PICA: Point-wise Instance and Centroid Alignment Based Few-shot Domain Adaptive Object Detection with Loose Annotations

Chaoliang Zhong[1], Jie Wang[1], Cheng Feng[1], Ying Zhang[1], Jun Sun[1] and Yasuto Yokota[2]
[1]Fujitsu R & D Center, Co., Ltd. [2]Fujitsu LTD
{clzhong, wwangjie, fengcheng, zhangying, sunjun, yokota.yasuto}@fujitsu.com

Abstract

In this work, we focus on supervised domain adaptation for object detection in few-shot loose annotation setting, where the source images are sufficient and fully labeled but the target images are few-shot and loosely annotated. As annotated objects exist in the target domain, instance level alignment can be utilized to improve the performance. Traditional methods conduct the instance level alignment by semantically aligning the distributions of paired object features with domain adversarial training. Although it is demonstrated that point-wise surrogates of distribution alignment provide a more effective solution in few-shot classification tasks across domains, this point-wise alignment approach has not yet been extended to object detection. In this work, we propose a method that extends the point-wise alignment from classification to object detection. Moreover, in the few-shot loose annotation setting, the background ROIs of target domain suffer from severe label noise problem, which may make the point-wise alignment fail. To this end, we exploit moving average centroids to mitigate the label noise problem of background ROIs. Meanwhile, we exploit point-wise alignment over instances and centroids to tackle the problem of scarcity of labeled target instances. Hence this method is not only robust against label noises of background ROIs but also robust against the scarcity of labeled target objects. Experimental results show that the proposed instance level alignment method brings significant improvement compared with the baseline and is superior to state-of-the-art methods.

1. Introduction

Thanks to the advancement of deep neural networks, the performance of object detection has been greatly improved [18]. However, the training of object detection models always requires a large amount of labeled training data. The collection and annotation of the training data are expensive and burdensome, because each instance in every image should be precisely annotated with a bounding box. Directly applying a model pre-trained in a label-rich source domain is attractive, but data bias often results in severe degradation of performance of the model on new data [19].

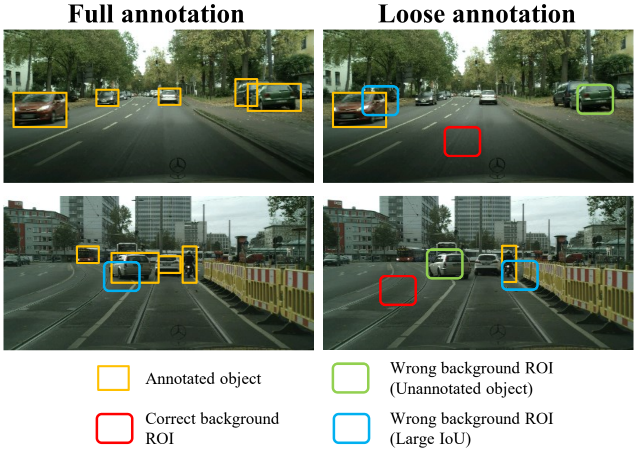
Full annotation
Loose annotation
Annotated object
Correct background ROI
Wrong background ROI (Unannotated object)
Wrong background ROI (Large IoU)

Figure 1. The difference between full annotation and loose annotation. In the fully annotated images (left-hand column), nearly all the objects of interest are annotated, while in the loosely annotated images (right-hand column), only a few objects of interest are annotated. (Best viewed in color)

Many unsupervised domain adaptation (UDA) methods have been proposed to tackle the data bias problem [4, 13, 29, 1, 20, 32]. Nevertheless, UDA methods still require the collection of large number of unlabeled target data and consume a lot of training time. To this end, we aim to develop a supervised domain adaptation method for object detection in few-shot loose annotation setting. The loose annotation means that only a few objects need to be annotated in each image of a very small target data set, thus it will not cost too much to collect and annotate many samples.

In few-shot loose annotation setting, instance level alignment can be used to improve the performance owing to the existence of annotated objects in target domain. However, the distribution alignment in instance level is difficult, due to the scarcity of labeled instances on the target domain. Inspired by FADA [14], traditional method FAFR-CNN [27] conducts the instance level alignment by semantically aligning the distributions of paired object features with domain adversarial training. Although it is demonstrated that point-wise surrogates [30] of distribution alignment provide a more effective solution in few-shot classification tasks across domains, this point-wise alignment approach has not yet been extended to object detection. In this work, we propose a method that extends the point-wise alignment from classification to object detection.

When conducting instance level alignment, we observed that the background ROIs suffer from severe label noises. As shown in Figure 1, in the fully annotated images (left-hand column), nearly all the objects of interest are annotated, while in the loosely annotated images (right-hand column), only a few objects of interest are annotated. In this case, the ROIs containing the unannotated objects (green boxes) are in the foreground, but they will be treated as background because the contained objects are not annotated, thus generating label noises. Another type of label noises (blue boxes) appear in both full annotation and loose annotation, which are caused by the large IoU. These label noises have severe negative impact on the performance and may make the point-wise alignment fail, as it is vulnerable to label noises.

To address this problem, in the proposed method, we compute the moving average centroid for each category, remove the ROIs of the background and only keep the centroid for the background, thus alleviating the problem of label noises of background ROIs. Moreover, we not only align the instances, but also align the instances and corresponding centroid across domains. Hence this method is not only robust against the scarcity of labeled target objects but also robust against label noises of background ROIs.

In addition, we found that the instance level alignment is better to be conducted on the feature of the branch of classification, rather than on the common feature of classification and localization. This is because the classification features of the same category are similar, but the features of bounding boxes of the same category may be dissimilar.

In summary, our contributions are three-fold:

1. We are the first to notice and mitigate the problem of label noises of background ROIs in loose annotation setting with moving average centroid method.

2. We propose a new instance level alignment method for object detection under few-shot loose annotation setting by improving and extending the point-wise alignment method from classification to object detection.

3. We propose to separate the classification and localization features and perform instance level alignment only on classification feature to improve the performance further.

2. Related Work

Object Detection Existing object detection architectures can be divided into two groups. The methods of the first group such as R-CNN [6], Fast R-CNN [5], Faster R-CNN [18] generate two- or multi-stage models, while methods of the second group such as YOLO [16], YOLOV2 [17], SSD [12] and Retinanet [11] learn single-stage models. However, all these methods require a large amount of fully and carefully annotated data to train a model with good performance, thus are costly and not applicable to object detection in unseen domains.

Domain Adaptation Many efforts are made to transfer knowledge from a source domain to target domains with domain adaptation to reduce the data preparation cost and generalize to unseen target domains. However, most of them are developed for classification under the unsupervised domain adaptation (UDA) setting, where the source domain has sufficient labeled data, but the target domain contains only unlabeled data. Dominant methods such as DANN [4] and CDAN [13] improve the performance on target domain by minimizing the classification loss on source domain and learning domain-invariant features simultaneously with adversarial training to align the marginal or conditional feature distributions across domains. A drawback of these methods is that they ignore the semantic information contained in samples. To address this problem, Xie *et al.* present moving semantic transfer network (MSTN), which learns semantic representations for unlabeled target samples by aligning labeled source centroid and pseudo-labeled target centroid [29]. The moving average centroid used by MSTN is robust against label noise introduced by pseudo labels.

Recently, UDA methods for object detection have attracted increasing attention. Chen *et al.* first noticed the image level shift and instance level shift and proposed two components to alleviate the domain discrepancy at image and instance levels respectively [1]. Saito *et al.* proposed SWDA [20] and further indicated that local feature alignment is more effective than global image-level feature alignment, since it does not change category level semantics. Zhu *et al.* improved the instance-level alignment by applying *k*-means clustering to group proposals and obtain the centroids of these clusters [32]. Zhuang *et al.* improved the instance-level alignment by conducting category-agnostic, category-aware and category-correlation instance alignment simultaneously [33]. Wu *et al.* explored the extraction of instance-invariant features by disentangling the domain-invariant features from domain-specific features. [28]. Other variants improve the performance by aligning features of multiple layers [7], forcing the network to focus on discriminative regions with context information at different scales [10] or performing center-aware feature

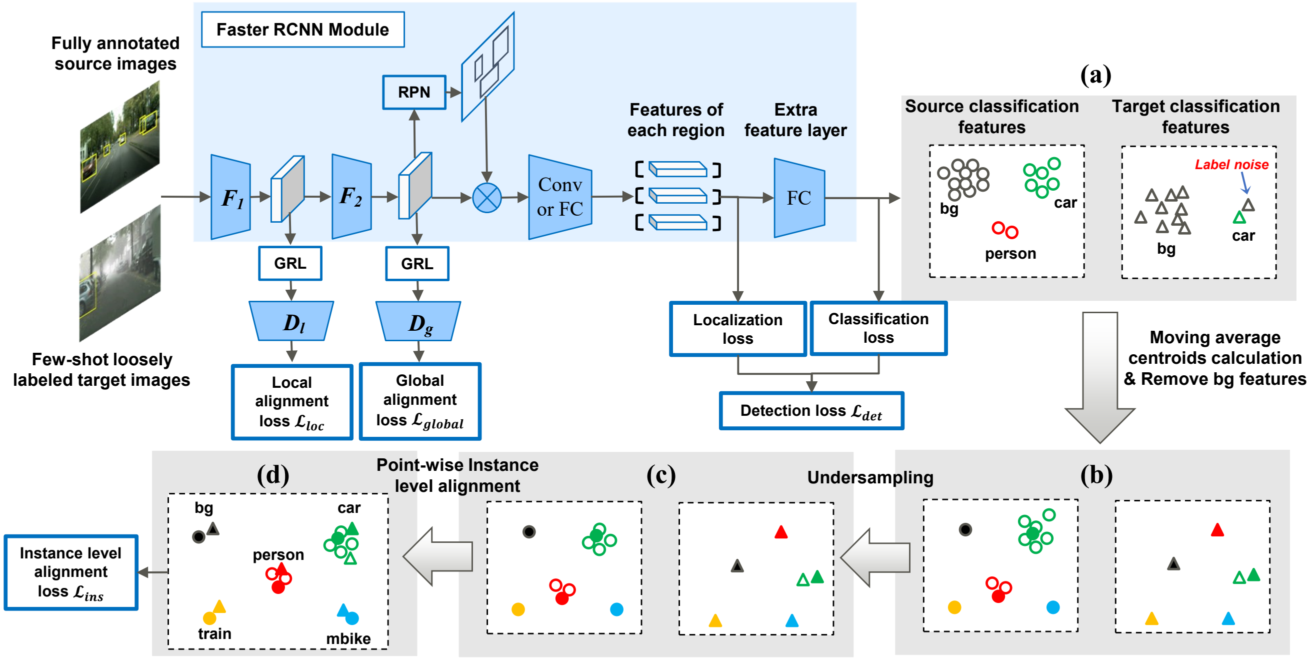
Fully annotated source images
Faster RCNN Module
RPN
F_1
F_2
Conv or FC
Features of each region
Extra feature layer
FC
GRL
GRL
D_l
D_g
Local alignment loss $\mathcal{L}_{loc}$
Global alignment loss $\mathcal{L}_{global}$
Localization loss
Classification loss
Detection loss $\mathcal{L}_{det}$
Few-shot loosely labeled target images
(a)
Source classification features
Target classification features
bg
car
person
Label noise
bg
car
Moving average centroids calculation & Remove bg features
(b)
Undersampling
(c)
Point-wise Instance level alignment
(d)
bg
car
person
train
mbike
Instance level alignment loss $\mathcal{L}_{ins}$

Figure 2. Framework of our proposed method for few-shot domain adaptive object detection. The framework integrates global alignment, local alignment and point-wise instance level alignment. To better implement instance level alignment, an extra feature layer is added into the classification branch. **(a)**→**(b)**: With the ROI features extracted by the extra feature layer, moving average centroids are computed and added to the batch of ROI features to mitigate the label noise problem of background (bg) ROIs and the problem that some classes are missing in current batch of features. Meanwhile, as the background ROI features are label noisy, we remove the background ROI features and only keep the background centroids. **(b)**→**(c)**: We conduct the instance balancing via undersampling. **(c)**→**(d)**: Finally, we exploit point-wise alignment over instances and centroids to solve the problem of scarcity of labeled target instances. The hollow points represent the features of instances, while the solid points refer to the centroid of each category. The circles represent the features of source domain, while the triangles represent the features of target domain. (Best viewed in color)

alignment by predicting pixel-wise objectness [8].

Few-shot Domain Adaptation Although UDA methods can reduce the cost of annotation, the cost of target data collection remains expensive. Moreover, in some scenarios, e.g., cases of rare disease, etc., we have only a few samples. In this case, the few-shot supervised domain adaptation (FDA) will be more feasible than UDA, since it utilizes little labeled target data with label-rich source data to train a target model, avoiding the need to collect and label many samples. Here, the term FDA has two meanings in different contexts. In some contexts, it means that the target domain has some new categories unseen in the source domain [24, 23, 31, 26]. In other contexts, it only means that the samples of each category in target domain are scarce, but the source and target domains still share the same label space [14, 15, 30, 27]. In this work, we will follow the protocol of [27] and focus on the second definition of FDA under loose annotation setting.

For classification tasks in FDA, Motiian *et al.* [15] proposed CCSA and found that alignment and separation of semantic probability distributions are difficult due to the lack of data and reverting to point-wise surrogates of distribution alignment provides an effective solution. They further proposed a more effective solution named FADA [14] by using adversarial learning to learn an embedded subspace that simultaneously maximizes the confusion between two domains while semantically aligning their embedding. Xu *et al.* [30] proposed a solution named d-SNE which is more effective than FADA by minimizing the largest distance between the samples of the same class and maximizing the smallest distance between the samples of different classes in feature space across domains.

To the best of our knowledge, FAFRCNN [27] is the only work for object detection in FDA, which is built on Faster R-CNN and introduces a pairing mechanism based on FADA over source and target features to alleviate the issue of insufficient target domain samples and implement instance level alignment. Differing from FAFRCNN, we implement the point-wise instance level alignment by extending and improving d-SNE, which is shown to be more effective than FADA. Moreover, our method can mitigate the problem of label noises of background ROIs.

3. Method

As shown in Figure 2, the framework of our method integrates global, local and instance level alignment. In this section, we first define the problem of FDA for object detection. Then, we present the integration with global and local alignment. Finally, we present how we extend the point-wise alignment from classification to object detection.

3.1. Preliminaries

In FDA, we are given a large data set of source domain $\{x_i^s, y_i^s\}_{i=1}^{n_s}$ as well as a very small data set of target domain $\{x_i^t, y_i^t\}_{i=1}^{n_t}$, where $n_t \ll n_s$, x_i^s and x_i^t denote an input image drawn from X_s and X_t respectively, y_i^s is drawn from Y_s and denotes complete bounding box annotation for x_i^s, and y_i^t is drawn from Y_t and denotes loose bounding box annotation for x_i^t. In particular, we assume Y_s and Y_t share the same label space of categories. With the label-rich source images and only a few objects annotated in target images, our goal is to learn a detection model with minimal degradation of performance in target domain, compared with the model trained with complete and fully labeled training set of target domain.

3.2. Integration with Global and Local Alignment

Although instance level alignment is effective to improve performance, using instance level alignment alone is not sufficient to guarantee the performance of FDA. Therefore, we develop our method upon SWDA [20], which is a UDA framework based on Faster R-CNN for object detection and consists of a local discriminator D_l, a global discriminator D_g and a feature extractor F, where F is decomposed as $F_2 \circ F_1$. Moreover, there is a network R that takes features from F and outputs bounding boxes with a class label. R consists of the Region Proposal Network (RPN), the extra feature layer (see Section 3.3.1) and other modules in Faster R-CNN [18]. Since the images of target domain are scarce and loosely annotated, minimizing the detection loss on target domain will not only lead to over-fitting, but also cause the training to fail because the loose annotations contain numerous label noises. Therefore, we only minimize the following detection loss on the source domain:

$$\mathcal{L}_{det}(F, R) = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathcal{L}(R(F(x_i^s)), y_i^s)) \tag{1}$$

where $\mathcal{L}$ denotes the loss function for object detection, which consists of the classification and bounding box regression loss. Then, we use weak global alignment to learn the domain invariant image level features. The loss of the weak global alignment for global-level discriminator D_g is denoted as follows,

$$\mathcal{L}_{global_s} = -\frac{1}{n_s} \sum_{i=1}^{n_s} (1 - D_g(F(x_i^s)))^{\gamma} \log(D_g(F(x_i^s))) \tag{2}$$

$$\mathcal{L}_{global_t} = -\frac{1}{n_t} \sum_{i=1}^{n_t} D_g(F(x_i^t))^{\gamma} \log(1 - D_g(F(x_i^t))) \tag{3}$$

$$\mathcal{L}_{global}(F, D_g) = \frac{1}{2}(\mathcal{L}_{global_s} + \mathcal{L}_{global_t}) \tag{4}$$

where bigger γ will make the model focus more on the hard-to-classify examples, and will achieve a weak alignment between domains [20]. Meanwhile, we use strong local alignment to learn domain invariant local level features, such as texture or color. The loss of the strong local alignment for local-level discriminator D_l is denoted as follows,

$$\mathcal{L}_{loc_s} = \frac{1}{n_s HW} \sum_{i=1}^{n_s} \sum_{w=1}^{W} \sum_{h=1}^{H} D_l(F_1(x_i^s))_{wh}^2 \tag{5}$$

$$\mathcal{L}_{loc_t} = \frac{1}{n_t HW} \sum_{i=1}^{n_t} \sum_{w=1}^{W} \sum_{h=1}^{H} (1 - D_l(F_1(x_i^t))_{wh})^2 \tag{6}$$

$$\mathcal{L}_{loc}(F_1, D_l) = \frac{1}{2}(\mathcal{L}_{loc_s} + \mathcal{L}_{loc_t}) \tag{7}$$

where W and H denote the width and height of a feature extracted by the feature extractor F_1. The adversarial loss for global and local alignment is summarized as follows,

$$\mathcal{L}_{adv}(F, D) = \mathcal{L}_{loc}(F_1, D_l) + \mathcal{L}_{global}(F, D_g) \tag{8}$$

3.3. Point-wise Instance and Centroid Alignment

Since it is demonstrated that the point-wise alignment, i.e., d-SNE [30] is more effective than the distribution alignment, i.e., FADA [14] for classification in FDA setting, our purpose is to extend the d-SNE to object detection. The d-SNE loss is deduced as follows,

$$\begin{aligned}\mathcal{L}_{d-SNE}(F, R) =& \frac{1}{|O_t|} \sum_{x_t \in O_t} \max(0, \sup_{x_s \in O_s^k} \{a | a \in d(x_s, x_t)\} \\ &- \inf_{x_s \in O_s^{\not{k}}} \{b | b \in d(x_s, s_t)\} + m)\end{aligned} \tag{9}$$

where $d(x_s, x_t)$ denotes the squared Euclidean distance between x_s and x_t in latent-space, k is the label of x_t, i.e., $k = y_t$, $O_s^k = \{\forall x_s | y_s = k\}$, $O_s^{\not{k}} = \{\forall x_s | y_s \neq k\}$ and m is a pre-defined margin for the efficiency of implementation. The d-SNE loss achieves point-wise alignment by minimizing the largest distance between the instances of the same class and maximizing the smallest distance between the instances of different classes.

However, this original d-SNE loss function cannot be applied to object detection in FDA setting directly for four reasons. (1) First, it is a loss function designed for classification. But for object detection, the features consist of classification and localization features. Directly conducting the point-wise alignment on the common feature of classification and localization may be not a good choice, since the classification features of the same category may be similar, but the localization features of the same category may be dissimilar. (2) Second, the d-SNE loss is very sensitive to label noises. In the loose annotations of target domain, there are a lot of label noises which may make the instance level alignment fail. (3) Third, the d-SNE loss does not consider the category imbalance problem. (4) Fourth, this d-SNE loss implementation only enlarges the relative difference between maximum intra-class distance and minimum inter-class distance for each target instance. It does not maximize the absolute minimum inter-class distance. Therefore, we propose our point-wise instance and centroid alignment method to address the problems of d-SNE method.

3.3.1 Extra Classification Feature Layer

To overcome the first problem, as shown in Figure 2, we add an extra feature layer into the classification branch and conduct the instance-level alignment on the features extracted by this extra feature layer across domains. Unlike traditional methods [32, 33, 27, 28] that only conduct instance-level alignment on the foreground ROI features, the features that our method aligns not only include foreground ROI features, but also background ROI features. This is because, with point-wise alignment, to obtain the instance level alignment loss, we need to calculate intra-class distances and inter-class distances; however, in some scenarios where there is only one foreground category, e.g., detecting cars and ignoring other objects, we are unable to calculate the inter-class distances if we only consider foreground ROI features. Thus, we take the background ROIs into consideration for computing instance level alignment loss.

3.3.2 Moving Average Centroid

As illustrated by MSTN [29], the moving average centroids are robust against label noises. Therefore, we employ the moving average centroid method to mitigate the second problem of d-SNE, as shown in Figure 2 (a)→(b). In each iteration, suppose we obtain the source classification features O_s and target classification features O_t from the extra feature layer. Since we are working under the supervised domain adaptation setting, we are aware of the label for $x_s \in O_s$ and $x_t \in O_t$, which are y_s and y_t, respectively. As aforementioned, O_s and O_t contain background ROIs which have many label noises and may make the instance alignment fail. Moreover, due to the scarcity of labeled instances in the target domain, O_t includes rare foreground ROI features and there may be no instances of some classes in a mini-batch. This makes the cross-domain intra-class distance calculation impossible for a source instance if there is no target instance with the same class label. Therefore, to address these problems, we first calculate the moving average centroid for each class as follows,

$$C^k_{S_{(t)}} \leftarrow \frac{1}{|O^k_s|} \sum_{x_s \in O^k_s} x_s \tag{10}$$

$$C^k_{T_{(t)}} \leftarrow \frac{1}{|O^k_t|} \sum_{x_t \in O^k_t} x_t \tag{11}$$

$$C^k_S \leftarrow \theta C^k_S + (1-\theta) C^k_{S_{(t)}} \tag{12}$$

$$C^k_T \leftarrow \theta C^k_T + (1-\theta) C^k_{T_{(t)}} \tag{13}$$

where $O^k_s = \{\forall x_s | y_s = k\}$, $O^k_t = \{\forall x_t | y_t = k\}$, $C^k_{S_{(t)}}$ and $C^k_{T_{(t)}}$ denote the average centroid of current iteration of category k for source and target domain respectively, C^k_S and C^k_T denote the moving average centroid of category k for source and target domain respectively, θ denote the moving average coefficient.

Then, the centroid of each category is added into O_s and O_t respectively, so that the intra-class distance and inter-class distance can be calculated for instances of any class. Since the background ROIs contain label noises, they are removed from O_s and O_t to alleviate the second problem of original d-SNE,

$$O_s \leftarrow O_s - \{\forall x_s | y_s = 0\} + \{C^k_S | k = 0 \cdots \mathcal{C}\} \tag{14}$$

$$O_t \leftarrow O_t - \{\forall x_t | y_t = 0\} + \{C^k_T | k = 0 \cdots \mathcal{C}\} \tag{15}$$

where $\mathcal{C}$ is the total number of classes and $y_s = 0$ or $y_t = 0$ denotes the background category.

3.3.3 Instance Balancing

The category imbalance problem would decrease the performance of models. Similarly, the imbalanced instance distribution would also have negative impact on the instance level alignment. In the data sets for object detection, e.g., Cityscapes [2], the instance distribution is extremely imbalanced, where *car* and *person* are dominant classes, which have 26,889 and 15,802 labeled instances respectively, whereas *train* and *bus* have only 339 and 166 instances respectively.

Therefore, to tackle the third problem, we perform undersampling on O_s and O_t to alleviate this problem, as shown in Figure 2 (b)→(c),

$$O_s \leftarrow undersample(O_s, \tilde{n}) \tag{16}$$

$$O_t \leftarrow undersample(O_t, \tilde{n}) \tag{17}$$

where $undersample()$ denotes a pre-defined function that limits the maximum number of instances of each category to be threshold $\tilde{n}$ by removing extra instances at random.

3.3.4 New Instance Level Alignment Loss

As depicted in Figure 2 (c)→(d), considering the d-SNE loss only enlarges the relative difference between maximum intra-class distance and minimum inter-class distance for each target instance, we propose an improved instance alignment loss as follows to overcome the fourth problem,

$$\begin{aligned}\mathcal{L}_{ins}(F, R) =& \frac{1}{|O_t|} \sum_{x_t \in O_t} \max(0, m_2 - \inf_{x_s \in O_s^{\not k}} \{b | b \in d(x_s, s_t)\}) \\ &+ \max(0, \sup_{x_s \in O_s^k} \{a | a \in d(x_s, x_t)\} \\ &- \inf_{x_s \in O_s^{\not k}} \{b | b \in d(x_s, s_t)\} + m)\end{aligned} \tag{18}$$

where m_2 denotes another pre-defined margin for maximizing the absolute minimum inter-class distance. Therefore, our improved instance alignment loss function can better separate classes from each other than original d-SNE, although we need one more hyper parameter m_2.

3.4. Overall Objective Function

The overall objective of our method is as follows,

$$\max_{D_l, D_g} \min_{F, R} \mathcal{L}_{det}(F, R) - \lambda_1 \mathcal{L}_{adv}(F, D) + \lambda_2 \mathcal{L}_{ins}(F, R) \tag{19}$$

where λ_1 and λ_2 are the weights of adversarial loss and instance level alignment loss respectively, which are used to control the trade-off between detection, adversarial training and instance level alignment losses. The mini-max loss function is achieved by a gradient reversal layer (GRL) [4].

4. Experiments

4.1. Datasets and Scenarios

Datasets Following [27], we utilize four datasets to establish and simulate the cross-domain adaptation scenarios for evaluating the adaptation performance of our model and comparing models. The first dataset is **Cityscapes** [2] which consists of around 5000 accurately labeled real world images. The second dataset **Foggy Cityscapes** [21] is derived from Cityscapes and constitutes a collection of synthetic foggy images. The third is the **SIM10K** [9] which contains 10K synthetic images with bounding box annotation for car, motorbike and person. And the last one is an open source dataset **Udacity self-driving** [25] (Udacity for short). The illumination, camera condition and surroundings of the images contained in Udacity are different from Cityscapes.

Scenarios In order to compare with the state-of-the-art method [27], we construct the following three scenarios: Cityscapes to Foggy Cityscapes (C→F), SIM10K to Cityscapes (S→C) and Udacity to Cityscapes (U→C). The first scenario (C→F) simulates the domain shift caused by the extreme weather change of normal to foggy condition. The second scenario (S→C) captures the domain shift between synthetic and real worlds. The last scenario (U→C) is designed for the domain shift between two real worlds, which is caused by illumination, camera condition, etc.

4.2. Baselines

Our method is compared with the following baselines: (1) **Source-only model**. This model is trained with source data and evaluated on a target test set. Its performance is considered as the lower bound of adaptation. (2) **Target-only model**. This model is trained with target training set and evaluated on target test set. Its performance is considered as the upper bound of adaptation. (3) **UDA models**. These models are trained with labeled source data and a large set of unlabeled target data. Besides SWDA and FAFRCNN, we also compare our method with two state-of-the-art UDA methods [33, 28]. (4) **FUDA models**. We directly apply the UDA (SWDA) method in FDA setting to assess its performance. (5) **FDA models**. The FAFRCNN model, which is the state-of-the-art method for FDA is used.

4.3. Implementation Details

In the experiments, we establish our method upon SWDA and use VGG16 [22] network pre-trained on ImageNet [3] as a backbone network. For the local discriminator D_l, global discriminator D_g, feature extractor F, region proposal network (RPN) and bounding box branch, we use the same network architecture as [20]. The extra feature layer in the classification branch is a fully connected layer with 128 hidden units.

All models are trained using mini-batch stochastic gradient descent (SGD) with a momentum of 0.9. Following the setting used in [20], we first train the networks with a learning rate of 0.001 for 50K iterations, then reduce the learning rate to 0.0001 and train for 20K more iterations. All models are trained with this scheduling and we report the performance trained after 70K iterations. We set γ to 5.0. For C→F, we set λ_1 to 1.0. For S→C and U→C, we set λ_1 to 0.1. For all experiments, we increase λ_2 gradu-

Method	Setting	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
Faster R-CNN [18]	Source-only	24.1	33.1	34.3	4.1	22.3	3.0	26.5	15.3	20.3
Faster R-CNN [18]	Target-only	33.2	45.9	49.7	35.6	50.0	37.4	36.2	34.7	40.3
SWDA CVPR'19 [20]	UDA	29.9	42.3	43.5	**24.5**	36.2	**32.6**	**30.0**	35.3	34.3
FAFRCNN CVPR'19 [27]		29.1	39.7	42.9	20.8	37.4	24.1	26.5	29.9	31.3
iFAN AAAI'20 [33]		-	-	-	-	-	-	-	-	35.3
Wu, et al. TPAMI'21 [28]		**33.12**	**43.41**	**49.63**	21.98	**45.75**	32.04	29.59	**37.08**	**36.57**
SWDA CVPR'19 [20]	FUDA	28.0±1.5	39.9±0.9	40.5±2.3	23.8±2.3	35.9±2.2	20.9±7.0	24.0±0.8	31.7±2.7	30.6±1.8
FAFRCNN CVPR'19 [27]	FDA	27.9±0.6	37.8±0.6	42.3±0.7	20.1±0.5	31.9±1.1	13.1±1.5	24.9±1.3	30.6±0.9	28.6±0.5
PICA (Ours)		**28.3±2.2**	**41.3±0.3**	**43.0±0.4**	**23.8±2.2**	**38.1±1.5**	**24.3±0.8**	**25.4±1.4**	**33.7±0.4**	**32.2±0.8**

Table 1. Quantitative results of our method (PICA) on C→F. The unsupervised setting (UDA) uses all the unlabeled images in target domain. In few-shot supervised setting (FDA), 8 images (1 image per class) are sampled for each experiment round, and 1 object bounding box is annotated for corresponding class per image. The few-shot unsupervised setting (FUDA) uses 8 images which are the same as FDA without using corresponding annotations. In Table 1 to 3, except for those results from original papers, all results are averaged over three random runs.

Method	Setting	S→C	U→C
Faster R-CNN [18]	Source-only	34.6	44.0
Faster R-CNN [18]	Target-only	53.1	53.1
SWDA CVPR'19 [20]	UDA	40.1	51.9
FAFRCNN CVPR'19 [27]		41.2	50.2
SWDA CVPR'19 [20]	FUDA	40.2±0.6	51.9±0.3
FAFRCNN CVPR'19 [27]	FDA	39.8±0.6	50.6±0.6
PICA (Ours)		**42.1±0.7**	**52.4±0.1**

Table 2. Quantitative results of our method (PICA) on S→C and U→C. The UDA uses all the unlabeled images in target domain. In few-shot supervised setting (FDA), for target domain, 8 images are sampled for each experiment round and 3 car objects are annotated per image. The FUDA uses 8 images which are the same as FDA without using corresponding annotations.

Method	S→C
SWDA CVPR'19 [20]	40.2±0.6
PICA	**42.1±0.7**
PICA w/o Instance balancing	41.2±0.7
PICA w/o Extra feature layer	40.6±0.6
PICA w/ Original d-SNE loss	40.7±0.6
PICA w/o Adding centroids	34.7±2.3
PICA w/o Removing background ROIs	36.3±2.8
PICA w/o SWDA	40.4±0.3

Table 3. Ablation study.

ally using the schedule $\lambda_2 = \min(0.1, p^2)$, where p denotes the training progress linearly changing from 0 to 1. In all experiments, we set the moving average coefficient θ to 0.5 and the undersampling threshold $\tilde{n}$ to 8. The batch size of input source and target images is set to 1 in all experiments. We use the default Faster R-CNN ROI sampling scheme to create training data for classification and regression heads, which separates foreground and background ROIs with an IoU threshold of 0.5 and samples them at a ratio 1:3. The batch size of source and target ROIs is set to 256. For instance level alignment loss, we set $m = 1$ and $m_2 = 30$.

4.4. Quantitative Results

As summarized in Table 1, our method (PICA) outperforms the state-of-the-art FDA method FAFRCNN significantly on C→F scenario. Compared with the SWDA in FUDA setting, our method provides 1.6 AP gain on mean value, which indicates the effectiveness of our method to make use of the scarce and loose target annotation for instance level alignment. Moreover, our method is superior to FAFRCNN even in UDA setting. For other UDA models, although they do not use the annotation of target samples, they obtain better performance than our model, due to the rich information hidden in diverse unlabeled target samples.

Our method shows the consistent results on S→C and U→C scenarios as well. As shown in Table 2, our method gives around 2.0 AP gain on the two scenarios compared with FAFRCNN. Compared with SWDA in FUDA setting, our method also achieves 1.9 AP gain on S→C and 0.5 AP gain on U→C. In these two scenarios, there is only one foreground category. Thus, the background ROIs must be considered when conducting point-wise alignment, otherwise the inter-class distance cannot be computed. Therefore, the AP improvement on these two scenarios can confirm the effectiveness of our proposed instance level alignment method.

4.5. Qualitative Results

Figure 3 shows the detection results of three images sampled from S→C scenario. The first row is the outputs from the source-only model, the second row is the outputs from the FAFRCNN model and the third row is the outputs from our model (PICA). From the results, we can see that our method can predict more accurate bounding boxes, such as the objects marked in red boxes. Moreover, our model can detect the objects that are difficult to detect using other models, such as the car marked in the purple box. A drawback of our model is that it tends to mis-detect more objects with small bounding boxes. This may be because our method does not use multi-scale feature extraction.

4.6. Ablation Study

We conduct an ablation study on S→C to evaluate the contribution of each component. As depicted in Table 3, our method can lead to 1.0 AP improvement compared with

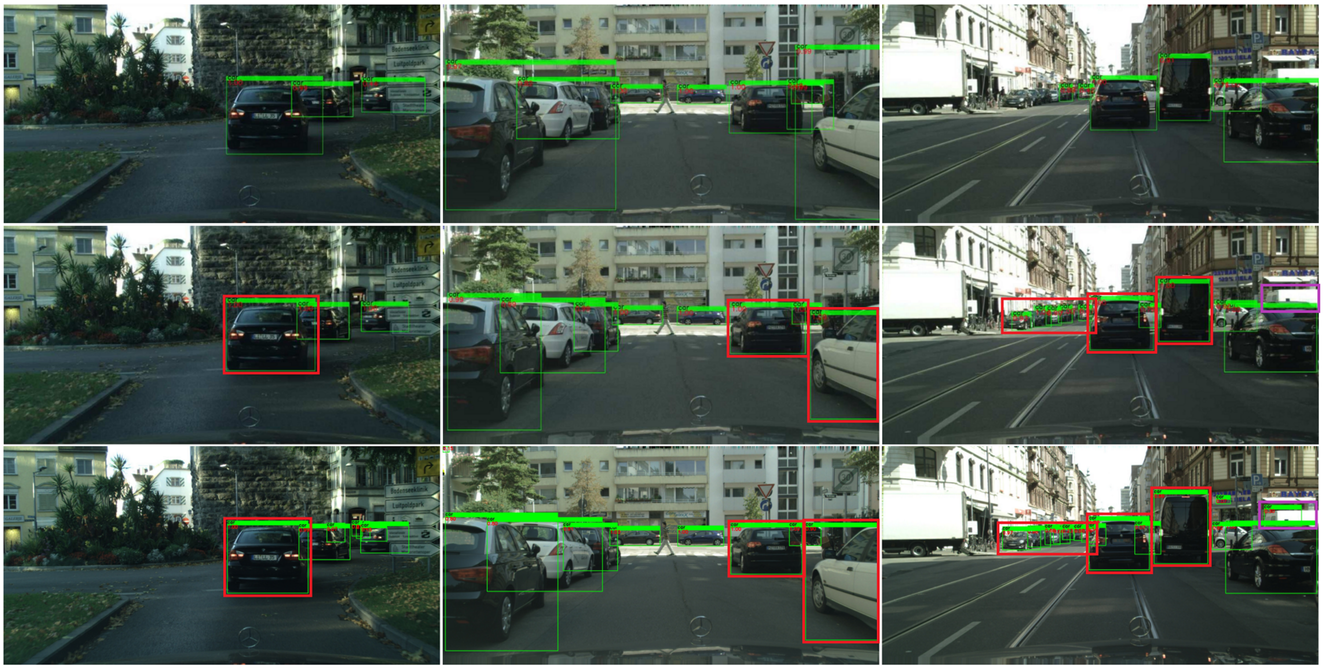

Figure 3. Qualitative results. We compare detection results of three images sampled from S→C scenario. The bounding box visualization threshold is set to 0.05. The first row shows the outputs from the source-only model, the second row shows the outputs from the FAFRCNN model, and the third row shows the outputs from our model (PICA).

SWDA even without instance balancing. The instance balancing can bring a further 0.9 AP improvement. We also find the extra feature layer and our proposed instance level alignment loss are critical. Without them, the improvements become marginal. However, if centroids are not added, the background ROIs must be kept for instance level alignment, as there is only one foreground category. In this case, the performance decreases from 42.1 to 34.7 due to label noises and missing classes. If the centroids are added but the background ROIs containing label noises are not removed, the performance also decreases severely. Finally, we find the performance of PICA alone is comparable with that of SWDA.

4.7. Alignment Effect Analysis

We compare the maximum intra-class distances and minimum inter-class distances of centroids and instances on S→C in the feature spaces learned with different instance level alignment loss functions. As shown in Figure 4, the results of our proposed instance alignment loss can better separate classes from each other than original d-SNE, since it not only enlarges the relative difference between minimum inter-class distance and maximum intra-class distance, but also enlarges the absolute minimum inter-class distance.

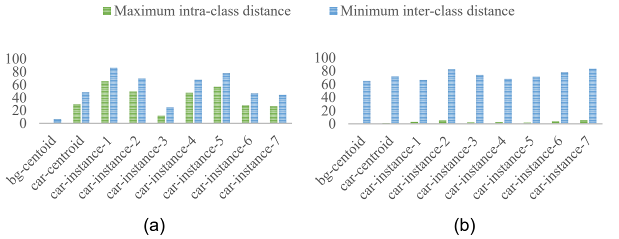

Figure 4. Comparison of the maximum intra-class distances and minimum inter-class distances of centroids and instances on S→C. (a) PICA w/ d-SNE loss. (b) PICA w/ our proposed instance alignment loss.

5. Conclusion

A new instance level alignment method is proposed by extending the point-wise alignment from classification to object detection. We conduct experiments in three typical scenarios and the results show that our method is superior to state-of-the-art methods and baseline model. Qualitative results show that our method can predict more accurate bounding boxes and detect objects that are difficult to detect when using other models. As for future work, this method will be improved by using multi-scale feature extraction to solve the problem of mis-detection of small bounding boxes.

References

[1] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. V. Gool. Domain adaptive faster r-cnn for object detection in the wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2018.

[2] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[4] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1180–1189, 2015.

[5] R. Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR '14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[7] Z. He and L. Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6668–6677, 2019.

[8] C.-C. Hsu, Y.-H. Tsai, Y.-Y. Lin, and M.-H. Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *European Conference on Computer Vision*, pages 733–748, 2020.

[9] 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? In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 746–753, 2017.

[10] C. Li, D. Du, L. Zhang, L. Wen, T. Luo, Y. Wu, and P. Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 481–497, 2020.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *14th European Conference on Computer Vision, ECCV 2016*, pages 21–37, 2016.

[13] M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, volume 31, pages 1640–1650, 2018.

[14] S. Motiian, Q. Jones, S. M. Iranmanesh, and G. Doretto. Few-shot adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, volume 30, pages 6670–6680, 2017.

[15] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto. Unified deep supervised domain adaptation and generalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5716–5726, 2017.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[17] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017.

[18] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[19] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV'10 Proceedings of the 11th European conference on Computer vision: Part IV*, pages 213–226, 2010.

[20] K. Saito, Y. Ushiku, T. Harada, and K. Saenko. Strong-weak distribution alignment for adaptive object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6956–6965, 2019.

[21] C. Sakaridis, D. Dai, and L. V. Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR 2015 : International Conference on Learning Representations 2015*, 2015.

[23] Y. Tan, Y. Li, and S.-L. Huang. Otce: A transferability metric for cross-domain cross-task representations. *arXiv preprint arXiv:2103.13843*, 2021.

[24] H.-Y. Tseng, H.-Y. Lee, J.-B. Huang, and M.-H. Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *ICLR 2020 : Eighth International Conference on Learning Representations*, 2020.

[25] Udacity. Udacity annotated driving data. `https://github.com/udacity/self-driving-car`, accessed 2021-07-21.

[26] H. Wang and Z.-H. Deng. Cross-domain few-shot classification via adversarial task augmentation. *arXiv: Computer Vision and Pattern Recognition*, 2021.

[27] T. Wang, X. Zhang, L. Yuan, and J. Feng. Few-shot adaptive faster r-cnn. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7173–7182, 2019.

[28] A. Wu, Y. Han, L. Zhu, and Y. Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[29] S. Xie, Z. Zheng, L. Chen, and C. Chen. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 5423–5432, 2018.

[30] X. Xu, X. Zhou, R. Venkatesan, G. Swaminathan, and O. Majumder. d-sne: Domain adaptation using stochastic neighborhood embedding. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2497–2506, 2019.

[31] A. Zhao, M. Ding, Z. Lu, T. Xiang, Y. Niu, J. Guan, J.-R. Wen, and P. Luo. Domain-adaptive few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1390–1399, 2021.

[32] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin. Adapting object detectors via selective cross-domain alignment. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 687–696, 2019.

[33] C. Zhuang, X. Han, W. Huang, and M. R. Scott. ifan: Image-instance full alignment networks for adaptive object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13122–13129, 2020.