Dual Bipartite Graph Learning: A General Approach for Domain Adaptive Object Detection

Chaoqi Chen[1], Jiongcheng Li[2], Zebiao Zheng[2], Yue Huang[2], Xinghao Ding[2], Yizhou Yu[1]*
[1]The University of Hong Kong [2]Xiamen University
cqchen1994@gmail.com, jiongchengli@stu.xmu.edu.cn, zbzheng@stu.xmu.edu.cn
huangyue05@gmail.com, dxh@xmu.edu.cn, yizhouy@acm.org

Abstract

Domain Adaptive Object Detection (DAOD) relieves the reliance on large-scale annotated data by transferring the knowledge learned from a labeled source domain to a new unlabeled target domain. Recent DAOD approaches resort to local feature alignment in virtue of domain adversarial training in conjunction with the ad-hoc detection pipelines to achieve feature adaptation. However, these methods are limited to adapt the specific types of object detectors and do not explore the cross-domain topological relations. In this paper, we first formulate DAOD as an open-set domain adaptation problem in which foregrounds (pixel or region) can be seen as the "known class", while backgrounds (pixel or region) are referred to as the "unknown class". To this end, we present a new and general perspective for DAOD named Dual Bipartite Graph Learning (DBGL), which captures the cross-domain interactions on both pixel-level and semantic-level via increasing the distinction between foregrounds and backgrounds and modeling the cross-domain dependencies among different semantic categories. Experiments reveal that the proposed DBGL in conjunction with one-stage and two-stage detectors exceeds the state-of-the-art performance on standard DAOD benchmarks.

1. Introduction

Object detection has gained unprecedented development in the past decade, owing to the renaissance in deep learning and the explosive increase of labeled training data. Nevertheless, the performance gains rely on an assumption that the training and test data are drawn from identical distribution, which is challenged to be satisfied in real-world applications. Moreover, collecting large-scale annotated data in various domains is impractical. An intuitive solution is to directly apply the off-the-shelf object detection models trained on the source domain to a new domain. However, domain shift [36] hinders the deployment of models and emerges as an inevitable challenge. Unsupervised Domain Adaptation (UDA) [31] serves as a plausible solution to tackle this problem by facilitating knowledge transfer and mitigating the distributional shift between domains. The design principle of UDA is to learn domain-invariant features and ensure that the learned features will preserve a low risk on the source domain. Existing UDA methods mainly fall into two types, *i.e.,* statistics matching [15, 12, 26, 49, 35] and adversarial learning [13, 42, 27, 5, 19]. In this paper, we aim to investigate the UDA techniques for object detection, namely, Domain Adaptive Object Detection (DAOD).

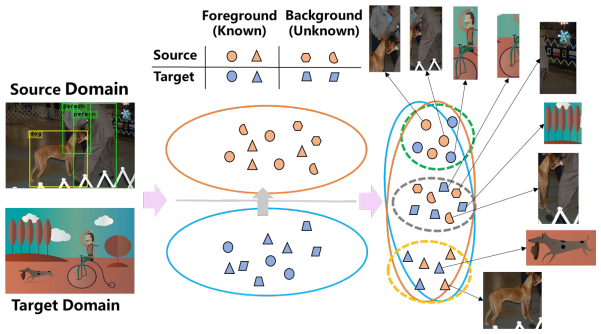

Figure 1: We formulate DAOD as an open-set domain adaptation problem, where foregrounds refer to the "known class" and backgrounds refer to the "unknown class".

Considering the local nature of detection tasks, most existing DAOD approaches strive to change the emphasis of adaptation from holistic to local in virtue of elaborate feature alignment modules regarding the foreground objects. However, they are highly model-related, that is to say, their adaptation process relies on the specific pipelines of detection models. For example, most of them [7, 52, 38, 4, 6, 47, 51, 46] resort to incorporate the adversarial training [13] within Faster R-CNN [37] based on the region proposal step to generate a sparse set of proposals (instance-level features). Given the dense prediction property of SSD [25], Kim *et al.* [21] propose to jointly reduce the false positives and false negatives during hard negative mining step. How

*Corresponding author

to bridge the gap between two-stage and one-stage DAOD is yet to be thoroughly studied. On the other hand, existing feature alignment techniques proposed by previous DAOD works focus on achieving one-to-one semantic matching while neglect the inherent topological structure regarding the relations among different foreground objects.

To tackle the above challenges, we first formulate DAOD as an Open Set Domain Adaptation (OSDA) problem [34]. Compared to closed set UDA problem, which assumes that the source and target domains share an identical label space, OSDA should additionally identify and isolate the unknown class before reducing the distributional shift of known classes between domains. In DAOD, as shown in Fig. 1, we found that backgrounds would be distinct across domains and thus can be seen as the "unknown class", *i.e., backgrounds* are non-transferable, while the *foregrounds* have more common features across domains. In this regard, strictly matching the whole distribution across domains will be risky and result in inferior performance. This motivates us to design DAOD algorithms in the following two steps: (1) Make a distinction between foreground and background feature representations in an unsupervised manner. (2) Apply adaptation to the foreground objects in both domains.

Motivated by this, we propose a general DAOD framework called Dual Bipartite Graph Learning (DBGL) to model the cross-domain topological relationships on *pixel-level* and *semantic-level* respectively, and learn fine-grained correspondence for knowledge transfer. The proposed DBGL can be seamlessly incorporated into any modern object detectors. To be specific, DBGL consists of two components, namely, Pixel-level Bipartite Graph Learning (PBGL) and Semantic-level Bipartite Graph Learning (SBGL). We search pixel-wise correspondence by only retaining mutual nearest neighbors that satisfy the mutual relation consistency requirement, and the pixel-level graph is constructed based on the searched pixel pairs that belong to the same foreground category across domains. Pixel prototype are introduced to reduce the influence of background pixels. Through message-passing, each foreground node aggregates the features from its neighbors of opposite domain, which naturally separates the foregrounds and backgrounds and strengthens the semantic correspondence. SBGL semantically models the cross-domain foreground object relations via bipartite graph learning. To identify and isolate the backgrounds, we first develop a cross-domain similarity regularization strategy to increase the similarity between foreground nodes and penalize the ones of nodes that are more likely to be backgrounds. To enhance the node features, we propose to utilize the internal node feature similarities to endow the node with context-aware ability and mitigate the negative influence of outlier nodes.

Our contributions can be summarized as follows:

- We formulate DAOD as an OSDA problem, which is not discussed by the literature and gives a hint to bridge the gap between theory and algorithm for DAOD. Then, we provide theoretical analysis on the upper bound of the expected target error under OSDA settings and reveal how to empirically optimize this upper bound in the context of our learning framework.
- We propose a new and general method that bridges the gap between one-stage and two-stage DAOD. The proposed DBGL, which jointly explores the cross-domain pixel-wise and semantic-wise topological relations, can discriminate the foreground-background and match foreground features in a more precise way.
- We conduct extensive experiments on three benchmarks based on two-stage (Faster R-CNN [37]) and one-stage (SSD [25]) object detectors. Experimental results reveal that our approach significantly outperforms the state-of-the-arts in DAOD.

2. Related Work

Unsupervised Domain Adaptation (UDA). A typical solution for UDA is to align the source and target feature representations in the shared latent space by incorporating well-defined divergence measures into deep architectures, such as Maximum Mean Discrepancy (MMD) [43, 26], Correlation Alignment (CORAL) [41], Central Moment Discrepancy (CMD) [49], and Optimal Transport (OT) distance [23, 48]. DANN [14] proposes a domain-adversarial training strategy to adversarially confuse a domain discriminator with the help of a Gradient Reversal Layer (GRL). ABG [29] develops an adversarial bipartite graph learning framework to model the source-target interactions for video-based UDA. Kang *et al.* [20] explore the pixel-level association (one-to-one) in the context of cross-domain semantic segmentation. However, they do not consider the topological correspondence between domains and thus fail to endow the adaptation model with cross-domain reasoning ability. More importantly, previous UDA works focus on the closed set setting and cannot be simply extended to OSDA [34]. Current OSDA methods [24, 1, 32, 30] are tailed for classification tasks and cannot generalize to detection task, where the foreground objects (positive samples) and backgrounds (negative samples) are naturally seen as the so-called known and unknown classes.

UDA for Object Detection. Domain Adaptive Faster R-CNN [7] is the first deep DAOD method that mitigates the domain disparity on both image-level and instance-level by domain adversarial training. Considering the local adaptation property of DAOD, most recent works [52, 38, 4, 16, 6, 47, 51, 46, 17, 40, 50] strive to change the emphasis of feature adaptation from global to local, and then explicitly align the derived local features on different levels. To be specific, Saito *et al.* [38] design a weak global alignment

module to avoid fully matching of the whole data distributions. Chen *et al.* [6] devise a hierarchical transferability calibration network to harmonize the contradiction between transferability and discriminability on different levels(*i.e.,* local-region, image, and instance). Xu *et al.* [47] and Zheng *et al.* [51] propose to perform fine-grained instance-level adaptation with respect to foreground objects based on prototype alignment [45, 5]. Zhao *et al.* [50] develop a collaborative self-training strategy to train RPN and RPC with high-confidence ROIs. On the other hand, Kim *et al.* [21], which is the only one-stage DAOD work, propose a weak self-training method to mitigate the negative effects of inaccurate pseudo-labels for adapting SSD [25]. Despite their strong capability on adapting certain detectors, current DAOD works can not be extended to distinct detection pipelines and thus fail to form a general adaptation framework. In addition, theoretical analysis regarding the statistical upper bound of DAOD is less investigated. And how to model the cross-domain topological relationships for capturing the interactions between two set of entities remains the boundary to explore.

3. Theoretical Motivation

We theoretically analyze the motivation of our approach with respect to the upper bound of OSDA, making using of statistical learning theory of domain adaptation [11, 2, 3]. Before introducing the generalization bound, we first provide the problem setting and definitions.

Definition 1. ***Open-Set Domain Adaptation (OSDA).*** *Suppose that we have a source domain $\mathcal{D}_s = \{(x_{s_i}, y_{s_i})\}_{i=1}^{n_s}$ of n_s labeled samples and a target domain $\mathcal{D}_t = \{x_{t_j}\}_{j=1}^{n_t}$ of n_t unlabeled samples. $\mathcal{D}_s$ and $\mathcal{D}_t$ are drawn from $P(\mathcal{X}_s, \mathcal{Y}_s)$ and $Q(\mathcal{X}_t, \mathcal{Y}_t)$, $P \neq Q$. The source and target label spaces share K known classes and individually include a unknown class u_s and u_t, which is different in both domains (i.e., $u_s \neq u_t$). The goal of OSDA is to learn an optimal target classifier $h : \mathcal{X}_t \rightarrow \mathcal{Y}_t$.*

Definition 2. ***Source and Target Risks.*** *The source risk $R_s(h)$ and target risk $R_t(h)$ of h w.r.t. $\mathcal{L}$ under source distribution P and target distribution Q are defined as*

$$R_s(h) \triangleq \mathbb{E}_{(x,y)\sim P}\mathcal{L}(h(x), y) = \sum_{i=1}^{K+1} \pi_i^s R_{s,i}(h)$$

$$R_t(h) \triangleq \mathbb{E}_{(x,y)\sim Q}\mathcal{L}(h(x), y) = \sum_{j=1}^{K+1} \pi_j^t R_{t,j}(h)$$

where $\pi_i^s = P(y = i)$ and $\pi_j^t = Q(y = j)$ are class-prior probabilities of P and Q. Then, we have

$$R_s(h) = \sum_{i=1}^{K} \pi_i^s R_{s,i}(h) + \pi_{K+1}^s R_{s,K+1}(h) = R_s^*(h) + \Delta_s$$

$$R_t(h) = \sum_{j=1}^{K} \pi_j^t R_{t,j}(h) + \pi_{K+1}^t R_{t,K+1}(h) = R_t^*(h) + \Delta_t$$

Given the hypothesis space $\mathcal{H}$ with a condition that constant function $K+1 \in \mathcal{H}$, for $\forall h \in \mathcal{H}$, the expected error on target samples $R_t(h)$ is bounded as,

$$\frac{R_t(h)}{1-\pi_{K+1}^t} \leq R_s^*(h) + d_{\mathcal{H}\Delta\mathcal{H}}(P_{X|Y\leq K}, Q_{X|Y\leq K}) + \lambda + \frac{\Delta_t}{1-\pi_{K+1}^t} \tag{1}$$

where the shared error $\lambda = \min_{h\in\mathcal{H}} \frac{R_t^*(h)}{1-\pi_{K+1}^t} + R_s^*(h)$, $R_s^*(h) = \sum_{i=1}^{K} \pi_i^s R_{s,i}(h)$, and $\Delta_t = \pi_{K+1}^t R_{t,K+1}(h)$. We show the derivation of Inequality (1) in the supplementary material. According to Inequality (1), the target error is bounded by four terms: (1) expected error on the known classes of source domain $R_s^*(h)$; (2) domain divergence $d_{\mathcal{H}\Delta\mathcal{H}}(P_{X|Y\leq C}, Q_{X|Y\leq C})$; (3) shared error λ of the ideal joint hypothesis h^*; (4) target open set risk Δ_t.

Remark 1. *$R_s^*(h)$ is expected to be small and can be easily minimize since we have source ground truth labels. $d_{\mathcal{H}\Delta\mathcal{H}}(P_{X|Y\leq C}, Q_{X|Y\leq C})$ is associated with domain disparity and can be minimize by domain alignment step. λ is associated with the class-wise conditional shift and can be minimize by category alignment, i.e., SBGL in our approach. The target open set risk Δ_t tends to be large when an approach does not regard the target backgrounds as an unknown class. In our approach, we optimize this term by the proposed DBGL to make a distinction between known and unknown classes. In a nutshell, our work aims to explicitly optimize the upper bound of expected target error by jointly minimizing the aforementioned four terms.*

4. Dual Bipartite Graph Learning

Framework Overview. As demonstrated in Figure 2, the proposed DBGL consists of two components, *i.e.,* PBGL and SBGL. PBGL builds the cross-domain pixel-wise correlations (based on low-level features) with respect to the possible foreground pixel pairs and explicitly enhances their connections via node classification, which enforce the separation between foreground and background pixels in an unsupervised manner. SBGL models the cross-domain inter-class interactions based on a set of instance-level (Faster R-CNN [37]) or per-anchor (SSD [25]) features, and hereby strengthens the context-aware ability and the semantic consistency of high-level features. Note that the proposed PBGL and SBGL are complementary to each other. Specifically, PBGL alleviates the negative influence of asymmetric semantic space for SBGL by making a clear distinction between foregrounds and backgrounds, and the corresponding class alignment learned by SBGL can boost the accuracy and robustness of separation guided by PBGL.

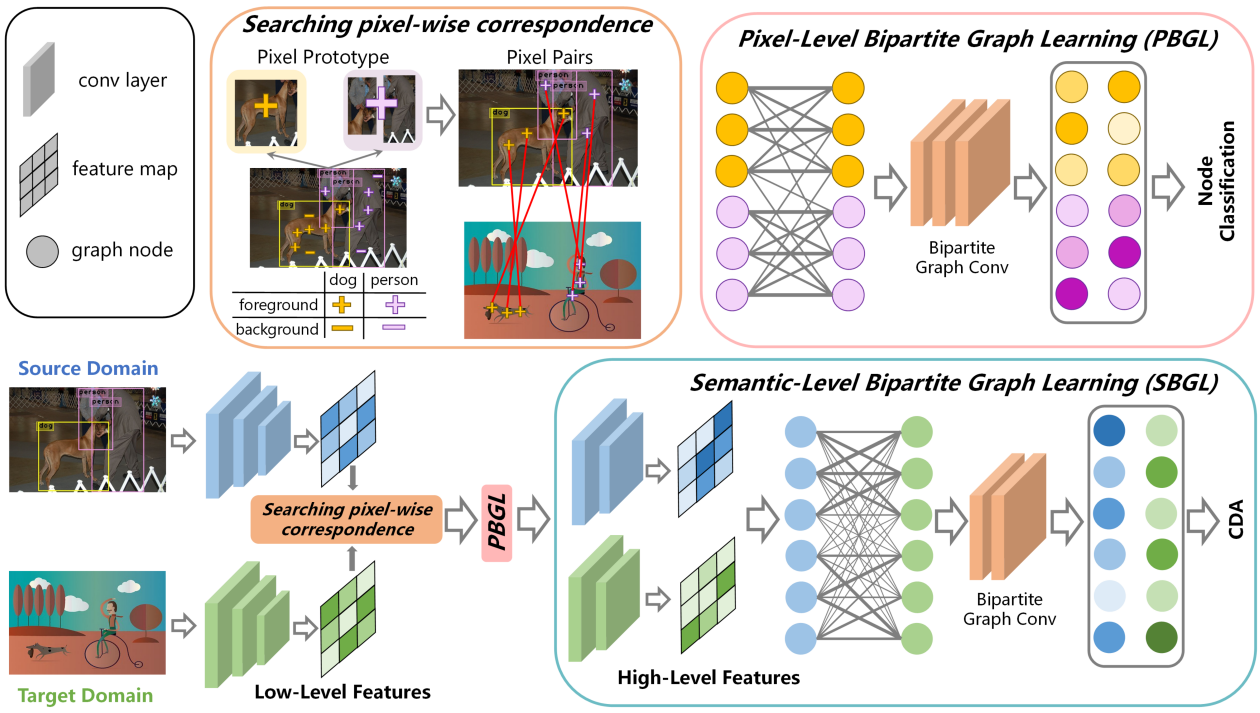

Figure 2: The overall architecture of DBGL, which mainly includes the pixel-level and semantic-level bipartite graph learning modules, *i.e.,* PBGL and SBGL. CDA denotes the category-aware domain alignment loss.

4.1. Pixel-Level Bipartite Graph Learning

For the low-level features, existing DAOD methods usually focus on strongly aligning them [7, 38, 16] or trying to capture the foreground objects via attention-like modules [6, 51, 17]. However, strong feature alignment will inevitably blend the foreground and background features, and thus cause negative transfer. Moreover, those attention-like modules extract the foreground features guided by source supervision, which makes adaptation process source-biased and error-prone. To discriminate the foregrounds and backgrounds, the proposed PBGL models the foreground pixel-level correlations between domains by message-passing and feature aggregation, which avoids "hard" separation or feature reweighing that were widely adopted by prior works.

Suppose that we are given the source and target 3D feature maps $F_s, F_t \in \mathbb{R}^{C\times H\times W}$ extracted from shallow layer of the backbone network. Then, we aim to project the spatial visual features F_s and F_t to node domain, *i.e.,* constructing a pixel-level bipartite graph $\mathcal{G}_P = \{\mathcal{V}_s^P, \mathcal{V}_t^P, \mathcal{E}^P\}$, where $\mathcal{V}_s^P$ and $\mathcal{V}_t^P$ denotes the source and target node sets. $\mathcal{E}^p$ stands for the set of edges, which measure the node affinity of pixel-level features between domains. An intuitive approach to construct edges is to linking all pixels across domains, which yet will incur redundant and bring in hefty computation. Thus, we propose a more efficient approach by only retaining mutual nearest neighbors that satisfy the mutual relation consistency requirement.

We first define the concept of *pixel prototype* for each source category, which denotes the mean feature of pixels belonging to the same object categories within a source image. Here, the category label of a source pixel is depended on the object annotation and bounding box, and the bounding box inevitably contains noisy background pixels. Thus, pixel prototype can alleviate the negative influence of backgrounds. The definition is formulated as,

$$c_s^k = \frac{1}{|I_s^k|}\sum_{F_s^i \in I_s^k} F_s^i,\ k = \{1, 2, \ldots, K\} \tag{2}$$

where i is the pixel index, and I_s^k is the set of pixels labeled with class k in a source feature map F_s. Then, we utilize c_s^k to select pixels in I_s^k that have higher similarity with c_s^k, *i.e.,* if $\cos(c_s^k, F_s^i) > \tau$, F_s^i is added into $\hat{I}_s^k$, where $\cos(\cdot,\cdot)$ denotes the cosine similarity, τ is a threshold, and $\hat{I}_s^k$ denotes the selected set.

For each source pixel i^* in $\hat{I}_s^k$, assume that j' is its nearest neighbor in the target domain. Similarly, i' is the nearest neighbor of target pixel j' in the source domain. If i' also belongs to the category k, we will assign the target pixel j' with pseudo-label k. By doing so, we can obtain two set of selected pixels in both domains, *i.e.,* $\mathcal{V}_s^P$ and $\mathcal{V}_t^P$. Bipartite graph edges $\mathcal{E}^p$ aim to represent the similarities between nodes. To mitigate the impact of noisy background pixels, we let the similarity be learnable,

$$\mathcal{E}_{ij}^p = \sigma([F_s^i, F_t^j]\theta_e^p) \tag{3}$$

where σ denotes the sigmoid function and θ_e^p is the learnable parameter. To conduct graph convolution on the constructed bipartite graph $\mathcal{G}_P$, we augment its original form,

$$\hat{\mathcal{V}}^P = [\mathcal{V}_s^P, \mathcal{V}_t^P], \tag{4}$$

$$\hat{\mathcal{E}}^P = \begin{pmatrix} \mathbf{0} & \mathcal{E}^P \\ (\mathcal{E}^P)^T & \mathbf{0} \end{pmatrix} \tag{5}$$

Then, the augmented bipartite graph $\hat{\mathcal{G}}^P = \{\hat{\mathcal{V}}^P, \hat{\mathcal{E}}^P\}$ can be learned by utilizing the modern Graph Convolutional Networks (GCN) techniques [22]. We stack multiple graph convolution layers in our implementation. Specifically, the graph convolution is recursively conducted as: $\boldsymbol{X}^{(l+1)} = \mathrm{ReLU}\left(\hat{\boldsymbol{A}}\boldsymbol{X}^{(l)}\boldsymbol{W}^{(l)}\right)$, where $\boldsymbol{W}^l$ is the parameter matrix, $\boldsymbol{X}^l$ are the hidden features of the l-th layer (where $1 \leq l \leq L$), and $\hat{\boldsymbol{A}}$ is the adjacency matrix. To further distinguish the foreground and background nodes, we conduct node classification in the bipartite graph. Note that the selected source and target pixels have ground-truth labels and pseudo labels, respectively. Formally, the last layer of pixel-level bipartite graph (GCN_1) predicts the label using a classifier and can be written as follows,

$$\hat{y} = \mathrm{softmax}(FC(GCN_1(x, \hat{\mathcal{G}}^P))), \tag{6}$$

where $\hat{y}$ is the predicted label, FC is a fully-connected layer, and x is the feature of source or target nodes. The node classification loss is denoted by $\mathcal{L}_{\mathrm{NC}}^{\mathrm{GCN}_1}$.

4.2. Semantic-Level Bipartite Graph Learning

Learning semantic correlations between domains is the central problem of domain adaptation. In this regard, numerous elaborate semantic alignment strategies have been proposed. Among them, prototype alignment [45, 5, 33] serve as the representative approach to achieve semantic consistency. Recent DAOD works [47, 51] also introduce this approach to align foreground objects with the same category labels based on a sparse set of object proposals.

Despite their general efficacy for various tasks ranging from classification to detection, these prototype alignment approaches are still confined by several limitations. (1) Prototype alignment only considers the one-to-one cross-domain correspondence without exploring the inter-class relations, which contain rich information with respect to the topological structure of the semantic space. (2) When computing the prototype of each foreground object category based on the embedded representations, it will inevitably include some negative samples (*i.e.,* backgrounds), which make the adaptation process risky and uncontrolled. (3) Prototype alignment is more suitable for adapting two-stage detectors since they have explicit instance-level features generated by the region proposal mechanism. By contrast, one-stage detectors usually require per-pixel prediction and thus including many negative candidates. (4) Vanilla prototype alignment cannot be simply applied to an OSDA problem since the source and target label spaces are asymmetric.

Inspired by the discussions above, we devise a semantic-level bipartite graph (GCN_2) to compensate for the lack of topological modeling *w.r.t.* inter-class relations between domains. Let the bipartite graph be $\mathcal{G}_\mathcal{S} = \{\mathcal{V}_s^S, \mathcal{V}_t^S, \mathcal{E}^S\}$. The source node set is $\mathcal{V}_s^S = \{v_{s_i}\}_{i=1}^{N_p} \in \mathbb{R}^{N_p \times d}$ and the target node set is $\mathcal{V}_t^S = \{v_{t_j}\}_{j=1}^{N_p} \in \mathbb{R}^{N_p \times d}$, where v_{s_i} and v_{t_j} denote the ROI-based instance-level features generated by RPN, N_p represents the number of proposals, and d denotes the node feature dimension. $\mathcal{E}^S$ denotes the set of edges. Note that we take Faster R-CNN as an exemplar to illustrate the technical details of SBGL and then generalize to one-stage detector (SSD) in experiments (cf. Section 5).

Cross-Domain Similarity Regularization. Firstly, we need to characterize the correspondence between two independent node sets, *i.e.,* define the adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N_p \times N_p}$, which associates each edge (v_{s_i}, v_{t_j}) with its element A_{ij}. An optional approach is to traverse all possible pairs between source and target proposals to compute their similarity. Intuitively, node pairs with higher similarity should be assigned larger weights. However, considering the asymmetry of OSDA problem, we need to make a distinction between known and unknown classes; otherwise, the message-passing process may make the target nodes aggregate biased semantic information.

To identify and isolate the backgrounds, we propose a Cross-Domain Similarity Regularization (CDSR) strategy to produce reliable node pairs between domains. Our motivation is to regularize the similarity measure such that the nearest neighbor of a source node, in the target domain, is more likely to have as a nearest neighbor this particular source node, *i.e.,* assign large weights to nodes from $\mathcal{V}_s$ and $\mathcal{V}_t$ that are mutual nearest neighbors. However, we found that v_t[1] being a K-NN of v_s does not indicate that v_s is a K-NN of v_t, which is also known as hubness problem [9, 8]. In high-level semantic space, some nodes are more likely to be the nearest neighbors of many other nodes (*e.g.,* easy positives), but some others may be not nearest neighbors of any node (*e.g.,* hard negatives). On bipartite graph $\mathcal{G}_S$, the neighborhood associated with a source node v_s is denoted by $\mathcal{N}_T(v_s)$. All K elements of $\mathcal{N}_T(v_s)$ are nodes from $\mathcal{V}_t$. Similarly, the neighborhood associated with a target node v_t is represented by $\mathcal{N}_S(v_t)$. The mean similarity of a source node v_s to its target neighborhood is denoted by,

$$r_T(v_s) = \frac{1}{K} \sum_{v_t \in \mathcal{N}_T(v_s)} \cos(v_s, v_t), \tag{7}$$

Likewise, the mean similarity of a target node v_t to its source neighborhood is represented by $r_S(v_t)$. Formally,

[1] We omit the subscripts i and j for simplicity.

we utilize these similarities to define a cross-domain similarity measure $\mathrm{CDSR}(\cdot,\cdot)$ between nodes,

$$\mathrm{CDSR}(v_s, v_t) = \sigma(2\cos(v_s, v_t) - r_T(v_s) - r_S(v_t)) \quad (8)$$

By doing so, we can obtain the adjacency matrix $\boldsymbol{A}$.

Node Feature Enhancement. In DAOD, the target high-level features are prone to be somewhat biased and inaccurate to represent an object under the presence of domain shift. For example, in the series of domain adaptive Faster R-CNN, the target region proposals are randomly generated and cannot be divided into positive or negative samples due to the absence of ground-truth labels. Thus, the constructed bipartite graph may be incapable of precisely modeling foreground object relations. To enhance the target node features, we draw motivation from non-local operations [44, 20] to model target intra-domain global dependencies by representing each node feature as a weighted sum of features from all the other target node features,

$$v_{t_j} = \theta v_{t_j} + (1-\theta)\sum_{k|k\neq j} w_k v_{t_k}, \quad (9)$$

$$w_k = \frac{e^{\mathrm{CDSR}(v_{t_j}, v_{t_k})}}{\sum_k e^{\mathrm{CDSR}(v_{t_j}, v_{t_k})}} \quad (10)$$

where θ is set to 0.5 in all experiments. Note that *this step is only used for initialization and does not be updated as training proceeds*. The enhanced target node features globally aggregate the feature of other positions over the semantic node space, which implicitly endow the node features with context-aware ability. In addition, by comparing the similarity of node features within the target domain, the relations of nodes belonging to the same category can be strengthened. We follow Eq. (4)-(5) to augment $\mathcal{G}^S$ as $\hat{\mathcal{G}}^S = \{\hat{\mathcal{V}}^S, \hat{\mathcal{E}}^S\}$ and then conduct graph convolution.

Category-Aware Domain Alignment. Based on $\hat{\mathcal{G}}^S_k$, we propose a Category-aware Domain Alignment (CDA) loss on top of $\hat{\mathcal{G}}^S$ to conduct domain alignment on all foreground categories. Technically, we contrastively align the source and target prototypes to achieve domain alignment. The source and target prototypes are defined as,

$$\begin{aligned} P^k_s &= \frac{1}{|\hat{\mathcal{G}}^S_k|}\sum_{x^i_s \in \hat{\mathcal{G}}^S_k} GCN_2(x^i_s, \hat{\mathcal{G}}^S_k) \\ P^k_t &= \frac{1}{|\hat{\mathcal{G}}^S_k|}\sum_{x^i_t \in \hat{\mathcal{G}}^S_k} GCN_2(x^i_t, \hat{\mathcal{G}}^S_k) \end{aligned} \quad (11)$$

where $|\hat{\mathcal{G}}^S_k|$ denotes the nodes in $\mathcal{G}^S$ belonging to class k ($k \in \{1, 2, \ldots, K\}$). We utilize the target pseudo-labels to cluster target nodes into K classes. Then, the CDA loss is formulated as follows,

$$\mathcal{L}^{\mathrm{GCN}_2}_{\mathrm{CDA}} = \sum_k \left\| P^k_s, P^k_t \right\|_2 + \sum_{m\neq n} (max\{0, \xi - \|P^m_s, P^n_t\|_2\}) \quad (12)$$

where ξ is the margin term and set to 1 in all experiments.

4.3. Overall Objective

Assume that the detection loss is denoted as $\mathcal{L}_{\text{det}}$, which contains the classification and regression losses. Since the proposed DBGL is capable of working in a plug-and-play manner, we incorporate DGBL into the domain adversarial training [13] framework by adding domain discriminators on low-level features. To this end, the overall objective function of DBGL is formulated as,

$$\mathcal{L}_{\mathrm{DBGL}} = \mathcal{L}_{\text{det}} + \alpha\mathcal{L}_{\text{adv}} + \beta\mathcal{L}^{\mathrm{GCN}_1}_{\mathrm{NC}} + \gamma\mathcal{L}^{\mathrm{GCN}_2}_{\mathrm{CDA}} \quad (13)$$

where α, β, and γ are hyper-parameters. $\mathcal{L}_{\text{adv}}$ denotes the vanilla adversarial training loss.

5. Experiments

5.1. Datasets

We evaluate the proposed DGBL on **Pascal VOC** [10], **Clipart1k**, **Watercolor2k**, and **Comic2k** [18] datasets, which form three DAOD tasks. Following prior DAOD works [38, 21, 16], we combine the Pascal VOC2007-trainval and VOC2012-trainval datasets as the source domain, and use Clipart1k, Watercolor2k, and Comic2k as the target domains respectively. The Pascal VOC [10] is a real-world image dataset, which contains 16,551 images with 20 object classes. Clipart1k, Watercolor2k, and Comic2k, which are collected from a website called Behance and annotated by Inoue *et al.* for cross-domain object detection tasks, consist of 1,000, 2,000, and 2,000 images respectively. Clipart1k has the same 20 object categories as Pascal VOC, and Watercolor2k and Comic2k share 6 identical object classes with the Clipart1k dataset, *i.e.,* bicycle, bird, cat, car, dog, and person. For Pascal VOC $\rightarrow$ Clipart, we use all images of Clipart1k as the target domain for both training and testing by following mainstream DAOD works [38, 6]. For Pascal VOC $\rightarrow$ Watercolor and Pascal VOC $\rightarrow$ Comic, we leverage the train set (1K images) for training and the test set (1K images) is held out for evaluation.

5.2. Implementation Details

For the two-stage detector based experiments, we follow the same setting in [38, 6] that choose Faster R-CNN framework with ResNet-101 architectures. The shorter side of each input image is resized to 600 and the batch size is set to 2 (one image per domain) to fit the GPU memory. For the one-stage detector based experiments, we follow the setting in [18, 21] that utilize SSD300 [25] framework with

Table 1: Results on PASCAL VOC → Clipart Dataset (%).

Methods	aero	bcycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hrs	bike	prsn	plnt	sheep	sofa	train	tv	mAP
Faster R-CNN + ResNet-101																					
Source Only [37]	**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 [7]	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 [38]	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 [6]	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
DBGL (Ours)	28.5	52.3	**34.3**	32.8	38.6	**66.4**	38.2	**25.3**	**39.9**	47.4	**23.9**	17.9	38.9	**78.3**	61.2	51.7	**26.2**	28.9	**56.8**	44.5	**41.6**
SSD + VGG-16																					
Source Only [25]	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 [14]	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 [18]	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 [21]	**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 [21]	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
BSR+WST [21]	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
DBGL (Ours)	23.2	**65.5**	**30.1**	18.3	24.6	**67.6**	**43.9**	15.1	38.7	36.4	31.3	**20.2**	25.0	**74.3**	**55.1**	38.2	12.5	**41.0**	49.1	**43.9**	**37.7**

Table 2: Results on Pascal VOC → Watercolor2k (%).

Methods	bike	bird	car	cat	dog	person	mAP
Faster R-CNN + ResNet-101							
Source Only [37]	68.8	46.8	37.2	32.7	21.3	60.7	44.6
BDC-Faster	68.6	48.3	47.2	26.5	21.7	60.5	45.5
DA-Faster [7]	75.2	40.6	48.0	31.5	20.6	60.0	46.0
SWDA [38]	82.3	**55.9**	46.5	32.7	35.5	**66.7**	53.3
DBGL (Ours)	**83.1**	49.3	**50.6**	**39.8**	**38.7**	61.3	**53.8**
SSD + VGG-16							
Source Only [25]	77.5	46.1	44.6	30.0	26.0	58.6	47.1
DANN [14]	73.4	41.0	32.4	28.6	22.1	51.4	41.5
BSR [21]	82.8	43.2	**49.8**	29.6	27.6	58.4	48.6
WST [21]	77.8	**48.0**	45.2	30.4	29.5	**64.2**	49.2
BSR+WST [21]	75.6	45.8	49.3	34.1	30.3	64.1	49.9
DBGL (Ours)	**84.0**	46.7	45.5	**36.2**	**35.7**	63.7	**52.0**

Table 3: Results on Pascal VOC → Comic2k (%).

Methods	bike	bird	car	cat	dog	person	mAP
Faster R-CNN + ResNet-101							
Source Only [37]	33.2	14.8	23.8	**19.5**	19.7	35.6	24.4
SWDA [38]	**36.0**	18.3	29.3	9.3	22.9	**48.4**	27.4
DBGL (Ours)	35.6	**20.3**	**33.9**	16.4	**26.6**	45.3	**29.7**
SSD + VGG-16							
Source Only [25]	43.3	9.4	23.6	9.8	10.9	34.2	21.9
DANN [14]	33.3	11.3	19.7	**13.4**	19.6	37.4	22.5
BSR [21]	45.2	15.8	26.3	9.9	15.8	39.7	25.5
WST [21]	45.7	9.3	30.4	9.1	10.9	46.9	25.4
BSR+WST [21]	**50.6**	13.6	**31.0**	7.5	16.4	41.4	26.8
DBGL (Ours)	45.4	**15.9**	24.8	11.5	**29.4**	**55.1**	**30.4**

VGG-16 [39] architectures. The input images are resized to 300×300 and the batch size is set to 32 (16 images per domain). We fine-tune ResNet-101 and VGG-16 pre-trained on ImageNet. In all experiments, we report mean average precision (mAP) with a IoU threshold of 0.5. We train the domain adaptive detection network using stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.001 and momentum 0.9. The learning rate is decreased to 0.0001 after 5 epochs. We set $\alpha = 1$ and $\beta = \gamma = 0.1$ in Eq. (13) for all experiments. We implement our experiments based on PyTorch deep learning framework.

5.3. Comparisons with State-of-the-Arts

We compare the proposed DGBL with the state-of-the-art DAOD methods, including Domain Adversarial Neural Networks **(DANN)** [14], Strong-Weak Distribution Alignment **(SWDA)** [38], adversarial Background Score Regularization + Weak Self-Training **(BSR+WST)** [21], and Hierarchical Transferability Calibration Network **(HTCN)** [6]. **Source Only** represents the baseline model that is trained on the source domain and directly applied to the target domain without adaptation procedure. We derive the quantitative results of DANN based on our reproduction. For other aforementioned methods, we cite the experimental results reported in their original papers.

Results on Clipart1k. We compare with the state-of-the-art methods in Table 1 based on Faster R-CNN and SSD detection frameworks respectively. The proposed DBGL substantially outperforms all compared methods in general and improves over state-of-the-art by +1.3% (40.3% to 41.6%) and +2.0% (35.7% 37.7%), indicating that our method can boost the adaptation ability of both one-stage and two-stage detectors. In addition, the results also reveal the importance of exploring the cross-domain topological relations and endowing the adaptation model with reasoning ability.

Results on Watercolor2k and Comic2k. Table 2 and Table 3 display the adaptation results on Pascal VOC → Watercolor2k and Pascal VOC → Comic2k respectively. The proposed DGBL exceeds all compared methods on most object categories and achieve the best mAP on average, demonstrating the efficacy and scalability of the proposed learning framework for modeling distinct DAOD scenarios. We can observe that DGBL shows impressive adaptation performance on the challenging DAOD task, *i.e.,* Pascal VOC → Comic2k (26.8% to 30.4%), where the distribu-

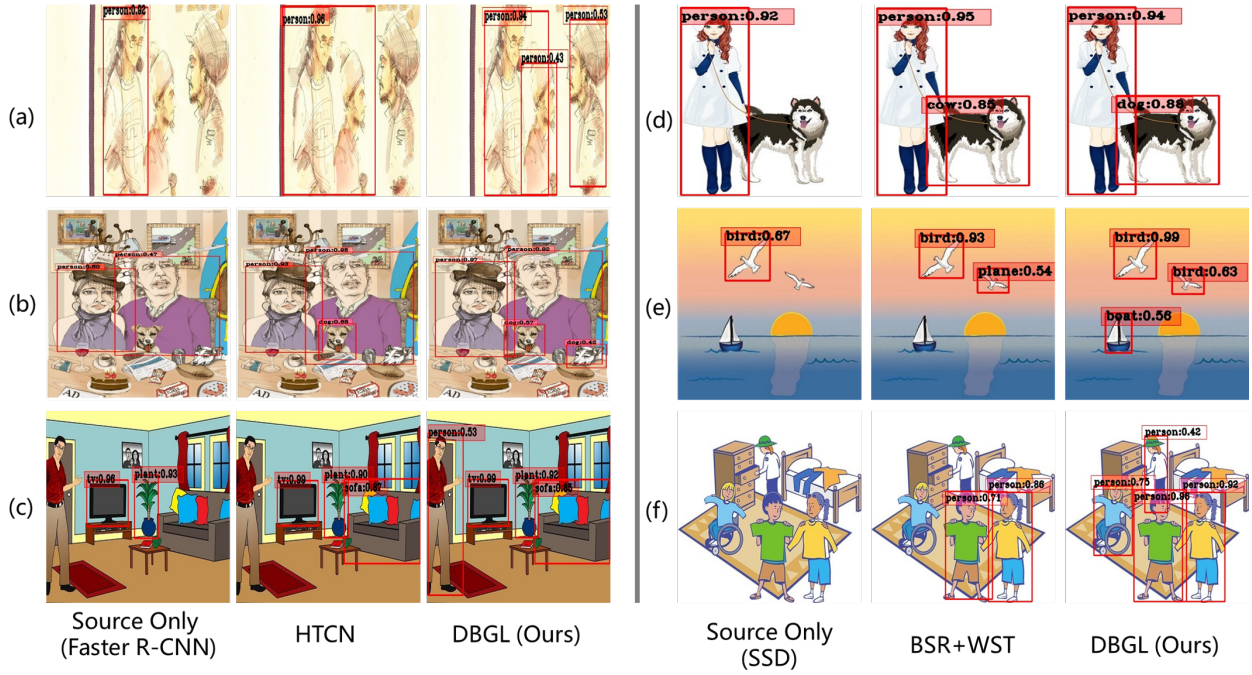

Figure 3: Qualitative detection results on Clipart1k, Watercolor2k, and Comic2k.

Table 4: Ablation of DBGL on three transfer tasks (%).

Source / Target	Clipart1k	Pascal VOC Watercolor2k	Comic2k
Faster R-CNN + ResNet-101			
w/o PBGL	39.5	52.0	28.3
w/o SBGL	39.1	51.7	27.6
PBGL w/ random link	37.2	48.1	26.4
SBGL w/o enhancement	41.0	53.1	28.8
DBGL (Full)	**41.6**	**53.8**	**29.7**
SSD + VGG-16			
w/o PBGL	36.1	50.5	28.0
w/o SBGL	35.3	50.1	27.1
PBGL w/ random link	33.4	45.6	26.4
SBGL w/o enhancement	37.0	50.9	29.6
DBGL (Full)	**37.7**	**52.0**	**30.4**

tional shift is considerably larger than other DAOD scenarios. The justification is that matching highly distinct distributions are error-prone, DBGL explicitly considers the topological information and thus achieve better alignment.

5.4. Further Empirical Analysis

Ablation Study. We delve into the individual effect and interaction of the proposed modules (*i.e.,* PBGL and SBGL) by conducting complete and in-depth ablation studies. The quantitative results are shown in Table 4. **(1)** w/o PBGL and w/o SBGL denote that we remove PBGL and SBGL from the full DBGL model respectively. **(2)** PBGL w/ random link denotes that we randomly select pixel-level graph nodes instead of using the proposed method to choose highly similar foreground pixel pairs. **(3)** SBGL w/o enhancement denotes that we remove the node feature enhancement step in the SBGL module. We can see that the performance drops accordingly when any one of the components modules is discarded, revealing the effectiveness and complementarity of all the proposed components in DBGL.

Qualitative detection results. Fig. 3 demonstrates some detection results of different methods on three target domains, *i.e.,* Clipart1k, Watercolor2k, and Comic2k. The proposed DBGL significantly outperforms Source Only, WST+BSR [21], and HTCN [6] models on different target domains. As can be seen, **(1)** DBGL detects the sample-scarce categories in a more precise way (*e.g.,* cow/dog in (d) and plane/bird in (e)). **(2)** DBGL is able to detect those obscured foreground objects and provide better bounding box regression (*e.g.,* dog in (b) and person in (c), (f)).

6. Conclusion

In this work, we propose a simple and general framework for DAOD problem by exploring the topology-aware and reasoning ability of detectors. Instead of relying on the ad-hoc detection pipelines, the key idea of our method is to model the cross-domain topological interactions and correlations on pixel-level and semantic level, and draw similar node features closer via message-passing and feature aggregation. Experiments on three DAOD benchmarks demonstrated the effectiveness of the proposed DBGL in conjunction with one-stage and two-stage detectors.

References

[1] Mahsa Baktashmotlagh, Masoud Faraki, Tom Drummond, and Mathieu Salzmann. Learning factorized representations for open-set domain adaptation. In *ICLR*, 2019. 2

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010. 3

[3] Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 129–136, 2010. 3

[4] 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. 1, 2

[5] 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. 1, 3, 5

[6] 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. 1, 2, 3, 4, 6, 7, 8

[7] 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. 1, 2, 4, 7

[8] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *ICLR*, 2018. 5

[9] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. In *ICLR, Workshop Track*, 2015. 5

[10] 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

[11] Zhen Fang, Jie Lu, Feng Liu, Junyu Xuan, and Guangquan Zhang. Open set domain adaptation: Theoretical bound and algorithm. *IEEE transactions on neural networks and learning systems*, 2020. 3

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

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015. 1, 6

[14] 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, 7

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

[16] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*, 2019. 2, 4, 6

[17] 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, 4

[18] 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

[19] Xiang Jiang, Qicheng Lao, Stan Matwin, and Mohammad Havaei. Implicit class-conditioned domain alignment for unsupervised domain adaptation. In *ICML*, 2020. 1

[20] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. *NeurIPS*, 33, 2020. 2, 6

[21] 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. 1, 3, 6, 7, 8

[22] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 5

[23] 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

[24] Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *CVPR*, pages 2927–2936, 2019. 2

[25] 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. 1, 2, 3, 6, 7

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

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

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

[29] Yadan Luo, Zi Huang, Zijian Wang, Zheng Zhang, and Mahsa Baktashmotlagh. Adversarial bipartite graph learning for video domain adaptation. In *ACM MM*, pages 19–27, 2020. 2

[30] Yadan Luo, Zijian Wang, Zi Huang, and Mahsa Baktashmotlagh. Progressive graph learning for open-set domain adaptation. In *ICML*, pages 6468–6478, 2020. 2

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

[32] Yingwei Pan, Ting Yao, Yehao Li, Chong-Wah Ngo, and Tao Mei. Exploring category-agnostic clusters for open-set domain adaptation. In *CVPR*, pages 13867–13875, 2020. 2

[33] 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

[34] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *ICCV*, pages 754–763, 2017. 2

[35] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 1

[36] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009. 1

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 1, 2, 3, 7

[38] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 1, 2, 4, 6, 7

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[40] 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

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

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

[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] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 6

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

[46] 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. 1, 2

[47] 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. 1, 2, 3, 5

[48] 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

[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] 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, 3

[51] 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. 1, 2, 3, 4, 5

[52] 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. 1, 2