Multi-adversarial Faster-RCNN for Unrestricted Object Detection

Zhenwei He Lei Zhang*
School of Microelectronics and Communication Engineering, Chongqing University
Shazheng street No.174, Shapingba District, Chongqing 400044, China
{hzw,leizhang}@cqu.edu.cn

Abstract

Conventional object detection methods essentially suppose that the training and testing data are collected from a restricted target domain with expensive labeling cost. For alleviating the problem of domain dependency and cumbersome labeling, this paper proposes to detect objects in unrestricted environment by leveraging domain knowledge trained from an auxiliary source domain with sufficient labels. Specifically, we propose a multi-adversarial Faster-RCNN (MAF) framework for unrestricted object detection, which inherently addresses domain disparity minimization for domain adaptation in feature representation. The paper merits are in three-fold: 1) With the idea that object detectors often becomes domain incompatible when image distribution resulted domain disparity appears, we propose a hierarchical domain feature alignment module, in which multiple adversarial domain classifier submodules for layer-wise domain feature confusion are designed; 2) An information invariant scale reduction module (SRM) for hierarchical feature map resizing is proposed for promoting the training efficiency of adversarial domain adaptation; 3) In order to improve the domain adaptability, the aggregated proposal features with detection results are feed into a proposed weighted gradient reversal layer (WGRL) for characterizing hard confused domain samples. We evaluate our MAF on unrestricted tasks including Cityscapes, KITTI, Sim10k, etc. and the experiments show the state-of-the-art performance over the existing detectors.

1. Introduction

Object detection is a computer vision task which draws many researchers' attentions. Inspired by the development of CNN [14, 17, 34], object detection has witnessed a great success in recent years [11, 21, 29, 28].

Although excellent results have been achieved, object detection in practical application still faces a bottleneck challenge, i.e., detecting objects in the wild where domain shifts always happen. Since the collected datasets [2, 6] are still domain restricted, the trained detectors are difficult to adapt to another domain due to the domain discrepancy between the training data and the testing data it will apply to. Most of conventional detection methods do not take into account the domain discrepancy, which leads to a prominent performance degradation in practice. The influence of domain disparity can be observed in the Figure 1, where we train a VGG16 based Faster-RCNN [29] with the Cityscapes [4] and test the model on Foggy Cityscapes [32]. The results in the second row in Figure 1 verify our idea that a considerable performance drop with many objects missing when the domain disparity exists.

Figure 1. Examples of unrestricted object detection. The first row denotes the pictures from the Cityscapes [4], while pictures of the last two rows are detected from the Foggy Cityscapes [32]. The results of the first two rows are detected by the traditional Faster-RCNN [29] trained on Cityscapes, and we can see that many objects are missing on the domain shifted Foggy Cityscapes (the second row). The third row shows the results of our approach, and the domain disparity between two datasets can be effectively removed.

Generally, it's difficult to quantitatively remove the do-

*Corresponding author

main discrepancy, therefore, for addressing unrestricted object detection challenge, we exploit the mind of domain adaptation and transfer learning [23, 25, 31, 38] in our detector. In our paradigm, we train the detector on the completely unlabeled target domain, by leveraging a semantic related but distribution different source domain with sufficient labels of bounding boxes. In this way, the domain-invariant features can be learned and there is no any annotation cast for target domain. An example of our proposed detector can be observed in Figure 1 (the third row), which shows much better performance than the results of the second row with conventional Faster-RCNN model.

Specifically, we propose a **m**ulti-**a**dversarial **F**aster-RCNN detector (MAF) for adversarial domain adaptation for hierarchical domain features and the proposal features. The hierarchical domain features from the convolutional feature maps progressively present the object position information in the whole image. The proposal features extracted in fully-connected layers can better characterize the semantic information of the generated proposals. In our MAF, we propose multiple adversarial submodules for both domain and proposal features alignment. With similar task, Chen *et al*. [3] proposed a domain adaptive Faster-RCNN (DAF) which demonstrate also that the detector was domain incompatible when image-level distribution difference exists. That is, if the domain feature is aligned, the detector will become domain invariant. Inspired by the wonderful Bayesian perspectives in [3], we focus on the hierarchical domain feature alignment module by designing multiple adversarial domain classifier on each block of the convolution layers for minimizing the domain distribution disparity.

In the proposed MAF, we take into account three important aspects: (1) The multiple domain classifier submodules are learnt to discriminatively predict the domain label, while the backbone network is trained to generate domain-invariant features to confuse the classifier. The multiple two-players adversarial games are implemented by gradient reversal layer (GRL) [9] based optimization in an end-to-end training manner. (2) In the hierarchical domain feature alignment module, the large convolutional feature maps formulate large training sets composed of pixel-wise channel features, which significantly slower the training efficiency. To this end, we propose a **s**cale **r**eduction **m**odule (SRM) without domain feature information loss for reducing the scale of the feature maps by increasing channel number in each convolution block. (3) In the proposal feature alignment module, we propose to aggregate the proposal features with the detection results (*i.e*., classification scores and regression coordinates) during training of the domain classifier. For further confusing hard samples between domains, we propose a weighted gradient reversal layer (WGRL) to down-weight the gradients of easily confused samples and up-weight the gradients of hard confused samples between domains. The contributions of this paper can be summarized as follows:

- A multi-adversarial Faster-RCNN (MAF) is introduced for unrestricted object detection tasks. Two feature alignment modules on both hierarchical domain features and aggregated proposal features are proposed with multi-adversarial domain classifier submodules for domain adaptive object detector.
- In adversarial domain classifier submodule, the scale reduction module (SRM) is proposed for down-scaling the feature maps without information loss, and the training efficiency of our MAF detector is improved.
- In the aggregated proposal feature alignment module, for improving the domain confusion of proposals, we propose a weighted gradient reversal layer (WGRL) which penalizes the hard confused samples with larger gradient weights and relax the easily confused samples with smaller gradient weights.
- Exhaustive experiments on Cityscapes [4], KITTI [10], SIM10K [16], etc. for unrestricted object detection tasks, which show the superior performance of our MAF over state-of-the-art detectors.

2. Related Work

Object Detection. The object detection is a basic task in computer vision and has been widely studied for many years. The earlier work [5, 7, 27] of the object detection were implemented with sliding windows and boost classifiers. Benefited by the success of CNN models [14, 17, 34], a number of CNN based object detection methods [1, 8, 20, 24, 33, 39] have been emerged. The region of interest (ROI) based two-stage object detection methods attracted a lot of attentions in recent years. R-CNN [12] is the first two-stage detector which classifies the ROIs to find the objects. Girshick *et al*. [11] further proposed a Fast-RCNN with ROI pooling layer that shares the convolution features, and both the detection speed and accuracy are promoted. After that, Faster-RCNN [29] was introduced by Ren *et al*., which integrate the Fast-RCNN and Region Proposal Network (RPN) together in an advanced structure. Faster-RCNN further improve the speed and accuracy of detection. In this paper, by taking the Faster-RCNN as backbone, we take into account the mind of domain transfer adaptation for exploring unrestricted object detection task across different domains.

Domain Adaptation. Domain adaptation aims to bridge different domains or tasks by reducing the distribution discrepancy, which has been a focus in various computer vision tasks [15, 22, 23, 37, 38]. The domain adaptation has been recently promoted by the powerful feature representation ability of deep learning. Long *et al*. [23] implemented the

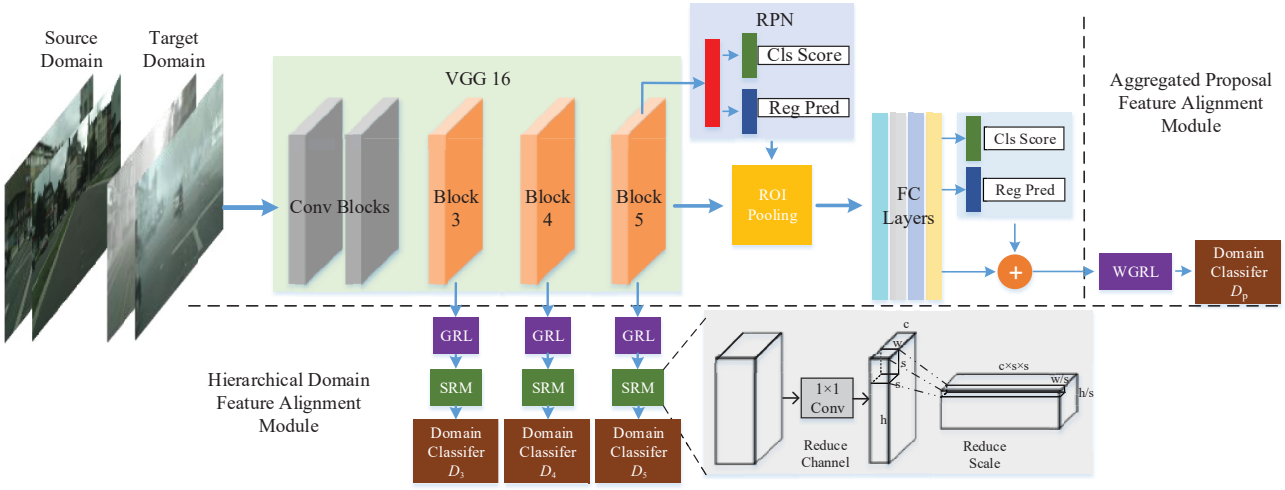

Figure 2. The network structure of our MAF. Inspired by the VGG16 based Faster-RCNN [29], our MAF applies the feature alignment modules on both domain features and proposal features. For the hierarchical domain feature alignment module, multiple adversarial domain classifier submodules are implemented on the block 3,4,5 of the VGG16. GRL layers [9] are used for the adversarial learning strategy and the size of the feature maps are reduced by SRMs. At the proposal feature alignment module, we concatenate the classification scores and bounding box regression results with corresponding features for the domain classifier while the WGRL is introduced for the adversarial learning strategy. SRM is composed of two parts, the first part is a 1×1 convolution layer which is applied to reduce the channel size. After that, a scale reduce part is used to concat $s \times s$ adjacent features, so that the size of the feature maps is reduced.

domain adaptation by minimizing the maximum mean discrepancy (MMD) between the two domain-specific fully-connected branches of the CNN. Besides that, domain confusion for feature alignment through two-player game adversarial learning between feature representation and domain classifier motivated by GAN [13] was extensive studied in transfer learning [18, 22, 26, 35, 40]. Tzeng *et al*. [36] proposed a two-step training scheme to learn a target encoder. Zhang *et al*. [40] take advantages of several domain classifiers to learn domain informative and domain uninformative features. These works focus on image classification tasks, however, for the object detection task, not only the object categories but also the bounding box location should be predicted, which makes the domain transfer of detectors more challenging. In our MAF detector, the mind of domain adaptation and transfer learning is taken into account for network design, and the adversarial optimization is implemented based on the Gradient Reversal Layer (GRL) [9]. Li *et al*. [19] proposed to transmit the knowledge of the strong categories to the weak categories. In [3], the domain disparity is tackled in both the image level and instance level. However, both works do not fully characterize the hierarchical domain feature alignment and the proposal feature alignment.

3. The Proposed MAF Detector

In this section, we will introduce our MAF detector. The source domain is marked by $\mathcal{D}_s$, and $\mathcal{D}_t$ is used for the target domain. In unrestricted setting, the source domain is fully labeled, and $\mathcal{D}_s = \{(x_i^s, b_i^s, y_i^s)\}_i^{n_s}$ stands for the n_s labeled data in the source domain, where the $b_i^s \in \mathcal{R}^{k\times 4}$ stands for the bounding box coordinates of the x_i^s, and $y_i^s \in \mathcal{R}^{k\times 1}$ is the category label for corresponding bounding boxes. $\mathcal{D}_t = \left\{(x_j^t)\right\}_j^{n_t}$ stands for n_t completely unlabeled image samples from target domain.

3.1. Network Structure

The proposed MAF detector is based on the Faster-RCNN [29] framework, and VGG16 [34] with five blocks of the convolution layers is utilized as the backbone of our MAF. The hierarchical domain feature alignment module is implemented on the convolutional feature maps, where the multi-adversarial domain classifier submodules are deployed on blocks 3, 4, and 5. On the top of the network, the aggregated proposal feature alignment module is deployed. With the combination of all feature alignment submodules on both convolution layers and fully collected layer, domain-confused features with domain discrepancy reduced are obtained. It's worth noting that the loss functions including classification loss and smooth L1 loss of the Faster-RCNN are only applied for the source domain. An overview of our network structure is illustrated in Figure 2. Two main modules including 1) hierarchical domain feature alignment module and 2) aggregated proposal feature alignment module formulate the MAF for domain adaptive detection. The former is formulated by multi-adversarial

domain classifier submodules, in which a scale reduction module (SRM) is designed on the top of the GRL [9] for down-scaling the feature maps and improving the training efficiency. The latter is formulated by an adversarial domain classifier, in which the aggregated proposal features with detection results are fed as input. For better characterizing the hard-confused samples between domains, the weighted GRL (WGRL) that adaptively re-weights the gradients of easily-confused and hard-confused samples is deployed, which can better improve the adversarial domain adaptation performance.

3.2. Hierarchical Domain Feature Alignment

The hierarchical domain feature alignment module aims to calibrate the distribution difference between source and target domain in convolution feature maps, which better characterize the image distribution than semantic layer. A intrinsic assumption is that if the image distribution between domains is similar, the distribution of object-level in the image between domains is basically similar also [3]. That is, the distribution difference in the whole image is the primary factor leading to domain discrepancy. In a deep network, the convolutional feature maps in middle level reflect the image information, such as shape, profile, edge, *etc*. Therefore, for domain discrepancy minimization between domains, we propose the hierarchical domain feature alignment module which is formulated by multi-adversarial domain classifier submodules in different convolution blocks. The adversarial domain classifier aims to confuse the domain features, with minimax optimization between the domain classifiers and the backbone network. We consider multi-adversarial domain classifiers instead of general single adversarial domain classifier, because hierarchical feature alignment is helpful to the final domain alignment.

Given an image x_i from the source domain or target domain, the domain features from the convolution layers of the mth block are represented as $C_m(x_i, w_m)$, where w_m stands for the network parameters. The adversarial classifier submodule at the mth block is denoted as D_m, which is learned to predict the domain label of x_i. Following the the adversarial learning strategy, the minimax learning of the adversarial classifier submodule in the mth convolution block can be written as:

$$\min_{\theta_m} \max_{w_m} \mathcal{L}_m \tag{1}$$

where $\mathcal{L}_m = \sum_{u,v} L_c(D_m(C_m(x_i, w_m)^{(u,v)}, \theta_m), d_i)$, in which L_c is the cross entropy loss, the $C_m(x_i, w_m)^{(u,v)}$ stands for the channel-wise feature at pixel (u, v) of the feature maps, and θ_m is the domain classifier parameters in the mth block. d_i is the domain label of sample x_i, which is labeled as 1 for the source domain and 0 for the target domain. In the Eq. (1), the parameters of the backbone network are learned to maximize the cross-entropy loss L_c, while the parameters of the domain classifier submodule struggle to minimize the loss function. By adversarial learning of the domain classifier with backpropagated gradient reverse (*i.e*. GRL [9]), the feature representations are characterized to be domain-invariant.

In order to efficiently train the hierarchical domain feature alignment module, inspired by [41], we introduce a scale reduction module (SRM) which aims at down-scaling the feature maps without information loss. Specifically, SRM contains two steps: 1) A 1×1 convolution layer is implemented to reduce the number of channels of feature maps in each block. This step can achieve domain informative features and reduce the dimensions of domain features for the effective training. 2) Re-align the features by reducing the scale while increasing channel number of the feature maps. This step aims to reduce the size of training set and increase feature dimensionality. In detail, the $s \times s$ adjacent pixels from the feature maps are collected end-to-end to generate a new pixel for the re-shaped feature maps. Obviously, this step is parameterless and easy to compute. The second step is formulated as follows.

$$F^S_{(u,v,c)} = F^L_{(u\times s+c\%s^2\%s, v\times s+\lfloor c\%s^2/s \rfloor, \lfloor c\%s^2 \rfloor)} \tag{2}$$

where the F^L stands for feature maps before the second component. The (u, v, c) presents the element on the cth feature map located at (u, v) and count from 0. F^S stands for the scale reduced feature maps, and s is the sampling factor, which means the adjacent $s \times s$ pixels of the feature maps are merged into one feature. $\%$ stands for the operation of mod and the $\lfloor . \rfloor$ presents the round down. Since SRM only has parameters in the first component, the number of parameters is reduced while the training efficiency is improved. The two components of our SRM can be clearly observed on the bottom of the Figure 2.

3.3. Aggregated Proposal Feature Alignment

The object classifier and bounding box regressor trained with the source domain samples can also not be domain adaptive. Therefore, the aggregated proposal feature alignment module aims to achieve semantic alignment while preserving the information for classification and regression. The proposals are obtained from the region proposal network (RPN), which represent the local parts of an image. In order to improve the semantic discriminative of the proposal features, we propose to aggregate the proposal features with the detection results, *i.e*., classification scores and bounding box regression coordinates, by using concatenation operator. The aggregation brings two kinds of advantages. First, the classification results enrich the information about the categories while the regression results are endowed with position knowledge of the bounding box. Second, the classification and the bounding box regression results improve

the discrimination of the features for easily and effectively training the domain classifier.

Given an input image x_i, the proposal features with respect to the image are represented as $F(x_i, w)$, where w is the CNN model parameters. D_p is the domain discriminator of the proposal feature alignment module. The loss function of the proposal feature alignment module can be written as

$$\min_{\theta_p} \max_{w} \mathcal{L}_p \tag{3}$$

where $\mathcal{L}_p = \sum_k L_c(D_p(F^k(x_i, w) \oplus c^k \oplus b^k, \theta_p), d_i)$, in which $F^k(x_i, w)$ is the feature of the kth proposal, and c^k and b^k are the softmax classification scores and the bounding box regression results of the $F^k(x_i, w)$, respectively. $L_c(\cdot)$ is the cross-entropy loss, θ_p is the domain classifier parameters, and $\oplus$ stands for the concatenation operation.

In order to apply the adversarial domain transfer strategy, in the proposal feature alignment module, we propose a weighted gradient reversal layer (WGRL) to relax the easily-confused samples and simultaneously penalize the hard-confused samples, such that better domain confusion can be achieved. An illustration of the proposed WGRL can be viewed in Figure 3. The samples close to the domain classifier decision boundary are recognised as the easily-confused samples, *i.e.*, they are not distinguishable by the classifier, while samples far from the decision boundary are hard-confused samples, *i.e.*, the domain discrepancy between these samples in both domains is still large. Thus, we should pay more attention to the distinguishable samples by penalizing these samples with larger weights on their gradients. Specifically, the proposed WGRL regards the scores of the domain classifier as the weights for the corresponding samples. Suppose the probability of one proposal in an image belonging to source domain predicted by the domain classifier to be p, the probability belonging to target domain is $1-p$, the gradient before reversal to be G, and the gradient after reversal to be G_{rev}, then WGRL is written as

$$G_{rev} = -\lambda(d \cdot p + (1-d)(1-p))G \tag{4}$$

where the λ is a hyper-parameter for the WGRL and d is the domain label of the image. According to the Eq. (4), the predicted scores are used as the weights for the gradients. The higher confidence of the domain classifier means that domain adaptation needs to be further improved, and the samples are automatically up-weighted. Otherwise, those samples with lower confidence of the domain classifier are considered to be indistinguishable and therefore down-weighted. Note that the minus of $-\lambda$ in Eq. (4) denotes the gradient reverse in optimization.

3.4. Overview of the MAF Detector

The overview of our model can be seen in the Figure 2. Besides the detection loss $\mathcal{L}_{det}$ of Faster-RCNN, *i.e.*, classification loss and regression loss, our MAF have another

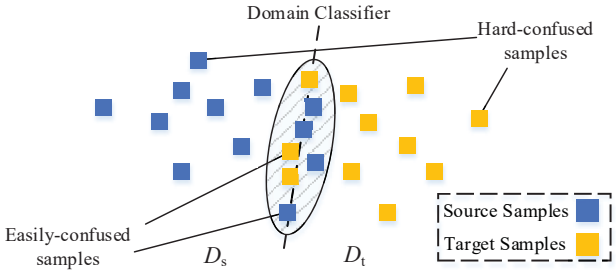

Figure 3. Illustration of WGRL. The blue color stands for the samples from source domain, while the yellow samples stand for the target domain. The samples close to the decision boundary of the domain classifier in the shadow region are recognised as easily-confused samples and up-weighted by our WGRL. The samples outside the shadow region are recognised as hard-confused (*i.e.*, distinguishable) samples that need to be down-weighted.

two extra minimax loss functions $\mathcal{L}_m$ and $\mathcal{L}_p$, *i.e.*, Eq.(1) and Eq.(3) for adversarial domain alignment.

Detection loss minimization. In training of the MAF detector, we utilize the source domain that are full of bounding box labels to train the Faster-RCNN detection loss for the object detection task. The features from the last block of the VGG16 [34] are fed into the RPN to generate a number of proposals for further detection. After that, the ROI pooling layer is used to generate the features with respect to the proposals. The fully-connected layers are trained to get the category labels of the proposals while refining the bounding box coordinates. Note that only the source domain has the annotations for the bounding boxes, the detection loss of the Faster-RCNN is trained on the source domain data.

Adversarial domain alignment loss. The domain alignment loss includes hierarchical domain feature alignment and aggregated proposal feature alignment, which is optimized in an adversarial manner. By jointly considering the Eq.(1) and Eq.(3), the proposed adversarial domain alignment loss in MAF can be written as:

$$\mathcal{L}_t = \mathcal{L}_p + \sum_{m=3}^{5} \mathcal{L}_m \tag{5}$$

Overall loss of MAF detector. With the combination of the detection loss and domain alignment loss, the final loss function of the proposed MAF detector can be written as:

$$\mathcal{L}_{MAF} = \mathcal{L}_{det} + \alpha \mathcal{L}_t \tag{6}$$

where $\mathcal{L}_{det}$ is the loss of Faster-RCNN [29] including softmax loss function and smooth l_1 loss [11], and α is a hyper-parameter between the detection loss and domain adaptation loss. The MAF is trained end-to-end with the Eq. (6). Standard SGD algorithm is implemented to optimize the parameters of the network.

	df.	pf.	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
Faster-RCNN	×	×	17.8	23.6	27.1	11.9	23.8	9.1	14.4	22.8	18.8
DAF	√	√	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
MAF*	√	√	25.3	36.7	41.9	23.5	38.2	**36.4**	18.3	28.0	30.9
MAF	×	√	25.6	36.8	39.9	18.8	32.0	24.1	21.3	29.2	28.5
	√	×	**29.0**	38.8	**43.9**	23.2	39.6	**36.4**	26.7	31.6	33.6
	√	√	28.2	**39.5**	**43.9**	**23.8**	**39.9**	33.3	**29.2**	**33.9**	**34.0**

Table 1. Results on the validation set of the Foggy Cityscapes. **df.** denotes domain feature alignment and **pf.** denotes proposal feature alignment. MAF* means that only one domain feature alignment in the block 5 and the proposal feature alignment are considered.

4. Experiments

In evaluation, we conduct unrestricted object detection experiments on several datasets including Cityscapes [4], Foggy Cityscapes [32], KITTI [10] and SIM10K [16]. We compare our results with the state-of-the-art domain adaptive Faster-RCNN [3] that we call DAF in experiments and the standard Faster-RCNN. To the best of our knowledge, DAF is the first work on the similar object detection task.

4.1. Implementation Details

The experiments in this paper follow the same setting in [3]. The source domain of our experiments is sufficiently annotated with bounding boxes and corresponding categories, while the target domain is completely unlabeled. In order to evaluate the performance of the unrestricted object detection, the testing performance of mean average precision (mAP) on the target domain is compared. The trade-off parameter α in Eq. (6) is set as 0.1 during the training phase. Besides that, for the detection part, we set the hyper-parameters by following [29]. We utilize the ImageNet [30] pre-trained VGG16 model for the initializing our MAF detector. Our model is trained for 50k iterations with the learning rate 0.001 and dropped to 0.0001 for another 20k iterations. Totally, 70k iterations are trained. The minibatch size is set as 2 and the momentum is set as 0.9.

4.2. Datasets

Four datasets including Cityscapes [4], Foggy Cityscapes [32], KITTI [10] and SIM10K [16] are adopted to evaluate the performance of our approach by following [3]. The details of these datasets are provided.

Cityscapes: Cityscapes [4] is designed to capture high variability of outdoor street scenes from different cities. The dataset is captured in common weather conditions and has 5000 images with dense pixel-level labels. These images are collected from 27 cities in different seasons which includes various scenes. Note that the dataset is not originally collected for the object detection task but semantic segmentation, therefore the bounding boxes were generated by the pixel-level annotations as shown in [3].

Foggy Cityscapes: All the images in the Cityscapes [4] were collected from good weather, the Foggy Cityscapes [32] are derived from the Cityscapes to simulate the foggy scenes and constitutes the images with the fog weather. The inherited pixel labels from Cityscapes are used for generation bounding boxes in experiments. Some examples of the Cityscapes and Foggy Cityscapes are illustrated in Figure 1.

KITTI: The KITTI [10] is a dataset produced based on an autonomous driving platform. The images of the dataset are captured in a mid-sized city. Totally 14999 images and 80256 bounding boxes are contained in the dataset for the object detection task. In our experiments, 7481 images in the training set are used for both adaptation and evaluation by following [3].

SIM10K: SIM10K [16] is a simulated dataset generated by the engine of the Grand Theft Auto V (GTA V). This dataset contains 10000 images with 58071 bounding boxes of the car. All images of the SIM10k are used as the source domain for training.

4.3. Experimental Results

In this section, we evaluate our approach on different datasets to simulate different domain shift scenes. Specially, we evaluate the influence of weather by the first. After that, SIM10k and Cityscapes are implemented to search the domain disparity of synthetic data and real data. Finally, the domain shift caused by different scenes is explored.

4.3.1 Detection From Cityscapes to Foggy Cityscapes

We implemented our approach on the Cityspaces [4] and Foggy Cityspaces [32] to evaluate our MAF under foggy weather condition. We take the Cityscapes as the source domain and the Foggy Cityscapes as the target domain. The VGG16 based Faster-RCNN [29] is implemented as the baseline of the experiments. The DAF [3], as a cross-domain detection method, is implemented as the competitor of our MAF. All categories in the Cityscapes are used for the experiments, including the person, rider, car, truck, bus, train, motorcycle and bicycle. The models are tested on the validation set of the Foggy Cityscapes. The results are shown in Table 1, where the **df.** stands for the hierarchical

domain feature alignment module and the **pf.** represents the proposal feature alignment module in all experiments.

According to the Table 1, our MAF achieves the best results among all compared methods. MAF with both domain and proposal feature alignment modules outperforms the DAF by 6.4%, which shows the significant effectiveness of our approach. Note that MAF with only proposal feature alignment module (*i.e.*, **pf.**) achieves 28.5% in mAP, which also outperforms the DAF and the performance of proposal feature alignment module is testified. Besides that, there are some other interesting conclusions can be observed with the results of the MAF* and our approach with only hierarchical domain feature alignment module used. MAF* is a model which contains only one adversarial domain classifier submodule on the block 5, with the submodules on block 3 and 4 removed. Obviously, multi-adversarial domain classifiers on more blocks of convolution layers can significantly improve the domain adaptation performance for better domain-invariant feature representation. With well-aligned domain features, our model achieves much better results and it also verifies our idea that the image distribution calibration in convolutional feature maps is more important than proposal feature alignment in the ultimate domain alignment for the unrestricted object detection task.

4.3.2 Detection from Synthetic Data to Real Data

The SIM10k [16] is a dataset composed of the synthetic data. In this experiment, the SIM10k is used as the source domain, while the Cityscapes is used as the target domain. Note that only the category of car is used for the unrestricted object detection task in the experiment. The results are tested on the validation set of the Cityscapes, which are shown in the Table 2.

	df.	**pf.**	AP of Car
Faster-RCNN	×	×	30.1
DAF	√	√	39.0
MAF	×	√	40.1
	√	×	40.7
	√	√	**41.1**

Table 2. The results on validation set of target domain Cityscapes, with the SIM10k as the source domain. The average precision (AP) of car is reported. Our MAF with different feature alignment modules (**df.** and **pf.**) added is analyzed in the experiment.

From the results of the Table 2, our MAF obtains the best results by comparing to others. Notably, our MAF under different settings can always achieve better performance than the classic Faster-RCNN [29]. Our approach also outperforms the DAF [3] by 2.1% in AP. The superiority of the proposed MAF is fully demonstrated for unrestricted object detection. Also, the proposed hierarchical domain feature alignment (**df.**) can effectively promote the detection performance.

4.3.3 Detection from One Scene to Another

Although the weather conditions are similar between Cityscapes and KITTI, there still exists domain disparity caused by different scenes, such as background, view, resolution, camera, *etc*. In this experiment, we apply the Cityscapes [4] and KITTI [10] as the datasets to study the cross-scene object detection. Specifically, the two datasets are implemented as source domain and target domain, alternately. We implement our MAF, DAF [3] and Faster-RCNN [29] in this experiment. The AP of car is reported for performance comparison. The results of the experiment is shown in Table 3.

	df.	**pf.**	K → C	C → K
Faster-RCNN	×	×	30.2	53.5
DAF	√	√	38.5	64.1
MAF	×	√	38.9	69.9
	√	×	39.7	71.4
	√	√	**41.0**	**72.1**

Table 3. The results of the unrestricted object detection task on the Cityspaces and KITTI. The performances of Cityscapes (C)→KITTI (K) and KITTI (K)→Cityspaces (C) are tested. The AP of car is reported for comparison.

In the Table 3, the K→C means that the KITTI [10] is used as the source domain while the Cityscapes [4] is the target domain and vice versa. Obviously, our MAF model gains the best performance under all conditions. The best performance is 8.1% higher than state-of-the-art DAF method. At this time, the performance of our MAF has been fully verified from hierarchical domain feature alignment to proposal feature alignment.

4.4. Analysis of Proposal Feature Alignment

In this section, we analyze the impact of the aggregated proposal feature and WGRL in the proposal feature alignment module. For fair comparison with DAF [3] that used one adversarial domain classifier for image-level adaptation, we also use one adversarial domain classifier in domain feature alignment, *i.e.* the MAF* with three settings. In this analysis, the Cityscapes [4] is used as the source domain and the Foggy Cityspaces [32] is the target domain, by following the same setting as Section 4.3.1. The analysis results of the experiments are shown in the Table 4.

In Table 4, the WGRL and proposal feature aggregation can be helpful to the final domain adaptation. The concatenation of the proposal features with the classification scores and regression results brings more semantic information for the proposal features, such that the domain classifier can be

	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
DAF	25.0	31.0	40.5	22.1	35.3	20.2	**20.0**	27.1	27.6
MAF* (w/o **WGRL**)	25.4	36.2	41.4	22.1	36.9	31.8	19.9	28.8	30.3
MAF* (w/o **Aggregate**)	**25.5**	35.6	**42.5**	20.7	38.1	31.0	19.5	**29.0**	30.2
MAF*	25.3	**36.7**	41.9	**23.5**	**38.2**	**36.4**	18.3	28.0	**30.9**

Table 4. Analysis of the proposal feature alignment module. The w/o **WGRL** denotes that the standard GRL is used in MAF* and w/o **Aggregate** denotes that the detection results are not concatenated with the proposal feature.

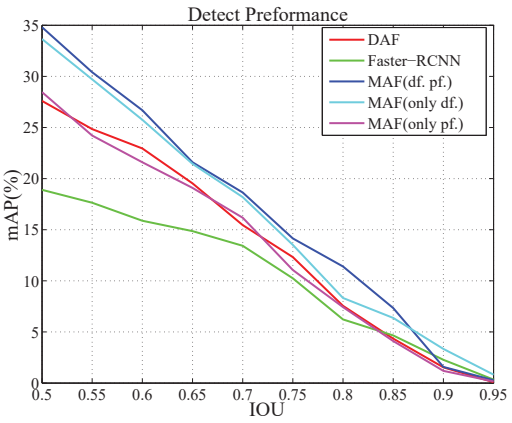

Figure 4. The mAP with different IOU thresholds. MAF, DAF and Faster-RCNN are tested and compared with different IOU thresholds and shown in different colors.

easily trained for feature confusion. WGRL assigns different weights for easily-confused and hard-confused samples, such that the model pays more attention to the samples that are hard to be confused and gains better training effect. Also, the combination of aggregated proposal feature and WGRL achieves the best mAP, therefore, the performance of the proposed proposal feature alignment module is testified.

4.5. Influence of IOU Threshold

The IOU threshold that controls the predicted bounding boxes can also impact the detection results of the testing data. In the previous experiments, the IOU threshold is set as 0.5. In this part, we tune the IOU threshold in the testing phase to study its impact. The Faster-RCNN [29], DAF [3], MAF and MAF with single feature alignment module are implemented with the Cityscapes as source domain and Foggy Cityscapes as target domain. The analysis results of all models are presented in the Figure 4.

From Figure 4, the mAP drops with the increasing of the IOU threshold for all models. The reason is explicit that a larger IOU threshold means that more predicted bounding boxes are excluded, such that insufficient bounding boxes results in a quick drop of the recall and accuracy. The slope of the curves approximately represents the number of predicted bounding boxes in the corresponding IOU range. Benefit from the multi-adversarial domain adaptation strategy with two feature alignment modules, our MAF achieves the best results under different IOU values. Besides, MAF with only hierarchical feature alignment module, *i.e.*, MAF (only df.) ranks the second place and the importance and effectiveness of the multi-adversarial domain feature alignment is shown. From Figure 4, our MAF gets the highest slope on the IOU range 0.8-0.9, the DAF gets the highest slope on range 0.75-0.85, and the Faster-RCNN achieves the highest slope at 0.7-0.8 of IOU range. With the comparison of the slopes, the results reveal that with the domain adaptation, the IOU for unrestricted object detection is increased on the target domain and our MAF with multi-adversarial domain feature alignment achieves the best IOU.

5. Conclusion

In this paper, we propose a multi-adversarial Faster-RCNN (MAF) detector for addressing unrestricted object detection problem. Our approach includes two important modules, *i.e.*, hierarchical domain feature alignment and aggregated proposal feature alignment. With an idea that the domain-adaptive object detection depends much on the alignment of image distribution between domains, we therefore propose multi-adversarial domain classifiers in different convolutional blocks for domain confusion of feature maps. For reducing the scale of the feature maps, we propose a SRM for improving the training efficiency of the adversarial domain classifiers. For domain-adaptive detector, we further deploy a proposal feature alignment module by aggregating the detection results for semantic alignment. The aggregated features are feed into the domain classifier with a weighted gradient reversal layer (WGRL), which can automatically focus on the hard confused samples. Our MAF detector can be trained end-to-end by optimizing the domain alignment loss function and the detection loss of Faster-RCNN. We test our model on several datasets with different domains and achieves state-of-the-arts results. The experiments testify the effectiveness of our model.

Acknowledgement: This work was supported by the National Science Fund of China under Grants (61771079), Chongqing Youth Talent Program, and the Fundamental Research Funds of Chongqing (No. cstc2018jcyjAX0250).

References

[1] Z. Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018. 2

[2] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *Computer Science*, 2015. 1

[3] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. V. Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 2, 3, 4, 6, 7, 8

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 1, 2, 6, 7

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[6] M. Everingham, S. Eslami, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 1

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9):1627–1645, 2010. 2

[8] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017. 2

[9] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014. 2, 3, 4

[10] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2, 6, 7

[11] R. Girshick. Fast r-cnn. *Computer Science*, 2015. 1, 2, 5

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 2

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 2

[15] J. Hoffman, S. Guadarrama, E. Tzeng, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. Lsda: Large scale detection through adaptation. *NeurIPS*, 4:3536–3544, 2014. 2

[16] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016. 2, 6, 7

[17] A. Krizhevsky, I. Sutskever, and E. G. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012. 1, 2

[18] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Poczos. Mmd gan: Towards deeper understanding of moment matching network. In *NeurIPS*, pages 2203–2213, 2017. 3

[19] Y. Li, J. Zhang, K. Huang, and J. Zhang. Mixed supervised object detection with robust objectness transfer. *IEEE TPAMI*, PP(99):1–1, 2018. 3

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. *IEEE TPAMI*, PP(99):2999–3007, 2017. 2

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and C. A. Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 1

[22] M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, pages 1640–1650, 2018. 2, 3

[23] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *NeurIPS*, pages 136–144, 2016. 2

[24] W. Ouyang, K. Wang, X. Zhu, and X. Wang. Chained cascade network for object detection. In *ICCV*, 2017. 2

[25] S. J. Pan and Q. Yang. A survey on transfer learning. *TKDE*, 22(10):1345–1359, 2010. 2

[26] Z. Pei, Z. Cao, M. Long, and J. Wang. Multi-adversarial domain adaptation. In *AAAI*, 2018. 3

[27] D. Piotr, A. Ron, B. Serge, and P. Pietro. Fast feature pyramids for object detection. *IEEE TPAMI*, 36(8):1532–1545, 2014. 2

[28] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[29] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1, 2, 3, 5, 6, 7, 8

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 6

[31] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226. Springer, 2010. 2

[32] C. Sakaridis, D. Dai, and L. V. Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, (11):1–20, 2017. 1, 6, 7

[33] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue. Dsod: Learning deeply supervised object detectors from scratch. In *CVPR*, pages 1919–1927, 2017. 2

[34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2, 3, 5

[35] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, 2017. 3

[36] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 7167–7176, 2017. 3

[37] S. Wang and L. Zhang. Lstn: Latent subspace transfer network for unsupervised domain adaptation. In *PRCV*, pages 273–284. Springer, 2018. 2

[38] L. Zhang, W. Zuo, and D. Zhang. Lsdt: Latent sparse domain transfer learning for visual adaptation. *IEEE TIP*, 25(3):1177–1191, 2016. 2

[39] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. Single-shot refinement neural network for object detection. In *CVPR*, pages 4203–4212, 2018. 2

[40] W. Zhang, W. Ouyang, W. Li, and D. Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *CVPR*, pages 3801–3809, 2018. 3

[41] P. Zhou, B. Ni, C. Geng, J. Hu, and Y. Xu. Scale-transferrable object detection. In *CVPR*, pages 528–537, 2018. 4