electronics

MDPI

Article

Style-Guided Adversarial Teacher for Cross-Domain Object Detection

Longfei Jia, Xianlong Tian *, Yuguo Hu, Mengmeng Jing, Lin Zuo and Wen Li

Qingshuihe Campus, University of Electronic Science and Technology of China, Chengdu 611731, China; jialongfei@uestc.edu.cn (L.J.); 202222090533@std.uestc.edu.cn (Y.H.); mmjing@uestc.edu.cn (M.J.); linzuo@uestc.edu.cn (L.Z.); liwen@uestc.edu.cn (W.L.)

* Correspondence: 202221081215@std.uestc.edu.cn

Abstract: The teacher–student framework is widely employed for cross-domain object detection. However, it suffers from two problems. One is that large distribution discrepancies will cause critical performance drops. The other is that the samples that deviate from the overall distributions of both domains will greatly mislead the model. To solve these problems, we propose a style-guided adversarial teacher (SGAT) method for domain adaptation. Specifically, on the domain level, we generate target-like images based on source images to effectively narrow the gaps between domains. On the sample level, we denoise samples by estimating the probability density ratio of the 'target-style' and target distributions, which could filter out the unrelated samples and highlight the related ones. In this way, we could guarantee reliable samples. With these reliable samples, we learn the domain-invariant features through teacher–student mutual learning and adversarial learning. Extensive experiments verify the effectiveness of our method. In particular, we achieve 52.9% mAP on Clipart1k and 42.7% on Comic2k, which are 6.4% and 5.0% higher than the compared baselines.

Keywords: target-style image generation; domain adaptation; teacher–student framework; adversarial learning; sample denoising

Citation: Jia, L.; Tian, X.; Hu, Y.; Jing, M.; Zuo, L.; Li, W. Style-Guided Adversarial Teacher for Cross-Domain Object Detection. *Electronics* **2024**, *13*, 862. https://doi.org/10.3390/electronics13050862

Academic Editors: Ioannis Yiannis Kompatsiaris, Stefanos Vrochidis, Giuseppe Amato and Sotiris Diplaris

Received: 18 January 2024
Revised: 18 February 2024
Accepted: 21 February 2024
Published: 23 February 2024

1. Introduction

Cross-domain object detection (CDOD) problems have received increasing attention in the computer vision community. Usually, models excel within the training data domain. In the real world, however, data always follow different distributions. When deploying trained models to new data, serious performance drops appear due to domain gaps. Many methods have been proposed to reduce the impacts of domain gaps, including adversarial learning [1,2], curriculum learning [3], and graph-based consistency [4]. Recently, the teacher–student framework has been employed for CDOD tasks, e.g., the unbiased mean teacher (UMT) [5] and the adaptive teacher (AT) [6]. As a semi-supervised method, the teacher model can output pseudo-labels for the target domain to supervise the student model without using target domain annotations.

CDOD methods based on the teacher–student framework have demonstrated an obvious advantage compared with previous methods. Nevertheless, despite their remarkable performances, these methods still face two serious problems. First, the teacher–student framework suffers from large domain gaps. When facing large domain gaps, it is difficult to transfer knowledge learned from the source domain to the target domain, and great performance drops will be observed. Second, due to the sample diversity, some outliers mislead the alignment of the two domains, leading to a negative transfer. As shown in Figure 1, the target dataset contains some easy samples that look similar to the source dataset, while it also contains samples that look dissimilar to the source dataset. Usually, the distribution of these dissimilar samples deviates much from the overall distributions, which we call unrelated samples. Forcing the model to align the distributions of unrelated samples also leads to performance drops [7].

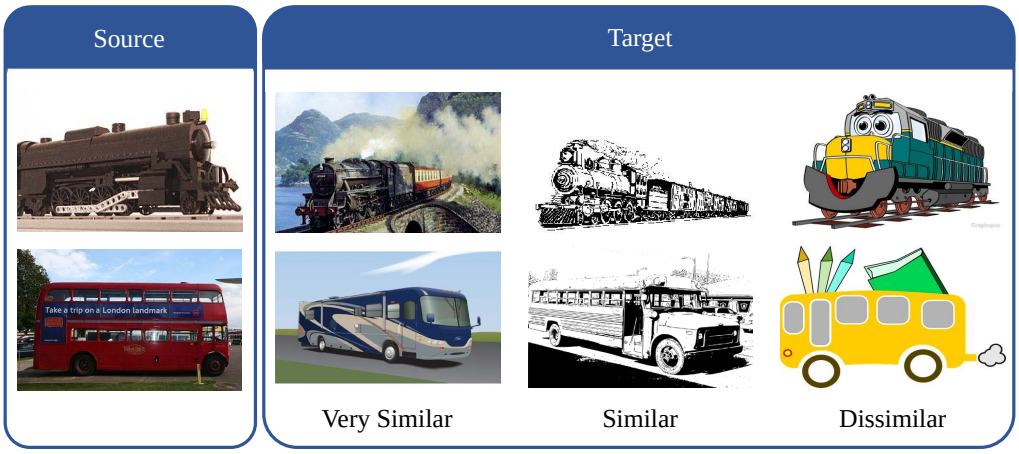
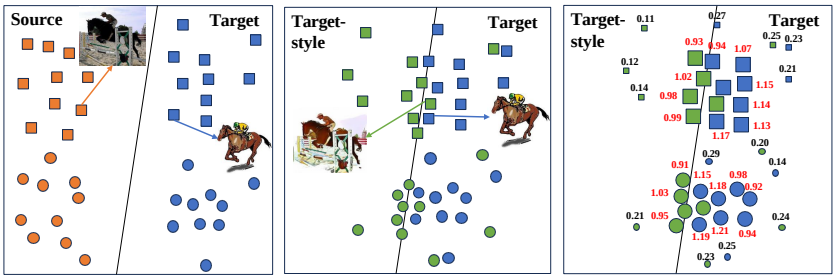


Figure 1. Examples of diverse target samples. Line 1 denotes source domain samples (real-world samples) and lines 2 to 4 are diverse target domain samples (clip art samples). As can be seen, from line 2 to line 4, the gaps between target domain samples and source domain samples are increasing, i.e., some target samples are similar to source samples while others are heavily dissimilar due to the sample diversity, which may bring challenges to CDOD.

In order to solve these two problems, we designed our style-guided adversarial teacher with three strategies. First, we introduce target-style image generation to process source images. As shown in Figure 2, we generate target-style images based on the content of source samples and styles of target samples. In this way, we can reduce the style mismatch so that domain gaps are initially alleviated. Second, we employ adversarial learning to further align the distributions across domains. Third, we adopt a sample denoising strategy to filter out the outliers to mitigate their negative impact and highlight the related ones to facilitate the alignment. The idea of our method is demonstrated in Figure 3. First, Figure 3a shows the distribution gaps between the source domain and the target domain. As shown in Figure 3b, the target-style source samples are much closer to the target samples, indicating that domain gaps are reduced after the target-like image generation and adversarial learning. Finally in Figure 3c, with the sample denoising strategy, we assign proper weights to samples to avoid the negative impact of the outliers.

(**a**) Source (**b**) Target (**c**) Target-style

Figure 2. Examples of transferred images; (**a**) the source example images from the PASCAL VOC dataset; (**b**) the target example images from the Clipart1k dataset; (**c**) the target-style images by transferring the source images (**a**) into the style of the target images (**b**).



(**a**) (**b**) (**c**)

Figure 3. (**a**) Distributions of the source domain and target domain. (**b**) Distributions of target-style domain and target domain, where the gaps between domains are mitigated. (**c**) Sample denoising has filtered out unrelated samples in both domains. Squares and circles represent different categories and different colors represent different domains.

The contributions of our work are summarized as follows:

- We propose to jointly employ target-like image generation and adversarial learning to alleviate the distribution shifts between domains. First, we transfer the styles of the source samples to generate target-like samples so that the target domain information can be expressed in the target-like samples, which could narrow the gaps initially. Then, we apply adversarial learning until the two domains cannot be distinguished clearly, which indicates that the two domains are aligned maximally.
- We propose using a sample denoising strategy to filter out unrelated samples and highlight related ones. We evaluate the density ratio of the source and target distributions and utilize it to denoise samples in both domains. In this way, we can filter out unrelated samples by assigning them lower weights and highlight related ones by assigning them higher weights.
- We carried out extensive experiments and observed that our style-guided adversarial teacher can achieve 52.9%, 57.0%, and 42.7% mAP on Clipart1k, Watercolor2k, and Comic2k, which are 6.4%, 1.8%, and 5.0% higher than previous methods, respectively.

2. Related Works

Object detection. Object detection is an important computer vision task that aims to localize the object and try to classify it at the same time. Recently, deep neural networks have been introduced in this field. Many researchers have proposed anchor-based methods such as R-CNN [8], Fast R-CNN [9], and Faster R-CNN [10]. For Faster R-CNN, it introduces region proposal networks (RPNs) to generate region proposals and it introduces regions of interest (ROI) to generate proposal feature maps [11–14]. After this, various anchor-based methods have been proposed to improve efficiency and effectiveness [15–18]. On the other hand, many anchor-free methods have also been proposed since two-stage methods are complex and computationally expensive. YOLO [19] is a widely used algorithm whose core lies in the model's smaller size and faster calculation speed. In our work, we employ Faster R-CNN as the backbone.

Unsupervised Domain Adaptation. Unsupervised domain adaptation (UDA) aims to transfer knowledge from a labeled source domain to an unlabeled target domain to enhance the performance of the target domain. Some researchers have used the maximum mean discrepancy (MMD) [20] to calculate the discrepancy between domains and various methods have been proposed to explore the potential of different versions of MMD. For example, JAN [21] explored joint MMD, which measures the Hilbert–Schmidt norm between the kernel mean embedding of empirical joint distributions of source and target data. WDAN [22] focused on the imbalanced data distribution and introduced an auxiliary weight for each class in the source domain. Other researchers have introduced adversarial learning to align two domains. Ref. [23] aligned the source domain and the target domain on both the image level and instance level using the gradient reverse layer (GRL module) [24]. Ref. [1] only focused on aligning globally similar images instead of globally dissimilar images by employing focal

loss [25]. Ref. [26] mined the discriminative regions with clustering methods and aligned these regions across domains. Recently, the teacher–student framework has been widely employed in UDA. Ref. [27] focused on the consistency of classification and localization predictions to improve the accuracy of pseudo-labels. Ref. [28] identified the alignment and synergy between the teacher–student framework and contrastive learning to maximize beneficial learning signals. In our work, we aim to further explore the effectiveness of adversarial learning and the teacher–student framework to improve the accuracy of pseudo-labels and better align the source domain and the target domain.

Object Detection with Domain Adaptation. The domain adaptation problem in object detection tasks has attracted more attention from the literature. Adversarial learning is used to align domain distributions in [29–32]. Some have proposed annotation-level methods [17,33,34], graph matching methods [35–37], and curriculum learning [3] for tasks. As an effective semi-supervised method, the teacher–student framework has also been introduced to this field from the mean teacher (MT) [4], as well as methods based on it [5,6,38]. Due to mutual learning between the teacher model and the student model, they have gained improved robustness against data variance [5]. Nevertheless, most of these teacher–student methods still suffer from large domain gaps and distribution discrepancies. Moreover, these methods also ignore the uniqueness of different samples and assign the same weight to all samples. In our work, we introduce target-like image generation to effectively narrow domain gaps. We also introduce a sample denoising strategy to assign every sample a proper weight to manifest the uniqueness of related samples and filter out unrelated ones.

3. Method

3.1. Overview

Problem Formulation. We label source domain images, $D_s = \{(X_s, B_s, C_s)\}$, where B_s and C_s denote the bounding box annotations and the corresponding class labels, respectively. We also have target domain images, $D_t = \{(X_t)\}$, without any annotations. The goal of UDA is to utilize D_s and D_t to train a robust object detector for the target domain.

Mean teacher. Mean teacher (MT) [4] was initially proposed for semi-supervised learning, and it was first introduced to UDA in [39]. MT consists of two models: a student model and a teacher model. The student model is trained with the labeled data and updated with backpropagation. The teacher model is initialized with the same weights as the student model and updated by the exponential moving average of the student model. Hence, the teacher model is more robust since it can be seen as an ensemble of the current and earlier versions of the student model. By applying the distillation loss to enforce consistency between the teacher and the student models, the student model is also guided to become more robust.

However, MT does not explicitly address DA problems. When applying the MT model to the cross-domain object detection task, the distillation will be less effective due to unreliable predictions from the teacher model caused by domain shifts. Hence, adversarial learning is introduced to MT to align two domains.

Vanilla Adversarial Learning. To apply adversarial learning, we place a domain discriminator, D, after the feature encoder, E. We use F_i and d_i to denote the derived feature and the output of D of the i-th sample, respectively. On the one hand, D tries to distinguish where F_i is from (source or target). We state that D outputs 0 for target-like images and 1 for target images. Corresponding to the output labels, we define the probability of an image that belongs to the target domain as $D(F_i)$ and the probability belonging to the target-like domain as $1 - D(F_i)$. In summary, the discriminator loss $\mathcal{L}_{adv}$ can be written as follows:

$$\mathcal{L}_{adv} = -\sum_i d_i log D(F_i) + (1 - d_i) log(1 - D(F_i)) \quad (1)$$

On the other hand, the feature encoder, E, is encouraged to extract features that can confuse the discriminator, D. Thus, the adversarial optimization function can be written as follows:

$$\mathcal{L}_{adv} = max_E min_D \mathcal{L}_{adv}. \quad (2)$$

In order to simplify the adversarial optimization, we apply a GRL module [24] between the feature encoder E and the discriminator D. During the backpropagation, the GRL module can reverse the gradient, and we only need to minimize the objective $\mathcal{L}_{adv}$.

3.2. Target-like Image Generation

Since only the source domain has the supervision information and domain gaps make it hard to transfer knowledge to the target domain, the MT model is inevitably biased toward the source domain. Therefore, the pseudo-labels on the unlabeled target samples could be of low quality. This issue will degrade the object detection performance for the unlabeled target domain.

The key to improving the quality of pseudo-labels lies in reducing domain gaps. Hence, we introduce target-like image generation to MT, as style mismatches widely exist between the source domain and target domain as part of an intuitive understanding of domain gaps. Target-like image generation tries to transfer a source image to the style of a target image without heavily changing the content. During this process, target domain information is transferred along with the style, and the domain gaps are narrowed accordingly.

Various methods can be employed to generate target-like images, e.g., CNN, Style Bank, and CycleGAN [40–43]. Here, we employ CycleGAN—a simple yet effective approach—as our target-like image generation module. CycleGAN contains a generator G_{cyc} that tries to generate source images that are in target styles. Meanwhile, it also contains a discriminator D_{cyc} that tries to discriminate whether source images are well-transferred to target styles.

In order to ensure the CycleGAN can transfer styles across domains, we simultaneously optimize two losses: generative loss and cycle consistency loss. Generative loss can guarantee the collaborative optimization of the generator, G_{cyc}, and the discriminator, D_{cyc}, so that generated pictures can be more realistic. Meanwhile, cycle consistency loss guarantees that generated pictures can reserve original content.

Following the above directions, we feed both X_s and X_t to generate target-like images, $D_{ts} = \{(X_{ts}, B_s, C_s)\}$ (our target-like image generation module does not require any kind of annotations). D_{ts} reserve their original content and annotations but are in the style of the target domain. In this way, domain gaps between X_{ts} and X_t are smaller than those between X_s and X_t. We then train our model with X_{ts} to effectively mitigate domain gaps and facilitate the adaptation.

Moreover, since style mismatches have been alleviated, it is more difficult for the discriminator to distinguish the target-like samples from the target samples, which is helpful in training a more discriminative model.

3.3. Denoised Adversarial Training

Although adversarial learning is helpful in aligning two domains, it still has limitations that come from ignoring the unrelated samples of both domains. Vanilla adversarial learning follows the assumption that all samples are equally useful. Nevertheless, unrelated samples deviate significantly from the target domain. These unrelated samples will mislead the alignment of both domains, causing negative transfer. Based on this observation, we propose a sample denoising strategy to optimize adversarial learning.

Firstly, for the target-like domain, to alleviate the impact of the unrelated samples, a natural idea is to reweight each target-like sample in a proper manner. Referring to Section 3.1 and [44], since the output of the discriminator represents the probability of a sample that belongs to the target domain, we reuse it to design the density ratio of each sample:

$$w_{x_{tsi}} = SG\left(\frac{D(x_{tsi})}{\overline{D(x_t)} + \epsilon}\right). \tag{3}$$

Here, we denote the number of target samples as N_t, and $\overline{D(x_t)} = \frac{1}{N_t}\sum_{i=1}^{N_t} D(x_{ti})$ denotes the average probability of all target samples. Meanwhile, ϵ is a small number applied to avoid the denominator being zero.

$\overline{D(x_t)}$ can be regarded as a representation of the target domain distribution so the density ratio $\frac{D(x_{tsi})}{\overline{D(x_t)}}$ naturally acts as an importance weight for the target-like data [44]. With N_{ts} denoting the number of target-like samples, the supervised loss can be rewritten as follows:

$$\mathcal{L}_{full-sup} = \sum_{i=1}^{N_{ts}} w_{x_{tsi}} \mathcal{L}_{full-sup}(x_{tsi}). \tag{4}$$

We can normalize the importance of weight to obtain our normalized supervised loss:

$$\mathcal{L}_{full-sup} = \sum_{i=1}^{N_{ts}} \frac{w_{x_{tsi}}}{W_{x_{ts}}} \mathcal{L}_{full-sup}(x_{tsi}), \tag{5}$$

where $W_{x_{ts}} = \sum_{i=1}^{N_{ts}} w_{x_{tsi}}$.

For a target-like sample, if it deviates from the target domain seriously, $D(x_{ts})$ will be much smaller than $\overline{D(x_t)}$. Under this circumstance, $w_{x_{ts}}$ will be very small, and this sample will be filtered out. On the contrary, if this target-like image contains enough target domain information, it can confuse the discriminator, and $D(x_{ts})$ can be larger than $\overline{D(x_t)}$. Under this circumstance, this sample can be highlighted ($w_{x_{ts}} > 1$).

Secondly, there are outliers in the target domain. To be specific, these outliers deviate from the overall distribution so that they mislead the alignment of the two domains. We should also alleviate their impact by applying the dense ratio to the semi-supervised loss, extending the supervised loss, as follows:

$$\mathcal{L}_{semi-sup} = \sum_{i=1}^{N_t} SG(\frac{D(x_{ti})}{\overline{D(x_{ts})} + \epsilon}) \mathcal{L}_{semi-sup}(x_{ti}). \tag{6}$$

Here, $\overline{D(x_{ts})} = \frac{1}{N_{ts}}\sum_{i=1}^{N_{ts}} D(x_{tsi})$ denotes the average probability of all target-like samples. $\overline{D(x_{ts})}$ can be regarded as a representation of the target-like domain. Similar to $\mathcal{L}_{full-sup}$, we can normalize the importance weight to obtain our normalized semi-supervised loss, as follows:

$$\mathcal{L}_{semi-sup} = \sum_{i=1}^{N_t} \frac{w_{x_{ti}}}{W_{x_t}} \mathcal{L}_{semi-sup}(x_{ti}). \tag{7}$$

where $w_{x_{ti}} = SG(\frac{D(x_{ti})}{\overline{D(x_{ts})}+\epsilon})$ and $W_{x_t} = \sum_{i=1}^{N_t} w_{x_{ti}}$.

For a target sample that disobeys the distributions of both domains, the discriminator will not regard it as a target sample, and $D(x_t)$ will be small. If $D(x_t)$ is lower than $\overline{D(x_{ts})}$, it suggests that this target sample deviates from the target domain more than target-like samples do. Under this circumstance, it will be regarded as an outlier and it will be filtered out ($w_{x_t} < 1$).

In summary, by estimating the density ratio of the target-like and target distributions, we can filter out unrelated samples to alleviate their negative impact on the domain adaptation and highlight related ones to facilitate the alignment.

3.4. Full Objective and Inference

The overview of our framework is presented in Figure 4. Our framework consists of three parts: a target-like image generation module, a teacher model for pseudo-labels, and a student model for feature extraction. The target-like image generation module takes in X_s and X_t. The student model takes in both D_{ts} and D_t, while the teacher model only takes in D_t. First, we obtain target-like images from the target-like image generation module.

Next, We train our student model with the labeled target-like images. Then we initialize the weights of the teacher model through the student model. Finally, at the stage of mutual learning, the teacher model generates pseudo-labels to train the student model, and the student model updates its weights to the teacher model via the exponential moving average (EMA). To further align the two domains, we optimize adversarial learning with our sample denoising strategy and employ it in the framework.

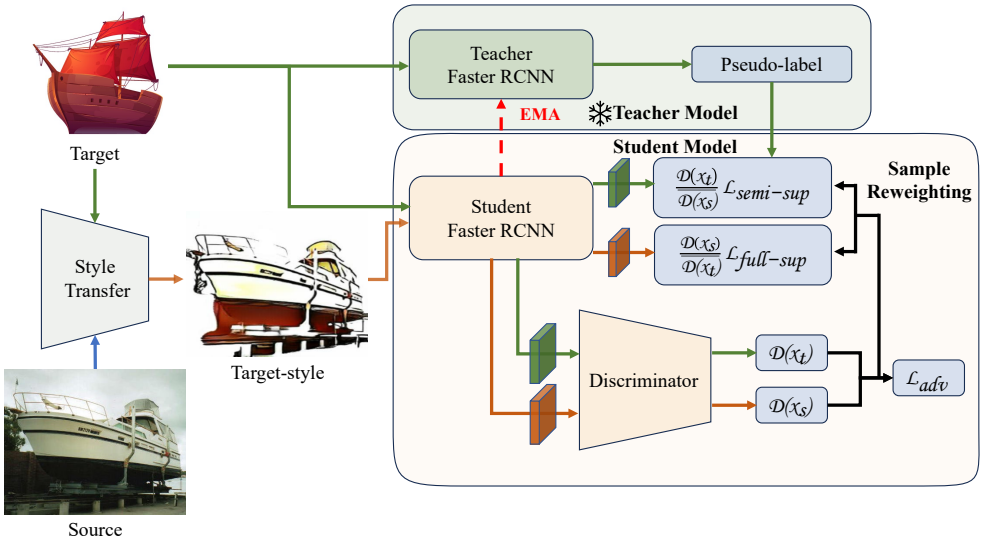


Figure 4. Overview of our proposed style-guided adversarial teacher (SGAT). We first transfer the source images to the target styles. Then, the target-like images and the real target images are fed into the student model to extract the features that cannot be distinguished by the discriminator. Moreover, the teacher model generates pseudo-labels to teach the student model while the student model updates the teacher model through the exponential moving average (EMA). Finally, we employ a sample denoising strategy to filter out the unrelated samples and highlight the related ones.

The total loss $\mathcal{L}$ for training our proposed style-guided adversarial teacher is summarized as follows:

$$\mathcal{L} = \mathcal{L}_{full-sup} + \lambda_{semi-sup}\mathcal{L}_{semi-sup} + \lambda_{adv}\mathcal{L}_{adv}, \tag{8}$$

where hyperparameters $\lambda_{semi-sup}$ and λ_{adv} are used to control the weights of losses. Note that $\mathcal{L}_{full-sup}$ and $\mathcal{L}_{semi-sup}$ are introduced to update the feature encoder and decoder while $\mathcal{L}_{adv}$ is introduced to update the feature encoder and the discriminator in the student model. The teacher model is updated only by EMA.

4. Experiments

4.1. Datasets

PASCAL VOC. PASCAL VOC [45] is a dataset consisting of real-world images, a corresponding bounding box, and class annotations of 20 categories. Following [1,46], we combine PASCAL VOC 2007 and 2012. PASCAL VOC contains 17,125 images in total.

Clipart1k. Clipart1k [47] shares the same categories with PASCAL VOC and consists of 1000 clip art images. We equally split these 1000 images into the training set and testing set, referring to the practice in [1,46].

Comic2k. Comic2k [47] contains comic-style images of 6 categories. Following [1,46], the dataset is divided into the training set and the testing set; each contains 1000 images in half.

Watercolor2k. Watercolor2k [47] contains watercolor-style images, consisting of images from 6 classes, and sharing the same classes with the Comic2k dataset. Following [1,46], the dataset is divided into the training set and the testing set, with each set containing 500 images (1000 images total).

4.2. Implementation Details

Following [1,23], we employ Faster R-CNN as the base object detection model in our method. The backbone of the Faster R-CNN is ResNet-101 [48] pre-trained on ImageNet. We scale input images by resizing their shorter sides to 600 without changing their original ratios. For hyperparameters, we set $\lambda_{semi-sup} = 1.0$ and $\lambda_{adv} = 0.1$. In order to warm up our model properly, we first train the student model for 20k iterations. Then we copy the weights of the student model to the teacher model—with every iteration using EMA—and train them together for another 20k iterations during the mutual learning. We set the learning rate as 0.01 and optimize our model using stochastic gradient descent (SGD). To improve the robustness of our model, we apply several data augmentation methods, including random horizontal flip, random color jittering, gray-scaling, Gaussian blurring, and cutting out patches. The weight smoothing coefficient of the EMA is set to 0.9996. We conduct our experiments on 2 NVIDIA TITAN RTX graphics cards. Experiments are all implemented in PyTorch torch 1.10.1+torchvision 0.11.2.

Real to synthetic adaptation. For this setting, we test our model on the domain shifts between the real images and the synthetic images. We use Pascal VOC as our source image dataset and Clipart1k, Watercolor2k, or Comic2k as our target image datasets. We report the average precision (AP) of each class for our experimental setting.

4.3. Domain Distance

Before we implement experiments with our teacher–student framework, we first measure the domain distances between the source domain, the target domain, and the target-style domain. Strictly speaking, we estimate the domain divergence using the $\mathcal{H}\Delta\mathcal{H}$-divergence [47]. Nevertheless, because of the complexity of calculating the $\mathcal{H}\Delta\mathcal{H}$-divergence, we use $\mathcal{A}$-distance as a proxy of the $\mathcal{H}\Delta\mathcal{H}$-divergence to calculate domain distances. The $\mathcal{A}$-distance is defined as follows:

$$\mathcal{A}(D_s, D_t) = 2(1 - 2\theta),$$

where θ denotes the domain classification error of a binary classifier for discriminating the source domain from the target domain. Two domains are closer when the $\mathcal{A}$-distance is smaller.

The results of $\mathcal{A}$-distance are shown in Table 1, and two conclusions can be drawn. For one, the $\mathcal{A}$-distance between the target-style domain and the target domain is smaller than that between the source domain and the target domain, indicating that after the target-style, image generation domain distances are truly alleviated. For another thing, the domain distance between the target-style domain and the target domain is still not completely mitigated, so it is essential to apply adversarial learning to further align two domains.

Table 1. The results of the $\mathcal{A}$-distance between the target-style domain, the source domain, and the target domain on Clipart1k, Watercolor2k, and Comic2k. A smaller $\mathcal{A}$-distance value indicates a smaller difference between two sample distributions (TS denotes the target-style domain).

Domains	$\mathcal{A}$-Distance
TS to Clipart1k	0.71620524
VOC to Clipart1k	1.11006462
TS to Watercolor2k	0.61042478
VOC to Watercolor2k	0.80800160
TS to Comic2k	0.77987614
VOC to Comic2k	1.30318528

4.4. Results and Comparisons

The results of the setting real to synthetic adaptation on PASCOL VOC , Clipart1k, Watercolor2k, and Comic2k are presented in Table 2, Table 3 and Table 4, respectively. We compare our method with the state-of-the-art methods and report the performance gaps between them.

Table 2. The results of cross-domain object detection on PASCAL VOC → Clipart1k. We use ResNet-101 pre-trained on ImageNet as the backbone.

Method	Aero	Bcycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Hrs	Mbike	prsn	plnt	Sheep	Sofa	Train	Tv	mAP
Source	**23.0**	**39.6**	**20.1**	**23.6**	**25.7**	**42.6**	**25.2**	**0.9**	**41.2**	**25.6**	**23.7**	**11.2**	**28.2**	**49.5**	**45.2**	**46.9**	**9.1**	**22.3**	**38.9**	**31.5**	**28.8 (−16.2)**
SCL [46]	44.7	50.0	33.6	27.4	42.2	55.6	38.3	**19.2**	37.9	69.0	30.1	26.3	34.4	67.3	61.0	47.9	21.4	26.3	50.1	47.3	41.5 (−3.5)
SWDA [1]	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 (−6.9)
DM [49]	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 (−3.2)
CRDA [30]	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 (−6.7)
HTCN [32]	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 (−4.7)
UMT [5]	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 (−0.9)
AT [6]	26.5	**72.3**	33.3	46.7	**63.4**	52.8	55.4	1.7	**60.6**	44.1	22.8	19.8	55.8	74.9	**81.6**	**61.5**	**33.8**	36.8	42.2	43.6	46.5 (+1.5)
SGAT	**49.6**	68.1	**42.5**	**52.3**	58.7	54.6	**56.3**	6.5	58.2	66.7	**37.3**	**42.9**	**67.9**	**93.3**	78.4	57.5	31.4	**46.2**	42.5	46.8	52.9 (+7.9)
Oracle	33.3	47.6	43.1	38.0	24.5	82.0	57.4	22.9	48.4	49.2	37.9	46.4	41.1	54.0	73.7	39.5	36.7	19.1	53.2	52.9	45.0

The bold of the text represents the highest mAP of this category among all methods. The red and blue colors represent the mAP difference between the method and Oracle.

Table 3. The results of cross-domain object detection on PASCAL VOC → Comic2k. We use ResNet-101 pre-trained on ImageNet as the backbone.

Method	Bike	Bird	Car	Cat	Dog	Person	mAP
Source	**32.5**	**12.0**	**21.1**	**10.4**	**12.4**	**29.9**	**19.7 (−34.0)**
SWDA [1]	36.4	21.8	29.8	15.1	23.5	49.6	29.4 (−24.3)
HTCN [32]	50.3	15.0	27.1	9.4	18.9	46.2	27.8 (−25.9)
DT+PL [47]	**53.0**	**23.7**	34.4	27.4	27.2	44.0	35.0 (−18.7)
AT [6]	34.0	9.1	**56.2**	24.6	38.8	**63.6**	37.7 (−16.0)
SGAT	36.7	22.3	50.0	**45.0**	**40.8**	61.4	42.7 (−11.0)
Oracle	61.9	38.9	50.8	48.9	45.2	76.6	53.7

The bold of the text represents the highest mAP of this category among all methods. The red and blue colors represent the mAP difference between the method and Oracle.

Table 4. The results of cross-domain object detection on PASCAL VOC → Watercolor2k. We use ResNet-101 pre-trained on ImageNet as the backbone.

Method	Bike	Bird	Car	Cat	Dog	Person	mAP
Source	**68.8**	**46.8**	**37.2**	**32.7**	**21.3**	**60.7**	**44.6 (−6.0)**
BDC-Faster [1]	68.6	48.3	47.2	26.5	21.7	60.5	45.5 (−5.1)
DA-Faster [23]	75.2	40.6	48.0	31.5	20.6	60.0	46.0 (−4.6)
WST-BSR [50]	75.6	45.8	49.3	34.1	30.3	64.1	49.9 (−0.7)
HTCN [32]	78.6	47.5	45.6	35.4	31.0	62.2	50.1 (−0.5)
SCL [46]	82.2	55.1	51.8	39.6	**38.4**	64.0	55.2 (+4.6)
AT [6]	83.8	34.3	**54.2**	36.1	31.6	**71.3**	51.9 (+1.3)
SGAT	**92.2**	47.2	52.0	**44.0**	37.9	68.9	57.0 (+6.4)
Oracle	61.9	38.9	50.8	48.9	45.2	76.6	50.6

The bold of the text represents the highest mAP of this category among all methods. The red and blue colors represent the mAP difference between the method and Oracle.

For PASCAL VOC to Clipart1k, we observe that our method achieves 52.9% mAP in Table 2, which is the highest among all compared methods. To be specific, our method can outperform the state-of-the-art method by 6.4% and even outperforms other methods by larger margins. Second, our method can achieve the best mAP in nine categories, showing the effectiveness of our model.

For PASCAL VOC to Comic2k, as indicated in Table 3, our model can perform well in all six categories. To be specific, our model surpasses the previous method by 5.0% and reaches the best mAP in two categories. For PASCAL VOC to Watercolor2k, similar results are shown in Table 4, and our model outperforms the previous method by 1.8%.

It is worth noting that the teacher–student framework methods have already had significant performance improvements compared to other types of methods since they

can gain improved robustness against data variance [4]. However, the teacher–student framework still suffers large domain shifts when it generates pseudo-labels. Moreover, the teacher–student framework is unable to filter out unrelated samples in either the source domain or the target domain, and these samples will lead to performance drops. With the employment of target-style image generation, adversarial learning, and sample denoising, our model can solve the above problems. For one, our model can mitigate domain gaps effectively with target-style image generation and adversarial learning. Moreover, we employ a sample denoising strategy that can filter out those unrelated samples and highlight related ones in both the target-style domain and the target domain. As a result, our model can achieve a largely improved performance compared to the previous teacher–student framework methods.

4.5. Ablation Studies

We conduct ablation studies on each of our important components to demonstrate their effectiveness.

Qualitative Results. We first provide some detection visualization results on an image from Clipart1k to demonstrate the benefit of our method in Figure 5. As we employ the target-style image generation and the adversarial learning separately, we can already effectively reduce the errors in pseudo-labels. When target-style image generation and adversarial learning are employed together, we can further improve the accuracy of pseudo-labels and make them much closer to the ground truth. To be specific, the detector can locate foreground objects and distinguish their categories better.

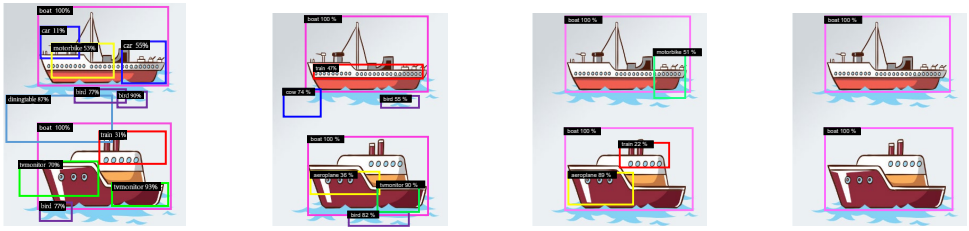
Target-style Image Generation. We benchmark the effectiveness of our target-style image generation module in Table 5, and large performance drops of more than 4.0% are observed on Clipart1k, Watercolor2k, and Comic2k, respectively, when we remove the target-style image generation module. Moreover, note that performance drops of removing the target-style image generation module are much larger than those of removing other modules on Clipart1k and Watercolor2k. These two phenomena can demonstrate that the target-style image generation module is of great importance in our method and it can alleviate domain gaps by reducing style mismatches.

Table 5. The ablation studies on our style-guided adversarial teacher. We report the mean average precision (mAP, %) on each of our important components.

Model	Clipart1k	Comic2k	WaterColor2k
SGAT	52.9	42.7	57.0
SGAT w/o sample denoising	51.7 (−1.2)	40.8 (−1.9)	56.0 (−1.0)
SGAT w/o adversarial learning	49.2 (−3.7)	37.1 (−5.6)	55.4 (−1.6)
SGAT w/o target-style image generation	47.3 (−5.6)	38.5 (−4.2)	52.5 (−4.5)

The red color represents the mAP difference between our method and our method without a certain module.

Adversarial Learning. We further analyze the importance of the adversarial learning module in our style-guided adversarial teacher from two aspects. For one, we remove the adversarial loss $\mathcal{L}_{adv}$ and report the corresponding performance in Table 5. We can see that large performance drops of up to 5.6% appear on Clipart1k, Watercolor2k, and Comic2k, respectively. Secondly, we analyze the weight λ_{adv} of the adversarial loss $\mathcal{L}_{adv}$ in Figure 6 and observe two phenomena. First, without employing the adversarial learning module, the performance of our model keeps dropping due to the domain distance (as shown with the green curve). Second, a proper weight of λ_{adv} results in proper guidance for our model. Thus, a better performance of our model is obtained (as shown by the rest three curves). From these two aspects, we can strongly demonstrate the importance and the effectiveness of the adversarial learning module in aligning two domains.



(**a**) Vanilla Mean Teacher (**b**) w/o IG (**c**) w/o $\mathcal{L}_{adv}$ (**d**) SGAT (ours)

Figure 5. Qualitative ablation studies on pseudo-labels generated on an image of clip art style. (**a**) A teacher–student framework with neither target-style image generation nor adversarial learning. (**b**) A teacher–student framework with adversarial learning but without target-style image generation (IG). (**c**) A teacher–student framework with target-style image generation but without adversarial learning. (**d**) Pseudo-labels outputted by our method.

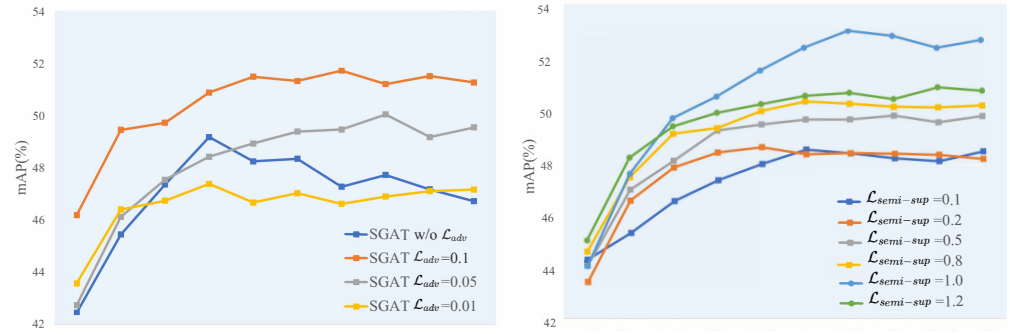


Figure 6. (**Left**) Performance curves of $\mathcal{L}_{adv}$ on the Clipart1k dataset. We run 4 identical experiments for our setting and plot the error bound accordingly in the figure. The results show that a proper weight of λ_{adv} can lead to a better and more stable performance. (**Right**) Performance curves of $\mathcal{L}_{semi-sup}$ on the Clipart1k dataset. We run 5 identical experiments for our setting and plot the error bound accordingly in the figure. The results show that in an appropriate range, weights of $\lambda_{semi-sup}$ that are larger than 0.5 can lead to better performances than the ones that are smaller than 0.5.

Semi-supervised loss with pseudo-labels. We analyze the weight $\lambda_{semi-sup}$ of the semi-supervised loss $\mathcal{L}_{semi-sup}$ in Figure 6. It can be observed that in an appropriate range, $\lambda_{semi-sup}$ larger than 0.5 can lead to better performances than $\lambda_{semi-sup}$ smaller than 0.5. With the employment of target-style image generation and adversarial learning, there are much fewer errors in pseudo-labels, and pseudo-labels are much closer to the ground truth. Under this circumstance, semi-supervised learning will be more reliable, so increasing the $\lambda_{semi-sup}$ of $\mathcal{L}_{semi-sup}$ can appropriately help align two domains more swiftly, leading to better performances.

Sample Denoising. We benchmark the effectiveness of our sample denoising strategy in Table 6. Performance improvements of more than 1.0% are observed when we add the sample denoising strategy to the teacher–student framework. Since the sample denoising strategy acts on both domains, we also benchmark the target-style sample denoising and the target sample denoising separately. The results are reported in Table 6 and improvements in both strategies can be observed, respectively. This demonstrates that unrelated samples exist in both domains and our sample denoising strategy can effectively filter them out. We also provide some visualization results to show the effectiveness of our sample denoising strategy in Figure 7. As shown in Figure 7–left, those well-transferred target-style samples are assigned large weights, and those poorly transferred target-style samples are assigned low weights. In Figure 7–right, target domain outliers are assigned low weights.

Table 6. The ablation studies on our sample denoising. We report the mean average precision (mAP, %) for each important component of our sample denoising.

Strategies	Clipart1k	Comic2k	WaterColor2k
w/o sample denoising	51.7	40.8	56.0
Only TS sample denoising	52.6 (+0.9)	42.2 (+1.4)	56.7 (+0.7)
Only Target sample denoising	52.4 (+0.7)	41.8 (+1.0)	56.6 (+0.6)
SGAT	52.9 (+1.2)	42.7 (+1.9)	57.0 (+1.0)

The blue color represents the mAP difference between our method without sample denoising and our method with certain sample denoising.

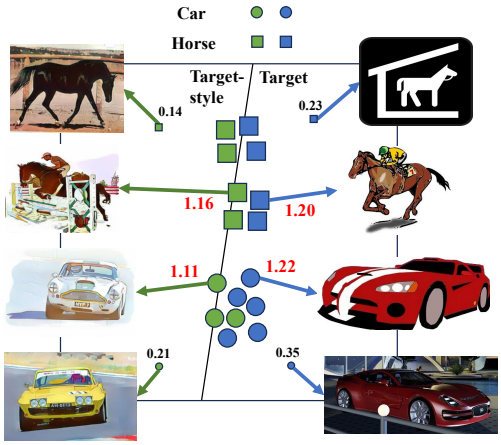


Figure 7. Visualization results of the sample denoising strategy. In this figure, the circles represent the category Car and the squares represent the category Horse. It has been observed that unrelated samples are filtered out by assigning them lower weights, and related ones are highlighted by assigning them higher weights.

4.6. Discussion

The proposed method is based on Faster R-CNN, instead of the prevalent YOLO. As a one-stage method, YOLO directly outputs predictions from feature maps. On the other hand, Faster R-CNN is a typical two-state method. It first generates region proposals with RPN, then it utilizes RoI pooling to extract fixed-size features from each region proposal for classification and bounding box regression. RoI optimizes proposals of RPN to obtain final predictions, so they are more accurate than predictions of YOLO that are directly obtained from feature maps. Therefore, based on the Faster R-CNN, the teacher model can output higher quality pseudo-labels to better guide the student model, as well as benefit the training process and the employment of our modules.

It is worth noting that the modules in our method are complementary to each other. The target-style image generation module reduces domain distances and the sample denoising strategy helps prevent negative transfer in adversarial learning to facilitate it, making our model more effective across multiple domains.

Our method has wide application perspectives. For example, autonomous driving will face various distributions in practical use, e.g., the weather may change from sunny to snowy or foggy, requiring the model to adapt snowy or foggy from sunny. Similarly, industrial quality inspection suffers from various industrial conditions, resulting in various distribution shifts. For instance, the lighting, background, and view angles of the image may exhibit significant differences. Under these circumstances, the model must be able to adapt to different distributions. Our model gains large improvements over the previous DA methods in all three datasets, demonstrating that it is effective and adaptive across domains. Hence, our model can be employed in fields that suffer from various distributions.

However, since source images need to be preprocessed by the target-style image generation module, our method cannot work in an end-to-end manner. Thus, an end-to-

end framework that integrates the online target-style image generation module and the domain adaptation module will be part of our future work.

5. Conclusions and Future Work

In this paper, we propose a new method for cross-domain object detection. First, by generating the target-like images, we can reduce the style mismatch between two domains to initially reduce domain gaps between them. Second, we employ adversarial learning to further align two domains. Moreover, we propose the sample denoising strategy to help the model filter out unrelated samples and highlight related ones in both the target-style domain and the target domain, effectively avoiding a negative transfer. The modules in our method are complementary to each other; hence, large improvements over previous DA methods are observed in all three datasets. The superior performance demonstrates that our method is effective and adaptive across domains so it can be applied to novel fields such as autonomous driving.

However, our model is unable to work in an end-to-end manner since source images need to be preprocessed by the target-style image generation module. Therefore, in future work, we may consider integrating the online target-style image generation module and the domain adaptation module for an end-to-end model.

Author Contributions: Methodology, L.J. and X.T.; validation, Y.H.; writing—original draft preparation, L.J. and X.T.; writing—review and editing, M.J., L.Z. and W.L.; visualization, X.T.; supervision, L.Z. and W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant nos. 62276054, 61877009), and the Sichuan Science and Technology Program (grant no. 2023YFG0156).

Data Availability Statement: The data presented in this study are available in The PASCAL Visual Object Classes Homepage and Aizawa-Yamakata-Matsui Laboratory at http://www.hal.t.u-tokyo.ac.jp/lab/en/index_1.xhtml, and http://host.robots.ox.ac.uk/pascal/VOC/, reference number [45,47]. These data were derived from the following resources available in the public domain: http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html#data. All data are accessed on 1 October 2023.

Acknowledgments: This work is completed with the help of my tutors, my seniors and my juniors. Thank my tutors for their suggestions on the method and paper writing and thank them for constantly encouraging and supporting me during this period. Thank my seniors and my juniors for collecting data for me and providing me with experience about submitting an article.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CDOD	cross-domain object detection
UDA	unsupervised domain adaptation
MMD	maximum mean discrepancy
EMA	exponential moving average
mAP	mean average precision

References

1. Saito, K.; Ushiku, Y.; Harada, T.; Saenko, K. Strong-weak distribution alignment for adaptive object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6956–6965.
2. Xu, M.; Wang, H.; Ni, B.; Tian, Q.; Zhang, W. Cross-domain detection via graph-induced prototype alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12355–12364.
3. Soviany, P.; Ionescu, R.T.; Rota, P.; Sebe, N. Curriculum self-paced learning for cross-domain object detection. *Comput. Vis. Image Underst.* **2021**, *204*, 103166. [CrossRef]
4. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.

5. Deng, J.; Li, W.; Chen, Y.; Duan, L. Unbiased mean teacher for cross-domain object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 4091–4101.
6. Li, Y.J.; Dai, X.; Ma, C.Y.; Liu, Y.C.; Chen, K.; Wu, B.; He, Z.; Kitani, K.; Vajda, P. Cross-domain adaptive teacher for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7581–7590.
7. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
8. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
9. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, QC, Canada, 7–12 December 2015.
11. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain, 5–10 December 2016.
12. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
13. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
14. Huang, H.; Zhou, H.; Yang, X.; Zhang, L.; Qi, L.; Zang, A.Y. Faster R-CNN for marine organisms detection and recognition using data augmentation. *Neurocomputing* **2019**, *337*, 372–384. [CrossRef]
15. Song, P.; Li, P.; Dai, L.; Wang, T.; Chen, Z. Boosting R-CNN: Reweighting R-CNN samples by RPN's error for underwater object detection. *Neurocomputing* **2023**, *530*, 150–164. [CrossRef]
16. Long, K.; Tang, L.; Pu, X.; Ren, Y.; Zheng, M.; Gao, L.; Song, C.; Han, S.; Zhou, M.; Deng, F. Probability-based Mask R-CNN for pulmonary embolism detection. *Neurocomputing* **2021**, *422*, 345–353. [CrossRef]
17. Sun, Y.; Su, L.; Luo, Y.; Meng, H.; Li, W.; Zhang, Z.; Wang, P.; Zhang, W. Global Mask R-CNN for marine ship instance segmentation. *Neurocomputing* **2022**, *480*, 257–270. [CrossRef]
18. Yi, D.; Su, J.; Chen, W.H. Probabilistic faster R-CNN with stochastic region proposing: Towards object detection and recognition in remote sensing imagery. *Neurocomputing* **2021**, *459*, 290–301. [CrossRef]
19. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
20. Gretton, A.; Borgwardt, K.M.; Rasch, M.J.; Schölkopf, B.; Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.* **2012**, *13*, 723–773.
21. Long, M.; Zhu, H.; Wang, J.; Jordan, M.I. Deep transfer learning with joint adaptation networks. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 2208–2217.
22. Yan, H.; Ding, Y.; Li, P.; Wang, Q.; Xu, Y.; Zuo, W. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2272–2281.
23. Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; Van Gool, L. Domain adaptive faster r-cnn for object detection in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3339–3348.
24. Ganin, Y.; Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 1180–1189.
25. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
26. Zhu, X.; Pang, J.; Yang, C.; Shi, J.; Lin, D. Adapting object detectors via selective cross-domain alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 687–696.
27. Deng, J.; Xu, D.; Li, W.; Duan, L. Harmonious Teacher for Cross-Domain Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 23829–23838.
28. Cao, S.; Joshi, D.; Gui, L.Y.; Wang, Y.X. Contrastive Mean Teacher for Domain Adaptive Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 23839–23848.
29. Su, P.; Wang, K.; Zeng, X.; Tang, S.; Chen, D.; Qiu, D.; Wang, X. Adapting object detectors with conditional domain normalization. In *Computer Vision—ECCV 2020; Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XI 16; Springer: Cham, Switzerland, 2020; pp. 403–419.
30. Xu, C.D.; Zhao, X.R.; Jin, X.; Wei, X.S. Exploring categorical regularization for domain adaptive object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11724–11733.
31. He, Z.; Zhang, L. Multi-adversarial faster-rcnn for unrestricted object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6668–6677.

32. Chen, C.; Zheng, Z.; Ding, X.; Huang, Y.; Dou, Q. Harmonizing transferability and discriminability for adapting object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8869–8878.
33. Bao, Z.; Luo, Y.; Tan, Z.; Wan, J.; Ma, X.; Lei, Z. Deep domain-invariant learning for facial age estimation. *Neurocomputing* **2023**, *534*, 86–93. [CrossRef]
34. Fan, C.; Liu, P.; Xiao, T.; Zhao, W.; Tang, X. Domain adaptation based on domain-invariant and class-distinguishable feature learning using multiple adversarial networks. *Neurocomputing* **2020**, *411*, 178–192. [CrossRef]
35. Li, W.; Liu, X.; Yao, X.; Yuan, Y. Scan: Cross domain object detection with semantic conditioned adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 1421–1428.
36. Li, W.; Liu, X.; Yuan, Y. Sigma: Semantic-complete graph matching for domain adaptive object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5291–5300.
37. Tian, K.; Zhang, C.; Wang, Y.; Xiang, S.; Pan, C. Knowledge mining and transferring for domain adaptive object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 9133–9142.
38. Cai, Q.; Pan, Y.; Ngo, C.W.; Tian, X.; Duan, L.; Yao, T. Exploring object relation in mean teacher for cross-domain detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11457–11466.
39. French, G.; Mackiewicz, M.; Fisher, M. Self-ensembling for visual domain adaptation. *arXiv* **2017**, arXiv:1706.05208.
40. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.
41. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
42. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision—ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings, Part II 14; Springer: Cham, Switzerland, 2016; pp. 694–711.
43. Chen, D.; Yuan, L.; Liao, J.; Yu, N.; Hua, G. Stylebank: An explicit representation for neural image style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1897–1906.
44. Wang, Z.; Dai, Z.; Póczos, B.; Carbonell, J. Characterizing and avoiding negative transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11293–11302.
45. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
46. Shen, Z.; Maheshwari, H.; Yao, W.; Savvides, M. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv* **2019**, arXiv:1911.02559.
47. Inoue, N.; Furuta, R.; Yamasaki, T.; Aizawa, K. Cross-domain weakly-supervised object detection through progressive domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5001–5009.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
49. Kim, T.; Jeong, M.; Kim, S.; Choi, S.; Kim, C. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 12456–12465.
50. Kim, S.; Choi, J.; Kim, T.; Kim, C. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6092–6101.