I^{3}Net: Implicit Instance-Invariant Network for Adapting One-Stage Object Detectors

Chaoqi Chen[1], Zebiao Zheng[2], Yue Huang[2*], Xinghao Ding[2], Yizhou Yu[1,3*]
[1] The University of Hong Kong
[2] School of Informatics, Xiamen University, China
[3] Deepwise AI Lab
cqchen1994@gmail.com, zbzheng@stu.xmu.edu.cn
huangyue05@gmail.com, dxh@xmu.edu.cn, yizhouy@acm.org

Abstract

Recent works on two-stage cross-domain detection have widely explored the local feature patterns to achieve more accurate adaptation results. These methods heavily rely on the region proposal mechanisms and ROI-based instance-level features to design fine-grained feature alignment modules with respect to the foreground objects. However, for one-stage detectors, it is hard or even impossible to obtain explicit instance-level features in the detection pipelines. Motivated by this, we propose an Implicit Instance-Invariant Network (I^{3}Net), which is tailored for adapting one-stage detectors and implicitly learns instance-invariant features via exploiting the natural characteristics of deep features in different layers. Specifically, we facilitate the adaptation from three aspects: (1) Dynamic and Class-Balanced Reweighting (DCBR) strategy, which considers the coexistence of intra-domain and intra-class variations to assign larger weights to those sample-scarce categories and easy-to-adapt samples; (2) Category-aware Object Pattern Matching (COPM) module, which boosts the cross-domain foreground objects matching guided by the categorical information and suppresses the uninformative background features; (3) Regularized Joint Category Alignment (RJCA) module, which jointly enforces the category alignment at different domain-specific layers with a consistency regularization. Experiments reveal that I^{3}Net exceeds the state-of-the-art performance on benchmark datasets.

1. Introduction

Object detection has achieved remarkable progress due to the unprecedented development of deep convolutional networks (CNNs) and the existence of large-scale annotated datasets. However, collecting large amounts of instance-level annotated data in various domains for object detection is prohibitively costly. An alternative would be applying the off-the-shelf detection model trained on the source domain to a new target domain. However, deep object detectors suffer from performance degradation when applied to a new domain under the presence of domain shift [41]. This problem has inspired the research on Unsupervised Domain Adaptation (UDA) [27], which aims to bridge the distribution discrepancy between source and target domains via knowledge transfer. Numerous approaches, such as moment matching [11, 6, 23, 25, 49] and adversarial learning [7, 42, 37, 24, 44], have been proposed for cross-domain image classification and semantic segmentation.

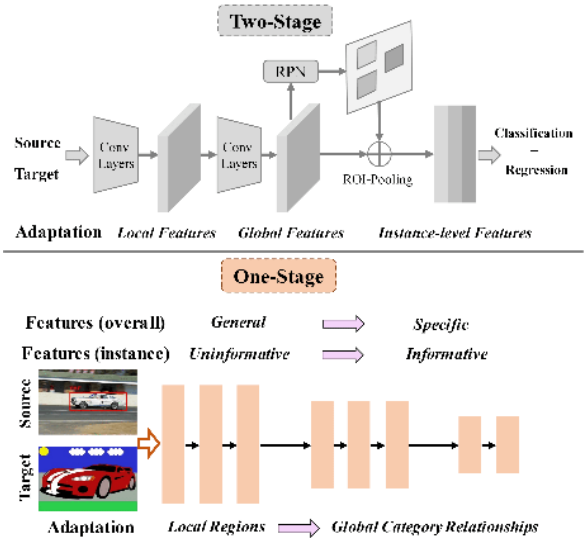

Figure 1: **Upper:** Illustration of previous two-stage cross-domain detection methods. **Lower:** Motivation of the proposed method based on the observation with respect to the characteristics of deep features in different layers.

*Corresponding authors

Compared to the conventional UDA problems, cross-domain object detection is a more sophisticated and challenging problem since the adaptation of classification and regression should be simultaneously considered. Current methods [4, 53, 34, 1, 14, 3, 46, 52, 45] mostly resort to the adversarial feature adaptation to explore discriminative feature patterns at local-level, global-level, and instance-level for adapting two-stage detectors (see top of Fig. 1), Faster R-CNN [33]. However, they heavily rely on the region proposal mechanisms and ROI-based instance-level features to design fine-grained feature alignment modules with respect to the foreground objects. For example, Zhu *et al.* [53] mine the target discriminative regions based on the region proposals derived from the RPN. Cai *et al.* [1] regularize the relational graphs by using the ROI-based features. Chen *et al.* [3] and Xu *et al.* [45] assist the instance-level feature alignment by the contextual or categorical regularization.

One-stage object detectors, such as SSD [22] and RetinaNet [21], have the merits of being faster and simpler in real-world applications. Unfortunately, it is unrealistic to obtain *explicit* instance-level features in the one-stage detectors due to the lack of region proposal step. Hence, how to adapt one-stage detectors is vital for practical scenarios but yet to be thoroughly studied. The motivation of this paper is shown in the bottom of Fig. 1. Deep features in the standard CNNs must eventually transition from general to specific along the network [48]. Inspired by this, in one-stage detectors, we can reasonably envision that the features at lower layers (*e.g.,* color, corner, edge, and illumination) are expected to be mostly *instance-uninformative*, while the features at higher layers (*e.g.,* object categories) are *instance-informative*. Therefore, we need to alleviate the negative influence of uninformative features and promote the alignment of informative features, *i.e.,* suppress redundant (such as background) information from the lower layers and enhance the cross-domain semantic correlation of foreground objects at the higher layers.

In this paper, we propose an Implicit Instance-Invariant Network (I^3Net) that removes the need for requiring explicit instance-level features. Instead, we implicitly learn instance-invariant features via the alignment of transferable regions and images while preserving the inter-domain class relationships. To be specific, we facilitate the adaptation of one-stage detectors from three aspects. Firstly, upon observing that there exist two conceptually orthogonal distribution variations hidden in the target data, *i.e.,* intra-domain and intra-class variations, we propose a Dynamic and Class-Balanced Reweighting (DCBR) strategy to dynamically reweight each target sample based on its *adaptation difficulty*, which is measured by the degree of class imbalance and the prediction uncertainty of a multi-label classifier. Secondly, considering that object with the same category label but from different domains will share similar object patterns, we design a Category-aware Object Pattern Matching (COPM) module to boost cross-domain foreground objects matching guided by the categorical information and suppress the uninformative background features at lower layers. Finally, we develop a Regularized Joint Category Alignment (RJCA) module to enable category alignment by considering complementary effect of different domain-specific layers and further incorporate a consistency regularization term with respect to the average prediction of different detection heads. Experimental results show that the proposed I^3Net significantly improves the state-of-the-art performance of one-stage cross-domain object detection on three benchmarks.

2. Related Work

Unsupervised Domain Adaptation (UDA) UDA methods have attracted much attention for alleviating the distributional variations between two distinct domains in image classification, semantic segmentation, and object detection. For UDA, a typical solution is to match the source and target feature distributions in the common space by embedding disparity measures into deep architectures, such as Maximum Mean Discrepancy (MMD) [43, 23], Correlation Alignment (CORAL) [40], Central Moment Discrepancy (CMD) [49], and transport distance [20, 47]. Inspired by the success of Generative Adversarial Nets (GAN) [12], a large amount works [8, 42, 35, 30, 44, 2, 50, 17] have been done by adversarially learning domain-invariant representations with extra categorical regularization.

Object Detection Object Detection is one of the most fundamental computer vision problems in the past few decades [54]. Our work focuses on how to adapt object detectors, so we only review several representative two-stage and one-stage detectors. The series of region-based convolutional networks (*i.e.,* R-CNN [10], Fast R-CNN [9], and Faster R-CNN [33]) have achieved compelling results in terms of detection accuracy. They count on the region proposal mechanisms to classify region of interest (ROI) independently [10], or share the convolution features with ROI pooling layer [9], or produce the region proposals based on a Region Proposal Network (RPN) [33]. On the other hand, one-stage detectors, such as SSD [22], YOLO [31, 32], and RetinaNet [21] have shown a clear superiority on the inference speed by directly carrying out the category confidence prediction and the bounding box regression.

UDA for Object Detection Domain Adaptive Faster R-CNN [4] is a pioneering two-stage cross-domain detection method that reduces the distributional shift by adversarially learning domain-invariant features on both image-level and instance-level. Considered the local nature of object detection task, most recent efforts [53, 34, 1, 14, 3, 46, 52, 45, 15, 39, 51] are devoted to capture the local feature patterns

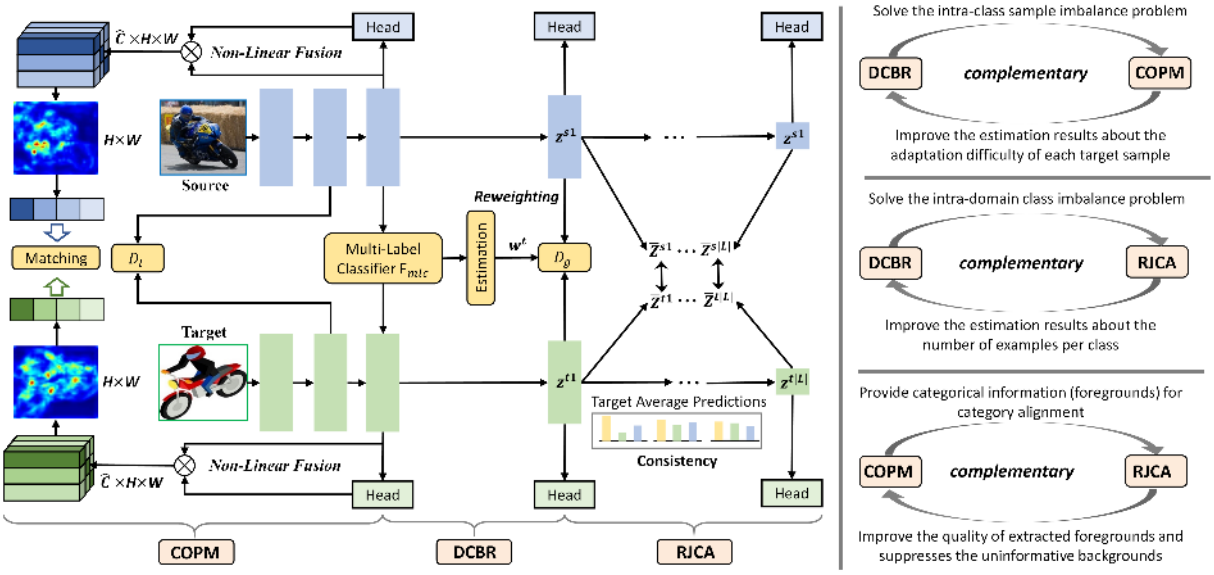

Figure 2: **Left:** The overall structure of the proposed I^{3}Net, where F_{mlc} is an image-level multi-label classifier, D_l and D_g are pixel-level and image-level domain discriminators respectively. Non-linear fusion stands for the tensor product operation. We adopt SSD as the base detection network. **Right:** DCBR, COPM, and RJCA are complementary to each other.

and explicitly align them at multiple levels. For instance, Chen *et al.* [3] propose to hierarchically calibrate the transferability of different level features (*i.e.,* local-region, image, and instance) to improve the discriminability of detectors; Xu *et al.* [46] and Zheng *et al.* [52] draw motivation from the cross-domain prototype alignment [44, 2, 28] to align the foreground objects with the same category between domains. However, these methods can not be simply extended to the one-stage detectors since they highly rely on the region proposals and pooled instance-level features. The study on adapting one-stage object detectors is very limited. A pioneering attempt [19] present a weak self-training strategy by simultaneously reducing the false positives and false negatives during the hard negative mining. However, self-training-based method may be vulnerable to the error accumulation problem, especially on the sophisticated cross-domain detection scenarios. Moreover, how to learn instance-invariant feature representations without the guidance of region proposal mechanisms, which is crucial for adapting one-stage detectors, still remains unclear.

3. Methodology

In the task of cross-domain object detection, we are given a source domain $\mathcal{D}_s = \{(x_i^s, y_i^s, b_i^s)\}_{i=1}^{N_s}$ ($y_i^s \in \mathcal{R}^{k\times 1}$, $b_i^s \in \mathcal{R}^{k\times 4}$) of N_s labeled samples, and a target domain $\mathcal{D}_t = \{x_j^t\}_{j=1}^{N_t}$ of N_t unlabeled samples. $\mathcal{D}_s$ and $\mathcal{D}_t$ are drawn from different data distributions, but share an identical group of classes (K classes in all). The objective of this paper is to transfer knowledge from $\mathcal{D}_s$ to $\mathcal{D}_t$ and achieve good detection results in $\mathcal{D}_t$.

Framework Overview. To this end, we propose an Implicit Instance-Invariant Network (I^{3}Net), which is comprised of three components, namely, Dynamic and Class-Balanced Reweighting (DCBR), Category-aware Object Pattern Matching (COPM), and Regularized Joint Category Alignment (RJCA). The overview of I^{3}Net is demonstrated in Figure 2. The basic idea is to utilize the inherent characteristics of representations at different layers of the detector to compensate for the lack of explicit instance-level features. DCBR reweights target samples based on the adaptation difficulty with respect to the intra-domain and intra-class variations, COPM captures the foreground object patterns and suppresses redundant background information, and RJCA promotes the cross-domain category alignment in different domain-specific layers (connected with detection heads) with a consistency regularization. Following the prior work on adapting one-stage detectors [19], our I^{3}Net is based on the SSD [22] framework.

3.1. Dynamic and Class-Balanced Reweighting

To date cross-domain detection methods [4, 34, 46, 52] mainly focus on the feature-level adaptation and treat all the target samples equally, while they neglect the distributional characteristics of the target data, which are crucial for the adaptation process. By contrast, the proposed DCBR strategy explicitly explores the intra-domain and intra-class variations within the unlabeled target domain to assign larger weights to those sample-scarce categories and easy-to-adapt samples. We analysis these two variations in the following.

Intra-Domain Variations. The class-imbalance prob-

lem [26], which refers to the inequality among the number of examples belonging to different classes, commonly exists in the object detection. Prior efforts, such as Focal Loss [21] and hard example mining [22, 36], are devoted to tackle the *foreground-background class imbalance*, which is irrelevant to the number of examples per class in a single domain. In cross-domain detection, we argue that the *foreground-foreground class imbalance*, which is dataset-relevant and may be different between domains, is prone to deteriorate the adaptation performance since the adaptation of each class will be affected by the number of examples per class in both domains, *i.e.,* the adaptation difficulty of different categories may be distinct.

Intra-Class Variations. Owing to the difference of background, object co-occurrence, and scene layouts across domains, excessively align the source and target features in the full dataset will result in negative transfer, *i.e.,* some target samples may be less transferable or even non-transferable. However, most leading cross-domain detection methods treat the target domain as a whole without considering the structures of intra-class data distributions. Motivated by this, we assume that the adaptation difficulty of different samples within the same class may be distinct. An intuitive solution is to utilize re-weighting techniques. However, this solution suffers a critical limitation in the context of cross-domain object detection. In contrast to the classification problem, where a single image usually contains only one semantic category, there exist multiple instances within the same image in the detection problem. Thus, how to measure the adaptation difficulty of an input target sample in cross-domain detection remains unclear.

Based on the above discussions, we formally provide the details of the proposed DCBR. The DCBR consists of two steps: (i) *estimate the adaptation difficulty of each individual target sample and each target class*; (ii) *reweight target samples based on the estimation results*. Technically, the adaptation difficulty of an target example x^t (for ease of denotation, we omit the subscript of x_i^s and x_j^t when they apply) is measured by using an image-level multi-label classifier (*i.e.,* F_{mlc} in Fig. 2). We first train F_{mlc} based on the labeled source samples for initialization. The multi-label classification loss on the source domain is formulated as:

$$\mathcal{L}_{\text{mlc}} = \sum_{k=1}^{K} y_k^s \cdot \log(\hat{y}_k^s) + (1 - y_k^s) \cdot \log(1 - \hat{y}_k^s) \quad (1)$$

where y_k^s is the k^{th} ($k = \{1, 2, ..., K\}$) element of y^s and $\hat{y}_k^s = F_{mlc}(G_1(x^s))^k$ (G_1 is a feature extractor connected to F_{mlc}). $y_k^s = 1$ means that there exists at least one object of class k in x^s; otherwise, $y_k^s = 0$ indicates that x^s does not contain the object of class k. For each target sample x^t, we denote the prediction of its multi-label classification as $\hat{y}_k^t = F_{mlc}(G_1(x^t))^k$. Then, we define the weight function of a target sample x^t *w.r.t.* the intra-class variation by using its multi-label classification output:

$$w_1^t = \frac{1}{K'} \sum_{k=1}^{K} \mathbb{1}(\hat{y}_k^t(x^t) > \tau) \cdot \hat{y}_k^t(x^t) + 1 \quad (2)$$

where $K' = \sum_{k=1}^{K} \mathbb{1}(\hat{y}_k^t(x^t) > \tau)$, and τ is a threshold. $\mathbb{1}(a)$ is an indicator function which is 1 if a is true and 0 otherwise. By doing so, target samples with higher classification confidence scores will be assigned larger weights since they are more similar with source domain. Note that the value of w_1^t increases continuously since the source and target distributions are getting closer as training proceeds.

To estimate the number of examples per class in $\mathcal{D}_t$, we resort to the classification output for roughly dividing $\mathcal{D}_t$ into K classes. x^t is added into the target domain of the class $\mathcal{D}_t^{k'}$ if $k' = \arg\max_k \hat{y}_t^k(x^t)$. Then, the unlabel target samples $\mathcal{D}_t$ are split into K classes, *i.e.,* $\mathcal{D}_t = \{\mathcal{D}_t^k\}_{k=1}^K$. To this end, we are able to assign larger weights to those sample-scarce categories. The weight function of x^t *w.r.t.* the intra-domain variation is formulated as,

$$w_2^t = e^{(1 - N_t^k / N_t)} \quad (3)$$

where N_t^k denotes the number of samples in class k.

Based on Eq. (2) and Eq. (3), the overall weight function of a target sample x_j^t is formulated as follows,

$$w^t = \theta w_1^t + (1 - \theta) w_2^t \quad (4)$$

where θ is a hyper-parameter to balance w_1^t and w_2^t. After adding the weights to all target samples, the adversarial loss of image-wise domain discriminator D_g can be written as:

$$\begin{aligned} \mathcal{L}_{\text{dcbr}} = &- \frac{1}{N_s} \sum_{i=1}^{N_s} \log(D_g(G_2(x_i^s))) \\ &- \frac{1}{N_s} \sum_{j=1}^{N_t} w_j^t \cdot \log(1 - D_g(G_2(x_j^t))) \end{aligned} \quad (5)$$

where G_2 is a feature extractor that is connected to D_g.

3.2. Category-Aware Object Pattern Matching

As we discussed in Section 1, the feature representations at lower layers contain various redundant information (*e.g.* background) and should not be fully aligned. Previous works [4, 34], which strictly matching the low-level features, may result in inferior performance especially on the one-stage detection. During the exploration, we observe that objects with the same category label but from different domains will own similar object patterns. Object pattern, which refers to the discriminative features of foreground objects, can provide rich semantic information *w.r.t.* the objects, such as object category, shape, size, *etc*. Driven

by this finding, we propose a Category-aware Object Pattern Matching (COPM) module to boost cross-domain foreground objects matching guided by the categorical information and suppress the uninformative background features.

Suppose that we have a CNN layer (*e.g.,* $Conv\ 4\_3$ in SSD300) and its corresponding activation tensor $A \in \mathbb{R}^{C\times H\times W}$, which consists of C feature planes and has height of H and width of W. An intuitive idea for local feature alignment is to extract attention maps from both domains and somehow match them. However, the target attention map tends to focus on the predominant foreground objects instead of the full foreground objects (cf. Fig. 3), which will impair the localization ability of detector for detecting those small or/and obscured objects. Thus, we resort to leverage classification output of the detection head (cf. Fig. 2), which is denoted by $\hat{p}_m$ ($\hat{p}_m \in \mathbb{R}^{K+1}$, m is the anchor index in A, and $m = \{1, 2, ..., H \times W\}$), to guide the object pattern matching. Specifically, the classification output $\hat{p}_m$ and the feature representation A_m ($A_m \in \mathbb{R}^C$) are nonlinearly fused via tensor product operation, *i.e.,* $\hat{A}_m = A_m \otimes \hat{p}_m$, where $\hat{A}_m$ is the fused feature vector. In order to prevent the dimension explosion, we draw motivations from the randomized multilinear map [18, 24] to estimate the tensor product via Hadamard product,

$$\hat{A}_m = (\boldsymbol{R}_1 A_m) \odot (\boldsymbol{R}_2 \hat{p}_m),\ A_m \in \mathbb{R}^{\hat{C}} \tag{6}$$

where $\odot$ denotes the Hadamard product. $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are random matrices and each of their element follows uniform distribution with univariance. $\hat{C}$ is the feature dimension after fusion ($\hat{C}$ is set to 1024 in our experiments). Based on the category-guided activation tensor $\hat{A}$, we output a spatial attention map via an activation-based mapping function: $\mathcal{F} : \mathbb{R}^{\hat{C}\times H\times W} \rightarrow \mathbb{R}^{H\times W}$, which can be written as follows:

$$(\mathcal{F}(\hat{A}))_m = \sum_{c=1}^{\hat{C}} |\hat{A}_m^c|^2 \tag{7}$$

To reduce computational cost, we flatten the source and target attention maps to vectors, which are denoted as f^s and f^t. Finally, We align the source and target object patterns by minimizing distance between the them,

$$\mathcal{L}_{\text{la}} = \sqrt{H \times W} \cdot \Phi\left(\frac{f^s}{\|f^s\|_2}, \frac{f^t}{\|f^t\|_2}\right) \tag{8}$$

where $\Phi(x, x') = \|x - x'\|_2$ is the Euclidean distance. Note that we incorporate a pixel-level domain discriminator (*i.e.,* D_l in Fig. 2) into COPM to further reduce the low-level feature disparity. Thus, the objective of COPM is formulated as: $\mathcal{L}_{\text{copm}} = \mathcal{L}_{\text{la}} + \mathcal{L}_{\text{adv}}$, where $\mathcal{L}_{\text{adv}}$ is a vanilla pixel-wise domain adversarial training loss.

3.3. Regularized Joint Category Alignment

Prototype[1]-based feature alignment has been widely explored to measure the category-level feature discrepancy in UDA [44, 2, 29] and been applied to the two-stage cross-domain detection [52, 46]. However, considering the dense prediction property of one-stage detectors, prototype alignment may be error-prone in this case compared to adapting two-stage detectors where most negative proposals will be filtered out. Moreover, prior efforts only implement the prototype alignment in a certain high-level feature layer without considering the potential complementary effect of different domain-specific layers. Motivated by this, we propose a Regularized Joint Category Alignment (RJCA) module to achieve the category alignment at different domain-specific layers and regularize the average prediction consistency of different layers with respect to the same category.

In the light of fully convolutional and multi-level prediction characteristics of one-stage detectors, we aims at jointly enforcing the cross-domain category alignment in different layers. First of all, assume that the deep networks will generate the activations in different layers as $\{(z_i^{s1}, ..., z_i^{s|L|})\}_{i=1}^{N_s}$ and $\{(z_j^{t1}, ..., z_j^{t|L|})\}_{i=1}^{N_t}$, where $l \in L$ and $z \in \mathbb{R}^{C\times H\times W}$. Then, we resort to the per-pixel prediction to compute the prototype of each source class in layer l, which can be written as:

$$\bar{Z}_k^{s|l|} = \frac{1}{n_s^k} \sum_{i=1}^{N_s} \sum_{m=1}^{H\times W} y_s^{imk} \cdot z_{im}^{s|l|} \tag{9}$$

where n_s^k denotes the number of source objects labeled with class k, m is the pixel index in z. $y_s^{imk} \in \{0, 1\}$ is an indicator for determining whether the current pixel is predicted as class k. The source global prototype of each class is computed at the beginning of training. Let the prediction of detection head *w.r.t.* a target object be represented by $\hat{p}(z_{jm}^{t|l|})$. The target local prototype is computed by:

$$\bar{z}_k^{t|l|} = \frac{1}{\hat{n}_t^k} \sum_{j=1}^{|B_t|} \sum_{m=1}^{H\times W} y_t^{jmk} \cdot z_{jm}^{t|l|} \tag{10}$$

where $\hat{n}_t^k$ denotes the number of objects that are assigned with pseudo label k and B_t is the mini-batch samples of the target domain. Similarly, we can obtain a set of source local prototypes $\{\bar{z}_k^{s|l|}\}_{k=1}^{K}$. The objective function of joint category alignment is formulated as follows:

$$\mathcal{L}_{\text{jca}} = \sum_l [\underbrace{\sum_k d(\bar{Z}_k^{s|l|}, \bar{Z}_k^{t|l|})}_{\text{Compactness}} + \underbrace{\sum_{m,n|m\neq n} h(\bar{Z}_m^{s|l|}, \bar{Z}_n^{t|l|})}_{\text{Separation}}] \tag{11}$$

where d and h are two different similarity functions to measure the distance between prototypes. In our case, we instantiate Eq. (11) by the contrastive loss as defined in [13].

[1]Prototype is the mean feature of the samples within the same class.

During training, the global prototype in Eq. (11) is updated by the local prototype in a moving average manner,

$$\bar{Z}_k^{|l|} \leftarrow \rho \bar{Z}_k^{|l|} + (1-\rho)\bar{z}_k^{|l|} \tag{12}$$

where ρ is set to 0.7 in all experiments. In addition, we regularize the prediction consistency of different layers *w.r.t.* the same class k by respectively minimizing their symmetrized Kullback−Leibler (KL) divergence, which is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{pr}} = \frac{1}{K}\sum_{l}\sum_{k=1}^{K}\frac{1}{2}[D_{\text{KL}}(\hat{p}(\bar{z}_k^{t|l_a|})\|\hat{p}(\bar{z}_k^{t|l_b|})) \\ + D_{\text{KL}}(\hat{p}(\bar{z}_k^{t|l_b|})\|\hat{p}(\bar{z}_k^{t|l_a|}))], \text{where } l_a, l_b \in L.\end{aligned} \tag{13}$$

where $\hat{p}(\bar{z}_k^{t|l_a|})$ and $\hat{p}(\bar{z}_k^{t|l_b|})$ stand for the average prediction *w.r.t.* the class k in different layers. Here, to smooth the prediction, we add a temperature variate T ($T = 2$ in all experiments) to the softmax function. To this end, the objective of the proposed RJCA can be written as: $\mathcal{L}_{\text{rjca}} = \mathcal{L}_{\text{jca}} + \gamma\mathcal{L}_{\text{pr}}$, where γ is set to 0.1 in all experiments.

3.4. Training Loss

Suppose that the detection loss is denoted as $\mathcal{L}_{\text{det}}$, which includes the classification and regression losses. Joint all the presented parts, the overall objective function of I^{3}Net is formulated as follows,

$$\mathcal{L}_{\text{I}^3\text{Net}} = \mathcal{L}_{\text{det}} + \lambda_1\mathcal{L}_{\text{dcbr}} + \lambda_2(\mathcal{L}_{\text{copm}} + \mathcal{L}_{\text{rjca}}) \tag{14}$$

where λ_1 and λ_2 are hyper-parameters for balancing different loss components.

4. Experiments

4.1. Datasets

We conduct experiments based on **Pascal VOC** [5], **Clipart1k**, **Watercolor2k**, and **Comic2k** [16] datasets. Following the previous one-stage method [19], we utilize the Pascal VOC2007-trainval and VOC2012-trainval datasets as the source domain, and Clipart1k, Watercolor2k, and Comic2k as the target domain respectively. The Pascal VOC [5], which is a real-world image dataset, contains 16,551 images with 20 distinct object categories. Clipart1k [16], which is a graphical image dataset with complex backgounds, consists of 1K images and has the same 20 categories as Pascal VOC. We utilize all images of Clipart1k as the target domain for both training and testing. Watercolor2k and Comic2k [16] contain 2K images respectively (*i.e.,* 1K as the train set and the other 1K as the test set). They share 6 identical categories with the Clipart1k dataset, *i.e.,* bicycle, bird, cat, car, dog, and person. Following the prior practice [19], we leverage the train set for training and the test set for evaluation.

4.2. Implementation Details

The base detection model in our experiments follows the same setting in [16, 19] that utilize SSD300 [22] framework with VGG-16 [38] architectures. The parameters of VGG-16 is fine-tuned from the model that has been pre-trained on ImageNet. In all experiments, the input images are resized to 300 $\times$ 300 and we conduct all augmentations used in [22, 19]. The batch size is selected as 32 (16 source images and 16 target images) to fit the GPU memory. We evaluate the cross-domain detection performance by reporting mean average precision (mAP) with a IoU threshold of 0.5 on the target domain. We adopt the stochastic gradient descent (SGD) optimizer for the detection network training with a momentum of 0.9, an initial learning rate of 0.001, weight decay of 5×10^{-4}. The learning rate is decreased to 0.0001 after 50 epochs. Note that the multi-label classifier F_{mlc} is pre-trained on the label source domain and keeps fixed when training our adaptation network. Without specific notation, we set $\tau = 0.5$ in Eq. (2) and $\theta = 0.5$ in Eq. (4). For the L in RJCA, we set $L = \{Conv7, Conv9\_2\}$ for the I^{3}Net model based on SSD. We set $\lambda_1 = 0.05$ and $\lambda_2 = 1$ in Eq. (14) for all experiments. Our experiments are implemented with the Pytorch deep learning framework.

4.3. Comparisons with State-of-the-Arts

State-of-the-arts. We make comparison to the state-of-the-art cross-domain object detection methods, including Domain Adversarial Neural Networks **(DANN)** [8], adversarial Background Score Regularization + Weak Self-Training **(BSR+WST)** [19], Strong-Weak Distribution Alignment **(SWDA†)** [34], and Hierarchical Transferability Calibration Network **(HTCN†)** [3]. The quantitative results of DANN, BSR, WST, and BSR+WST are cited from the original paper [19]. We reproduce the complete SWDA model on our one-stage scenarios. Moreover, we remove the context-aware instance-level alignment component from the HTCN model and re-implement the rest modules in our experiments. Note that mainstream cross-domain detection methods (*e.g.,* [4, 53, 1, 14, 46, 52, 45]) are tailored for two-stage detector and cannot be simply extended to one-stage-based experiments since they highly count on the region proposal mechanisms.

Results on Clipart1k. Table 1 displays the adaptation results on Pascal VOC $\rightarrow$ Clipart1k. Source Only denotes that the baseline SSD is trained on the source domain and directly tested on the target domain without any adaptation. The proposed I^{3}Net significantly outperforms all the compared methods in terms of mAP and improves over state-of-the-art by +2.0% (35.8% to 37.8%). It is noteworthy that all components of the proposed I^{3}Net are designed appropriately and when we remove any one of these components,

Table 1: Results of adapting PASCAL VOC to Clipart1k (%). mAP is reported on Clipart1k.

Methods	aero	bcycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hrs	bike	prsn	plnt	sheep	sofa	train	tv	mAP
Source Only [22]	27.3	60.4	17.5	16.0	14.5	43.7	32.0	10.2	38.6	15.3	24.5	16.0	18.4	49.5	30.7	30.0	2.3	23.0	35.1	29.9	26.7
DANN [8]	24.1	52.6	27.5	18.5	20.3	59.3	37.4	3.8	35.1	32.6	23.9	13.8	22.5	50.9	49.9	36.3	11.6	31.3	48.0	35.8	31.8
DT+PL w/o label [16]	16.8	53.7	19.7	**31.9**	21.3	39.3	39.8	2.2	**42.7**	**46.3**	24.5	13.0	**42.8**	50.4	53.3	38.5	14.9	25.1	41.5	37.3	32.7
WST [19]	**30.8**	65.5	18.7	23.0	24.9	57.5	40.2	10.9	38.0	25.9	**36.0**	15.6	22.6	66.8	52.1	35.3	1.0	34.6	38.1	39.4	33.8
BSR [19]	26.3	56.8	21.9	20.0	24.7	55.3	42.9	11.4	40.5	30.5	25.7	17.3	23.2	66.9	50.9	35.2	11.0	33.2	47.1	38.7	34.0
SWDA† [34]	29.0	60.7	25.0	20.4	24.6	55.4	36.1	13.1	41.2	38.3	30.3	17.0	21.2	55.2	50.4	36.6	10.6	38.4	49.2	41.2	34.7
BSR+WST [19]	28.0	64.5	23.9	19.0	21.9	**64.3**	**43.5**	16.4	42.2	25.9	30.5	7.9	25.5	67.6	54.5	36.4	10.3	31.2	**57.4**	43.5	35.7
HTCN† [3]	28.7	67.7	25.3	16.1	28.7	56.0	38.9	12.5	41.0	33.0	29.6	12.9	22.9	69.0	**55.9**	36.1	11.8	34.1	48.8	46.8	35.8
I^{3}Net w/o DCBR	30.5	66.9	25.6	17.9	24.0	47.8	35.7	13.8	40.6	36.3	27.8	16.5	24.5	**71.4**	56.6	38.2	10.5	**39.9**	50.7	44.5	36.0
I^{3}Net w/o COPM	28.7	66.8	28.4	23.1	25.3	58.4	42.8	**19.2**	40.4	33.6	32.7	18.1	23.5	53.8	52.5	35.6	13.4	37.3	52.4	46.0	36.6
I^{3}Net w/o RJCA	28.8	**67.8**	25.4	16.2	28.9	56.1	39.0	12.6	41.1	33.1	29.7	13.0	22.9	69.1	**55.9**	36.3	11.9	34.2	48.9	**46.9**	35.9
I^{3}Net (Full)	30.0	67.0	**32.5**	21.8	**29.2**	62.5	41.3	11.6	37.1	39.4	27.4	**19.3**	25.0	67.4	55.2	**42.9**	**19.5**	36.2	50.7	39.3	**37.8**

Table 2: Results on adaptation from Pascal VOC to Watercolor2k (%). mAP is reported on the Watercolor2k test set.

Methods	bike	bird	car	cat	dog	person	mAP
Source Only [22]	77.5	46.1	44.6	30.0	26.0	58.6	47.1
DANN [8]	73.4	41.0	32.4	28.6	22.1	51.4	41.5
BSR [19]	**82.8**	43.2	**49.8**	29.6	27.6	58.4	48.6
WST [19]	77.8	48.0	45.2	30.4	29.5	64.2	49.2
SWDA† [34]	73.9	48.6	44.3	36.2	31.7	62.1	49.5
BSR+WST [19]	75.6	45.8	49.3	34.1	30.3	64.1	49.9
HTCN† [3]	78.6	47.5	45.6	35.4	31.0	62.2	50.1
I^{3}Net w/o DCBR	78.7	49.2	42.6	37.4	32.4	62.5	50.5
I^{3}Net w/o COPM	75.6	49.2	45.9	**37.9**	33.2	63.6	50.9
I^{3}Net w/o RJCA	81.8	46.3	40.4	33.3	**34.0**	65.1	50.2
I^{3}Net (Full)	81.1	**49.3**	46.2	35.0	31.9	**65.7**	**51.5**

Table 3: Results on adaptation from Pascal VOC to Comic2k (%). mAP is reported on the Comic2k test set.

Methods	bike	bird	car	cat	dog	person	mAP
Source Only [22]	43.3	9.4	23.6	9.8	10.9	34.2	21.9
DANN [8]	33.3	11.3	19.7	**13.4**	19.6	37.4	22.5
BSR [19]	45.2	15.8	26.3	9.9	15.8	39.7	25.5
WST [19]	45.7	9.3	30.4	9.1	10.9	46.9	25.4
BSR+WST [19]	**50.6**	13.6	31.0	7.5	16.4	41.4	26.8
SWDA† [34]	47.4	12.9	29.5	12.7	19.1	44.1	27.6
HTCN† [3]	50.3	15.0	27.1	9.4	18.9	46.2	27.8
I^{3}Net w/o DCBR	44.2	14.0	**35.1**	6.5	19.3	51.7	28.5
I^{3}Net w/o COPM	47.1	14.5	32.3	7.1	**20.3**	**51.8**	28.9
I^{3}Net w/o RJCA	45.0	12.1	33.9	8.0	20.1	50.5	28.3
I^{3}Net (Full)	47.5	**19.9**	33.2	11.4	19.4	49.1	**30.1**

the final performance will drop accordingly.

Results on Watercolor2k and Comic2k. Results on the tasks of Pascal VOC → Watercolor2k and Pascal VOC → Comic2k are reported on Table 2 and Table 3 respectively. I^{3}Net achieves better performance on most object categories, indicating that I^{3}Net is capable of learning more transferable representations and scalable for different cross-domain detection scenarios. It is noteworthy that I^{3}Net substantially exhibits better adaptation performance on the challenging transfer task (27.8% to 30.1%), *i.e.,* Pascal VOC → Comic2k, where the domain discrepancy is substantially large between source and target data.

4.4. Further Empirical Analysis

Ablation Study. We verify the effect of the proposed DCBR, COPM, and RJCA by evaluating variants of I^{3}Net. The results are reported in Table 4. **(1)** DCBR w/o Dynamic and DCBR w/o CB denote that we remove w_1^t and w_2^t from Eq. (4) respectively. **(2)** COPM w/o C denotes that we remove the non-linear fusion step (Eq. (6)) and directly match the source and target vectorized attention maps. COPM w/ MMD and COPM w/ Adv denote that we replace the L_2 distance in Eq. (8) by MMD [23] and domain adversarial loss [7] respectively. **(3)** RJCA w/o J is the variant that only conducts the category alignment in one layer. RJCA w/o PR is the variant without prediction regularization (Eq. (13)). The results of COPM w/ MMD and COPM w/ Adv reveal that L_2 distance is able to better preserve the structured information (*i.e.,* object patterns). The results of RJCA w/o J verify the significance of considering the complementary effect of different domain-specific layers.

Table 4: Ablation of I^{3}Net on three transfer tasks (%).

Source	Pascal VOC		
Target	Clipart1k	Watercolor2k	Comic2k
DCBR w/o Dynamic	37.3	51.4	29.2
DCBR w/o CB	37.1	51.0	29.3
COPM w/o C	36.8	51.1	29.0
COPM w/ MMD	34.9	48.4	27.0
COPM w/ Adv	37.0	50.7	29.8
RJCA w/o J	36.6	50.8	29.1
RJCA w/o PR	37.4	51.5	29.4
I^{3}Net (Full)	**37.8**	**51.8**	**30.1**

Visualization of COPM. Figure 3 visualizes the attention maps generated by Source Only [22], HTCN† [3], and I^{3}Net (Ours). The brighter the color is, the larger the weight value is. It is notable that the proposed I^{3}Net is capable of (i) capturing the discriminative regions which contain rich semantic information, (ii) highlighting the foreground objects

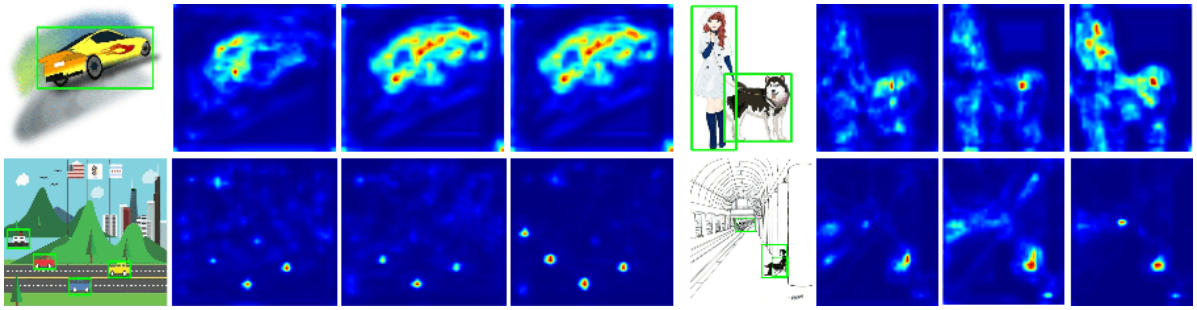

Figure 3: Illustration of the target attention maps generated by Source Only, HTCN†, and I^{3}Net. From left to right: input target images with ground-truth bounding boxes, Source Only, HTCN†, I^{3}Net.

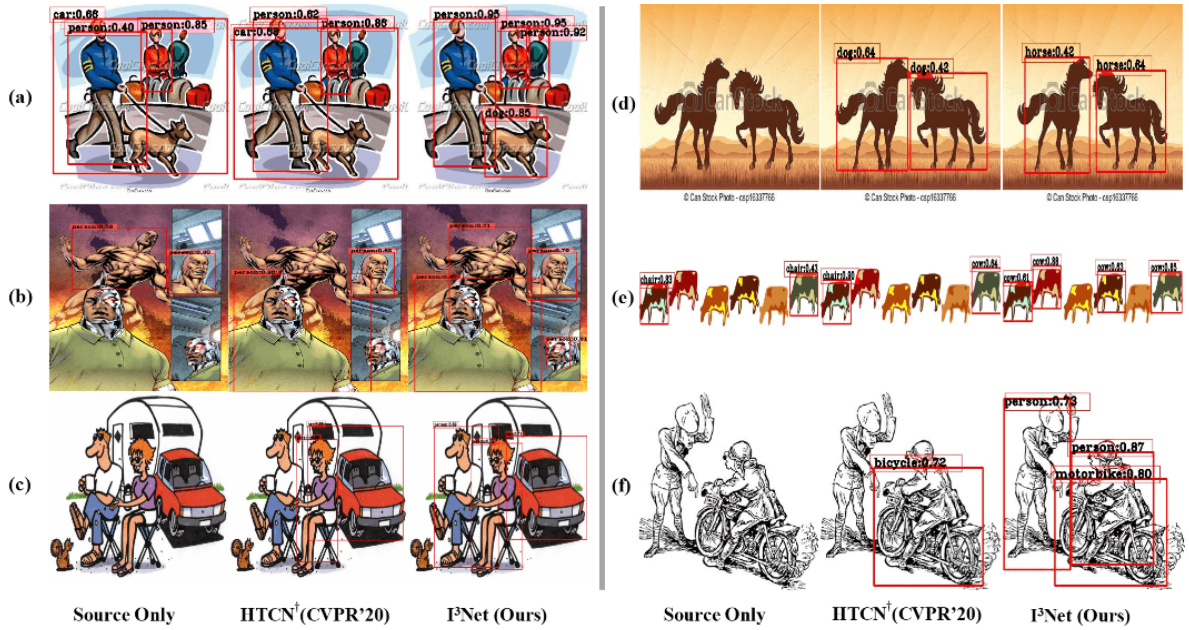

Figure 4: Qualitative detection results on Clipart1k, Watercolor2k, and Comic2k.

even with small object size, and (iii) suppressing the redundant background information.

Qualitative detection results. Figure 4 demonstrates the example of detection results on the three target domains, *i.e.,* Clipart1k, Watercolor2k, and Comic2k. The proposed I^{3}Net consistently and significantly outperforms both Source Only [22] and HTCN† [3] models in different transfer tasks. Owing to the introduction of DCBR, I^{3}Net is capable of precisely detecting the sample-scarce categories (*e.g.,* (a), (d), and (e)). I^{3}Net is able to detect those obscured objects and provide accurate bounding box predictions since we explicitly encourage the alignment of cross-domain object patterns via the proposed COPM (*e.g.,* (a), (b), (c), and (f)). In addition, due to the presence of RJCA, I^{3}Net is able to ensure the cross-domain semantic consistency, and thus significantly reduce the false positive results and enhance the classification accuracy (*e.g.,* (d) and (e)).

5. Conclusion

In this paper, we proposed the Implicit Instance-Invariant Network (I^{3}Net) to solve the cross-domain object detection problem based on the one-stage detectors without requiring explicit instance-level features. The key idea of our method is to implicitly learn instance-invariant features via exploiting the natural characteristics of deep features in different layers, *i.e.,* suppressing redundant information from the lower layers and enhancing the cross-domain semantic correlation of foreground objects at the higher layers. Experiments on three standard cross-domain detection benchmarks verified the effectiveness of our method.

Acknowledgement This work was partially supported by National Key Research and Development Program of China (No.2020YFC2003900) and the National Natural Science Foundation of China under Grants U19B2031, 61971369.

References

[1] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *CVPR*, pages 11457–11466, 2019. 2, 6

[2] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *CVPR*, pages 627–636, 2019. 2, 3, 5

[3] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *CVPR*, pages 8869–8878, 2020. 2, 3, 6, 7, 8

[4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 2, 3, 4, 6

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, pages 303–338, 2010. 6

[6] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, pages 2960–2967, 2013. 1

[7] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015. 1, 7

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016. 2, 6, 7

[9] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 2

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 2

[11] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073, 2012. 1

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 2

[13] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 5

[14] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*, 2019. 2, 6

[15] 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 *ECCV*, 2020. 2

[16] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, pages 5001–5009, 2018. 6, 7

[17] Xiang Jiang, Qicheng Lao, Stan Matwin, and Mohammad Havaei. Implicit class-conditioned domain alignment for unsupervised domain adaptation. In *ICML*, 2020. 2

[18] Purushottam Kar and Harish Karnick. Random feature maps for dot product kernels. In *Artificial Intelligence and Statistics*, pages 583–591, 2012. 5

[19] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *ICCV*, pages 6092–6101, 2019. 3, 6, 7

[20] Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *CVPR*, pages 13936–13944, 2020. 2

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 2, 4

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37, 2016. 2, 3, 4, 6, 7, 8

[23] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015. 1, 2, 7

[24] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NIPS*, pages 1640–1650, 2018. 1, 5

[25] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, pages 2208–2217, 2017. 1

[26] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 4

[27] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010. 1

[28] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *CVPR*, pages 2239–2247, 2019. 3

[29] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *CVPR*, 2019. 5

[30] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *AAAI*, 2018. 2

[31] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 2

[32] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, pages 7263–7271, 2017. 2

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 2

[34] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive ob-

ject detection. In *CVPR*, pages 6956–6965, 2019. 2, 3, 4, 6, 7

[35] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, pages 3723–3732, 2018. 2

[36] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, pages 761–769, 2016. 4

[37] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. In *ICLR*, 2018. 1

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[39] Peng Su, Kun Wang, Xingyu Zeng, Shixiang Tang, Dapeng Chen, Di Qiu, and Xiaogang Wang. Adapting object detectors with conditional domain normalization. In *ECCV*, 2020. 2

[40] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, pages 443–450. Springer, 2016. 2

[41] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528, 2011. 1

[42] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 1, 2

[43] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 2

[44] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *ICML*, pages 5419–5428, 2018. 1, 2, 3, 5

[45] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, pages 11724–11733, 2020. 2, 6

[46] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, pages 12355–12364, 2020. 2, 3, 5, 6

[47] Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, and Jindong Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *CVPR*, pages 4394–4403, 2020. 2

[48] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NIPS*, pages 3320–3328, 2014. 2

[49] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. In *ICLR*, 2017. 1, 2

[50] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *CVPR*, 2019. 2

[51] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *ECCV*, 2020. 2

[52] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *CVPR*, pages 13766–13775, 2020. 2, 3, 5, 6

[53] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *CVPR*, pages 687–696, 2019. 2, 6

[54] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019. 2