iFAN: Image-Instance Full Alignment Networks for Adaptive Object Detection

Chenfan Zhuang, Xintong Han, Weilin Huang*, Matthew R. Scott
Malong Technologies, Shenzhen, China
Shenzhen Malong Artificial Intelligence Research Center, Shenzhen, China
{fan,xinhan,whuang,mscott}@malong.com

Abstract

Training an object detector on a data-rich domain and applying it to a data-poor one with limited performance drop is highly attractive in industry, because it saves huge annotation cost. Recent research on unsupervised domain adaptive object detection has verified that aligning data distributions between source and target images through adversarial learning is very useful. The key is when, where and how to use it to achieve best practice. We propose Image-Instance Full Alignment Networks (**iFAN**) to tackle this problem by precisely aligning feature distributions on both image and instance levels: 1) Image-level alignment: multi-scale features are roughly aligned by training adversarial domain classifiers in a hierarchically-nested fashion. 2) Full instance-level alignment: deep semantic information and elaborate instance representations are fully exploited to establish a strong relationship among categories and domains. Establishing these correlations is formulated as a metric learning problem by carefully constructing instance pairs. Above-mentioned adaptations can be integrated into an object detector (e.g. Faster R-CNN), resulting in an end-to-end trainable framework where multiple alignments can work collaboratively in a coarse-to-fine manner. In two domain adaptation tasks: synthetic-to-real (SIM10K $\rightarrow$ Cityscapes) and normal-to-foggy weather (Cityscapes $\rightarrow$ Foggy Cityscapes), iFAN outperforms the state-of-the-art methods with a boost of $10\%+$ AP over the source-only baseline.

Introduction

Training neural networks on one domain that generalizes well on another domain can significantly reduce the cost for human labeling, making domain adaptation a hot research topic. Researchers have studied the effectiveness of domain adaptation in various tasks, including image classification (Kumar et al. 2018; Saito, Ushiku, and Harada 2017; Shu et al. 2018; Long et al. 2018; 2017), object detection (Chen et al. 2018; Saito et al. 2019; Wang et al. 2019a; RoyChowdhury et al. 2019; Cai et al. 2019; Zhu et al. 2019) and semantic segmentation (Wu et al. 2018; Zhang, David, and Gong 2017; Hoffman et al. 2018). In this paper, we aim to train a high-performance unsupervised domain adaptive object detector on a fully-annotated source domain and apply it to an unlabeled target domain. For example, an object detector, trained on synthesized images generated from a game engine such as SIM10K (Johnson-Roberson et al. 2017) where object bounding boxes are readily available, can be applied to real-world images from a target domain, such as Cityscapes (Cordts et al. 2016) or KITTI (Geiger et al. 2013).

Recently, many efforts have been devoted into developing cross-domain models with unsupervised domain adaption. Existing approaches mainly focus on aligning deep features directly between source domain and target domain. In the context of object detection, the alignment is usually achieved by domain-adversarial training (Ganin and Lempitsky 2015; Tzeng et al. 2017) at different stages of object detectors. For example, based on a Faster R-CNN framework (Ren et al. 2015), (Chen et al. 2018) aligned the feature maps in backbone with an image-level adaptation module; then aligned the ROI-pooled features before feeding them into the final classifier and box regressor. (Saito et al. 2019) strongly aligned patch distributions of the low-level features (e.g. `conv3` layer) to enhance local consistence and weakly aligned the global image-level features before RPN.

We follow this line of research to develop multi-level domain alignments for cross-domain object detection, as shown in Figure 1. Unlike previous approaches that merely concern image-level alignment at a single convolutional layer (e.g. (Wang et al. 2019a; Chen et al. 2018)), we design hierarchically-nested domain discriminators to reduce domain discrepancies in accordance with various characteristics in the network hierarchies (Figure 1a); meanwhile the instance-level features are carefully aligned, making use of the ROI-level representations. Notice that the traditional instance-level alignments, such as (Chen et al. 2018), attempt to learn domain-invariant features without fully exploring the semantic category-level information. This inevitably leads to a performance drop due to the misalignment of objects within the same category. To address this problem, we develop a category-aware instance-level adaptation by leveraging object classification results of the detector (Figure 1b). Finally, we propose a novel category-

*Corresponding author: whuang@malong.com

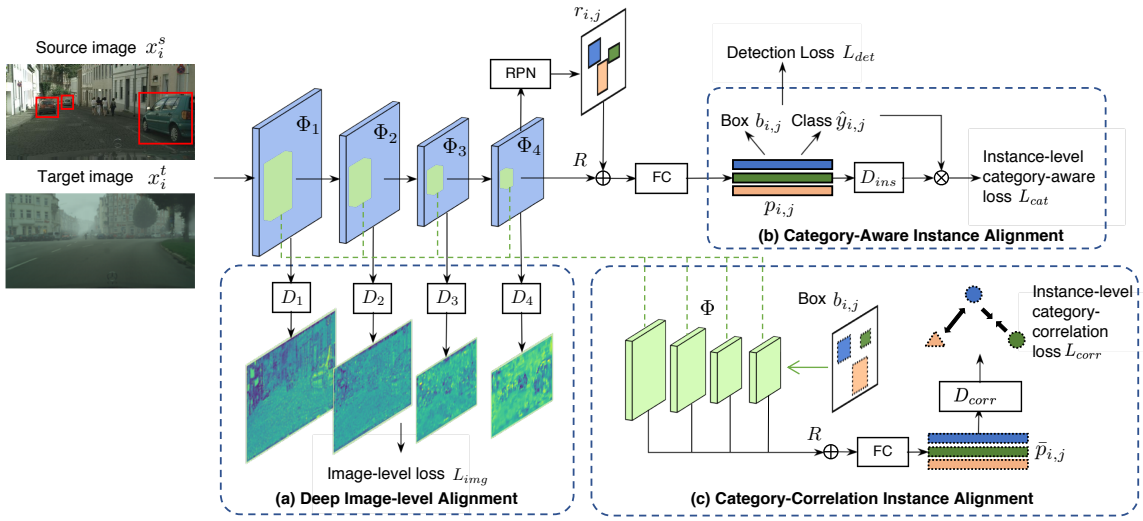

Figure 1: **Overview of the proposed iFAN with a Faster R-CNN detector**. (a) **Image-level Adaptation**: Hierarchical domain classifiers align image-level features at different semantic levels. (b) **Category-Aware Instance Adaptation**: ROI-pooled instance features are aligned in a category-aware fashion guided by the corresponding predicted class labels. (c) **Category-Correlation Instance Adaptation**: The predicted bounding boxes are utilized to extract accurate representations for the instances, then paired as input to a metric learning framework to learn the correlations across domains and categories.

correlation instance alignment: utilize the predicted bounding boxes to attain the refined instance representations and precisely align them using deep metric learning - establish the correlations among domains and categories, as shown in Figure 1c.

Contributions. The main contributions of this work are four-fold: 1) To mitigate the domain shift occurred in multiple semantic levels, we apply domain-adversarial training on multiple intermediate layers, allowing us to align multi-level features more effectively; 2) A category-aware instance-level alignment is then proposed to align ROI-based features, subtly incorporating deep category information; 3) We formulate the category-correlation instance-level alignment to a metric learning problem, further study the cross-domain category correlations; 4) Our approach surpasses the state-of-the-art unsupervised domain adaptive object detectors, e.g. (Wang et al. 2019a; Chen et al. 2018; Saito et al. 2019) on synthetic-to-real (SIM10K $\rightarrow$ Cityscapes) and normal-to-foggy weather (Cityscapes $\rightarrow$ Foggy Cityscapes) tasks.

Related Work

Unsupervised Domain Alignment. Unsupervised domain adaptation (UDA) refers to train domain invariant models on images with annotated images from source domain and images from target domain without any annotation. Many UDA methods show the effectiveness of distribution matching by reducing the domain gap. (Saito et al. 2017) focused on generating features to minimize the discrepancies between two classifiers which are trained to maximize the discrepancies on target samples. (Liu et al. 2019) generated transferable examples to fill in the gap between the source and target domain by adversarially training deep classifiers to output consistent predictions over the transferable examples. However, since object detectors often generate numerous region proposals, many of which could be background or beyond the given classes, these methods can not fit very well in object detection. Another series of solutions (Wu et al. 2018; Inoue et al. 2018; Hoffman et al. 2018), following the success of unsupervised image-to-image translation networks (Zhu et al. 2017; Liu, Breuel, and Kautz 2017; Huang and Belongie 2017), directly aligned pixel-level distributions by transferring source images into the target style, and then trained models on the transferred images. Inspired by Generative Adversarial Networks (Goodfellow et al. 2014), training a domain discriminator to identify the source from the target and then reacting on the feature extractor to deceive the discriminator, has been frequently used and proven efficient (Ganin and Lempitsky 2015; Saito et al. 2019; Chen et al. 2018; Long et al. 2018; Tzeng et al. 2017; Liu and Tuzel 2016).

Adaptive Object Detection. Object detectors with deep architectures (Girshick 2015; Girshick et al. 2014; Ren et al. 2015) play an important role in myriad computer vision applications. To eliminate the dataset bias, a number of methods have been developed to UDA object detection problems (Inoue et al. 2018; Chen et al. 2018; Saito et al. 2019; Cai et al. 2019; Zhu et al. 2019). (Inoue et al. 2018) sequentially fine-tuned an object detector with an image-to-image domain transfer and weakly supervised pseudo-labeling. (Chen et al. 2018) developed a Faster R-CNN (Ren et al. 2015) detector with feature alignment on both image-level and instance-level. Along this direction, (Saito et al. 2019) forced the image-level features to be strongly aligned in the lower layers and weakly aligned in the higher layers and concatenate them together. (Zhu et al. 2019) grouped instances into discriminatory regions and aligned region-level

features across domains. (Cai et al. 2019) learned relation graphs, which regularizes the teacher and student models to learn consistent features. Among these methods, inherent feature hierarchies and deep semantic information are not exhaustively exploited, which motivates our method.

Metric Learning. Our method also relates to metric learning aims to (Chopra et al. 2005; Schroff, Kalenichenko, and Philbin 2015; Oh Song et al. 2016; Wang et al. 2019b) as we construct pairs as input to the category-correlation adaptation. Metric learning approaches learn an embedding space where similar samples are pulled closer, while dissimilar ones are pushed apart from each other. In this paper, we train a metric learning model to draw two instances closer if they share the same category, or push them apart otherwise despite domains. The idea of learning with paired samples has also been utilized in few-shot domain adaptation approaches (Wang et al. 2019a; Motiian et al. 2017) for handling scarce annotated target data. Unlike these approaches, our category-correlation adaptation works without any supervision from the target domain. Instead, we use the predictions of classifiers as pseudo labels.

Image-Instance Full Alignment Networks

The whole pipeline of our proposed iFAN is presented in Figure 1. Given images with annotated bounding boxes from the source domain and unlabeled images from the target, our goal is to align the distributions of two domains via image-level (Figure 1a) and full instance-level alignments, including category-aware (Figure 1b) and category-correlation (Figure 1c), step by step, to boost the performance of a detector, without charging anything extra on inference. Formally, let $\{x_i^s, y_i^s\}_{i\in[N_s]}$ denote a set of N_s images $x_i^s \in \mathbb{R}^{H\times W\times 3}$ from the source domain, with corresponding annotations y_i^s. For the target domain, we only have images $\{x_i^t\}_{i\in[N_t]}$ without any annotation. An object detector (e.g. Faster R-CNN (Ren et al. 2015) in this paper) can be trained in the source domain by minimizing:

$$L_{det} = \frac{1}{N_s}\sum_{i=1}^{N_s} L(x_i, y_i), \tag{1}$$

where L is the loss function for object detection. Generally, such a detector is difficult to generalize well to a new target domain due to the large domain gap.

Deep Image-Level Alignment

Recent domain adaptive object detectors (Chen et al. 2018; Shan, Lu, and Chew 2018; Wang et al. 2019a) commonly align image-level features to minimize the effect of domain shift, by applying a patch-based domain classifier on the intermediate features drawn from a single convolutional layer (typically the global features before the RPN). Since the receptive field of each activation corresponds to a patch of the input image, a domain classifier can be trained to guide the networks to learn a domain-invariant representation for the image patch, and thus reduces the global image domain shift (e.g. image style, illumination, texture, etc.). This patch-based discriminator has also been proved to be effective on cross-domain image-to-image translation task (Isola et al. 2017; Zhu et al. 2017).

Nevertheless, these methods just focus on features extracted from a certain layer; they may miss the rich domain information contained in other intermediate layers, such as, the domain displacement of different scales. Recent works of object detection (Lin et al. 2017) and image synthesis (Zhang, Xie, and Lin 2018; Yang et al. 2018), which explore the inherent multi-scale pyramidal hierarchy of convolutional neural networks to achieve meaningful deep multi-scale representations, have greatly inspired our work.

We propose to build a hierarchically-nested domain classifier bank at the multi-scale intermediate layers, as shown in Figure 1a. Let Φ denote the backbone of an object detector. For the feature maps $\Phi_l \in \mathbb{R}^{H_l\times W_l\times C_l}$ from the l^{th} intermediate layer, a domain classifier D_l is constructed in a fully convolutional fashion (e.g. 3 convolutional layers with 1×1 kernels) to distinguish source (domain label = 0) and target (domain label = 1) samples, by minimizing a mean square-error loss as (Saito et al. 2019; Zhu et al. 2017):

$$\begin{aligned} L_{D_l} = &\frac{1}{N_s H_l W_l}\sum_{i=1}^{N_s}\sum_{m=1}^{H_l W_l} D_l(\Phi_l(x_i^s))_m^2 + \\ &\frac{1}{N_t H_l W_l}\sum_{i=1}^{N_t}\sum_{m=1}^{H_l W_l} (1 - D_l(\Phi_l(x_i^t)))_m^2. \end{aligned} \tag{2}$$

In this paper, we use `pool2`, `pool3`, `pool4`, `relu5_3` in VGG-16 backbone (Simonyan and Zisserman 2015) or `res2c_relu`, `res3d_relu`, `res4f_relu`, `res5c_relu` for ResNet50 (He et al. 2016) as the intermediate layers. Then the loss for our hierarchically-nested domain classifier bank forms:

$$L_{img} = \sum_{l=1}^{4} \lambda_l L_{D_l}, \tag{3}$$

where λ_l denotes a balancing weight, empirically set to reconcile each penalty. By minimizing L_{img}, the domain classifiers are forced to discriminate the multi-scale features of the source domain from the target; meanwhile, the detector is trying to generate domain-ambiguous features to deceive these domain discriminators via reversal gradient (Ganin and Lempitsky 2015) , yielding domain-invariant features that generalize well to the target domain.

Compared to the single global domain classifier developed in (Chen et al. 2018; Shan, Lu, and Chew 2018), our hierarchically-nested image-level alignment enjoys the following merits: 1) Receptive field with various sizes on the hierarchy enable the model to align image patches in a bigger range of scales in the spirit of feature pyramid network (Lin et al. 2017).

2) Our multi-layer alignment is designed to capture multi-granularity characteristics of domains at a time, from low-level features (e.g. texture and color) to high-level (e.g. shape). This voracious strategy can effectively reduce domain discrepancies of various kind.

3) Unlike existing domain-adversarial frameworks which might suffer from unstable training (Saito et al. 2019; Wang et al. 2019a), our proposed hierarchical supervisions could guide the alignment gradually and moderately, from shallow to deep, leading to better convergence.

Full Instance-Level Alignment

Category-Agnostic Instance Alignment Recent adaptive object detectors, e.g. (Chen et al. 2018), also integrate a category-agnostic instance domain classifier D_{ins} on the top of ROI-based features to mitigate domain shift between local instances, e.g. the appearance and the shape of objects. Following this line, we extend the image-level alignment to instance level. Let $p = R(f, r)$ denote the output of ROI-Align operation (He et al. 2017), conditioned on feature maps f and a region proposal r. $p_{i,j} = R(\Phi_4(x_i), r_{i,j})$ is the instance feature of the j^{th} region proposal of image x_i, as shown in Figure 1b. Our loss function of a naive instance alignment formulates:

$$L_{ins} = \frac{1}{N_s}\sum_{i=1}^{N_s}\frac{1}{N_i^s}\sum_{j=1}^{N_i^s} D_{ins}(R(\Phi_4(x_i^s), r_{i,j}^s))^2 + \frac{1}{N_t}\sum_{i=1}^{N_t}\frac{1}{N_i^t}\sum_{j=1}^{N_i^t}(1 - D_{ins}(R(\Phi_4(x_i^t), r_{i,j}^t)))^2, \tag{4}$$

where N_i^s and N_i^t denote the numbers of instances in x_i^s and x_i^t, respectively. To clarify the effectiveness of instance-level adaptation, we simply adopt the same architecture used in the image-level alignment, which consists of 1×1 convolutions.

However, we observed that applying such an instance-level alignment from the beginning of training may not be the best practice. At early stage of the training, the predictions of detector for both source and target images are inaccurate. With the supervision of ground-truth, knowledge from the source data can be steadily learned and simultaneously transferred to the target data. Intuitively, aligning nonsensical patches instead of the valid instances could bring negative effects: the instance alignment can make positive impact only when the detector becomes relatively stable in both source and target domains. This challenge was discussed and validated in (Saito et al. 2019) as well. To tackle this problem, we propose a technique called late launch: an activate instance-level alignment at one third of the whole training iterations. Note that the total of training iterations remains unchanged.

Category-Aware Instance Alignment Our image-level and category-agnostic instance-level alignments are able to blend the features from two domains together. However, category information has not been taken into consideration, and instances from two different domains are possible to be aligned incorrectly into different classes. For example, the feature of a car in the source may be aligned with a bus in the target, resulting in undesired performance drop.

To this end, we propose to incorporate category information into instance-level alignment by modifying D_{ins} to C-way output instead of the original single-way. In other words, each category owns a domain discriminator. Thus the c^{th} dimension of $D_{ins}(p_{i,j})$ indicates a domain label (source = 0 or target = 1) for the instance (with corresponding features $p_{i,j}$) from category c.

However, category labels for the instances from the target domain are not provided; thus the methodology to assign the category labels to target proposals is pending. Enlightened by the pseudo-labeling approach described in (Inoue et al. 2018), we directly use the classifier output of the detector $\hat{y}_{i,j}$ as *soft pseudo-labels* for target instances, as shown in Figure 1b. The classifier output indicates the probability distribution of how likely an instance belongs to the C classes. According to the possibility, domain classifiers of each category independently update their own parameters. As a result, the category-aware instance alignment loss takes the following form:

$$L_{cat} = \frac{1}{N_s}\sum_{i=1}^{N_s}\frac{1}{N_i^s}\sum_{j=1}^{N_i^s}\sum_{c=1}^{C}\hat{y}_{i,j,c}^s D_{ins}(p_{i,j}^s)_c^2 + \frac{1}{N_t}\sum_{i=1}^{N_t}\frac{1}{N_i^t}\sum_{j=1}^{N_i^t}\sum_{c=1}^{C}\hat{y}_{i,j,c}^t(1 - D_{ins}(p_{i,j}^t)_c)^2, \tag{5}$$

where the loss of domain classifier is weighted by the predicted category probability. Notice that, for source instances, we use the predicted labels in the same way as we found that this soft assignment policy factually works better than using ground truth.

Category-Correlation Instance Alignment Due to the location misalignment between a coarse region proposal and its accurate bounding box, ROI-based features may fail to precisely characterize the instances. Popular object detectors prefer to refine the bounding boxes in an iterative (Gidaris and Komodakis 2015) or cascaded (Cai and Vasconcelos 2018) manner to reduce such misalignment for higher accuracy. Similarly, in this paper, we propose to enhance instance representations by mapping the predicted bounding boxes back to the backbone feature maps, and crop the selective features out for further alignment. This process is illustrated in Figure 1c. Formally, for the j^{th} predicted object in image x_i, we use $b_{i,j}$ to denote its predicted bounding box, then pool the corresponding feature maps $\Phi_l(x_i)$ at the l^{th} layer with ROI-Align (He et al. 2017), finally shaping a group of representations for this instance: $p_{i,j,l} = R(\Phi_l(x_i), b_{i,j})$.

Moreover, following the principle of image-level alignment, we fuse these representations to combine all possible information together. The feature maps $p_{i,j,l}$ $(l = 1, 2, 3, 4)$ are individually passed into 1×1 convolutions to generate features with 256 channels; then element-wise summation is applied, yielding the refined features $\bar{p}_{i,j}$. Compared with the ROI-pooled features computed from a single layer ($\Phi_1 \sim \Phi_4$), the summation operation can improve $0.5\%+$ mAP.

We then project the refined instance features into an embedding space through a fully-connected layer denoted as D_{corr}. Given a pair of instances f_i and f_j, it should belong

to one of the four groups according to its domain and category: 1) same-domain and same-category $\mathcal{S}_{sdsc}$; 2) same-domain and different-category $\mathcal{S}_{sddc}$; 3) different-domain and same-category $\mathcal{S}_{ddsc}$; 4) different-domain and different-category $\mathcal{S}_{dddc}$. We found that minimizing the distances in S_{sdsc} and maximizing in S_{dddc} are two simplistic tasks, thanks to the previous alignments in the object detector. Therefore we only focus on S_{sddc} and S_{ddsc} to optimize the correlations of domains and categories via metric learning.

With D_{corr} used as a *metric discriminator*, we can minimize the following contrastive loss (Chopra et al. 2005):

$$L_{corr} = \frac{1}{|\mathcal{S}_{sddc}|} \sum_{(f_i,f_j)\in S_{sddc}} d(f_i,f_j)^2 + \frac{1}{|\mathcal{S}_{ddsc}|} \sum_{(f_i,f_j)\in S_{ddsc}} \max(0, m - d(f_i,f_j))^2, \tag{6}$$

where $d(f_i,f_j) = ||D_{corr}(f_i) - D_{corr}(f_j)||_2$ denotes the Euclidean distance, and m is a fixed margin. Remember that we are under an adversarial training; hence, D_{corr} ought to pull together the instance pairs in the same domain even from different categories, while pushing apart the pairs from different domains but of the same category. On the contrary, the Faster R-CNN is trying to confuse this metric discriminator by maximizing L_{corr}. As a result, the object detector can generate features that encourage: 1) Different categories are well separated within the same domain (S_{sddc}); 2) Features are domain-invariant for instances of the same class (S_{ddsc}). Both of them are the desired properties for an ideal domain adaptive classifier.

Since the category labels of target instances are not available, we again use the predicted labels of the instances to construct pairs. Similarly, late launch technique is used here too, as mentioned in the category-aware instance alignment.

Training and Inference

The full training objective of our method is:

$$\min_G \max_D L_{det}(G) - \lambda_{adv}(L_{img}(G,D) + L_{cat}(G,D) + L_{corr}(G,D)), \tag{7}$$

where G denotes a Faster R-CNN object detector, and D indicates one of the three domain classifiers: D_l, D_{cat}, or D_{corr}. λ_{adv} is the weight of adversarial loss to balance the penalty between the detection and adaptation task. The minimax loss function is implemented by a gradient reverse layer (GRL) (Ganin and Lempitsky 2015).

No worries to increase the burden on inference stage, because all alignments are only carried out during training and the modules can be easily peeled off. Although iFAN increases the computational cost during training, luckily not too much, the inference speed is identical to a vanilla Faster R-CNN as (Chen et al. 2018; Wang et al. 2019a; Cai et al. 2019) , and is faster than (Saito et al. 2019), whose inference involves computing the outputs of domain classifiers.

Method	Backbone	S → C	C → F
Oracle	VGG16	61.1	38.9
Source-only Faster R-CNN	VGG16	34.9	16.9
ADDA (Tzeng et al. 2017)	VGG16	36.1	24.9
DT (Zhu et al. 2017) + FT	VGG16	36.8	26.1
DA-Faster (Chen et al. 2018)	VGG16	40.0	27.6
SW (Saito et al. 2019)	VGG16	40.1	34.3
Few-shot (Wang et al. 2019a)	VGG16	41.2	31.3
SelectAlign (Zhu et al. 2019)	VGG16	43.0	33.8
iFAN	VGG16	**46.9**	**35.3**
Oracle	ResNet50	66.4	45.2
Source-only Faster R-CNN	ResNet50	35.1	21.0
MTOR (Cai et al. 2019)	ResNet50	46.6	35.1
iFAN	ResNet50	**47.1**	**36.2**

Table 1: Comparison with other methods. Mean average precision (mAP, %) on SIM10K → Cityscapes (S → C) and Cityscapes → Foggy Cityscapes (C → F).

	img	ins	corr	AP	gain
Source only				34.9	-
iFAN	✓			43.0	8.1
	✓	✓		46.1	11.2
	✓		✓	45.3	10.4
	✓	✓	✓	**46.9**	**12.0**

Table 2: Ablations on SIM10K → Cityscapes. img, ins, corr denote our image-level, category-agnostic and category-correlation instance alignment respectively. No category-aware alignment in this scenario, since only car is evaluated.

Experiments and Results

Experimental Setup

Datasets We evaluate **iFAN** on two domain adaptation scenarios: 1) train on SIM10K (Johnson-Roberson et al. 2017) and test on Cityscapes (Cordts et al. 2016) dataset (SIM10K → Cityscapes); 2) train on Cityscapes (Cordts et al. 2016) and test on Foggy Cityscapes (Sakaridis, Dai, and Van Gool 2018) (Cityscapes → Foggy). Rendered by the Grand Theft Auto game engine, the SIM10K dataset consists of 10,000 images with 58,701 bounding boxes annotated for cars. The Cityscapes dataset has 3,475 images of 8 object categories taken from real urban scenes, where 2,975 images are used for training and the remaining 500 for evaluation. We follow (Saito et al. 2019; Chen et al. 2018) to extract bounding box annotations by taking the tightest rectangles of the instance masks. The Foggy Cityscapes (Sakaridis, Dai, and Van Gool 2018) dataset was created by applying fog synthesis on the Cityscapes dataset and inherit the annotations. In the SIM10K → Cityscapes scenario, only the car category is used for training and evaluation, while for Cityscapes → Foggy, all 8 categories are considered. We use an average precision with threshold = 0.5 (mAP_{50}) as the evaluation metric for object detection.

Implementation Details. To make a fair comparison with existing approaches, we strictly follow the implementation details of (Saito et al. 2019). We adopt Faster R-CNN (Ren

	img	ins	cat	corr	person	car	moto	rider	bicycle	bus	train	truck	mAP	gain
Source only					21.5	28.8	13.6	21.9	21.4	16.0	5.0	7.0	16.9	-
iFAN	✓				**33.0**	47.2	25.2	**41.3**	**33.3**	41.1	15.2	23.6	32.5	15.6
	✓	✓			32.3	48.4	**28.1**	41.0	32.7	41.4	23.0	22.6	33.1	16.2
	✓		✓		32.4	**48.9**	23.9	38.3	32.5	44.8	28.5	27.5	34.6	17.7
	✓			✓	32.3	47.8	20.5	38.6	32.9	43.5	**33.0**	27.3	34.5	17.6
	✓		✓	✓	32.6	48.5	22.8	40.0	33.0	**45.5**	31.7	**27.9**	**35.3**	**18.4**

Table 3: Ablations on Cityscapes → Foggy Cityscapes. img, ins, cat and corr denote our image-level, category-agnostic, category-aware and category-correlation instance alignment respectively.

et al. 2015) + ROI-alignment (He et al. 2017) and implement all with maskrcnn-benchmark (Massa and Girshick 2018). The shorter side of training and test images are set to 600. The detector is first trained with a learning rate of $lr = 0.001$ for 50K iterations, and then $lr = 0.0001$ for another 20K iterations. The category-agnostic/aware instance-level alignment *late launches* at 30K-th iteration and category-correlation alignment at 50K-th. The *late launches* timing is empirically set according to the loss curve: a new alignment starts when the previous ones go stable. We set $\lambda_{adv} = 0.1$ in Eqn. 7 and $\lambda_l = 1.0$ in Eqn. 3. The embedding dimension of category-correlation alignment is set to 256, with a margin of $m = 1.0$. VGG-16 is used as the backbone if not specifically indicated.

Competing Methods. We compare **iFAN** with the following baselines and recent state-of-the-art methods: 1) Faster R-CNN (Ren et al. 2015): a vanilla Faster R-CNN trained only on the source domain. 2) ADDA (Tzeng et al. 2017): the deep features from the last layer of the detector backbone are aligned with a global domain classifier. 3) Domain Transfer + Fine-Tuning (DT (Zhu et al. 2017)+FT): a CycleGAN (Zhu et al. 2017) is used to transfer the source images to the target style, and then a Faster R-CNN detector is trained on the transferred images. A similar approach is also described in (Inoue et al. 2018). 4) Domain adaptive Faster R-CNN (DA-Faster) (Chen et al. 2018): an image-level domain classifier used to align global features, an instance domain classifier for aligning the instance representations and a consistency loss for regularizing the image-level and instance-level loss to consistency. 5) Strong-Weak Alignment (SW) (Saito et al. 2019): strong local alignment on the top of `conv3_3` and weak global alignment on `relu5_3`. The outputs of the alignment modules are later concatenated to the instance features which are then fed into the classifier and box regressor. 6) Few-shot adaptive Faster R-CNN (Few-shot) (Wang et al. 2019a): multi-scale local features are paired for image-level alignment, with semantic instance-level alignment of object features. 7) Selective cross-domain alignment (SelectAlign) (Zhu et al. 2019): discover the discriminatory regions by clustering instances and align two domains at the region level. 8) Mean Teacher with Object Relations (MTOR) (Cai et al. 2019): capture the object relations between teacher and student models by proposing graph-based consistency losses, with 50-layer ResNet (He et al. 2016) as backbone (we follow its implementation details for comparison).

Main Results

Our method is compared with state-of-the-art UDA object detectors in Table 1. As can be found, all methods can improve the performance of baseline (Faster R-CNN trained only on the source domain) by learning domain-invariant features at various stages in the networks. Particularly, in the SIM10K → Cityscapes scenario, our method obtains more than 10% AP improvement ($34.9\% \rightarrow 46.9\%$) over the source model, achieving a higher accuracy than state-of-the-art: 46.0% (iFAN) vs 43.0% (Zhu et al. 2019). For Cityscapes → Foggy Cityscapes, our method doubles the mAP of the source-only model ($16.9\% \rightarrow 35.3\%$) with VGG16 backbone, outperforming the other approaches by at least 1% on mAP.

In Figure 2, we illustrate two example results from source-only baseline and iFAN. Clearly, iFAN generalizes better to the novel data by detecting more challenging cases. Figure 3 shows that in Cityscapes → Foggy Cityscapes task, how instances move from original chaos to domain-invariant state with category cohesion, conforming the advances of iFAN.

We also report oracle results by training a Faster R-CNN detector directly on the fully-annotated training images on target domain. We can see that here still exists a performance gap between iFAN and the oracle result, especially on SIM10K → Cityscapes, which indicates that more sophisticated UDA methods are yet required to match the performance.

Discussions

Ablation Study. We conduct ablation study by isolating each component in iFAN. The results are presented in Table 2 and 3. Here are our observations: 1) With image-level alignment alone, we achieve a significant performance gain; 2) Instance-level alignment further reduces the domain discrepancies for objects; 3) For multi-class dataset like Cityscapes → Foggy, category-aware and category-correlation instance-level alignments obtain a higher accuracy than category-agnostic alignment, suggesting that the exploration on richer semantic information of instances can work better. 4) Integrating deep image-level with full instance-level alignments reaches the best results.

Layers Used in Image-Level Alignment. Table 4a shows mAP of disparate combinations of intermediate features in image-level alignment. Our hierarchically-nested discriminators are designed for characterizing domain shift at different semantic levels, and thus yield higher performance

Ground-truth | Faster R-CNN (source only) | Ours

Figure 2: Qualitative results. Top: SIM10K → Cityscapes. Bottom: Cityscapes → Foggy Cityscapes.

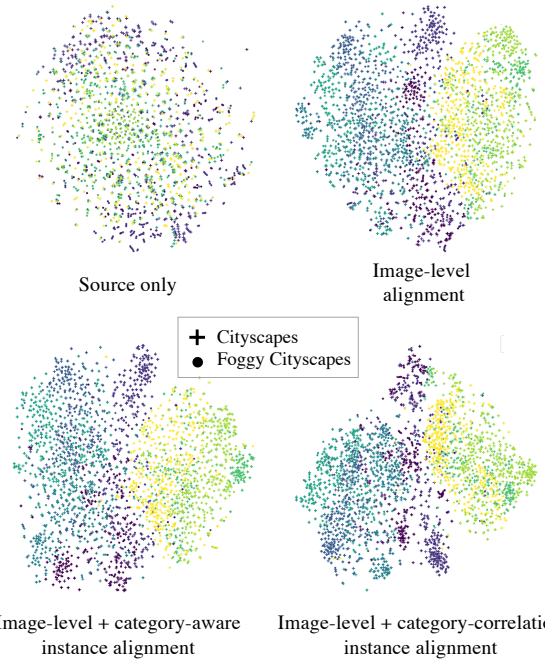

Figure 3: Visualization of ROI features from iFAN trained on Cityscapes → Foggy Cityscapes. Colors represent categories. We can see that intra- and inter-class relations are gradually optimized when deeper semantic information is encoded.

than individual layer. Moreover, we found that the lower layers work better than the higher ones, indicating that domain discrepancies are caused more heavily by low-level features like texture, color or illumination.

Timing for late launch. The instance-level alignment is activated in the middle of the training procedure. In Table 4b, we report AP@Car on SIM10K → Cityscapes with various *late launch* timings for category-agnostic instance alignment. As expected, starting instance-level alignment too early causes performance degradation: 42.1% (start at 10K-th iters) vs 43.0% (image-level alignment only); while too late, the instance discriminators fail to fully converge. Similarly, timing of the late launch is pivotal to the joint category-aware and category-correlation alignment.

Layers	Φ_1	Φ_2	Φ_3	Φ_4	$\Phi_1 \sim \Phi_4$
AP (%)	41.3	41.8	39.4	38.3	**43.0**

(a) Comparisons of different image-level alignment strategies. Multi-level features outperform individuals.

Start Step	10K	20K	30K	40K	50K
AP (%)	42.1	43.6	**46.1**	45.2	45.1

(b) Effect on which training iteration to start instance-level category-agnostic alignment.

Table 4: More results on SIM10K → Cityscapes.

Conclusion

We have presented a new domain alignment framework **iFAN** for unsupervised domain adaptive object detection. Two granularity levels of alignments are introduced: 1) Image-level alignment is implemented by aggregating multi-level deep features; 2) Full instance-level alignment is at first improved by explicitly encoding category information of the instances, and then enhanced by learning cross-domain category correlations using a metric learning formulation. The proposed iFAN achieves new state-of-the-art performance on two domain adaptive object detection tasks: synthetic-to-real (SIM10K → Cityscapes) and normal-to-foggy weather (Cityscapes → Foggy Cityscapes), with a boost of more than 10% AP over the source-only baseline.

References

Cai, Z., and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *CVPR*.

Cai, Q.; Pan, Y.; Ngo, C.-W.; Tian, X.; Duan, L.; and Yao, T. 2019. Exploring object relation in mean teacher for cross-domain detection. In *CVPR*.

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2018.

Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*.

Chopra, S.; Hadsell, R.; LeCun, Y.; et al. 2005. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *CVPR*.

Ganin, Y., and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *IJRR*.

Gidaris, S., and Komodakis, N. 2015. Object detection via a multi-region and semantic segmentation-aware cnn model. In *ICCV*.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*.

Girshick, R. 2015. Fast r-cnn. In *ICCV*.

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*.

Huang, X., and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*.

Inoue, N.; Furuta, R.; Yamasaki, T.; and Aizawa, K. 2018. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*.

Johnson-Roberson, M.; Barto, C.; Mehta, R.; Sridhar, S. N.; Rosaen, K.; and Vasudevan, R. 2017. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *ICRA*.

Kumar, A.; Sattigeri, P.; Wadhawan, K.; Karlinsky, L.; Feris, R.; Freeman, B.; and Wornell, G. 2018. Co-regularized alignment for unsupervised domain adaptation. In *NeurIPS*.

Lin, T.-Y.; Dollár, P.; Girshick, R. B.; He, K.; Hariharan, B.; and Belongie, S. J. 2017. Feature pyramid networks for object detection. In *CVPR*.

Liu, M.-Y., and Tuzel, O. 2016. Coupled generative adversarial networks. In *NeurIPS*.

Liu, H.; Long, M.; Wang, J.; and Jordan, M. I. 2019. Transferable adversarial training: A general approach to adapting deep classifiers. In *Proceedings of the 36th International Conference on Machine Learning*.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *NeurIPS*.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *ICML*.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. In *NeurIPS*.

Massa, F., and Girshick, R. 2018. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark.

Motiian, S.; Jones, Q.; Iranmanesh, S.; and Doretto, G. 2017. Few-shot adversarial domain adaptation. In *NeurIPS*.

Oh Song, H.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep metric learning via lifted structured feature embedding. In *CVPR*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.

RoyChowdhury, A.; Chakrabarty, P.; Singh, A.; Jin, S.; Jiang, H.; Cao, L.; and Learned-Miller, E. 2019. Automatic adaptation of object detectors to new domains using self-training. *arXiv preprint arXiv:1904.07305*.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2017. Maximum classifier discrepancy for unsupervised domain adaptation. *arXiv preprint arXiv:1712.02560*.

Saito, K.; Ushiku, Y.; Harada, T.; and Saenko, K. 2019. Strong-weak distribution alignment for adaptive object detection. In *CVPR*.

Saito, K.; Ushiku, Y.; and Harada, T. 2017. Asymmetric tri-training for unsupervised domain adaptation. In *ICML*.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Semantic foggy scene understanding with synthetic data. *IJCV*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*.

Shan, Y.; Lu, W. F.; and Chew, C. M. 2018. Pixel and feature level based domain adaption for object detection in autonomous driving. *arXiv preprint arXiv:1810.00345*.

Shu, R.; Bui, H. H.; Narui, H.; and Ermon, S. 2018. A dirt-t approach to unsupervised domain adaptation. In *ICLR*.

Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *CVPR*.

Wang, T.; Zhang, X.; Yuan, L.; and Feng, J. 2019a. Few-shot adaptive faster r-cnn. *arXiv preprint arXiv:1903.09372*.

Wang, X.; Han, X.; Huang, W.; Dong, D.; and Scott, M. R. 2019b. Multi-similarity loss with general pair weighting for deep metric learning. In *CVPR*.

Wu, Z.; Han, X.; Lin, Y.-L.; Uzunbas, M. G.; Goldstein, T.; Lim, S. N.; and Davis, L. S. 2018. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *ECCV*.

Yang, H.; Huang, D.; Wang, Y.; and Jain, A. K. 2018. Learning continuous face age progression: A pyramid of gans. In *CVPR*.

Zhang, Y.; David, P.; and Gong, B. 2017. Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*.

Zhang, Z.; Xie, Y.; and Lin, Y. 2018. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *CVPR*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.

Zhu, X.; Pang, J.; Yang, C.; Shi, J.; and Lin, D. 2019. Adapting object detectors via selective cross-domain alignment. In *CVPR*.