more reliable compared with the student model.

Due to the noise predictions existed in $\boldsymbol{p}$, we need to construct a category relation mask M to filter the noise with the help of pseudo labels. A confidence threshold h is set to exploit the pseudo label of each region proposal as

$$\boldsymbol{l}^r = \begin{cases} \arg\max_c \boldsymbol{p}^{r,c} & \text{if } \max_c \boldsymbol{p}^{r,c} \geq h, \\ 0 & \text{otherwise.} \end{cases}, \tag{7}$$

where $\boldsymbol{l}^r$ denotes the label of the r region, 0 represents the background and h is a hyperparameter. In this way, unreliable labels are filtered. Then, we further define M as follows,

$$M^{i,m} = \begin{cases} 1 & \text{if } \quad \boldsymbol{l}^i = \boldsymbol{l}^m, \text{ and } \quad \boldsymbol{l}^i \neq 0, \\ 0 & \text{otherwise.} \end{cases}. \tag{8}$$

By requiring the category consistency, M not only filters out the low confidence features but also reduces the redundant alignment between features.

Finally, we obtain graph matching matrix as follows,

$$T = I + \beta M \otimes \Gamma, \tag{9}$$

where $\otimes$ denotes the element-wise multiplication. β is a hyperparameter. The first term I is a unit matrix due to the correspondence between the features $\tilde{\boldsymbol{f}}_x$ and $\tilde{\boldsymbol{f}}_{\hat{x}}$ at the same region. While, the second term filters out the noise in Γ through M and enhances matching degree between the features with the same category.

Remark. Compared with the existing graph matching methods [35, 38], instead of learning a matching matrix, we directly construct a class-wise matching matrix based on the category probability. It is more suitable for objection detection which requires feature discrimination.

Graph based image-level alignment. Domain adaptation of object detector includes not only instance-level transferring, but also scene adaptation. Similar to instance-level alignment, we define another graph matching based alignment constrain. First, the image-level features $\boldsymbol{g}_x$ and $\boldsymbol{g}_{\hat{x}}$ are divided into $H * W$ patches. Then, we define two graphs $G'(\mathcal{V}', C')$ and $\hat{G}'(\hat{\mathcal{V}}', \hat{C}')$, respectively. $\mathcal{V}'$ and $\hat{\mathcal{V}}'$ are the corresponding patch-level features; C' and $\hat{C}'$ are cosine similarity matrices for these features respectively. By requiring the correspondence of features at the same patch, we construct graph matching matrix $Q = I$ and obtain the loss for *Graph based Image-level Alignment* (GIA) as

$$\mathcal{L}_{GIA} = \sum_{i,j,m,n} L((C')^{i,j}, (\hat{C}')^{m,n}) Q^{i,m} Q^{j,n}. \tag{10}$$

Remark. There are also some graph based works for UDAOD which can access source data. Compared with [2], our method uses a class-wise graph matching matrix instead of point-to-point matching. As for [39], we construct two graphs and then align them instead of prototypes, which is more suitable to our situation, i.e, overlooking domain style.

3.3.2 Overall loss function

In the Mean-Teacher framework, the parameters of the student model θ_{stu} are updated by gradient descent via the following objective function:

$$\min_{\theta_{stu}} \mathcal{L}_{ce} + \gamma \mathcal{L}_{GCIA} + \lambda \mathcal{L}_{GIA}, \tag{11}$$

where γ and λ are hyperparameters to balance loss components which are fixed as 0.1. The cross-entropy loss L_{ce} is used which utilizes the generated pseudo labels in Eq. (7) to increase the discrimination of student model.

While the parameters of the teacher model θ_{tea} are the exponential moving average of the historical parameters of the student model:

$$\theta_{tea} = \eta \cdot \theta_{tea} + (1 - \eta) \cdot \theta_{stu}, \tag{12}$$

where η is a hyperparameter to control the updating of teacher parameters which is fixed as 0.999. As a consequence, the alignment from style enhanced feature to the original feature will also guide the teacher detector overlooking domain style.

Our proposed strategy achieves a *two-way knowledge distillation*. Knowledge flows from the teacher model to the student model when achieving alignment, while the learned knowledge by the student model flows to the teacher model by parameter updating. Both detectors have learned to overlook domain style and thus achieved similar performance.

4. Experiment

Datasets. For the transfer scenario **Pascal**→ **Clipart**, **Pascal**[1] [8] is a dataset containing 20 categories of natural images. Similarly, **Clipart**[2] [15] also includes the same 20 categories as Pascal and 1K clipart-style images. According to [4, 28], we employ approximately 15K images from

[1] http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

[2] https://naoto0804.github.io/cross_domain_detection

Table 1. Detection results on **Pascal** → **Clipart**. The mean Average Precision (mAP, in %) of UDAOD methods is compared.

Methods	aero	bcycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hrs	bike	prsn	plnt	sheep	sofa	train	tv	mAP
Source Only	24.4	38.8	24.9	21.4	32.0	38.5	33.7	12.8	27.9	21.0	16.3	12.3	25.1	42.3	31.6	27.8	10.5	20.8	40.0	29.8	26.6
SWDA [28]	26.2	48.5	32.6	33.7	38.5	54.3	37.1	18.6	34.8	58.3	17.0	12.5	33.8	65.5	61.6	52.0	9.3	24.9	54.1	49.1	38.1
ICR-CCR [37]	28.7	55.3	31.8	26.0	40.1	63.6	36.6	9.4	38.7	49.3	17.6	14.1	33.3	74.3	61.3	46.3	22.3	24.3	49.1	44.3	38.3
HTCN [4]	33.6	58.9	34.0	23.4	45.6	57.0	39.8	12.0	39.7	51.3	21.1	20.1	39.1	72.8	63.0	43.1	19.3	30.1	50.2	51.8	40.3
DBGL [3]	28.5	52.3	34.3	32.8	38.6	**66.4**	38.2	**25.3**	39.9	47.4	23.9	17.9	38.9	78.3	61.2	51.7	**26.2**	28.9	**56.8**	44.5	41.6
DM [17]	25.8	63.2	24.5	**42.4**	47.9	43.1	37.5	9.1	47.0	46.7	26.8	24.9	48.1	78.7	63.0	45.0	21.3	**36.1**	52.3	53.4	41.8
PD [33]	41.5	52.7	34.5	28.1	43.7	58.5	41.8	15.3	40.1	54.4	26.7	28.5	37.7	75.4	63.7	48.7	16.5	30.8	54.5	48.7	42.1
SAPNet [20]	27.4	**70.8**	32.0	27.9	42.4	63.5	47.5	14.3	**48.2**	46.1	31.8	17.9	43.8	68.0	68.1	49.0	18.7	20.4	55.8	51.3	42.2
UMT [7]	39.6	59.1	32.4	35.0	45.1	61.9	**48.4**	7.5	46.0	**67.6**	21.4	**29.5**	**48.2**	75.9	**70.5**	**56.7**	25.9	28.9	39.4	43.6	44.1
SOAP [36]	34.6	46.7	26.8	23.2	34.9	33.5	39.3	16.5	29.1	33.6	17.9	12.0	26.9	41.2	37.1	34.5	14.3	23.4	36.3	35.7	29.9
Our method	**43.1**	61.4	**40.1**	36.8	**48.2**	45.8	48.3	20.4	44.8	53.3	**32.5**	26.1	40.6	**86.3**	68.5	48.9	25.4	33.2	44.0	**56.5**	**45.2**

Table 2. Detection results on **Pascal** → **Watercolor**.

Methods	Bike	Bird	Car	Cat	Dog	Person	mAP
Source Only	85.6	46.8	43.1	24.5	21.9	54.8	46.1
SWDA [28]	82.3	55.9	46.5	32.7	35.5	66.7	53.3
AFAN [32]	87.0	46.4	47.3	33.1	30.0	60.1	50.6
DBGL [3]	83.1	49.3	50.6	39.8	38.7	61.3	53.8
ATF [10]	78.8	**59.9**	47.9	41.0	34.8	66.9	54.9
SAPNet [20]	81.1	51.1	**53.6**	34.3	39.8	**71.3**	55.2
VDD [34]	90.0	56.6	49.2	39.5	38.8	65.3	56.6
PD [33]	**95.8**	54.3	48.3	42.4	35.1	65.8	56.9
UMT [7]	88.2	55.3	51.7	39.8	43.6	69.9	58.1
SOAP [36]	79.3	44.3	41.4	**45.7**	39.3	55.9	51.0
Our method	95.2	53.1	46.9	37.2	**47.6**	69.3	**58.2**

the training and validation sets of the PASCAL VOC 2007 and 2012 to pre-train the source model. **Pascal**→ **Watercolor** scenario has a dataset **Watercolor**[3] [15] which contains 2K watercolor-style images and 6 categories in common with Pascal. As the previous works [17, 28], we utilize its training and testing images to train and test our model correspondingly. **Cityscapes** → **Foggy-Cityscapes.** Captured under normal weather, **Cityscapes**[4] [6], consisting of 2,975 training images and 500 testing images, have a total of 8 categories. **Foggy-Cityscapes**[5] [29] applys images of Cityscapes to simulate foggy as well as inherits the annotations of Cityscapes. Following the general setting [22, 39], we utilize the training set of Cityscapes to pre-train the source model, and test our model on the test set of Foggy-Cityscapes. For the transfer scenario **KITTI** → **Cityscapes**, **KITTI**[6] [9] contains 7,481 urban images which are different from Cityscapes. Following the general setting [5, 23], we merely detect the category of car and pre-train the source model using all data.

Implementation details. For the sake of fairness, we follow the experimental setting of [5, 23, 28], where Faster R-CNN is used as the base detector. We first utilize Adam to train the style enhancement module with an initial learning rate of 0.0001. Then we keep it fixed and use a fixed learning rate of 0.0001 and SGD to train the overlooking style module. We report mean average percision (mAP) with an IoU threshold of 0.5 during test. We set $\beta = 0.5$, $h = 0.8$ as default. Particularly, $h = 0.6$ for KITTI → Cityscapes. The student model is used for test because it learns faster.

4.1. Comparison with state-of-the-art methods

We compare our LODS with the state-of-the-art SFOD and UDAOD methods. SFOD methods are SED [23] and SOAP [36]. UDAOD methods have distribution alignment based DA-Faster [5], SWDA [28], HTCN [4], SSA-DA [43], PD [33], SAPNet [20], iFAN [45], ATF [10], VDD [34], MeGA-CDA [31], SGA-S [41], CST-DA [42], DBGL [3]; NL [16] based on pseudo labels; sample generation based DM [17], AFAN [32]; auxiliary model based MTOR [2], UMT [7], ICR-CCR [37]. Source Only and Oracle represent Faster R-CNN [27] trained on source domain data and target domain data respectively. The results in the tables are cited from their papers.

Pascal → **Clipart**. In this scenario, we transfer object detector from real images to clipart-style images, where there is a huge domain shift. Table 1 shows the detection results after adaptation, from which our proposed method LODS achieves the state-of-the-art performance with a mAP of 45.2%, which means we gain the mAP of SFOD method by +15.3% (from 29.9% to 45.2%). Compared with the state-of-the-art methods which can access data, our method significantly boosts the mAP by +1.1% (from 44.1% to 45.2%), which not only strongly demonstrates the effectiveness of our proposed method, but also indicates that the knowledge in source data is not fully explored and trans-

[3]https://naoto0804.github.io/cross_domain_detection
[4]https://github.com/tiancity-NJU/da-faster-rcnn-PyTorch
[5]https://github.com/tiancity-NJU/da-faster-rcnn-PyTorch
[6]http://www.cvlibs.net/datasets/kitti/

Table 3. Detection results on **Cityscapes** → **Foggy-Cityscapes**.

Methods	Pson	Rder	Car	Tuck	Bus	Tain	Mcle	Bcle	mAP
Source Only	25.8	33.3	35.2	13.0	26.4	9.1	19.0	32.3	24.3
DA-Faster [5]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
SWDA [28]	29.9	42.3	43.5	24.5	36.2	32.6	35.3	30.0	34.3
DM [17]	30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
MTOR [2]	30.6	41.4	44.0	21.9	38.6	**40.6**	28.3	35.6	35.1
iFAN [45]	32.6	40.0	48.5	27.9	**45.5**	31.7	22.8	33.0	35.3
SED [23]	21.7	44.0	40.4	32.6	11.8	25.3	34.5	34.3	30.6
SED(Mosaic) [23]	25.5	44.5	40.7	**33.2**	22.2	28.4	34.1	39.0	33.5
HCL [13]	26.9	**46.0**	41.3	33.0	25.0	28.1	**35.9**	**40.7**	34.6
SOAP [36]	**35.9**	45.0	48.4	23.9	37.2	24.3	31.8	37.9	35.5
Our method	34.0	45.7	**48.8**	27.3	39.7	19.6	33.2	37.8	**35.8**
Oracle	37.2	48.2	52.7	35.2	52.2	48.5	35.3	38.8	43.5

ferred to target domain which was also stated in [23].

Pascal → **Watercolor**. In this scenario, we adapt the detector from real images to the watercolor-style images. As shown in Table 2, our method LODS achieves the state-of-the-art performance for both tasks with a mAP of 58.2% after adaptation where the domain shift is also large. Compared with SFOD method, the performance is improved by +7.2%, which demonstrates the strong applicability of our method for different styles of images. While SOAP [36] does not work because of bad quality of generated samples due to large domain gap.

Cityscapes → **Foggy-Cityscapes**. Compared with the before mentioned scenarios, this scenario, across different weather conditions, is not so hard. Since the backgrounds are similar, the mean of target domain images is utilized as a style image for better representing the style of this domain. As shown in Table 3, our method achieves the state-of-the-art performance for the SFOD task. Compared with popular domain adaptive object detection methods, competitive performance is also achieved. Furthermore, we can also see from Table 3 that the improvement of our method is not large, because these two domains are very similar.

KITTI → **Cityscapes**. In this scenario, we evaluate the adaptation performance of our method across different cameras, as shown in Table 4. Cause we do not utilize Mosaic [1] in all experiments, for a fairer comparison, here we do not compare the performance of SED(Mosaic) [23]. As shown in Table 4, our method LODS achieves 43.9% in this adaptation scenario and comparable performance compared to many recent methods which can access source data. Compared with the SFOD methods, our method also achieves better performance.

Qualitative Comparison. In Fig. 4, we visualize the detection results from **KITTI** → **Cityscapes** and **Cityscapes** → **Foggy-Cityscapes**. We compare our method with Faster R-CNN [27], SED [23], and SED(Mosaic) [23]. Obviously, our method is capable of detecting more objects while ensuring accuracy. In particular, on **Foggy-Cityscapes**, even if some objects are heavily obscured by fog, our method is still able to detect them accurately.

Table 4. Detection results on **KITTI** → **Cityscapes**.

Methods	AP on car	Methods	AP on car
Source Only	39.2	SSA-DA [43]	43.3
ATF [10]	42.1	SAPNet [20]	43.4
MeGA-CDA [31]	43.0	SGA-S [41]	43.5
NL [16]	43.0	CST-DA [42]	43.6
SED [23]	43.6	SOAP [36]	42.7
Our method	**43.9**	Oracle	49.9

Table 5. Ablation study on **Pascal**→ **Watercolor** and **Pascal**→ **Clipart**. ENH, TRA, and RAN stand for style enhancement, style transfer and random augmentation, respectively.

Methods	Enhancement			Removal		mAP	
	ENH	TRA	RAN	GIA	GCIA	Water	Clipart
Source Only	×	×	×	×	×	46.1	26.6
			√	√	√	53.1	33.2
		√		√	√	55.4	39.8
	√				√	56.6	44.5
LODS	√			√	√	**58.2**	**45.2**

4.2. Further Analysis

Ablation Study. To explore the effectiveness of different modules during adaptation, as shown in Table 5, we conduct ablation study on the transfer scenarios **Pascal** → **Watercolor** and **Pascal** → **Clipart**. (1) With the same alignment modules, to demonstrate the superiority of our style enhanced module, we utilize a random image augmentation, denoted as RAN. Its mAP is 53.1% for Watercolor (33.2% for Clipart). While directly using style transfer (TRA) outperforms the random augmentation by +2.3% for Watercolor (+6.6% for Clipart); using style enhancement (ENH) improves more +2.8% for Watercolor (+5.4% for Clipart). It clearly shows that our style enhancement technique works. (2) With the same style enhancement module, we demonstrate the effects of the modules GIA and GCIA. As shown in Table 5, by only using GCIA, its mAP is 56.6% for Watercolor (44.5% for Clipart). While using both modules GIA+GCIA, the performance is improved +1.6% for Watercolor (+0.7% for Clipart). This ablation study demonstrates that not only our proposed style enhancement model is highly effective, but also the two alignment constrains really help the object detector to overlook domain style.

Hyper-parameters Sensitivity. We perform sensitivity analysis on h and β under the adaptation scenario of

Figure 4. Illustration of the detection results on the target domain compared our method with SED and SED(Mosaic). The first and second lines represent the scenarios of **KITTI** → **Cityscapes** and **Cityscapes** → **Foggy-Cityscapes**, respectively. *Zoom in for best view.*

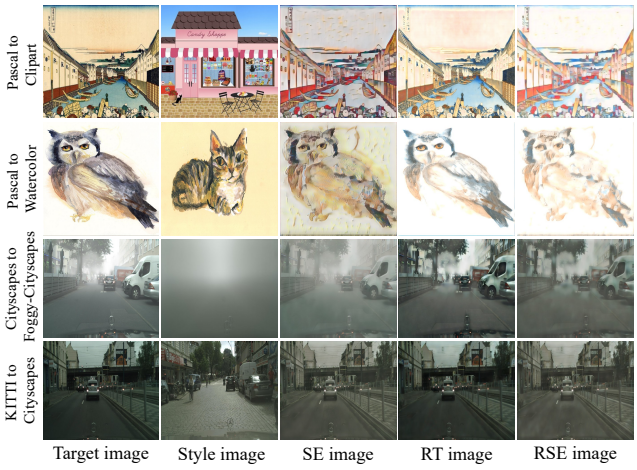

Figure 5. Validation of the learned ability to overlook target domain style. The SE, RT and RSE images stand for style enhanced image, the reconstructed image with target feature and the reconstructed image with style enhanced feature. *Zoom in for best view.*

Pascal→Watercolor. As shown in Fig. 6, our model can keep a relatively stable result in a wide range of h and β. As we expected, a high h leads to a decrease in the knowledge exploration ability, and a low h leads to the explosion of the number of misclassified samples which bring excessive noises to the model. Note that 0.8 and 0.5 are the most appropriate hyperparameters for h and β respectively, so we fix these two hyperparameters for most of experiments.

Visualization of overlooking domain style. In order to demonstrate that the object detector does learn the ability to overlook target domain style, based on the source model, we first use target data to train a decoder, which is stated in the style enhancement module. After that, the target feature and the corresponding style enhanced version are extracted from the adapted model and then are fed into the decoder to reconstruct the corresponding images. As shown in Fig. 5, the reconstructed images have significantly less target domain style compared with the original target images (e.g. Target image vs RT image; SE image vs RSE image). It indicates that the object detector has indeed learned the ability to overlook target domain style, and adequately demonstrates the correctness of the proposed method. The semantic inconsistency in RSE image indicates that the detector is able to ignore the unimportant details. It can be observed that style enhancement is also not enough. A more efficient style enhancement method is worthy to be further explored.

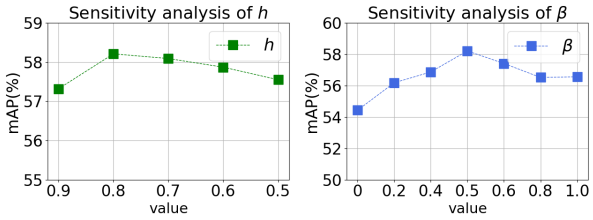

Figure 6. Hyperparameter analysis with respect to h and β on **Pascal→ Watercolor**.

5. Conclusion

We proposed a novel strategy which endows the detector an ability of learning to overlook domain style. In this way, the object detector can be adapted to the new scenario. Based on this strategy, a style enhancement module and overlooking style module are proposed. Compared with the sample generation line, style enhancement to the target image is easier. The overlooking style module also employs all target samples to help detector adaptation while self-supervised learning strategy only uses high-confidence samples. Experiments confirm the effectiveness of our method. Moreover, not limited to object detection, theoretically the proposed method can also be applied to other tasks, e.g. classification or semantic segmentation.

Acknowledgement

This work was supported in part by the National Key R&D Program of China (2018YFE0203900) and Sichuan Science and Technology Program (2020YFG0476). Mao Ye and Xiatian Zhu are corresponding authors.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 1, 7

[2] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 5, 6, 7

[3] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2703–2712, October 2021. 2, 6

[4] 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 (CVPR)*, June 2020. 2, 5, 6

[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 (CVPR)*, June 2018. 1, 2, 6, 7

[6] 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 (CVPR)*, June 2016. 6

[7] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4091–4101, June 2021. 2, 6

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 5

[9] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 6

[10] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 309–324, Cham, 2020. Springer International Publishing. 6, 7

[11] Yunzhong Hou and Liang Zheng. Visualizing adapted knowledge in domain transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13824–13833, June 2021. 1, 2

[12] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6891–6902, June 2021. 4

[13] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 7

[14] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3

[15] 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 (CVPR)*, June 2018. 5, 6

[16] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G. Macready. A robust learning approach to domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2, 6, 7

[17] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 4, 6, 7

[18] Youngeun Kim, Donghyeon Cho, Kyeongtak Han, Priyadarshini Panda, and Sungeun Hong. Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence*, pages 1–1, 2021. 1, 2

[19] Vinod K. Kurmi, Venkatesh K. Subramanian, and Vinay P. Namboodiri. Domain impression: A source data free domain adaptation method. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 615–625, January 2021. 1, 2

[20] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 481–497, Cham, 2020. Springer International Publishing. 6, 7

[21] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2

[22] Shuai Li, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Category dictionary guided unsupervised domain adaptation for object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):1949–1957, May 2021. 2, 6

[23] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8474–8481, May 2021. 1, 2, 6, 7

[24] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In Hal Daumé III and Aarti

Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6028–6039. PMLR, 13–18 Jul 2020. 1, 2

[25] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2

[26] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2664–2672, New York, New York, USA, 20–22 Jun 2016. PMLR. 4

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 1, 6, 7

[28] 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 (CVPR)*, June 2019. 2, 5, 6, 7

[29] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 6

[30] Jiayi Tian, Jing Zhang, Wen Li, and Dong Xu. Vdm-da: Virtual domain modeling for source data-free domain adaptation. *arXiv preprint arXiv:2103.14357*, 2021. 1, 2

[31] 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 (CVPR)*, pages 4516–4526, June 2021. 6, 7

[32] Hongsong Wang, Shengcai Liao, and Ling Shao. Afan: Augmented feature alignment network for cross-domain object detection. *IEEE Transactions on Image Processing*, 30:4046–4056, 2021. 2, 6

[33] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 6

[34] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. Vector-decomposed disentanglement for domain-invariant object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9342–9351, October 2021. 2, 6

[35] Haifeng Xia and Zhengming Ding. Structure preserving generative cross-domain learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 5

[36] Lin Xiong, Mao Ye, Dan Zhang, Yan Gan, Xue Li, and Yingying Zhu. Source data-free domain adaptation of object detector through domain-specific perturbation. *International Journal of Intelligent Systems*, 36(8):3746–3766, 2021. 1, 2, 6, 7

[37] 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 (CVPR)*, June 2020. 2, 6

[38] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-Wasserstein learning for graph matching and node embedding. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6932–6941. PMLR, 09–15 Jun 2019. 5

[39] 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 (CVPR)*, June 2020. 5, 6

[40] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 4

[41] Chong Zhang, Zongxian Li, Jingjing Liu, Peixi Peng, Qixiang Ye, Shijian Lu, Tiejun Huang, and Yonghong Tian. Self-guided adaptation: Progressive representation alignment for domain adaptive object detection. *IEEE Transactions on Multimedia*, pages 1–1, 2021. 2, 6, 7

[42] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 86–102, Cham, 2020. Springer International Publishing. 2, 6, 7

[43] Zhen Zhao, Yuhong Guo, and Jieping Ye. Bi-dimensional feature alignment for cross-domain object detection. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 671–686, Cham, 2020. Springer International Publishing. 2, 6, 7

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[45] Chenfan Zhuang, Xintong Han, Weilin Huang, and Matthew Scott. ifan: Image-instance full alignment networks for adaptive object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13122–13129, Apr. 2020. 6, 7