WHEN FEW-SHOT MEETS CROSS-DOMAIN OBJECT DETECTION: LEARNING INSTANCE-LEVEL CLASS PROTOTYPES FOR KNOWLEDGE TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial learning is typically used to tackle cross-domain object detection, which transfers a well-trained model from a source domain to a new target domain, assuming that extensive unlabeled training data from the new domain can be easily obtained. However, such an assumption may not hold in many access-constrained target scenarios such as biomedical applications. To this end, we study the Few-Shot Domain Adaptation Object Detection (FSDAOD) problem, where only a few labeled instances from the target domain are available, making it difficult to comprehensively represent the target domain data distribution, and causing the adversarial feature alignment using only a few instances hard to transfer complete knowledge from source to target. Benefiting from the success of prototype-based meta-learning in the few-shot learning community, we propose an Instance-level Prototype learning Network (IPNet) for addressing the FSDAOD problem. The IPNet first develops an Instance-level Prototypical Meta-alignment (IPM) module, which fuses instances from both domains to learn the domain-invariant prototypical representations, for boosting the adaptation model's discriminability between different classes. Then a Feature-level Spatial Attention Transfer (FSAT) module is further developed, which employs the instance-level prototypes to discriminate various features' salience in one domain to make the model attend to foreground regions, then transfers such attention extraction knowledge to the target images via adversarial learning. Extensive experiments are conducted on several datasets with domain variances, including cross-weather changes, cross-sensor differences, and cross-style variances, and results show the consistent accuracy gains of the IPNet over state-of-the-art methods, *e.g.*, 9.7% mAP increase on Cityscapes-to-FoggyCityscapes setting and 3.0% mAP increase on Sim10k-to-Cityscapes setting.

1 INTRODUCTION

The purpose of Domain Adaptive Object Detection (DAOD) (Chen et al., 2018; Inoue et al., 2018) is to generalize a well-trained detection model from the source domain to a new target domain, which is crucial in many real applications such as autonomous driving and surveillance that often encounter severe train and test environment variations. The main challenge for DAOD is how to find the discriminative spatial regions or object instances from both the train and test images, and reduce the instance-level inter-domain distribution gap during the model adaptation process.

Generally, the typical solution to alleviate the train-test domain gap is to employ adversarial learning methods (Ganin & Lempitsky, 2014; Long et al., 2018) that can align the source-to-target cross-domain features. Several attempts (Chen et al., 2018; Deng et al., 2021; VS et al., 2021; Jiang et al., 2022; Hsu et al., 2020; Zhu et al., 2019) have been made to leverage adversarial learning for tackling the DAOD problem. The first one (Chen et al., 2018) is to align the high-level features extracted by the backbone and instance-level features obtained from the region proposal module (Ren et al., 2015). However, it is found that adversarially aligning the high-level backbone features may hurt the detector's discriminability (Chen et al., 2019; 2020). Hence D-adapt framework (Jiang et al., 2022) decouples the adversarial adaptation process from the detector training stage, to separately perform adaptation and detector learning. It should be noted that these DAOD methods (Saito et al.,

2019; Hsu et al., 2020; Deng et al., 2021; VS et al., 2021; Jiang et al., 2022) all assume that the target domain has a large number of unlabeled training samples, which is often difficult to guarantee in many real applications such as biomedical scenarios. As a result, the study of DAOD given only few-shot target samples becomes meaningful and urgent.

However, due to the limited representation capability of few-shot target samples, achieving the above-mentioned Few-Shot Domain Adaptation Object Detection (FSDAOD) has two more challenges than a typical DAOD task as follows.

Few-shot worsens between-class confusion in target domain: Between-class confusion in the target domain, meaning that some instances in one class are incorrectly classified as another class, becomes worse for few-shot scenario, as illustrated in Fig. 1(b). The reasons include the insufficient representation capability for different classes in the target domain, and the model overfitting and adaptation bias for the specific few-shot target samples in the adversarial learning process. Therefore in this work, we propose to fuse the limited number of target samples and plenty of source samples together to generate a set of discriminative instance prototypes via meta-learning, which helps to enhance the feature representation for each class in the target domain to make within-class features compact and inter-class features faraway. These prototypes also help to achieve better cross-domain feature alignment for the same class when integrated with the adversarial adaptation process, as shown in Fig. 1(c).

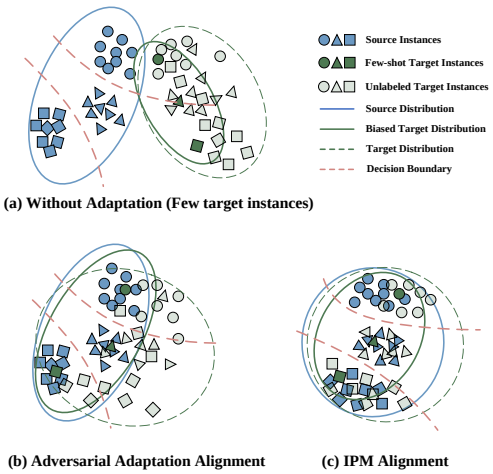

Figure 1: Since few-shot target instances are difficult to represent the overall target domain distribution, performing adversarial alignment using these few instances thus causes insufficient adaptation with between-class confusion for the target domain data.

Few-shot weakens foreground attention extraction in target domain: Extracting important foreground regions and instances is important for achieving successful DAOD. However, even though a well-trained source model has good foreground attention extraction capability, under the few-shot scenario, it is still hard to transfer such capability from the source domain to the target domain. This may become worse when target images are loosely annotated, where some very small or severely occluded objects in one image may not be annotated, causing the learning of object semantics to be ambiguous. Therefore in this work, we propose to bridge the source and target domain attention extraction gap by leveraging the learned prototypes from fused domains as mentioned above, and utilizing such prototypes to achieve cross-domain attention knowledge transfer.

In view of these, we propose an Instance-level Prototype learning Network (IPNet) for tackling the FSDAOD task. Unlike traditional DAOD works that adopt adversarial adaptation only for achieving source-to-target feature alignment, we revisit the FSDAOD problem and solve it from the perspective of prototype-based meta-learning, a typical solution in FSL community. In particular, the IPNet first develops a key Instance-level Prototypical Meta-alignment (IPM) module, which aims to meta-learn a set of representative prototypes by fusing samples of both domains (source and target), and enforces sample features of the same class but from different domains to be as close as possible for feature alignment. This effectively reduces the cross-domain feature gap for the same class, and meanwhile enhances the between-class discrimination in the target domain. Further, a Feature-level Spatial Attention Transfer (FSAT) module is developed, which computes the semantic similarity between each learned instance-level prototype and various proposal representations, to identify the spatially discriminative foreground regions and transfer such attention knowledge from source to target.

The main contributions of this paper can be summarized as follows:

1) From a new angle of meta-learning, we investigate the source-to-target cross-domain knowledge transfer under the few-shot scenario, and propose to embed the meta-learning into the typical adversarial adaptation process to relieve the issue of adaptation overfitting to only a few specific samples in the target domain.
2) We propose an IPNet framework to solve the FSDAOD. The IPNet consists of a developed IPM module to enhance the between-class discrimination during the target domain adaptation process, and a FSAT module to make the adapted model attend to the loosely distributed target objects in the input image.
3) We conduct experiments on six open benchmarks including Cityscapes, FogyCityscapes, Sim10k, PASCAL VOC, Clipart, and Udacity. The experimental results demonstrate that our IPNet can bring consistent detection accuracy gains under the FSDAOD scenario, even outperforming the unsupervised DA (UDA) methods that can access a large number of target samples.

2 Related Works

2.1 Domain Adaptation for Object Detection (DAOD)

Generally, the domain adaptation methods for object detection task can be categorized into pixel-level based (Russo et al., 2018; Kim et al., 2019) and feature-level based (Chen et al., 2018; Deng et al., 2021; VS et al., 2021; Jiang et al., 2022; Hsu et al., 2020; Zhu et al., 2019). The former first generates images or instances approximating target domain distribution using the well designed generative model such as CycleGAN (Zhu et al., 2017), then retrains the source-only detector on these synthesized instances. This kind of method usually imposes a heavy computational burden on the final model adaptation. On the other hand, the feature-level adaptation employs the patch-level or instance-level alignment by means of adversarial learning to reduce the inter-domain discrepancy of features. For example, DA-Faster (Chen et al., 2018) is the first work that tackles the DAOD problem. UMT (Deng et al., 2021) utilizes the consistency constraints between the teacher and student models via knowledge distillation. MeGA (VS et al., 2021) introduces object class information into the adaptation process. D-adapt (Jiang et al., 2022) decouples the adversarial adaptation from the detector training process, to better adapt the localization task for DAOD. Considering that all these works are based on adversarial learning or GRL module to achieve the cross-domain feature alignment, we also follow the adversarial adaptation pipeline for solving Few-Shot Domain Adaptation Object Detection (FSDAOD), but found that only adversarial-based technique **is difficult to** well adapt the detection model. This is mainly due to that only few-shot target instances cannot well represent the overall target domain data distribution, which causes the adversarial learning easy to be overfitting to the limited target domain data. Therefore, we embed the meta-learning that is a successful solution to few-shot learning problem, into the typical adversarial adaptation process to solve the FSDAOD task.

2.2 Few-Shot Object Detection (FSOD)

Since there are very few cross-domain object detection works using only a few target samples, we review the typical Few-Shot Object Detection (FSOD) works (Kang et al., 2019; Yang et al., 2020; Fan et al., 2020; Sun et al., 2021; Qiao et al., 2021; Li et al., 2021) to help for solving our task. FSDet (Kang et al., 2019) first tries to tackle the FSOD problem by re-scaling features from the channel dimension. RepMet (Yang et al., 2020) effectively encodes negative proposals for knowledge transfer from the query to support set. FSCE (Fan et al., 2020) aims to learn contrastive object proposal encodings which facilitate the classification of instance objects. DeFRCN (Qiao et al., 2021) develops a Gradient Decoupled Layer for multi-task decoupling. CME (Li et al., 2021) designs a fully connection layer to decouple localization features for boosting few-shot detection accuracy. We find that these FSOD works construct support and query sets to learn prototypes for representing sample-scarce scenarios. Thus, we similarly develop a new instance-level prototypical meta-alignment module to construct more representative prototypes. Meanwhile, since this work still follows the DA pipeline, the developed prototypical meta-alignment module is further integrated with the adversarial adaptation to attain a state-of-the-art FSDAOD accuracy.

3 The Proposed Method

The overall framework is illustrated in Fig. 2. For easy understanding, we first present the problem formulation and the selected baseline model. Then we introduce the proposed IPNet framework and its key modules. Finally, we give the overall loss function and few-shot domain adaptation strategy.

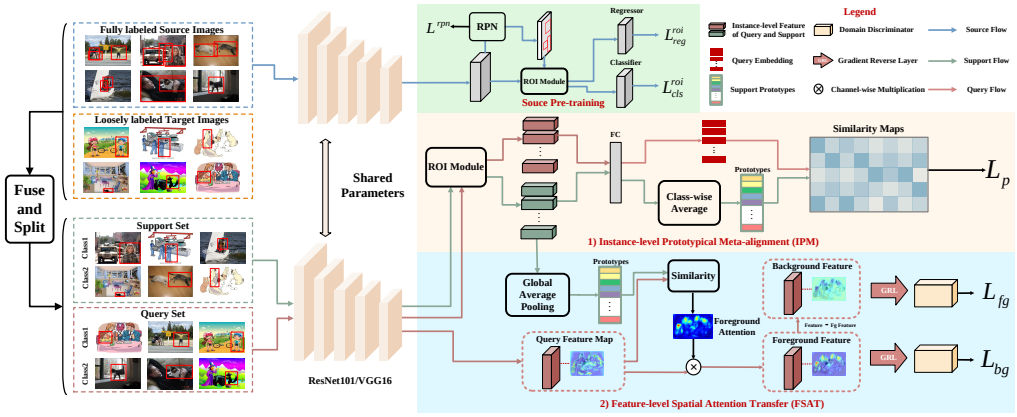

Figure 2: The IPNet consists of an IPM module and a FSAT module to tackle the FSDAOD problem. Given the source domain samples, the upper pre-training branch first trains a source-only detection baseline. Next, the source and target samples are fused to construct the query set and support set, which are first fed into the shared backbone followed by the IPM to meta-learn a set of class prototypes, and the FSAT to achieve cross-domain attention transfer. In the inference stage, the IPM and FSAT are removed and only the adapted Faster R-CNN is used for cross-domain detection.

3.1 PRELIMINARIES

Problem Definition. Suppose $(x_i^s, b_i^s, c_i^s)_{i=1}^{n_s}$ and $(x_i^t, b_i^t, c_i^t)_{i=1}^{n_t}$ are image annotation triplets from the source domain s and target domain t, where n_s and n_t denote the source and target domain sample number, and F and I denote the learned feature-level and instance-level representations of a well-trained detector, respectively. Given a large number of annotated instances from the source domain $(x_i^s, b_i^s, c_i^s)_{i=1}^{n_s}$ and very few instance objects from the target domain $(x_i^t, b_i^t, c_i^t)_{i=1}^{n_t}$, where $n_t \ll n_s$, the purpose of Instance-level Prototype learning Network (IPNet) is from extensive source data and only a few target samples, to learn generalized representations F and I which can well adapt to the target domain. Note that in our task, each image from the target domain often only contains loose annotation information, meaning that only a part of instances in the whole image are annotated.

Baseline Introduction. Following previous DAOD works (Chen et al., 2018; Deng et al., 2021; VS et al., 2021; Jiang et al., 2022; Hsu et al., 2020; Zhu et al., 2019), we use the Faster R-CNN as the baseline detector, which is a typical two-stage detection framework where instance-level features can be effectively encoded via Region-of-Interest (RoI) module. The overall loss function of Faster R-CNN can be written as follows:

$$L_{det} = L_{cls}^{rpn} + L_{reg}^{rpn} + L_{cls}^{roi} + L_{reg}^{roi}, \tag{1}$$

where L_{cls}^{rpn} and L_{reg}^{rpn} denote the classification and bounding box regression loss of Region Proposal Network (RPN), while L_{cls}^{roi} and L_{reg}^{roi} are the instance-level classification and regression loss calculated by RoI head. We employ Eq. 1 to pre-train the model on the source domain.

3.2 THE PROPOSED IPNET FRAMEWORK

The proposed IPNet framework consists of an Instance-level Prototypical Meta-alignment (IPM) module and a Feature-level Spatial Attention Transfer (FSAT) module, which work together to tackle the few-shot cross-domain object detection problem.

3.2.1 INSTANCE-LEVEL PROTOTYPICAL META-ALIGNMENT MODULE

The goal of instance-level domain alignment is to make the learned instance features domain-invariant and meanwhile discriminative for different target categories. Unfortunately, under the few-shot target domain setting, it is hard to align inter-domain features using traditional adversarial learning methods (Chen et al., 2018; Saito et al., 2019) as explained before. Here we also validate this by the comparison experiment results of SWDA (Saito et al., 2019) under different domain adaptation settings as in Table 1.

Considering this, we endeavor to solve the FSDAOD problem from the Few-Shot Object Detection (FSOD) view. To be specific, the source domain instances are regarded as the base classes while

the few-shot target samples are considered as the novel classes. Through this, the typical few-shot learning approach such as TFA method (Wang et al., 2020) for solving FSOD can be applied to handle the FSDAOD problem. The preliminary experimental results are shown in Table 1, and results show a significant cross-domain **performance drop** when the meta-learning based FSOD method is directly used to solve the FSDAOD problem. Such performance degradation is mainly due to that these FSOD methods assume similar train and test data distribution. However, in the DAOD setting, the training and test sets often present a substantial domain gap.

To address this, we design the Instance-level Prototypical Meta-alignment (IPM) module, by learning multiple prototypes from fused domains with the aid of meta-learning, to strengthen the model's representation capability for the few-shot target domain. In particular, given the input set $(x_i^s, b_i^s, c_i^s)_{i=1}^{n_s}$ and $(x_i^t, b_i^t, c_i^t)_{i=1}^{n_t}$ from different domains, the first step is to build sub-tasks for meta-learning. Suppose that a N-way K-shot task set is adopted, we first randomly sample N classes among all c^s classes (If c^s contains only one class, background is considered as an extra class). Then we construct a fusion set $\mathbf{U}$ consisting of NK source domain and NK target domain images. For both domains, K instances from different images of each class are randomly sampled using the given annotations. If the number of given instances is smaller than K, we simply duplicate target images and their annotations to satisfy the required instance number. Through this, the number of images from the source and target domains in the fusion set $\mathbf{U}$ keeps a balance. If background regions are needed, we get samples from proposals with low confidence scores and suitable sizes, generated by the RPN module of the source pre-trained Faster R-CNN. For each of N categories, $2K$ samples are randomly divided into **support set** $\left(x_i^S, b_i^S, c_i^S\right)\left(c_i^S = 1 \cdots N\right)$ and **query set** $\left(x_i^Q, b_i^Q, c_i^Q\right)\left(c_i^Q = 1 \cdots N\right)$ with K images each as follows:

$$\{(x_i^S, b_i^S, c)\} \cup \{(x_i^Q, b_i^Q, c)\} = \{(x_i^s, b_i^s, c) | x_i^s \in \mathbf{U}\} \cup \{(x_i^t, b_i^t, c) | x_i^t \in \mathbf{U}\} (c = 1 \cdots N). \tag{2}$$

Based on the obtained support and query sets, we calculate class prototypes using annotated instances in the support set. The feature extractor and RoI module of the detection model are used to generate instance features, which are flattened and fed into a new fully-connected (FC) layer for feature embedding. Then the averaging strategy as in ProtoNet (Snell et al., 2017) is employed to calculate support class prototypes. The prototype of the j^{th} class is thus formulated as follows:

$$P_j = \frac{1}{K} \sum_{c_i^S = j} w(I(x_i^S, b_i^S)), \tag{3}$$

where $w(\cdot)$ is the FC layer for the feature embedding. Then we calculate the semantic similarity between the embedded features of query instances and support prototypes as follows:

$$Cos(w(I(x_i^Q, b_i^Q)), P_j) = \frac{P_j \cdot w(I(x_i^Q, b_i^Q))}{\|P_j\| \times \|w(I(x_i^Q, b_i^Q))\|}. \tag{4}$$

In order to enhance the between-class discrimination meanwhile keep within-class compactness between query instances, we introduce a triplet loss used for training on all pairs of query instances within a mini-batch. The triplet loss reduces the intra-class divergence between the prototype and its positive instances, and increases the inter-class distance between the prototype and negative instances, formulated as follows:

$$L_p = \sum_{x \in \mathbb{B}} ReLU(Cos(w(I(x_j^Q, b_j^Q)), P_{c_i^Q}) - Cos(w(I(x_i^Q, b_i^Q)), P_{c_i^Q}) + \alpha), \tag{5}$$

where $\mathbb{B}$ denotes a mini-batch of query images, and α is the margin constant. For an input pair, we use $ReLU$ to ensure that the distance of samples belonging to the same class is minimized while that of samples from different classes is maximized. Through enhancing the model between-class discriminability in the target domain, the triplet loss also helps to alleviate the cross-domain feature discrepancies for the same class.

3.2.2 Feature-level Spatial Attention Transfer (FSAT) Module

As mentioned before, to make the source-domain model well attend to the foreground regions in the target domain, we develop the FSAT module to decouple generated feature maps into foreground

and background parts with the learned **class prototype**, and make the model focus more on the foreground during the adaptation process.

In particular, we first sample the support and query sets as in the IPM module 3.2.1. In an N-way K-shot sub-task (N is set to the total number of categories), K support images $\left(x_i^S, b_i^S, c_i^S\right)\left(c_i^S = 1 \cdots N\right)$ and K query images $\left(x_i^Q, b_i^Q, c_i^Q\right)\left(c_i^Q = 1 \cdots N\right)$ are selected from each class. In order to align the channel number between the learned prototypes and feature maps to be enhanced, global average pooling is used to obtain embeddings of all support instances from the features aligned by IPM. Then these extracted embeddings are averaged in each of N support classes to obtain class prototypes, which are concatenated to get a prototype-level matrix $\mathbf{P} \in t^{N \times C}$.

After getting the matrix $\mathbf{P} \in t^{N \times C}$, a patch-wise similarity matrix $\mathbf{S} \in t^{N \times H \times W}$ is computed between feature maps $\mathbf{F} = F(x) \in t^{H \times W \times C}$ and $\mathbf{P}$, with each item indicating the probability of each pixel of the feature map belonging to a foreground object. The similarity matrix is computed as follows,

$$\mathbf{S}_{i,j,k} = \frac{\mathbf{P}_{i,:} \cdot \mathbf{F}_{j,k,:}}{\|\mathbf{P}_{i,:}\| \times \|\mathbf{F}_{j,k,:}\|}, \tag{6}$$

where $\mathbf{P}_{i,:}$ denotes the i^{th} class prototype. Then a foreground spatial attention $\mathbf{A}_{fg} \in f^{1 \times H \times W}$ can be obtained by channel-wise maximum of $\mathbf{S}$. The foreground feature $\mathbf{F}_{fg} \in t^{H \times W \times C}$ can be acquired by multiplying $\mathbf{A}_{fg}$ with each channel of $\mathbf{F}$. The background feature $\mathbf{F}_{bg}$ is simply obtained by subtracting $\mathbf{F}_{fg}$ from $\mathbf{F}$.

Finally, a domain discriminator is employed to perform the cross-domain alignment for the decoupled foreground and background representations respectively as follows:

$$\begin{aligned} L_{fg} = &-\frac{1}{HW}\sum_{w=1}^{W}\sum_{h=1}^{H}(1 - D(\mathbf{F}_{fg})_{wh})^{\gamma} log(D(\mathbf{F}_{fg})_{wh}) \\ &-\frac{1}{HW}\sum_{w=1}^{W}\sum_{h=1}^{H} D(\mathbf{F}_{fg})_{wh}^{\gamma} log((1 - D(\mathbf{F}_{fg})_{wh})), \end{aligned} \tag{7}$$

where D denotes the domain discriminator, γ is the focal loss parameter, and L_{fg} denotes the foreground domain discrimination loss, and the L_{bg} can be easily calculated by feeding the background features $\mathbf{F}_{bg}$ into Eq. 7.

3.3 The Overall Objective and Adaptation Strategy

Overall Objective. The overall objective of the proposed IPNet can be formulated as follows:

$$L = L_{det}^s + L_p + L_{fg} + L_{bg}, \tag{8}$$

where L_{det}^s denotes the original detection loss calculated on the source domain images.

Few-shot Adaptation Strategy. IPNet contains three training steps: source pre-training, instance-level prototypical few-shot alignment, and feature-level spatial attention transfer. **1) Source pre-training:** In this step, the detector is pre-trained on labeled source data $(x_i^s, b_i^s, c_i^s)_{i=1}^{n_s}$, and learns knowledge that is useful for source image detection. **2) Instance-level prototypical meta-alignment (Section 3.2.1):** We fuse training images from both domains and split them into support and query sets to build the training sub-tasks, and train the detector with the triplet loss in Eq. 5. **3) Feature-level prototypical spatial attention transfer (Section 3.2.2):** Based on the learned class prototypes above, we perform the domain alignment of dense representations for the foreground and background regions using Eq. 7.

4 Experiments

4.1 Datasets and Scenarios

Datasets. The following six datasets are used to conduct cross-domain object detection experiments: Cityscapes (Cordts et al., 2016), FoggyCityscapes (Sakaridis et al., 2018), SIM10K (Johnson-Roberson et al., 2017), Udacity (Udacity, 2018), Pascal VOC (Everingham et al., 2010) and Clipart

Table 1: Results for the cross-weather scenario (Cityscapes to FoggyCityscapes). Unsupervised domain adaptation (UDA) denotes the traditional setting where all unlabeled images from the target domain are available. In the few-shot supervised setting (FDA), only a few loosely annotated target images are used for training. Few-shot UDA provides the same number of target images as that in FDA but with no annotations. We report the mean value and standard deviations of the mAP over three runs.

	Setting	Backbone	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
F-RCNN	Source-only	VGG16	25.1	32.7	31.0	12.5	23.9	9.1	23.7	29.1	23.4
		ResNet101	33.8	34.8	39.6	18.6	27.9	6.3	18.2	25.5	25.6
DA-Faster (CVPR'18)	UDA	VGG16	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.7
SWDA (CVPR'19)		VGG16	36.2	35.3	43.5	30.0	29.9	42.3	32.6	24.5	34.3
CADA (ECCV'20)		VGG16	41.9	38.7	**56.7**	22.6	41.5	36.8	24.6	35.5	36.0
UMT (CVPR'21)		VGG16	**56.5**	37.3	48.6	30.4	33.0	**46.7**	**46.8**	34.1	41.7
MeGA (CVPR'21)		VGG16	25.4	**52.4**	49.0	**37.7**	**46.9**	34.5	39.0	**49.2**	**41.8**
D-adapt (ICLR'22)		VGG16	44.3	48.1	54.6	28.6	34.4	28.5	33.8	41.1	39.2
CADA (ECCV'20)		ResNet101	41.5	43.6	**57.1**	29.4	**44.9**	**39.7**	29.0	36.1	40.2
D-adapt (ICLR'22)		ResNet101	**42.8**	**48.4**	56.8	**31.5**	42.8	37.4	**35.2**	**42.4**	**42.2**
SWDA (CVPR'19)	FUDA	VGG16	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) (1-shot)	FDA	VGG16	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 (WACV'22) (1-shot)		VGG16	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
TFA (ICML'20) (1-shot)		ResNet101	35.8±0.1	50.6±0.1	36.1±0.0	25.1±0.3	35.8±0.9	18.2±0.0	41.0±1.4	44.7±0.1	35.9±0.1
IPNet (1-shot)		VGG16	39.4±0.1	49.6±0.4	52.4±0.1	32.9±1.0	47.3±0.8	30.5±1.9	38.2±1.2	45.7±0.4	41.9±0.2
IPNet (1-shot)		ResNet101	46.7±0.9	59.9±0.4	53.2±0.2	**42.8±1.0**	54.0±2.0	59.9±1.5	47.3±1.8	58.9±0.4	52.8±0.4
IPNet (3-shot)		ResNet101	47.3±0.1	59.3±0.1	**53.4±0.1**	42.6±0.9	55.9±3.0	**60.7±1.4**	47.5±1.2	59.4±0.1	53.2±0.1
IPNet (5-shot)		ResNet101	**47.6±0.4**	**60.1±0.1**	53.3±0.1	42.1±1.5	**59.1±0.3**	60.3±0.5	**49.7±0.8**	**59.8±0.4**	**54.2±0.2**

Table 2: Results for fake-to-real scenario.

	Setting	Backbone	mAP
F-RCNN	Source-only	VGG16	34.6
		ResNet101	41.8
DA-Faster (CVPR'18)	UDA	VGG16	38.9
SWDA (CVPR'19)		VGG16	40.1
CADA (ECCV'20)		VGG16	49.0
UMT (CVPR'21)		VGG16	43.1
MeGA (CVPR'21)		VGG16	44.8
D-adapt (ICLR'22)		VGG16	**50.3**
CADA (ECCV'20)		ResNet101	51.2
D-adapt (ICLR'22)		ResNet101	**53.2**
FAFRCNN (CVPR'19) (24-shot)	FDA	VGG16	39.8±0.6
PICA (WACV'22) (24-shot)		VGG16	42.1±0.7
IPNet (24-shot)		VGG16	45.1±0.7
IPNet (5-shot)		ResNet101	43.3±0.3
IPNet (8-shot)		ResNet101	44.9±0.7
IPNet (24-shot)		ResNet101	46.5±0.5
IPNet (1/64 target images)		ResNet101	**55.8±0.5**

Table 3: Results for cross-camera scenario.

	Setting	Backbone	mAP
F-RCNN	Source-only	VGG16	43.1
		ResNet101	44.5
SWDA (CVPR'19)	UDA	VGG16	**51.9**
FAFRCNN (CVPR'19) (24-shot)	FDA	VGG16	50.6±0.6
PICA (WACV'22) (24-shot)		VGG16	52.4±0.1
IPNet (24-shot)		VGG16	52.8±0.1
IPNet (5-shot)		ResNet101	52.4±0.7
IPNet (8-shot)		ResNet101	53.0±0.8
IPNet (24-shot)		ResNet101	**53.5±0.5**

(Inoue et al., 2018). The Cityscapes contains 2,975 training images and 500 validation images with 8 object categories. FoggyCityscapes is a synthetic dataset generated from Cityscapes with three different levels of fog (0.005, 0.01, 0.02). We choose the level of 0.02 in our experiments. SIM10K consists of 10K images rendered by the game engine and has 58,701 car annotations. The Udacity self-driving dataset (Udacity for short) contains more than 20,000 frames collected from driving cars in California. In our experiments, we choose the 'car' category from the 4 different categories. The Pascal VOC covers 20 categories of common real-world objects and 16551 images, while Clipart contains 1000 comic images with the same 20 categories as Pascal VOC.

Our Scenarios. To compare with UDA and FDA methods, we establish the following four cross-domain scenarios: 1) cross-weather scenario from Cityscapes to FoggyCityscapes, 2) fake-to-real scenario from SIM10K to Cityscapes, 3) cross-sensor scenario from Udacity to Cityscapes, and 4) cross-style scenario from Pascal VOC to Clipart. In the first scenario, there are domain shifts from normal to extreme weather. The second scenario denotes a fake-to-real adaptation case. The third scenario denotes the sensor adaptation between different cameras. The first three scenarios are important for autonomous driving, where only a few target data can be obtained due to the expensive cost of data collection. In the last scenario, the domain gap between two datasets is greater than that of the other three scenarios.

4.2 Implementation Details

In the first three scenarios, all images are rescaled by setting the shorter side of each image to 600 pixels while keeping the aspect ratios unchanged. In the last scenario, we utilize CycleGAN (Zhu et al., 2017) to perform source-to-target translation due to the large domain shift. We use Faster R-CNN as the base detection model, and VGG16 and ResNet101 as the backbone network. Three rounds of experiments are conducted for each scenario. In each round, we randomly sample different few-shot target images and annotations. The training process is given in Section 3.3.

Table 4: Results for cross-style scenario. Due to space limitation, we only report mean mAP over three rounds.

	Setting	Backbone	aero	bcycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hrs	bike	prsn	plnt	sheep	sofa	train	tv	mAP
F-RCNN	Source-only	ResNet101	35.6	52.5	24.3	23.0	20.0	43.9	32.8	10.7	30.6	11.7	13.8	6.0	36.8	45.9	48.7	41.9	16.5	7.3	22.9	32.0	27.8
DA-Faster	UDA		15.0	34.6	12.4	11.9	19.8	21.1	23.2	3.1	22.1	26.3	10.6	10.0	19.6	39.4	34.6	29.3	1.0	17.1	19.7	24.8	19.8
SWDA			26.2	48.5	32.6	33.7	38.5	54.3	37.1	18.6	34.8	58.3	17.0	12.5	33.8	65.5	61.6	52.0	9.3	24.9	54.1	49.1	38.1
HTCN			33.6	58.9	34.0	23.4	45.6	57.0	39.8	12.0	39.7	51.3	21.1	20.1	39.1	72.8	63.0	43.1	19.3	30.1	50.2	51.8	40.3
UMT			39.6	59.1	32.4	**35.0**	45.1	61.9	48.4	7.5	46.0	**67.6**	21.4	**29.5**	**48.2**	**75.9**	**70.5**	**56.7**	**25.9**	28.9	39.4	43.6	44.1
D-adapt			**49.3**	**63.8**	**40.1**	34.4	**49.5**	**87.3**	**51.2**	**33.3**	**47.5**	59.1	**27.9**	22.4	42.3	73.9	68.2	49.1	24.9	**35.1**	**58.9**	**64.6**	**49.1**
IPNet (1-shot)	FDA		34.5	59.0	37.2	32.0	40.1	64.1	41.0	19.6	45.4	48.4	**36.2**	29.0	34.1	76.6	67.3	45.1	24.5	36.0	46.0	43.9	43.0
IPNet (3-shot)			**46.9**	68.3	33.4	34.3	**43.4**	64.1	**48.0**	30.1	47.4	56.7	31.0	33.6	43.7	80.8	67.8	**46.0**	28.2	**39.8**	**60.4**	57.5	48.0
IPNet (5-shot)			44.6	**71.0**	**37.4**	**38.9**	42.4	**74.6**	**48.0**	**34.8**	**51.5**	**67.0**	31.6	**36.7**	**47.4**	**94.2**	**69.3**	44.6	**29.6**	36.5	52.5	**66.9**	**51.0**

For **source pre-training**, we train the basic Faster R-CNN model using stochastic gradient descent with a momentum of 0.9 in all scenarios. We adopt a learning rate of 0.005 for VGG16 backbone and 0.02 for ResNet101 backbone. **The IPM** adopts 5-way-5-shot setting in the first scenario, 2-way-5-shot in the second and third scenarios, and 8-way-5-shot in the last scenario. The **FSAT** adopts 8-way-5-shot setting in the first scenario, 2-way-5-shot in the second and third scenarios, and 20-way-5-shot in the last scenario. Our implementation is built upon Detectron2 (Wu et al., 2019). Codes will be available soon.

4.3 Experimental Results

We compare our method with state-of-the-art methods on UDA task including DA-Faster (Chen et al., 2018), SWDA (Saito et al., 2019), CADA (Hsu et al., 2020), UMT (Deng et al., 2021), MeGA (VS et al., 2021), D-adapt (Jiang et al., 2022) and FDA task including FAFRCNN (Wang et al., 2019), PICA (Zhong et al., 2022). We also implement FSOD method such as TFA (Wang et al., 2020) under the FSDAOD setting.

Results for cross-weather scenario. We first show results on cross-weather scenario using the Cityscapes dataset as the source domain and FoggyCityscapes as the target domain. In this scenario, k-shot means that k images from each class with 1 bounding box annotation are randomly sampled. Table 1 shows that the IPNet outperforms the state-of-the-art FDA methods by 9.7% on mAP under the same 1-shot setting. When only a few images are available on the target domain, UDA methods will face significant performance drop. For example, D-adapt (Jiang et al., 2022) obtains a relatively high accuracy of 42.2% under the UDA setting, while SWDA (Saito et al., 2019) only achieves 30.6% mAP under the few-shot UDA setting. As a comparison, our IPNet even outperforms many UDA methods by a large margin.

Results for fake-to-real scenario. We use SIM10K as the source domain and Cityscapes as the target domain. Following previous works (Wang et al., 2019; Zhong et al., 2022), we sample 8 images from the target domain and randomly annotate 3 car objects per image as the 24-shot setting. Further, we report the mAP of target validation set on car category. Extra experiments for 5/8-shot (5/8 target images and one annotation each) and 1/64 target images (44 images with full annotation) are conducted to show the effectiveness of the IPNet. Table 2 shows that the IPNet outperforms all FDA methods and achieves significant accuracy improvement.

Results for cross-sensor scenario. We perform adaptation from Udacity to Cityscapes and report results in Table 3. In this scenario, the FDA setting for the target domain is the same as that in the second scenario. Results illustrate that the cross-domain detection accuracy of IPNet is higher than previous UDA and FDA methods.

Results for cross-style scenario. In this scenario, the source-to-target domain gap is large. For the target domain dataset Clipart, we randomly sample k bounding boxes per class as k-shot setting. One image may contain multiple annotated instances belonging to the same class. We have 52 images for 3-shot setting and 82 images for 5-shot setting. Note that previous FDA works have not conducted experiments in this scenario. As shown in Table 4, under the 5-shot setting, our IPNet outperforms the state-of-the-art UDA method by 1.9% when using about 8% target images and 3% target annotations.

4.4 Ablation Studies

Comparisons between prototypical-based and adversarial-based alignment. Table 5 shows the comparison results between the proposed method and other adversarial-based methods (Chen et al., 2018; Saito et al., 2019). First, the IPM performs much better than the corresponding instance-level and feature-level adversarial adaptation methods. Compared with adversarial adaptation methods that fail to model the overall target data distribution, our IPM leverages the domain-fused prototypes

Table 5: Comparison between IPM, FSAT and other adversarial learning methods.

Modules	mAP
Source-only	25.6
Adversarial-instance	33.8 ± 0.0
IPM	47.2 ± 0.8
IPM+Adversarial-feature	50.0 ± 0.5
IPM+FSAT	52.8 ± 0.4
IPM+FSAT (0.5 threshold)	49.1 ± 0.4

Table 6: Ablation study of the IPM and FSAT modules.

	Triplet loss	CE loss	Fg align	Bg align	mAP
Source-only					25.6
IPM	✓				47.2 ± 0.8
		✓			44.9 ± 0.5
	✓	✓			47.5 ± 1.6
IPM+FSAT	✓	✓	✓		52.4 ± 0.1
	✓	✓		✓	48.5 ± 0.1
	✓	✓	✓	✓	52.6 ± 0.2

to enhance the model discriminability. Second, we also design a naive baseline, namely, **a threshold of 0.5** is used to decouple the foreground and background features from the normalized feature map. By comparing IPM+FSAT (0.5 threshold) and IPM+FSAT in Table 5, it can be concluded that the learned prototype from IPM is effective in decoupling the foreground features from feature representations.

Study of IPM and FSAT. Table 6 shows the effectiveness of the designed triplet loss and background alignment in IPM and FSAT. It can be seen that the triplet loss has better performance than CE loss, due to that the triplet loss considers relations between different query samples to enhance the compactness of query features from the same class, while the CE loss only focuses on relations between query samples and support prototypes, ignoring the model discriminability for different classes' query samples. It is also shown that simultaneously aligning the decoupled foreground and background features is the most effective way, demonstrating the importance of both background and foreground features for domain adaptation.

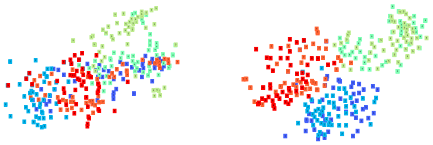

Figure 3: The tSNE results. Different colors denote different classes, and the class number $1, 4, 7$, and $2, 5, 8$ represent the source classes and target classes, respectively.

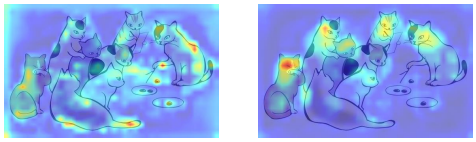

Figure 4: Visualization of the learned spatial attention features before and after applying FSAT module.

Instance feature visualization. We visualize the instance features learned for cross-weather scenario (Cityscapes to FoggyCityscapes) using t-SNE (Maaten & Hinton, 2008). The visualization demonstrates that our IPM module can enhance the discriminability of different classes' features from the target domain.

Spatial attention visualization. In Fig. 4, we visualize the high-level dense representations of the backbone learned from cross-style scenario (VOC to Clipart), to show the learned spatial attention by FSAT module. It is seen that after doing attention transfer, the model can attend more to the important foreground instances. For more visualization results, please see Figs. 5-9 in Section A

5 Conclusion

This work reveals that when only a few target samples are available, traditional adversarial learning-based alignment methods cannot work well, and thus inspires us to propose an Instance-level Prototype learning Network (IPNet) for tackling the few-shot DAOD problem. The proposed IPNet consists of an Instance-level Prototypical Meta-alignment (IPM) module to enhance the target domain between-class discriminability, and a Feature-level Spatial Attention Transfer (FSAT) module to make the adapted model attend to important foreground areas for target image detection. Experiments are conducted on six common cross-domain detection benchmarks and detection results show the consistent superiority of the proposed IPNet over state-of-the-art methods under different few-shot cross-domain scenarios.

REFERENCES

Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8869–8878, 2020.

Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International Conference on Machine Learning*, pp. 1081–1090, 2019.

Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3339–3348, 2018.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schie. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pp. 3213–3223, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4091–4101, 2021.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4013–4022, 2020.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 2014.

K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. In *International Conference on Computer Vision*, 2017.

Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *European Conference on Computer Vision*, pp. 733–748, 2020.

Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5001–5009, 2018.

Junguang Jiang, Baixu Chen, Jianmin Wang, and Mingsheng Long. Decoupled adaptation for cross-domain object detection. In *ICLR*, 2022.

Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *IEEE International Conference on Robotics and Automation*, pp. 746–753, 2017.

Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *ICCV*, 2019.

Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12456–12465, 2019.

Bohao Li, Boyu Yang, Chang Liu, Feng Liu, Rongrong Ji, and Qixiang Ye. Beyond max-margin: Class margin equilibrium for few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7363–7372, 2021.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 1640–1650, 2018.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008.

Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8681–8690, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *International Conference on Neural Information Processing Systems*, pp. 91–99, 2015.

Paolo Russo, Fabio M Carlucci, Tatiana Tommasi, and Barbara Caputo. From source to target and back: symmetric bi-directional adaptive gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8099–8108, 2018.

Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6956–6965, 2019.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, 2018.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7352–7362, 2021.

Udacity. Udacity annotated driving data. 2018. URL https://github.com/udacity/self-driving-car.

Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4516–4526, 2021.

Tao Wang, Xiaopeng Zhang, Li Yuan, and Jiashi Feng. Few-shot adaptive faster r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7173–7182, 2019.

Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E. Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9919–9928, 2020.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

Yukuan Yang, Fangyu Wei, Miaojing Shi, and Guoqi Li. Restoring negative information in few-shot object detection. In *NeurIPS*, 2020.

Chaoliang Zhong, Jie Wang, Cheng Feng, Ying Zhang, Jun Sun, and Yasuto Yokota. Pica: Point-wise instance and centroid alignment based few-shot domain adaptive object detection with loose annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2329–2338, 2022.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 687–696, 2019.

A APPENDIX

A.1 MORE DETAILS OF THE BASELINE MODEL

Network architecture. Following previous Domain Adaptation Object Detection (DAOD) works (Chen et al., 2018; Deng et al., 2021; VS et al., 2021; Jiang et al., 2022; Hsu et al., 2020; Zhu et al., 2019), we use the backbone of ResNet101 and VGG16 as the feature extraction. The first two blocks are frozen during the training phase. Also, we employ the RoI Align (He et al., 2017) as the RoI module in Faster R-CNN baseline with a pooled spatial resolution of 7×7. Domain discriminator consists of three convolution layers, where each layer employ a kernel size of 3×3 and stride of 1. Channel numbers of the three convolution layers are 1024, 256, and 1, respectively. Focal loss with Sigmoid activation function is employed as the loss of domain discriminator.

Training details. Our model is trained on 2 NVIDIA V-100 GPUs. To stabilize the training process, we apply the warm-up strategy in the early stage of the training phase. The source-only model is trained based on an Imagenet (Deng et al., 2009) pre-trained backbone. As for data augmentations, random flip is applied for all training processes and multi-scale training is used for Pascal VOC to Clipart scenario.

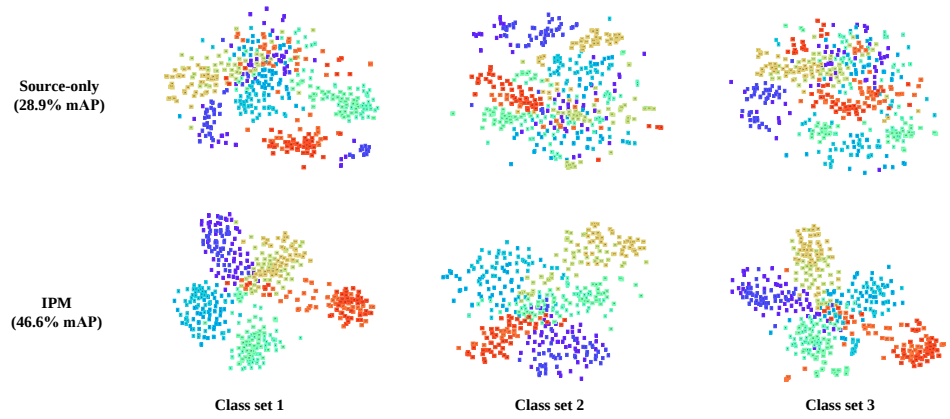

Figure 5: More tSNE visualization: we visualize the tSNE results before and after applying IPM module. Each class set contains five randomly selected classes. Class number 1,5,9,13,17 and 2,6,10,14,18 represent the source classes and target classes, respectively.

Figure 6: More features visualization: we visualize the learned spatial attention before and after applying FSAT module.

A.2 QUALITATIVE ANALYSIS

tSNE and attention map results. We have visualized the tSNE results of instance-level features learned by the source-only model and our IPNet, respectively. In this part, we give more tSNE visualization results for Pascal VOC to Clipart cross-domain scenario. As illustrated in Fig. 5, these visualization results demonstrate that our IPM module can largely boost the model discriminability for the target domain instances, enhancing the cross-domain detection accuracy under the few-shot setting.

Besides, In Fig. 6, we show more attention maps to verify the effectiveness of the learned spatial attention features by the designed FSAT. These visualized attention maps show our FSAT can successfully transfer the spatial attention to the foreground regions of the whole image.

Detection results. Furthermore, we show some visual detection results for adapting the detector under the cross-weather, fake-to-real and cross-style scenarios in Figs. 7-9. Compared with the baseline model trained only using source domain data, our IPNet can detect more instance objects only using few-shot loosely-annotated images from the target domain.

Figure 7: Detection results of adapting the source-only model from Cityscapes to FoggyCityscapes.

Figure 8: Detection results of adapting the source-only model from SIM10K to Cityscapes.

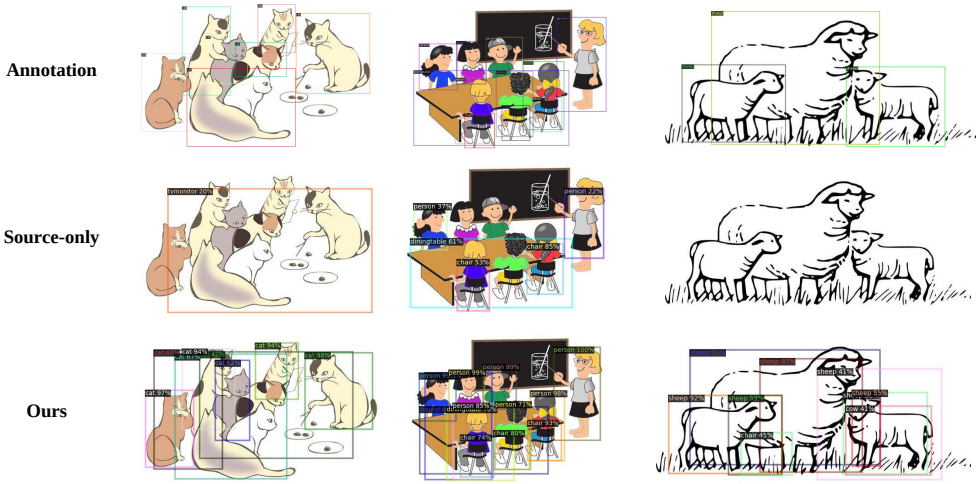

Figure 9: Detection results of adapting the source-only model from PASCAL VOC to Clipart.