Domain Adaptive Object Detection via Asymmetric Tri-way Faster-RCNN

Zhenwei He[0000−0002−6122−9277] and Lei Zhang(✉)[0000−0002−5305−8543]

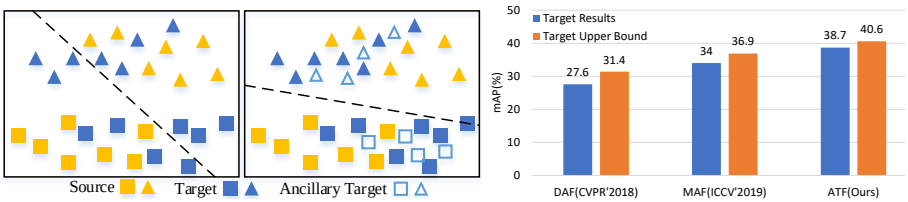
Learning Intelligence & Vision Essential (LiVE) Group
School of Microelectronics and Communication Engineering, Chongqing University
{hzw,leizhang}@cqu.edu.cn
http://www.leizhang.tk/

Abstract. Conventional object detection models inevitably encounter a performance drop as the domain disparity exists. Unsupervised domain adaptive object detection is proposed recently to reduce the disparity between domains, where the source domain is label-rich while the target domain is label-agnostic. The existing models follow a parameter shared siamese structure for adversarial domain alignment, which, however, easily leads to the collapse and out-of-control risk of the source domain and brings negative impact to feature adaption. The main reason is that the labeling unfairness (asymmetry) between source and target makes the parameter sharing mechanism unable to adapt. Therefore, in order to avoid the source domain collapse risk caused by parameter sharing, we propose an asymmetric tri-way Faster-RCNN (ATF) for domain adaptive object detection. Our ATF model has two distinct merits: 1) A ancillary net supervised by source label is deployed to learn ancillary target features and simultaneously preserve the discrimination of source domain, which enhances the structural discrimination (object classification vs. bounding box regression) of domain alignment. 2) The asymmetric structure consisting of a chief net and an independent ancillary net essentially overcomes the parameter sharing aroused source risk collapse. The adaption safety of the proposed ATF detector is guaranteed. Extensive experiments on a number of datasets, including Cityscapes, Foggy-cityscapes, KITTI, Sim10k, Pascal VOC, Clipart and Watercolor, demonstrate the SOTA performance of our method.

Keywords: Object detection; Transfer learning; Deep learning

1 Introduction

Object detection is one of the significant tasks in computer vision, which has extensive applications in video surveillance, self-driving, face analysis, medical imaging, etc. Motivated by the development of CNNs, researchers have made the object detection models fast, reliable, and precise [13,15,30,26,8,24]. However, in real-world scenarios, object detection models face enormous challenges due to the diversity of application environments such as different weathers, backgrounds, scenes, illuminations, and object appearances. The unavoidable envi-



(a) Enhancement of Classifier (b) Accuracy on Target Domain

Fig. 1. The motivation of our approach. (a) shows the learning effect of the classification decision boundary for both source and target data without/with the ancillary target features. (b) presents the target domain detection performance (mAP) based on different transfer models and the upper bound of the target performance based on their deep features and groundtruth target labels.

ronment change causes the domain shift circumstance for object detection models. Nevertheless, conventional object detection models are domain constrained, which can not take into account the domain shift happened in open conditions and noticeable performance degradation is resulted due to the environmental changes. One way to avoid the influence of domain shift is to train the detector by domain/scene specific data, but labeling the training data of each domain is time-consuming and impractical. For reducing the labeling cast and getting a domain adaptive detector, in this paper, the unsupervised domain adaptive object detection task is addressed by transferring the label-rich source domain to the label-agnostic target domain. Since the training label of the target domain is unnecessary, there is no extra annotation cast for the target domain. More importantly, detectors are more generalizable to the environmental change benefitting from the co-training between domains.

Very recently, unsupervised domain adaptive object detection is proposed to mitigate the domain shift problem by transferring the knowledge from the sematic related source domain to target domain [4,17,22,20]. Most of cross-domain object detection models learn the domain invariant features with transfer learning ideas. Inspired by the pioneer of this field [4,17], the detector can be domain-invariant if only the source and target domains are sufficiently confused. No suspicion that a domain-invariant feature representation can enable the object detector to be domain adaptive. However, the domain-invariant detector does not guarantee good object classification and bounding box (bbox) regression. This is mainly due to the lack of domain specific data. Relying solely on labeled source data and unlabeled target data to entirely eliminate the domain disparity is actually not an easy task. This is motivated in Fig. 1(a) (left), where we use a toy classifier as an example. We can see that as the features from different domains are not fully aligned, the decision boundary learned with labeled source data and unlabeled target data at hand can not correctly classify the samples from the target domain. Thus, we have an instinct thought to implicitly learn ancillary target data for domain-invariant class discriminative and bbox regres-

sive features. Specifically, to better characterize the domain-invariant detector, as is shown in Fig. 1(a) (right), we propose to learn the ancillary target features with a specialized module, which we call "ancillary net". We see that the target process features contribute to the new decision boundary of the classifier such that the target features are not only domain invariant but class separable.

Most of the CNN based domain adaption algorithms aim to learn a transferable feature representation [33,40,5,10]. They utilize the feature learning ability of CNN to extract domain-invariant representation for the source and target domains. A popular strategy to learn transferable features with CNN is adversarial training just like the generative adversarial net (GAN) [14]. The adversarial training strategy is a two-player gaming, in which a discriminator is trained to distinguish different domains, while a generator is trained to extract domain-invariant features to fool the discriminator. However, adversarial learning also has some risks. As indicated by [25], forcing the features to be domain-invariant may inevitably distort the original distribution of domain data, and the structural discrimination (intra-class compactness vs. inter-class separability) between two domains may be destroyed. This is mainly because the target data is completely unlabeled.

Similarly, distribution distortion also occurs in cross-domain object detection. Since the target data is unlabeled and the model is trained only with source labels, the learned source features can be discriminative and reliable, while the discrimination of the target features is vulnerable and untrustworthy. However, most existing models such as DAF [4] and MAF [17] default that the source and target domains share the same network with parameter sharing. A forthcoming problem of parameter sharing network is that aligning the reliable source features toward the unreliable target features may enhance the risk of source domain collapse and eventually deteriorate the structural discrimination of the model. It will inevitably bring a negative impact to object classification and bbox regression of the detector. According to the domain adaption theory in [1], the expected target risk $\epsilon_T(h)$ is upper bounded by the empirical source risk $\epsilon_S(h)$, domain discrepancy $d_{\mathcal{A}}$ and shared error $\lambda = \epsilon_T(h^*) + \epsilon_S(h^*)$ of the ideal hypothesis h^* for both domains. Therefore, effectively controlling the source risk and avoiding the collapse of the source domain is particularly important for the success of a domain adaptive detector. In this paper, we propose an **A**symmetric **T**ri-way structure to enhance the transferability of **F**aster-RCNN, which is called ATF and consists of a chief net and an ancillary net, as is shown in Fig. 2. The asymmetry originates from that the ancillary net is independent of the parameter shared chief net. Because the independent ancillary net is only trained by the labeled source data, the asymmetry can largely avoid source collapse and feature distortion during transfer.

Our model inclines to preserve the discrimination of source features and simultaneously guide the structural transfer of target features. One evidence is shown in Fig. 1 (b), in which we implement the domain adaptive detectors, such as DAF [4], MAF [17] and ATF from Cityscapes [6] dataset to the Foggy Cityscapes [34], respectively. Additionally, in order to observe the upper target

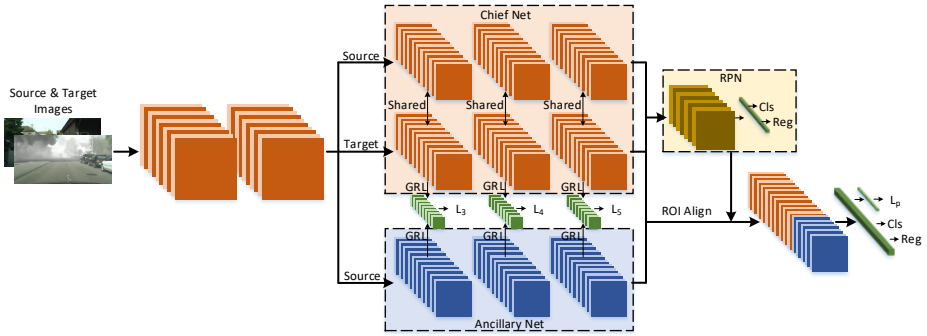


Fig. 2. The net structure of our ATF. Our ATF has three streams upon the backbone. The first two streams with shared parameters are the Chief net (orange color). Another stream is the ancillary net (blue color) which is independent of the chief net. All streams are fed into the same RPN. A ROI-Align layer is deployed to get the pooled features of all streams. The pooled features are used for final detection. The chief net is trained by ancillary data guided domain adversarial loss and source label guided detection loss. The ancillary net is trained with source label guided detection loss.

performance, we also use the features from the networks and train the detector with the ground truth target label. We can see that our ATF achieves both higher performance (11.1% and 4.7% resp.) than DAF (CVPR'18) and MAF (ICCV'19). In summary, this paper has two distinct merits. 1) We propose a source only guided ancillary net in order to learn ancillary target data for reducing the domain bias and model bias. 2) We propose an asymmetric tri-way structure, which overcomes the adversarial model collapse of the parameter shared network, i.e. the out-of-control risk in the source domain.

2 Related Work

Object detection. Object detection is an essential task of computer vision, which has been studied for many years. Boosted by the development of deep convolutional neural networks, object detection has recently achieved significant advances [23,3,29,38]. Object detection models can be roughly categorized into two types: one-stage and two-stage detection. One-stage object detection is good at computational efficiency, which attains real-time object detection. SSD [26] is the first one-stage object detector. Although SSD has a similar way to detect objects as RPN [30], it uses multiple layers for various scales. YOLO [28] outputs sparse detection results with high computation speed. Recently, RetinaDet [24] addresses the unbalance of foreground and background with the proposed focal loss. Two-stage detectors generate region proposals for the detection, for example, Faster-RCNN [30] introduced the RPN for the proposal generation. FPN [23] utilized multi-layers for the detection of different scales. NAS-FPN [12] learn the net structure of FPN to achieve better performance. In this paper, we select the Faster-RCNN as the base network for domain adaptive detection.

Domain Adaption. Domain adaption is proposed for bridging the gap between different domains. The domain adaption problem has been investigated for several different computer vision tasks, such as image classification and object segmentation [33,32,39,5]. Inspired by the accomplishment of deep learning, early domain adaptive models minimize the disparity estimate between different domains such as maximum mean discrepancy (MMD) [27,37]. Recently, several domain adaptive models were proposed based on adversarial learning. The two-player game between the feature extractor and domain discriminator promotes the confusion between different domains. Ganin *et al.* [9] proposed the gradient reverse layer (GRL), which reverses the gradient backpropagation for adversarial learning. For feature regularization, Cicek *et al.* introduced a regularization method base on the VAT. Besides the adversarial of extractor and discriminator, Saito *et al.* [32] employed two distinct classifiers to reduce the domain disparity. In our ATF, an unsupervised domain adaption mechanism is exploited.

Domain Adaptive Object Detection. Domain adaptive or cross-domain object detection has been raised very recently for the unconstrained scene. This task was firstly proposed by Chen *et al* [4], which addresses the domain shift on both image-level and instance-level. This work indicated that if the features are domain-invariant, the detector can achieve better performance on the target domain. After that, several models are also proposed for domain adaptive detection. In [31], low-level features and high-level features are treated with strong and weak alignment, respectively. Kim *et al.* [21] introduces a new loss function and exploits the self-training strategy to improve the detection of the target domain. He and Zhang proposed a MAF [17] model, in which a hierarchical structure is designed to reduce the domain disparity at different scales. In addition to the domain adaption guided detectors, the mean teacher is introduced to get the pseudo labels for the target domain [2]. Besides that, Khodabandeh *et al.* [20] train an extra classifier and use KL-divergence to learn a model for correction.

3 The Proposed ATF Approach

In this section, we introduce the details of our Asymmetric Tri-way Faster-RCNN (ATF) model. For convenience, the fully labeled source domain is marked as $D_s = \{(x_i^s, b_i^s, y_i^s)\}_i^{n_s}$, where x_i^s stands the image, b_i^s is the coordinate of bounding boxes, y_i^s is the category label and n_s is the number of samples. The unlabeled target domain is marked as $D_t = \{(x_i^t)\}_i^{n_t}$, where n_t denotes the number of samples. Our task is to transfer the semantic knowledge from D_s to D_t and achieve successful detection in the target domain.

3.1 Network Architecture of ATF

Our proposed ATF model is based on the Faster-RCNN [30] detection framework. In order to overcome the out-of-control risk of source domain in the conventional symmetric Siamese network structure with shared parameters, we introduce an asymmetric tri-way network as the backbone. Specifically, the images

from the source or target domain are fed into the first two convolution blocks. On top of that, we divide the structure into three streams. As shown in Fig. 2, the first two streams with the shared parameters in orange color are the Chief net. Features from the source and target data are fed into the two streams, respectively. The third stream with blue color is the proposed Ancillary net, which is parametrically independent of the chief net. That is, the ancillary net has different parameters from the chief net. The source only data is fed into the ancillary net during the training phase. Three streams of the network share the same region proposal network (RPN) as Faster-RCNN does. We pool the features of all streams based on the proposals with the ROI-Align layer. Finally, we get the detection results on top of the network with the pooled features.

An overview of our network structure is illustrated in Fig. 2. For training the ATF model, we design the adversarial domain confusion strategy, which is established between the chief net and ancillary net. The training loss of our ATF consists of two kinds of losses: domain adversarial confusion (**Dac**) loss in the chief net for bounding the domain discrepancy $d_{\mathcal{A}}$ and the source labeled guided detection loss (**Det**) in the ancillary net for bounding the empirical source risk $\epsilon_S(h)$. So, the proposed model is easily trained end-to-end.

3.2 Principle of the Chief Net

Domain discrepancy is the primary factor that leads to performance degradation in cross-domain object detection. In order to reduce the domain discrepancy $d_{\mathcal{A}}$, we introduce the domain adversarial confusion (Dac) mechanism which bridges the gap between the chief net (target knowledge) and the ancillary net (source knowledge). The features from the ancillary net should have a similar distribution to the target stream features from the chief net. Considering that object detection refers to two stages, i.e., image-level feature learning (global) and proposal-level feature learning (local), we propose two alignment modules based on the Dac mechanism, i.e. global domain alignment with Dac and local domain alignment with Dac.

Global domain alignment with Dac. Obviously, the Dac guided global domain alignment focuses on the low-level convolutional blocks between the chief net (target stream) and the ancillary net in ATF. Let x_i^s and x_i^t be two images from the source and target domains, respectively. The feature maps of the $k^{th}(k = 3, 4, 5)$ block of the chief net and ancillary net as shown in Fig. 2 are defined as $F_c(x_i^t, \theta_c^k)$ and $F_a(x_i^s, \theta_a^k)$, respectively. d is the binary domain label, and $d = 1$ for source domain and 0 for target domain. The Dac based global domain alignment loss for k^{th} block ($\mathcal{L}_{G-Dac}^k$) can be written as:

$$\mathcal{L}_{G-Dac}^k = -\sum_{u,v}((1-d)\log(D_k(F_c(x_i^t, \theta_c^k)^{(u,v)}, \theta_d^k) + d\log(D_k(F_a(x_i^s, \theta_a^k)^{(u,v)}, \theta_d^k)) \tag{1}$$

where the (u, v) stands for the pixel coordinate of the feature map. D_k is the discriminator of k^{th} block. θ_c^k, θ_a^k, and θ_d^k are the parameters of chief net, ancillary

net and discriminator in the k^{th} block, respectively. In principle, the discriminator D tries to minimize $\mathcal{L}^k_{G-Dac}(D)$ to distinguish the features from different domains. With the gradient reversal layer (GRL) [9], the chief net F_c and ancillary net F_a try to maximize $\mathcal{L}^k_{G-Dac}(F_c, F_a)$. Then, the global features between the chief net (target stream) and the ancillary net are confused (aligned).

Local domain alignment with Dac. The domain alignment on the global image level is still not enough for the local object based detector. Therefore, we propose to further align the local object-level features across domains. As shown in Fig. 2, the features pooled by the ROI-Align layer stand for the local part of an image, including foreground and background. Similar to the global domain alignment module, we confuse (align) the local object-level features pooled from the target stream of the chief net and the ancillary net. Suppose the pooled features from the ancillary net to be f_a and features from the chief net to be f_c, the Dac based local domain alignment loss ($\mathcal{L}_{L-Dac}$) on the local object-level features is formulated as:

$$\mathcal{L}_{L-Dac} = -\frac{1}{N}\sum_{n}((1-d)\log(D_l(F_l(f_c^n, \theta_f), \theta_d) + d\log(D_l(F_l(f_a^n, \theta_f), \theta_d)) \quad (2)$$

where the D_l is the local domain discriminator, F_l is the local backbone network, θ_l and θ_d are the parameters of the backbone and the discriminator, respectively. We implement the adversarial learning with GRL [9]. The discriminator tries to minimize $\mathcal{L}_{L-Dac}(D_l)$ while the local backbone network tries to maximize $\mathcal{L}_{L-Dac}(F_l)$ for local domain confusion.

3.3 Principle of the Ancillary net

As discussed above, the chief net aims to bound the domain discrepancy d_A. In this section, we propose to bound the empirical source risk ϵ_S by using the ancillary net. The reason why we construct a specialized ancillary net for bounding the source risk rather than using the source stream of the chief net has been elaborated. The main reason is that the chief net is parameter shared, so the empirical risk of source stream in the chief net is easily out-of-control due to the unlabeled problem of the target domain. Because the source domain is sufficiently labeled with object categories and bounding boxes in each image, the source risk ϵ_S of the ancillary net is easy to be bounded by the classification loss and regression loss of detector. In our implementation, the detection loss for the chief net is reused for the supervision of the ancillary net.

From the principles of the chief net and the ancillary net, we know that the ancillary net is trained to generate features that have a similar distribution to the target stream in the chief net as shown in Eqs.(1) and (2). That is, the ancillary net adjusts the features learned by the target stream of the chief net to adapt to the source data trained detector. Meanwhile, the ancillary net is restricted by the classifier and regressor of the source detector, such that the structural discrimination is preserved in the source domain. Therefore, with the domain alignment and source risk minimization, the expected task risk can be effectively bounded for domain adaptive object detection.

3.4 Training Loss of Our ATF

The proposed asymmetric tri-way Faster-RCNN (ATF) contains the two loss functions, the detection based source risk loss and the domain alignment loss. The detection loss function for both chief and ancillary nets is shown as:

$$\mathcal{L}_{Det} = \mathcal{L}_{cls}(x_i^s, b_i^s, y_i^s) + \mathcal{L}_{reg}(x_i^s, b_i^s, y_i^s) \tag{3}$$

where $\mathcal{L}_{cls}$ is the softmax based cross-entropy loss and $\mathcal{L}_{reg}$ is the smooth-L_1 loss, which are standard detection losses for bounding the empirical source risk. In summary, by revisiting the Eqs.(1), (2) and (3)), the total loss function for training our model can be written as:

$$\mathcal{L}_{ATF} = \mathcal{L}_{Det} + \alpha(\mathcal{L}_{L-Dac} + \sum_{k=3}^{5} \mathcal{L}_{G-Dac}^{k}) \tag{4}$$

Where the α is a hyper-parameter to adjust the weight of domain alignment loss. The model is easily trained end-to-end with Stochastic Gradient Descent (SGD). Overview of our ATF can be observed in Fig. 2.

4 Experiments

In this section, we evaluate our approach on several different datasets, including Cityscapes [6], Foggy Cityscapes [34], KITTI [11], SIM10k [19], Pascal VOC [7], Clipart [18] and Watercolor [18]. We compare our results with state-of-the-art methods to show the effectiveness of our model.

4.1 Implementation Details

The base network of our ATF model is VGG-16 [36] or ResNet-101 [16] in the experiments. We follow the same experimental setting as [4], where the source domain is fully labeled, and the target domain is completely unlabeled. The trade-off parameter α in Eq.(3) is set as 0.7 in our implementation. We use the ImageNet pre-trained model for the initialization of our model. For each iteration, one labeled source sample and one unlabeled target sample are fed into ATF for training. In the test phase, the well-trained chief net is used to get the detection results. For all datasets, we report the average precisions (AP, %) and mean average precisions (mAP, %) with a threshold of 0.5.

4.2 Datasets

Cityscapes: Cityscapes [6] captures high-quality video for the outdoor scenes in different cities for automotive vision. The dataset includes 5000 manually selected images from 27 cities, which are collected with a similar weather condition. These images are annotated with dense pixel-level image annotation. Although

the Cityscapes dataset is labeled for the semantic segmentation task, we generate the bounding box based on the pixel-level label in the experiment as [4] did for the cross-domain detection task.

Foggy Cityscapes: Foggy Cityscapes [34] dataset simulates the foggy weather based on the Cityscapes. The pixel-level labels of Cityscapes can be inherited by the Foggy Cityscapes such that we can generate the bounding box for the dataset. In the experiments, the validation set of the Foggy Cityscapes is used as the testing set in our experiment.

KITTI: KITTI [11] dataset is collected by the autonomous driving platform, which includes two color and two grayscale PointGrey Flea2 video cameras. Images of the dataset are manually selected in several different scenes in a mid-sized city. The dataset includes 14999 images and 80256 bounding boxes for the detection task. Only the training set of KITTI is used for our experiment.

SIM10K: Images of SIM10K [19] are generated by the engine of Grand Theft Auto V (GTA V). The dataset simulates different scenes, such as different time or weather. SIM10K contains 10000 images with 58701 bounding boxes of car. All images of the dataset are used for training.

Pascal VOC: Pascal VOC [7] is a famous object detection dataset. This dataset contains 20 categories with bounding boxes. The image scale of the dataset is diverse. In our experiment, the training and validation split of VOC07 and 12 are used as the training set, which results in about 15k images.

Clipart and Watercolor: The Clipart and Watercolor [18] are constructed by the Amazon Mechanical Turk, which is introduced for the domain adaption detection task. Similar to the Pascal VOC, the Clipart contains 1000 images and 20 categories. Watercolor has 2000 images of 6 categories. Half of the datasets are introduced for training while the remaining is used for the test.

4.3 Cross-domain Detection in Different Visibility and Cameras

Domain adaption across different visibility. Visibility change caused by weather can shift the data distribution of the collected images. In some application scenarios, such as autonomous driving, the detection model has to adapt to different weather conditions. In this section, we evaluate our ATF with the cityscapes [6] and the foggy cityscapes [34] datasets. We treat the Cityscapes as the source domain and the Foggy Cityscapes as the target domain. Our model uses VGG16 as the base net in the experiment. We introduce the source only trained Faster-RCNN (without adaptation), DAF [4], MAF [17], Strong-Weak [31], Diversify&match(D&match) [22], Noisy Labeling(NL) [20], and SCL [35] for the comparison. Our ATF is trained for 18 epochs in the experiment, where the learning rate is set as 0.001 and changes to 0.0001 in the 12^{th} epoch.

The results are presented in Table 1. We can see that our ATF achieves 38.7% mAP, which outperforms all the compared models. Due to the lack of domain specific data, the models which only concentrate on the feature alignment can not work well, such as MAF [17], Strong-Weak [31] and SCL [35]. With the ancillary target features from the ancillary net to reduce the domain shift and

Table 1. The cross-domain detection results from Cityscapes to Foggy Cityscapes.

Methods	person	rider	car	truck	bus	train	mcycle	bcycle	mAP
Faster-RCNN	24.1	33.1	34.3	4.1	22.3	3.0	15.3	26.5	20.3
DAF(CVPR'18) [4]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
MAF(ICCV'19) [17]	28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
Strong-Weak [31]	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
D&match [22]	30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
NL /w res101 [20]	**35.1**	42.2	49.2	30.1	45.3	27.0	26.9	36.0	36.5
SCL [35]	31.6	44.0	44.8	**30.4**	41.8	**40.7**	**33.6**	36.2	37.9
ATF(1-block)	33.3	43.6	44.6	24.3	39.6	10.5	27.2	35.6	32.3
ATF(2-blocks)	34.0	46.0	49.1	26.4	**46.5**	14.7	30.7	37.5	35.6
ATF(ours)	34.6	**47.0**	**50.0**	23.7	43.3	38.7	33.4	**38.8**	**38.7**
ATF*	34.6	46.5	49.2	23.5	43.1	29.2	33.2	39.0	37.3

bias, our model gets preferable performance. Additionally, our ATF model also outperforms the pseudo label based model [20], in which it has to generate and update the pseudo labels with features extracted by a source only trained extractor, where the target features are untrustworthy. The unreliable feature based pseudo labels can not lead to a precise target model. In order to prove the effectiveness of ancillary net, we conduct ablation studies that reduce the convolutional blocks of ancillary net. The results of 1-block (the 3^{rd} and 4^{th} blocks are removed) and 2-blocks (the 3^{rd} blocks is removed) from the ancillary net are shown in Table 1. As we reduce the convolutional block of the ancillary net, the performance drops. Otherwise, we freeze the parameter of the backbone and fine-tune the regressors and classifiers of the ATF detector with source labels. The performance of the fine-tuned model denoted as ATF* is also presented in Table 1. We find that the performance of ATF* is inferior to the ATF because the effect of ancillary target features computed by the ancillary net is removed. The merit of the proposed ancillary net is validated.

Table 2. The results of domain adaptive object detection on Cityscapes and KITTI.

Tasks	Faster-RCNN	DAF [4]	MAF [17]	S-W [31]	SCL [35]	ATF(ours)
K to C	30.2	38.5	41.0	37.9	41.9	**42.1**
C to K	53.5	64.1	72.1	71.0	72.7	**73.5**

Domain adaption across different cameras. The camera change is another important factor leading to the domain shift in real-world application scenarios. In this experiment, we employ the Cityscapes (C) [34] and KITTI (K) [11] as the source and target domains, respectively. The source only trained Faster-RCNN (without adaption), DAF [4], MAF [17], and Strong-Weak [31] are

Table 3. The cross-domain detection results from Pascal VOC to Clipart.

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	
Faster-RCNN	35.6	52.5	24.3	23.0	20.0	43.9	32.8	10.7	30.6	11.7	
DAF [4]	15.0	34.6	12.4	11.9	19.8	21.1	23.2	3.1	22.1	26.3	
BDC-Faster	20.2	46.4	20.4	19.3	18.7	41.3	26.5	6.4	33.2	11.7	
WST-BSR [21]	28.0	64.5	23.9	19.0	21.9	**64.3**	**43.5**	16.4	**42.2**	25.9	
Strong-Weak [31]	26.2	48.5	32.6	33.7	38.5	54.3	37.1	18.6	34.8	58.3	
MAF [17]	38.1	61.1	25.8	**43.9**	40.3	41.6	40.3	9.2	37.1	48.4	
SCL [35]	**44.7**	50.0	**33.6**	27.4	**42.2**	55.6	38.3	**19.2**	37.9	69.0	
ATF(ours)	41.9	**67.0**	27.4	36.4	41.0	48.5	42.0	13.1	39.2	**75.1**	
Methods	table	dog	horse	mbike	prsn	plant	sheep	sofa	train	tv	mAP
Faster-RCNN	13.8	6.0	36.8	45.9	48.7	41.9	16.5	7.3	22.9	32.0	27.8
DAF [4]	10.6	10.0	19.6	39.4	34.6	29.3	1.0	17.1	19.7	24.8	19.8
BDC-Faster	26.0	1.7	36.6	41.5	37.7	44.5	10.6	20.4	33.3	15.5	25.6
WST-BSR [21]	30.5	7.9	25.5	**67.6**	54.5	36.4	10.3	**31.2**	**57.4**	43.5	35.7
Strong-Weak [31]	17.0	12.5	33.8	65.5	**61.6**	52.0	9.3	24.9	54.1	**49.1**	38.1
MAF [17]	24.2	13.4	36.4	52.7	57.0	**52.5**	18.2	24.3	32.9	39.3	36.8
SCL [35]	30.1	**26.3**	34.4	67.3	61.0	47.9	21.4	26.3	50.1	47.3	41.5
ATF(ours)	**33.4**	7.9	**41.2**	56.2	61.4	50.6	**42.0**	25.0	53.1	39.1	**42.1**

implemented for comparisons. The AP of car on the target domain is computed for the test. The experimental results are presented in Table 2.

In Table 2, K to C represents that the KITTI is used as the source domain, while the Cityscapes is used as the target domain and vice versa. We can observe that our ATF achieves the best performance on both K to C and C to K tasks among all the compared models, which testify the effectiveness of our model in alleviating the domain shift problem caused by the change of cameras.

4.4 Cross-domain Detection on Large Domain Shift

In this section, we concentrate on the domains with large domain disparity, especially, from the real image to the comical or artistic images. We employ the Pascal VOC [7] dataset as the source domain, which contains the real image. The Clipart or Watercolor [18] is exploited as the target domain. The backbone for the experiments is the ImageNet pretrained ResNet-101. We train our model for 8 epochs with the learning rate of 0.001 and change the learning rate to 0.0001 in the 6^{th} epoch to ensure convergence.

Transfer from Pascal VOC to Clipart. The Clipart [18] contains the comical images which has the same 20 categories as Pascal VOC [7]. In this experiment, we introduce the source only Faster RCNN, DAF [4], WST-BSR [21], MAF [17], Strong-Weak [31] and SCL [35] for the comparison. The results are shown in Table 3. Our ATF achieves 42.1% mAP and outperforms all models.

Table 4. The cross-domain detection results from Pascal VOC to Watercolor.

Methods	bike	bird	car	cat	dog	person	mAP
Faster-RCNN	68.8	46.8	37.2	32.7	21.3	60.7	44.6
DAF [4]	75.2	40.6	48.0	31.5	20.6	60.0	46.0
BDC-Faster	68.6	48.3	47.2	26.5	21.7	60.5	45.5
WST-BSR [21]	75.6	45.8	**49.3**	34.1	30.3	64.1	49.9
MAF [17]	73.4	55.7	46.4	36.8	28.9	60.8	50.3
Strong-Weak [31]	**82.3**	55.9	46.5	32.7	**35.5**	66.7	53.3
ATF(ours)	78.8	**59.9**	47.9	**41.0**	34.8	**66.9**	**54.9**

Transfer from Pascal VOC to Watercolor. The Watercolor [18] dataset contains 6 categories which are the same as the Pascal VOC [7]. In this experiment, the source only trained Faster-RCNN, DAF [4], WST-BSR [21], MAF [17] and Strong-Weak [31] are introduced for comparison. The results are shown in Table 4, from which we can observe that our proposed ATF achieves the best performance among all compared models and the advantage is further proved.

4.5 Cross-domain Detection from Synthetic to Real

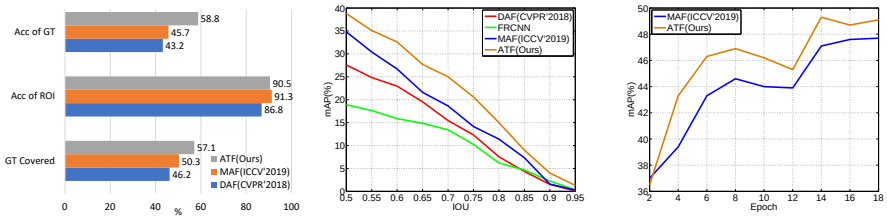
Domain adaption from the synthetic scene (auxiliary data) to the real scene is an important application scenario for domain adaptive object detection. If the domain adaption from synthetic data to real scene data is effective, the burden for the annotating the real images will be eased. In this section, we introduce the SIM10k [19] as the synthetic scene and the Cityscapes [34] as the real scene. The source only trained Faster-RCNN, DAF [4], MAF [17], Strong-Weak [31], and SCL [35] are introduced for comparisons. The AP of the car is computed for the evaluation of the experiment which is presented in Table 5. We observe that our model outperforms all the compared models. The superiority of the proposed ATF is further demonstrated for the cross-domain object detection.

Table 5. The results of domain adaptive object detection on SIM10k and Cityscapes.

Methods	F-RCNN	DAF [4]	MAF [17]	S-W [31]	SCL [35]	ATF(ours)
AP(%)	34.6	38.9	41.1	40.1	42.6	**42.8**

4.6 Analysis and Discussion

In this section, we will implement some experiments to analyze our ATF model with four distinct aspects, including parameter sensitivity, accuracy of classifiers, IOU v.s. detection performance and visualization.



(a) Accuracy of Classifiers (b) Performance v.s IOUs (c) Performance v.s Epoch

Fig. 3. Analysis of our model. (a) We test the accuracy of the classifier in the detector. **Acc of GT**: We crop the ground truth of the image and use the classifier on R-CNN to classify them. **Acc of ROI**: The accuracy of RCNN's classifier with the ROIs generated by RPN. **GT Covered**: The proportion of ground truth seen by RPN. (b) The performance change with different IOUs. Better viewed in color version. (c) The performance of source domain with different epoches.

Parameter sensitivity on α. In this part, we show the sensitivity of parameter α in Eq.(3). α controls the power of domain adaption. We conduct the cross-domain experiments from Cityscapes to Foggy Cityscapes. The sensitivity of α is shown in Table 6. When $\alpha = 0.7$, our model achieves the best performance.

Table 6. Parameter Sensitivity on α.

α	0.3	0.5	0.7	0.9	1.1	1.3
mAP(%)	36.7	37.5	38.7	38.7	38.4	38.0

The accuracy of classifiers. Our model enhances the training of the detector with the ancillary target feature. In this part, we analyze the classifier in the detector with different models. First, we static the number of ground truth boxes covered by the ROIs. If the IOU between the ROI and ground truth is higher than 0.5, we think the corresponding ROI is predicted by the RPN (region proposal network). Second, we compute the accuracy of the RCNN classifier with the ROIs from the RPN. Last, we crop the ground truth of the testing set and use the RCNN classifier to predict their label and get the accuracy. The DAF [4], MAF [17] and our model are tested with Cityscapes and Foggy Cityscapes datasets. The results are shown in Fig. 3(a). We observe that the RPN of our model finds 57.1% of the ground truth, which is the best result of all compared models. The RCNN classifiers from all three tested models achieve above 90% accuracy when classifying the generated ROIs. However, when the cropped ground truth samples are fed into the model, the accuracy of the RCNN classifier sharply drops. The ground truth samples missed by RPN are also misjudged by the RCNN in the experiments. Therefore, we experimentally find that

Fig. 4. The visualization results on the target domain (Foggy Cityscapes [34]).

it is very important to improve the recall of RPN in the cross-domain object detection task. In our ATF, benefited by the asymmetric structure which enhances the structural discrimination for the detector and the new decision boundary contributed by ancillary target features, our model achieves a preferable recall in RPN. This motivation is experimentally proved.

Detection performance w.r.t. different IOUs. The IOU threshold is an important parameter in the test phase. In the previous experiments, the IOU threshold is set as 0.5 by default. In this part, we test our model with different IOUs. The source only Faster-RCNN, DAF [4], MAF [17], and our ATF are implemented for comparison. We conduct the experiments on the Cityscapes [6] and Foggy Cityscapes [34] datasets. The results are shown in Fig. 3(b), where the IOU threshold is increased from 0.5 to 0.95. The performance drops as the IOU threshold increases in the experiment. Our model achieves the best performance on all tested IOU thresholds.

Source performance w.r.t. epoches in monitoring source risk. In this part, we monitor the training process of adaption from Cityscapes [6] to Foggy Cityscapes [34]. The mAP(%) on the test set of Cityscapes is shown in Fig. 3(c). Our ATF achieves higher mAP compared to the parameter shared MAF [17] during the training phase. Benefited by the asymmetric structure, our ATF can well prevent the collapse of the source domain and preserve the structural discrimination of source features. This experiment fully proves that parameter sharing of network will deteriorate the empirical source risk $\epsilon_S(h)$, which then leads to high target risk $\epsilon_T(h)$. Thus, a safer adaption of our ATF than the parameter shared MAF is verified.

Visualization of domain adaptive detection. Fig. 4 shows some qualitative object detection results of several models on the Foggy Cityscapes dataset [34], i.e. target domain. The state-of-the-art models, DAF [4] and MAF [17], are also presented. We can clearly observe that our ATF shows the best domain adaptive detection results and better matches the ground-truth.

5 Conclusions

In this paper, we propose an asymmetric tri-way network (ATF) to address the out-of-control problem of parameter shared Siamese transfer network for unsupervised domain adaptive object detection. In ATF, an independent network, i.e. the ancillary net, supervised by source labels, is proposed without parameter sharing. Our model has two contributions: 1) Since the domain disparity is hard to be eliminated in parameter shared siamese network, we propose the asymmetric structure to enhance the training of the detector. The asymmetry can well alleviate the labeling unfairness between source and target. 2) The proposed ancillary net enables the structural discrimination preservation of source feature distribution, which to a large extent promotes the feature reliability of the target domain. Our model is easy to be implemented end-to-end for training the chief net and ancillary net. We conduct extensive experiments on a number of benchmark datasets and state-of-the-art results are obtained by our ATF.

References

1. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. In: NeurIPS (2006)
2. Cai, Q., Pan, Y., Ngo, C.W., Tian, X., Duan, L., Yao, T.: Exploring object relation in mean teacher for cross-domain detection. In: CVPR. pp. 11457–11466 (2019)
3. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: CVPR. pp. 6154–6162 (2018)
4. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: CVPR. pp. 3339–3348 (2018)
5. Chen, Y.C., Lin, Y.Y., Yang, M.H., Huang, J.B.: Crdoco: Pixel-level domain transfer with cross-domain consistency. In: CVPR. pp. 1791–1800 (2019)
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016)
7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88**(2), 303–338 (2010)
8. Fu, C.Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C.: Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659 (2017)
9. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495 (2014)
10. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR **17**(1), 2096–2030 (2016)
11. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR. pp. 3354–3361. IEEE (2012)
12. Ghiasi, G., Lin, T.Y., Le, Q.V.: Nas-fpn: Learning scalable feature pyramid architecture for object detection. In: CVPR. pp. 7036–7045 (2019)
13. Girshick, R.: Fast r-cnn. In: ICCV. pp. 1440–1448 (2015)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS. pp. 2672–2680 (2014)

15. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. TPAMI (2018)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
17. He, Z., Zhang, L.: Multi-adversarial faster-rcnn for unrestricted object detection. In: ICCV. pp. 6668–6677 (2019)
18. Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-domain weakly-supervised object detection through progressive domain adaptation. In: CVPR. pp. 5001–5009 (2018)
19. Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., Vasudevan, R.: Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? arXiv preprint arXiv:1610.01983 (2016)
20. Khodabandeh, M., Vahdat, A., Ranjbar, M., Macready, W.G.: A robust learning approach to domain adaptive object detection. In: ICCV. pp. 480–490 (2019)
21. Kim, S., Choi, J., Kim, T., Kim, C.: Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In: ICCV. pp. 6092–6101 (2019)
22. Kim, T., Jeong, M., Kim, S., Choi, S., Kim, C.: Diversify and match: A domain adaptive representation learning paradigm for object detection. In: CVPR. pp. 12456–12465 (2019)
23. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017)
24. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017)
25. Liu, H., Long, M., Wang, J., Jordan, M.: Transferable adversarial training: A general approach to adapting deep classifiers. In: ICML. pp. 4013–4022 (2019)
26. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV. pp. 21–37. Springer (2016)
27. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Unsupervised domain adaptation with residual transfer networks. In: NeurIPS. pp. 136–144 (2016)
28. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. pp. 779–788 (2016)
29. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
30. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. TPAMI (6), 1137–1149 (2017)
31. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: CVPR. pp. 6956–6965 (2019)
32. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR. pp. 3723–3732 (2018)
33. Saito, K., Yamamoto, S., Ushiku, Y., Harada, T.: Open set domain adaptation by backpropagation. In: ECCV. pp. 153–168 (2018)
34. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. IJCV **126**(9), 973–992 (2018)
35. Shen, Z., Maheshwari, H., Yao, W., Savvides, M.: Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. arXiv preprint arXiv:1911.02559 (2019)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
37. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: ECCV. pp. 443–450. Springer (2016)

38. Vu, T., Jang, H., Pham, T.X., Yoo, C.: Cascade rpn: Delving into high-quality region proposal network with adaptive convolution. In: NeurIPS. pp. 1430–1440 (2019)
39. Wang, Q., Breckon, T.P.: Unsupervised domain adaptation via structured prediction based selective pseudo-labeling. arXiv preprint arXiv:1911.07982 (2019)
40. Xu, R., Li, G., Yang, J., Lin, L.: Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In: ICCV. pp. 1426–1435 (2019)