CIGAR: Cross-Modality Graph Reasoning for Domain Adaptive Object Detection

Yabo Liu[1,2] Jinghua Wang[1*] Chao Huang[3] Yaowei Wang[2] Yong Xu[1,2*]
[1]Harbin Institute of Technology, Shenzhen [2]Peng Cheng Laboratory
[3]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
yaboliu.ug@gmail.com wangjh2012@foxmail.com huangchao_08@126.com
wangyw@pcl.ac.cn yongxu@ymail.com

Abstract

Unsupervised domain adaptive object detection (UDA-OD) aims to learn a detector by generalizing knowledge from a labeled source domain to an unlabeled target domain. Though the existing graph-based methods for UDA-OD perform well in some cases, they cannot learn a proper node set for the graph. In addition, these methods build the graph solely based on the visual features and do not consider the linguistic knowledge carried by the semantic prototypes, e.g., *dataset labels. To overcome these problems, we propose a cross-modality graph reasoning adaptation (CIGAR) method to take advantage of both visual and linguistic knowledge. Specifically, our method performs cross-modality graph reasoning between the linguistic modality graph and visual modality graphs to enhance their representations. We also propose a discriminative feature selector to find the most discriminative features and take them as the nodes of the visual graph for both efficiency and effectiveness. In addition, we employ the linguistic graph matching loss to regulate the update of linguistic graphs and maintain their semantic representation during the training process. Comprehensive experiments validate the effectiveness of our proposed CIGAR.*

1. Introduction

Object detection is a fundamental technique in computer vision tasks, and it has been widely explored in many applications, *e.g.,* self-driving and public safety. A variety of works [31,38,39,56,57] have achieved improvements in detection performance due to the development of deep neural networks. However, a detector significantly degrades if we deploy it in a novel domain due to the problem of domain shift. The domain shift can be induced by many factors, such as the variation of the capture condition from sunny to foggy weather, from virtual to real-world, and from one camera to another.

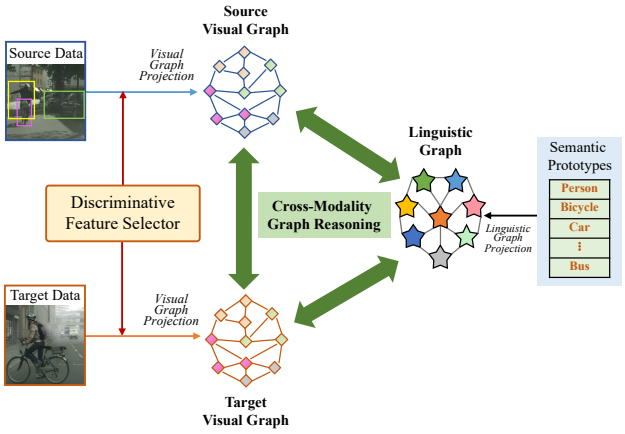


Figure 1. Illustration of the proposed Cross-modality Graph Reasoning Adaptation (CIGAR) framework.

To deal with the problem of domain shift [9], researchers have proposed many unsupervised domain adaptive object detection (UDA-OD) methods to bridge the domain gap between the source and target domains. Among them, the self-training based methods [5, 10, 26, 35] have shown excellent performances. However, they cannot be easily extended to real applications because they are computationally expensive and inefficient. Feature alignment-based methods [3,4,6,22,25] have also been extensively studied. These methods are structurally elegant and can be categorized into three groups, including *global-level* alignment, *instance-level* alignment, and *category-level* alignment. The *global-level* alignment methods [6,41] align the whole shallow feature maps produced by a backbone network. The *instance-level* alignment methods [4, 17] extract the feature maps for all the instances and learn to achieve cross-domain align-

* Corresponding authors.

ment in the deep feature space. The *category-level* alignment methods [58, 61] normally first use the detectors or an additional classification model for pseudo label assignment to the target samples, then align the category-wise instance features of two domains based on the ground-truth source labels and the pseudo target labels.

Numerous works [3, 24, 25] achieve feature alignment using graph-based approaches. These methods take dense features as nodes to construct the graphs and investigate the relationship between nodes. They use the knowledge carried by the graphs to enhance the features for the purpose of cross-domain alignment and thus improve the detection performance. Though the graph-based methods have significantly improved the feature alignment, they still have two inherent limitations. First, the existing methods do not construct the graph with the proper node set. Many graph-based methods randomly or uniformly [25] select features to construct graphs, resulting in the missing of some discriminative features. Some other methods [3] are not robust against noise and are computationally expensive, as they take all features as the graph nodes. Secondly, they only explore the visual knowledge extracted from the images. In this way, they ignore the critical knowledge of the linguistic modality, which carries the semantic prototypes of the domains, *e.g.,* linguistic dataset labels. Linguistic modality knowledge is very effective in regulating visual knowledge, and its absence severely reduces the representative ability of the resulting features. Some existing works [18, 63] have focused on using semantic category information to enhance performances in vision tasks. Singh et al. [46] also used a language model for the semi-supervised domain adaptive task and achieved improved performance.

To overcome the two limitations mentioned above, we propose a **C**ross-modal**I**ty **GrA**ph **R**easoning Adaptation (CIGAR) framework for category-level alignment via graph-based learning, as shown in Fig. 1. To enhance efficiency and improve the robustness against noise, we propose a Discriminative Feature Selector (DFS) for finding discriminative features and constructing the visual graph using only the discriminative features. In particular, we first conduct a procedure of singular value decomposition (SVD) and drop the small singular values, then evaluate the information richness of each feature via the summation of the absolute value of the elements. We can improve the representation ability of visual graphs by only taking these discriminative features as the nodes. Our method is more computationally efficient than previous methods, which use all image features to construct graphs. Our CIGAR also explores the graph in the linguistic modality and performs cross-modality graph reasoning between the linguistic modality and the visual modality. The linguistic modality knowledge can guide the mapping of visual modality knowledge from different domains to the same feature space. Our CIGAR can build a graph not only for the tasks with multiple categories and capture the relationship between different categories but also for the tasks with a single category and capture the relationship between different components of a single category. We maintain the semantic representation of the knowledge in linguistic modality and use it to guide the training procedure.

We summarize our contributions as follows:

- We propose a Cross-modality Graph Reasoning Adaptation (CIGAR) method for the domain adaptive object detection problem. To the best of our knowledge, this is the first work to tackle the UDA-OD task by graph reasoning across different modalities.
- We propose a Discriminative Feature Selector for finding discriminative image features and efficiently constructing the representative visual graph.
- Extensive experiments are conducted on four adaptation tasks, and our CIGAR achieves state-of-the-art performance, outperforming existing works by a large margin.

2. Related Works

2.1. Unsupervised Domain Adaptation

Unsupervised Domain adaptation methods aim to transfer knowledge from a labeled source domain to an unlabeled target domain. Many domain adaptation methods are widely investigated in computer vision tasks, *e.g.,* image classification [43, 51, 52], image clustering [29, 30], semantic segmentation [7], and object detection [6]. Inspired by Generative Adversarial Nets [14], DANN [12] proposes to learn the domain-invariant features by a Gradient Reversal Layer (GRL). ADDA [50] takes similar adversarial learning but uses two feature extractors to generate image features for the source domain and target domain. CDAN [32] performs the adversarial learning process within different categories. Bousmalis et al. [1] use an image-to-image translation method to generate source-like and target-like images to reduce the domain gap of image styles. Ma et al. [34] transform the images into graphs and use the graph convolution network to align their features in the feature space. Luo et al. [33] employ the bipartite graph method to perform domain interactions for the video adaptation problem. Kang et al. [21] formulate the domain adaptive semantic segmentation problem as a pixel-wise matching problem.

2.2. Domain Adaptive Object Detection

To deal with the domain shift problem in the object detection tasks, many UDA-OD methods [4, 17, 58] have been proposed. DA-Faster [6] uses the gradient reversal layer to perform adversarial learning for globally aligning features

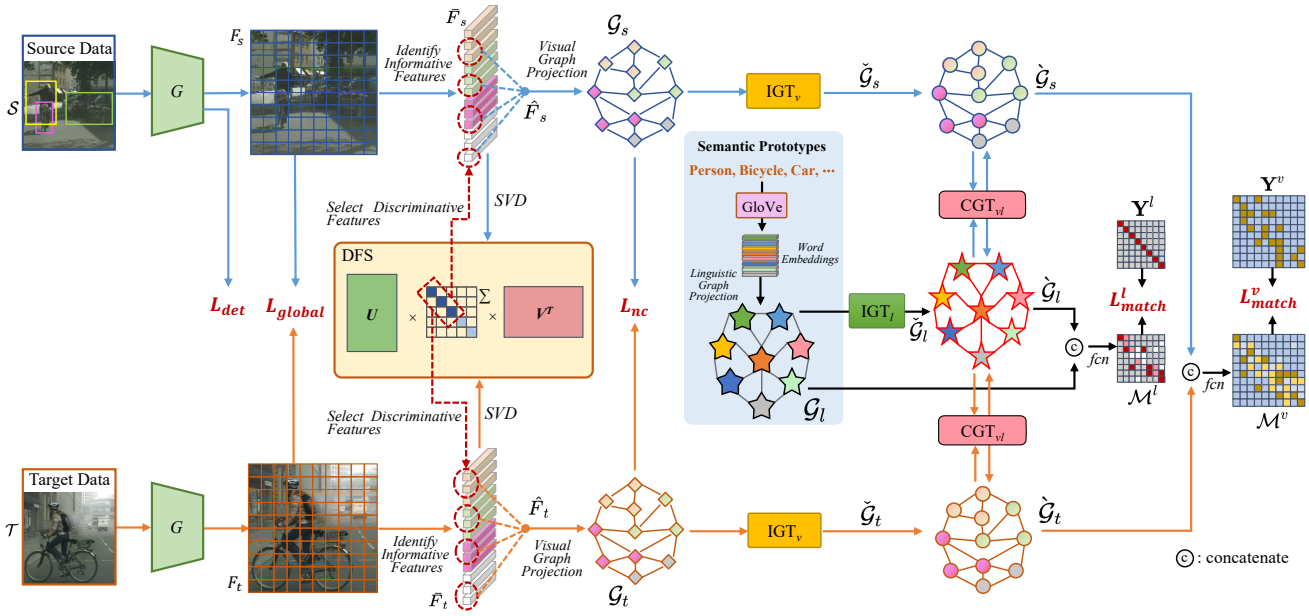


Figure 2. Overview of the proposed Cross-modality Graph Reasoning Adaptation (CIGAR) framework.

in different domains. SWDA [41] employs weak alignment on global shallow feature maps and strong alignment on local instance feature maps. Zheng et al. [61] highlight the importance of instance features and employ a coarse-to-fine method to perform adaptation. Wu et al. [54] propose a feature-decoupled method by vector decomposition, dividing the extracted features into domain-specific features and domain-invariant features. SCAN [24] uses graphs to represent the semantic relations of features and introduces the graph neural network to complete information between different domains. FGRR [3] uses all foreground instance features as nodes to construct graphs and perform intra- and inter-relational reasoning between source and target graphs in both pixel and semantic spaces. SIGMA [25] uniformly samples features within foreground features to construct category-wise graphs. They also use a memory bank to complete the graphs in each batch. Besides, SIGMA proposes to tackle adaptive detection as the graph matching problem. However, the existing graph-based methods do not build graphs with the proper node set, as they simply use all features or uniformly sample features to construct graphs. To deal with this issue, we propose a discriminative feature selector to mine discriminative features.

3. Proposed Method

Let $\mathcal{S} = \{(x_s, y_s)\}$ be the source domain, where x_s is a source image and y_s is the corresponding source label. Similarly, we denote the target domain as $\mathcal{T} = \{(x_t, y_t)\}$. Due to the domain shift, the data distributions of $\mathcal{S}$ and $\mathcal{T}$ are different. Note that the samples in both domains belong to $\{1, 2, \cdots, C\}$ categories. Given the labeled source data $\{(x_s, y_s)\}$ and the unlabeled target images $\{x_t\}$, a UDA-OD task aims to learn a detector and apply it to the target domain to predict the $\{y_t\}$.

Fig. 2 illustrates the structure of our proposed CIGAR. As the source domain and the target domain share the same semantic knowledge space, we can use the relationships between semantic prototypes to enhance the visual features (*i.e.,* the features of the images) in both domains. To be specific, we construct visual graphs based on the image features of both domains and use them to explore the knowledge in the visual modality. In addition, we extract the semantic prototypes to construct linguistic graphs and explore the knowledge of linguistic modality. We use both intra-graph and cross-graph knowledge reasoning to accomplish the interaction between the visual modality and the linguistic modality. To enhance the feature representation of the visual graph, we propose a new feature selection component, *i.e.,* Discriminative Feature Selector (DFS). This DFS can analyze the feature space via subspace learning and select the most representative features to construct visual graphs. To the best of our knowledge, our method is the first to tackle the UDA-OD task by the graph reasoning of the linguistic modality knowledge and the visual modality knowledge.

Our method is different from the previous methods in three points. First, we introduce the Singular Value Decomposition (SVD) in the DFS to select feature vectors with rich information. Second, we utilize the knowledge from two modalities for graph reasoning so that the visual and linguistic graphs can interact with each other and enhance

the resulting representations. Third, we propose the linguistic graph matching loss to regulate the update of linguistic graphs and maintain their semantic representation during the training process.

3.1. Visual Graph Construction

We use a shared backbone network (*i.e.,* G in Fig. 2) to extract features from the labeled source sample x_s and the unlabeled target sample x_t, *i.e.,* $F_{s/t} = G(x_{s/t}) \in \mathbb{R}^{h\times w\times c}$, where (h, w) and c are the size and channel dimension of a feature map. We construct a visual graph with the feature maps via three steps. The first step identifies the informative features associated with the foreground or background regions and denotes them as $\bar{F}_{s/t}$. The second step defines a module DFS to evaluate the information richness of each feature and selects the most discriminative features $\hat{F}_{s/t}$ from $\bar{F}_{s/t}$. The third step constructs visual graphs $\mathcal{G}_{s/t}$ based on the representative features in $\hat{F}_{s/t}$.

Informative Feature Identification. Not all features are equally important in a detection task. In this work, we select the informative features and use them as nodes to build our visual graphs. While the informative features are easily obtained in the labeled source domain, we cannot directly obtain them in the target domain, as the ground-truth bounding boxes in the target domain are unavailable. Following SIGMA [25], we first obtain the classification score map of a target image via a source detector, then consider a pixel to be the informative foreground feature of the j-category if its j-th classification score is higher than a threshold θ_{fg}, and the informative background feature if all of its classification scores are smaller than the threshold θ_{bg}. Here, we have $\theta_{bg} < \theta_{fg}$. The pixels associated with the classification score in $[\theta_{bg}, \theta_{fg}]$ are not trustful, and we do not determine whether they are foreground or background. In the source/target domain, we take the informative feature vectors belonging to the j-category to form a feature matrix $\bar{F}^j_{s/t} \in \mathbb{R}^{N^j_{s/t}\times c}$, where $N^j_{s/t}$ is the number of informative features associating with the j-category.

Discriminative Feature Selection. For efficiency and robustness, we expect to build our visual graph using only the most discriminative features instead of all the informative features. For this purpose, we design the Discriminative Feature Selector (DFS) to estimate the discriminative abilities of the features and use it to select the most discriminative features from the informative ones.

Inspired by HRank [28], we use singular value decomposition (SVD) to decouple $\bar{F}^j_{s/t}$. It can be formulated as follows:

$$\begin{aligned}\bar{F}^j_{s/t} &= \sum_{i=1}^{R} r_i \mathbf{u}_i \mathbf{v}_i^T = \sum_{i=1}^{L} r_i \mathbf{u}_i \mathbf{v}_i^T + \sum_{i=L+1}^{R} r_i \mathbf{u}_i \mathbf{v}_i^T \\ &= \mathbf{SVD}_{LS}(\bar{F}^j_{s/t}) + \mathbf{SVD}_{SS}(\bar{F}^j_{s/t}),\end{aligned} \tag{1}$$

where r_i is the i-th largest singular value, its left singular vector and right singular vector are u_i and v_i, respectively. R is the rank of the feature matrix $\bar{F}^j_{s/t}$ and $L \le R$. Thus, a feature matrix with R singular values can be disentangled into the matrix $\mathbf{SVD}_{LS}(\bar{F}^j_{s/t}) \in \mathbb{R}^{N^j_{s/t}\times c}$ containing rich information and the matrix $\mathbf{SVD}_{SS}(\bar{F}^j_{s/t}) \in \mathbb{R}^{N^j_{s/t}\times c}$ containing little information. we use $\mathcal{I}_m = \sum_{n=1}^{c} |\mathbf{SVD}_{LS}(\bar{F}^j_{s/t})|_{m,n}$ to denote the information richness of the m-th feature in $\bar{F}^j_{s/t}$. For each category $j \in \{0, 1, \cdots, C\}$, we select the top N_v^{class} feature vectors in terms of information richness from $\bar{F}^j_{s/t}$ and use them to build a feature matrix $\hat{F}^j_{s/t}$.

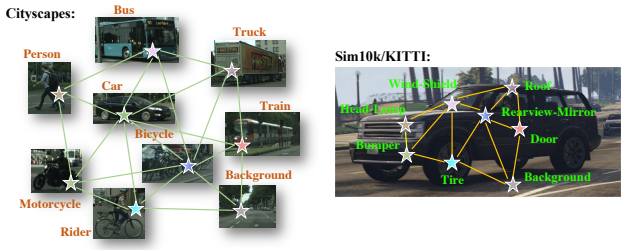
This new feature selection method has two advantages. First, we drop the noisy features and improve the representation ability. Second, we reduce the scale of graph structures and decrease the computational burden by only considering the most discriminative features.

Visual Graph Initialization. Let $\hat{F}_{s/t} = \{\hat{F}^0_{s/t}, \hat{F}^1_{s/t}, \cdots, \hat{F}^C_{s/t}\} \in \mathbb{R}^{N_v\times c}$ be the visual features of all categories, where $N_v = (N_v^{class} \times (C + 1))$ is the total number of all selected feature vectors. We use a fully connected network to learn the embeddings of $\hat{F}_{s/t}$ and take them as the visual graph node features $V_{s/t} \in \mathbb{R}^{N_v\times c_g}$, where c_g is the channel dimension for all graph nodes. We randomly initialized a learnable visual adjacency matrix $A_{s/t} \in \mathbb{R}^{N_v\times N_v}$ to represent edge relations between nodes within each domain and denote the visual graph of the source/target domain as $\mathcal{G}_{s/t} = \{V_{s/t}, A_{s/t}\}$.

3.2. Linguistic Graph Construction

We build a linguistic graph $\mathcal{G}_l$ based on the word embeddings to describe the relationship between different semantic prototypes. In a detection task with multiple categories, we take each category as a semantic prototype and use the linguistic graph to capture the cross-category relationship. To build the linguistic graph in the detection task with one single category, we take each component of the object as a semantic prototype and use the linguistic graph to capture the intra-category relationship.

Multi-category Linguistic Graph. If the detection task has $C(C > 1)$ categories, we naturally use the category labels as the semantic prototypes and feed them into GloVe [37] to generate the foreground word embeddings $F_l^{fg} \in \mathbb{R}^{C\times c_l}$, where c_l is the channel dimension of each word embedding. Besides the C foreground categories, we also use a randomly initialized embedding $F_l^{bg} \in \mathbb{R}^{1\times c_l}$ to represent the knowledge of the background. We use a fully connected network to transform both F_l^{fg} and F_l^{bg} into linguistic graph nodes $V_l \in \mathbb{R}^{N_l\times c_g}$, where $N_l = C+1$ is the number of embeddings. We randomly initialize the knowledge adjacency matrix $A_l \in \mathbb{R}^{N_l\times N_l}$, which represents the de-



(a) Multi-Category Domain Linguistic Graph.

(b) Single-Category Domain Linguistic Graph.

Figure 3. The linguistic graphs of the multi-category domain (Cityscapes) and the single-category domain (Sim10k/KITTI).

pendencies between different categories. As shown in Fig. 3 (a), we define the linguistic graph as $\mathcal{G}_l = \{V_l, A_l\}$.

Single-Category Linguistic Graph. Some detection tasks only have a single category. For example, the benchmark datasets *Sim10k* [36] and *KITTI* [13] only have one foreground category, *i.e., Car*. In this case, it is inappropriate to take the category label as the semantic prototype to build the linguistic graph. To mine the linguistic knowledge inside the category, we take C_{fine} components inside the category as semantic prototypes. For example, we use seven semantic prototypes to describe a car, including *Roof*, *Tire*, *Rearview mirror*, *Wind shield*, *Head lamp*, *Bumper*, and *Door*. With these semantic prototypes and a random initialized background feature, we can obtain the graph embeddings $V_l \in \mathbb{R}^{N_l \times c_g}$, where $N_l = C_{fine} + 1$ is the number of embeddings, *i.e.,* fine-grained semantic prototypes. We also use a randomly initialized matrix $A_l \in \mathbb{R}^{N_l \times N_l}$ as the graph adjacency matrix. As shown in Fig. 3 (b), we can also define a single-category linguistic graph $\mathcal{G}_l = \{V_l, A_l\}$ to capture the relationship between different components.

3.3. Cross-Modality Graph Reasoning

We formulate the feature alignment task into a graph matching problem following SIGMA [25]. Our learning procedure involves not only the optimization of each modality graph independently but also the interaction reasoning between the visual modality graphs and the linguistic modality graph. In this section, we propose a linguistic knowledge matching loss to regulate the updating of the linguistic graph to maintain its semantic representation during the training process.

Intra-graph Reasoning. We introduce the visual/linguistic intra-graph transformer ($\text{IGT}_{v/l}$) to perform graph reasoning within each graph. First, we use a graph convolution network (GCN) to aggregate information about each node from its neighborhoods. The GCN process is formulated as follows:

$$\tilde{V}_{s/t/l} = \text{Relu}((A_{s/t/l} + I) \cdot V_{s/t/l} \cdot W_{v/l}), \tag{2}$$

where $\tilde{V}_{s/t/l}$ is the enhanced features, I is an identity matrix, $W_v \in \mathbb{R}^{N_v \times N_v}$ is the learnable parameters for processing $\mathcal{G}_{s/t}$, and $W_l \in \mathbb{R}^{N_l \times N_l}$ is the learnable parameters for processing $\mathcal{G}_l$. Then, we feed the enhanced graph $\tilde{\mathcal{G}}_{s/t/l} = \{\tilde{V}_{s/t/l}, A_{s/t/l}\}$ into a self-attention based transformer to encode their node embeddings. The transformation process can be formulated as follows:

$$\begin{aligned} &Q_T = \tilde{V}_{s/t/l} \cdot \check{W}^q_{v/l},\ K_T = \tilde{V}_{s/t/l} \cdot \check{W}^k_{v/l},\ V_T = \tilde{V}_{s/t/l} \cdot \check{W}^v_{v/l}, \\ &\check{V}_{s/t/l} = \text{Softmax}(Q_T K_T^T) \cdot V_T + \tilde{V}_{s/t/l}, \end{aligned} \tag{3}$$

where $\check{V}_{s/t/l}$ is the enhanced graph nodes after intra-graph reasoning. $\check{W}^q_{v/l}$, $\check{W}^k_{v/l}$, and $\check{W}^v_{v/l}$ are the learnable parameters in $\text{IGT}_{v/l}$ for projecting embeddings. The obtained graphs can be formulated as $\check{\mathcal{G}}_{s/t/l} = \{\check{V}_{s/t/l}, A_{s/t/l}\}$.

Cross-graph Reasoning. We introduce the visual-linguistic cross-graph transformer (CGT_{vl}) to achieve the cross-modality reasoning between the two different types of graphs. In this way, we can use the linguistic knowledge extracted from semantic prototypes to improve the representations of the visual graphs. In the CGT_{vl}, we feed the $\check{V}_{s/t/l}$ into a cross-attention based transformer to perform reasoning between visual and linguistic graphs. Inspired by SIGMA [25] and GINet [55], The cross-attention operation is formulated as follows:

$$\begin{aligned} &\grave{V}_{s/t} = \text{Softmax}[(\check{V}_{s/t} \cdot \grave{W}^q)(\check{V}_l \cdot \grave{W}^k)^T] \cdot (\check{V}_l \cdot \grave{W}^v) + \check{V}_{s/t}, \\ &\grave{V}_l = \text{Softmax}[(\check{V}_l \cdot \grave{W}^q)(\check{V}_{s/t} \cdot \grave{W}^k)^T] \cdot (\check{V}_{s/t} \cdot \grave{W}^v) + \check{V}_l, \end{aligned} \tag{4}$$

where $\grave{V}_{s/t/l}$ are the enhanced graph nodes after cross-graph reasoning. $\grave{W}^q$, $\grave{W}^k$, and $\grave{W}^v$ are learnable parameters of CGT_{vl}. We use $\grave{\mathcal{G}}_{s/t/l} = \{\grave{V}_{s/t/l}, A_{s/t/l}\}$ to denote the visual and linguistic graphs after cross-graph reasoning.

Visual Graph Matching Loss. To enhance the alignment of the visual features at the category level, we follow SIGMA [25] to formulate the domain adaptation into a graph matching problem between $\grave{\mathcal{G}}_s$ and $\grave{\mathcal{G}}_t$. Given the graph nodes $\grave{V}_{s/t}$ enhanced by cross-graph reasoning, we learn the affinity matrix $\mathcal{M}^v \in \mathbb{R}^{N_v \times N_v}$ to denote the alignment between every pair of nodes from two different domains. The element $\mathcal{M}^v_{i,j}$ represents the affinity representation between the i-th node in $\grave{V}_s$ and the j-th node in $\grave{V}_t$. Here, we derive $\mathcal{M}^v_{i,j}$ via a fully connected network, which receives the concatenation of the features of two involving nodes. The positive element in $\mathcal{M}^v$ indicates that the corresponding two nodes from different domains are matched, and vice versa. The matching loss is detailed as follows:

$$\begin{aligned} L^v_{match} = &\sum_i \frac{1}{N_v} [\max_j (\mathcal{M}^v \odot \mathbf{Y}^v)_{i,j} - 1]^2 \\ &+ \sum_{i,j} \frac{1}{|1 - \mathbf{Y}^v|} [\mathcal{M}^v \odot (1 - \mathbf{Y}^v)]^2_{i,j}, \end{aligned} \tag{5}$$

where $\mathbf{Y}^v$ is the matching label matrix and $\mathbf{Y}^v_{i,j} = 1$ if and only if the i-th node in $\grave{\mathcal{G}}_s$ and the j-th node in $\grave{\mathcal{G}}_t$ have

Table 1. Results on Cityscapes→Foggy Cityscapes. 'Bb' and 'prsn' denote 'Backbone' and 'person'.

Method	Bb	prsn	rider	car	truck	bus	train	motor	bike	mAP
CFFA [61]	VGG-16	34.0	46.9	52.1	30.8	43.2	29.9	34.7	37.4	38.6
EPM [17]		41.9	38.7	56.7	22.6	41.5	26.8	24.6	35.5	36.0
RPNPA [60]		33.6	43.8	49.6	32.9	45.5	46.0	35.7	36.8	40.5
UMT [10]		33.0	46.7	48.6	**34.1**	**56.5**	46.8	30.4	37.4	41.7
MeGA [62]		37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8
VDD [54]		33.4	44.0	51.7	33.9	52.0	34.7	34.2	36.8	40.0
KTNet [48]		46.4	43.2	60.6	25.8	41.2	40.4	30.7	38.8	40.9
SSAL [35]		45.1	47.4	59.4	24.5	50.0	25.7	26.0	38.7	39.6
D-adapt [20]		44.9	**54.2**	61.7	25.6	36.3	24.7	**37.3**	**46.1**	41.3
SCAN [24]		41.7	43.9	57.3	28.7	48.6	48.7	31.0	37.3	42.1
FGRR [3]		34.4	47.6	51.3	30.0	46.8	42.3	35.1	38.9	40.8
SIGMA [25]		**46.9**	48.4	**63.7**	27.1	50.7	35.9	34.7	41.4	43.5
CIGAR(ours)		45.3	45.3	61.6	32.1	50.0	**51.0**	31.9	40.4	**44.7**
GPA [59]	ResNet-50	32.9	46.7	54.1	24.7	45.7	41.1	32.4	38.7	39.5
EPM [17]		39.9	38.1	57.3	28.7	50.7	37.2	30.2	34.2	39.5
DIDN [27]		38.3	44.4	51.8	28.7	53.3	34.7	32.4	40.4	40.5
DSS [53]		42.9	**51.2**	53.6	**33.6**	49.2	18.9	**36.2**	**41.8**	40.9
SDA [40]		38.8	45.9	57.2	29.9	50.2	**51.9**	31.9	40.9	43.3
SIGMA [25]		44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2
CIGAR(ours)		**46.1**	47.3	**62.1**	27.8	**56.6**	44.3	33.7	41.3	**44.9**

the same category, otherwise 0. The L^v_{match} promotes the model to match pair-wise nodes correctly, thus aligning visual features at the category level.

Linguistic Graph Matching Loss. To maintain the semantic representation of the linguistic graph during the training process, we propose the linguistic graph matching loss to regulate the update from $\mathcal{G}_l$ to $\grave{\mathcal{G}}_l$. We concatenate the features of the i-th node in V_l and the j-th node in $\grave{V}_l$ and feed the concatenation into a fully connected network to generate the linguistic affinity matrix $\mathcal{M}^l$. The linguistic graph matching loss is formulated as follows:

$$L^l_{match} = \sum_i \frac{1}{N_l}[\max_j(\mathcal{M}^l \odot \mathbf{Y}^l)_{i,j} - 1]^2 + \sum_{i,j} \frac{1}{|1-\mathbf{Y}^l|}[\mathcal{M}^l \odot (1-\mathbf{Y}^l)]^2_{i,j}, \tag{6}$$

where $\mathbf{Y}^l$ is the matching label matrix whose definition is similar to $\mathbf{Y}^v$. By minimizing L^l_{match}, we match the pair-wise nodes in $\mathcal{G}_l$ to $\grave{\mathcal{G}}_l$ and avoid semantic bias when updating the linguistic graph in the training procedure.

The overall training loss is as follows:

$$L_{total} = L_{det} + L_{global} + L_{nc} + \lambda_v L^v_{match} + \lambda_l L^l_{match}, \tag{7}$$

where L_{det} is the detection loss of FCOS [49], L_{global} is the global alignment loss [17], L_{nc} is the cross-entropy classification loss of each node for visual graphs, L^v_{match} is the visual graph matching loss, L^l_{match} is the linguistic graph matching loss, λ_v and λ_l are the weights of visual and linguistic graph matching losses.

4. Experiments

4.1. Datasets and Evaluation

To verify our proposed method, we conduct extensive experiments on four different types of domain shifts. We use FCOS as our baseline detector and the mean Average Precision (mAP) with a threshold of 0.5 to evaluate the detection performance on target domains. In addition, we use GAIN to calculate the improvement of the method in comparison with the models trained using only the source samples. Six benchmark datasets are used: Cityscapes [8], Foggy Cityscapes [42], Pascal VOC [11], Clipart [19], Sim10k [36], and KITTI [13].

Cityscapes→Foggy Cityscapes. These two datasets involve the adaption from sunny to foggy weather. Cityscapes is a city landscape dataset captured in clear weather. It contains 2975 training images and 500 validation images with bounding box annotations. Foggy Cityscapes is a synthesized dataset generated from Cityscapes.

Pascal VOC→Clipart. Pascal VOC is a real-world dataset consisting of twenty categories of objects. We use its 2007 and 2012 versions, a total of 16551 samples, as training images. Clipart contains 1k images with the same twenty categories as Pascal VOC.

Sim10k→Cityscapes. These two datasets involve the adaption from synthetic to real-world. Sim10k is rendered from the gaming engine of Grand Theft Auto. It contains 10k images and 58701 objects with bounding box annotations of the car category.

KITTI→Cityscapes. These two datasets involve the adaptation of different capture cameras. KITTI is a city street scene dataset from vehicle-mounted cameras. It consists of 7481 images of the car category.

4.2. Implementation Details

We employ VGG-16 [45], ResNet-50 [15], and ResNet-101 [15] as backbones to extract image features. We adopt the Stochastic Gradient Descent(SGD) [47] optimizer with a learning rate of 0.002 and a batch size of 2 for 50k iterations to train our models. We set θ_{fg} and θ_{bg} to be 0.5 and 0.05. Both λ_v and λ_l in the total loss function are set to 0.1. We set L to be one-half of rank R in SVD. Besides, we set N_v^{class} to be 40, *i.e.,* we sample up to 40 graph nodes in each category. If the number of all category nodes is less than $40 \times (C+1)$ in a batch, we use the memory bank in SIGMA to complete the visual graphs. We set both the dimensions of visual and linguistic node embeddings as 256. To obtain the word embeddings of semantic prototypes, we use the GloVe word embedding model to process the linguistic words. The original word embedding dimension c_l is set to 300. All experiments are performed with NVIDIA 3090 GPUs.

Table 2. Comparison results on Pascal VOC→Clipart with the ResNet-101 backbone.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	prsn	plant	sheep	sofa	train	tv	mAP	SO/GAIN
SWDA [41]	26.2	48.5	32.6	33.7	38.5	54.3	37.1	18.6	34.8	58.3	12.5	12.5	33.8	65.5	54.5	52.0	9.3	24.9	54.1	49.1	38.1	27.8/ 6.9
CR [58]	28.7	55.3	31.8	26.0	40.1	**63.6**	36.6	9.4	38.7	49.3	17.6	14.1	33.3	74.3	61.3	46.3	22.3	24.3	49.1	44.3	38.3	27.8/ 11.3
ATF [16]	41.9	67.0	27.4	36.4	41.0	48.5	42.0	13.1	39.2	75.1	33.4	7.9	41.2	56.2	61.4	50.6	**42.0**	25.0	53.1	39.1	42.1	14.3/ **27.8**
SCL [44]	**44.7**	50.0	33.6	27.4	42.2	55.6	38.3	19.2	37.9	**69.0**	30.1	26.3	34.4	67.3	61.0	47.9	21.4	26.3	50.1	47.3	41.5	27.8/ 10.3
HTCN [4]	33.6	58.9	34.0	23.4	45.6	57.0	39.8	12.0	39.7	51.3	20.1	20.1	39.1	72.8	61.3	43.1	19.3	30.1	50.2	51.8	40.3	27.8/ 12.5
SAPN [23]	27.4	**70.8**	32.0	27.9	42.4	63.5	47.5	14.3	48.2	46.1	31.8	17.9	43.8	68.0	68.1	49.0	18.7	20.4	55.8	51.3	42.2	27.8/ 14.4
UMT [10]	39.6	59.1	32.4	**35.0**	45.1	61.9	**48.4**	7.5	46.0	67.6	21.4	**29.5**	48.2	75.9	**70.5**	**56.7**	25.9	28.9	39.4	43.6	44.1	27.8/ 16.3
DBGL [2]	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	27.8/ 13.8
FGRR [3]	30.8	52.1	35.1	32.4	42.2	62.8	42.6	21.4	42.8	58.6	**33.5**	20.8	37.2	**81.4**	66.2	50.3	21.5	29.3	**58.2**	47.0	43.3	27.8/ 15.5
IIDA [27]	41.5	52.7	34.5	28.1	43.7	58.5	41.8	15.3	40.1	54.4	26.7	28.5	37.7	75.4	63.7	48.7	16.5	30.8	54.5	48.7	42.1	27.8/ 14.3
CIGAR(ours)	35.2	55.0	**39.2**	30.7	**60.1**	58.1	46.9	**31.8**	**47.0**	61.0	21.8	26.7	**44.6**	52.4	68.5	54.4	31.3	**38.8**	56.5	**63.5**	**46.2**	25.3/ 20.9

Table 3. Comparison on Sim10k→Cityscapes (S→C) and KITTI→Cityscapes (K→C) with VGG-16 as backbone. GAIN indicates the detection gains compared with Source-only (SO) trained models.

Method	S→C	SO/GAIN	K→C	SO/GAIN
EPM [17]$_{ECCV'20}$	49.0	39.8/ 9.2	43.2	34.4/ 8.8
DSS [53]$_{CVPR'21}$	44.5	34.7/ 9.8	42.7	34.6/ 8.1
MEGA [62]$_{CVPR'21}$	44.8	34.3/ 10.5	43.0	30.2/ 12.8
RPNPA [60]$_{CVPR'21}$	45.7	34.6/ 11.1	-	-
UMT [10]$_{CVPR'21}$	43.1	34.3/ 8.8	-	-
KTNet [48]$_{ICCV'21}$	50.7	39.8/ 10.9	45.6	34.4/ 11.2
SSAL [35]$_{NeurIPS'21}$	51.8	38.0/ 13.8	45.6	34.9/ 10.7
D-adapt [20]$_{ICLR'22}$	50.3	34.6/15.7	-	-
FGRR [3]$_{TPAMI'22}$	44.5	34.6/ 9.9	-	-
SCAN [24]$_{AAAI'22}$	52.6	39.8/ 12.8	45.8	34.4/ 11.4
SIGMA [25]$_{CVPR'22}$	53.7	39.8/ 13.9	45.8	34.4/ 11.4
CIGAR(ours)	**58.5**	39.8/ **18.7**	**48.5**	34.4/ **14.1**

Table 4. Results on Sim10k→Cityscapes with different methods to select discriminative features when constructing visual graphs. LSV denotes our proposed large singular value-aware method. We compare two methods: random selection and uniform selection.

Selection Method	$mAP_{0.5}$	$mAP_{0.75}$	$mAP_{0.5:0.95}$
random	56.3	32.4	31.9
uniform	55.6	34.5	33.9
LSV(ours)	**58.5**	**35.7**	**35.2**

4.3. Comparison with State-Of-The-Arts

Cityscapes→Foggy Cityscapes. As shown in Tab. 1, we present the performances of several benchmark detectors using ResNet-50 and VGG-16 as backbones. CIGAR achieves 44.7 and 44.9 mAP, outperforming the existing methods. Compared with other latest graph-based alignment methods with VGG-16 backbone, e.g., SCAN [24] and FGRR [3], CIGAR achieves 2.6 and 3.9 mAP improvements. We compare CIGAR with D-adapt, the latest self-training method, and achieve 3.4 mAP improvement.

Pascal VOC→Clipart. Tab. 2 lists the results of several methods. Our CIGAR achieves a 46.2 mAP performance which outperforms other methods. In comparison with the graph-based FGRR [3] and decoupling IIDA [27], our CIGAR surpasses them by 2.9 and 4.1 mAP increments.

Sim10k→Cityscapes. Tab. 3 (left) lists the comparison results. Our CIGAR achieves a 58.5 mAP, surpassing the existing methods by a large margin. In comparison with FGRR [3], SCAN [24] and SIGMA [25], CIGAR improves the mAP by a margin 14, 5.9 and 4.8, respectively. Our method also surpasses the self-training D-adapt by 8.2 mAP, showing its advantage over existing adaptation approaches.

KITTI→Cityscapes. Tab. 3 (right) shows the experiment comparison. CIGAR achieves a 48.5 mAP, and the value of GAIN can be as large as 14.1 in terms of mAP, surpassing the existing best method by a competitive improvement. Compared with another graph-based method SIGMA [25], CIGAR achieves 2.7 mAP improvement.

4.4. Ablation Studies

We present detailed ablation studies in this section to show the effectiveness of our CIGAR. All ablation experiments are conducted on two tasks (Sim10→Cityscapes and Cityscapes→Foggy Cityscapes) with VGG-16 backbone.

Ablation on the discriminative feature selection. Tab. 4 shows the effectiveness of the discriminative feature selector. Compared with randomly and uniformly sample discriminative features, our method gives 2.2 and 2.9 mAP improvements. These results demonstrate that the features corresponding to large singular values are discriminative and can represent the fine-grained feature of objects.

Ablation on the cross-modality graph reasoning. Tab. 5 shows the impact of different graph transformers. The IGT_v, IGT_l, and CGT_{vl} all can enhance the detection performance of the baseline method. When all kinds of modules are used to perform knowledge reasoning intra-graph and cross-graphs, our CIGAR achieves the best performance, showing that all components are helpful in improving the representations of visual and linguistic graphs.

Ablation on the linguistic graph. As shown in Tab. 6, we investigate the influences of the linguistic graph and linguis-

Figure 4. Qualitative comparison results on the Cityscapes→Foggy Cityscapes task between EPM [17] and our proposed CIGAR.

Table 5. Comparison results on the cross-modality graph reasoning. C→F and S→C indicate Cityscape→Foggy Cityscapes and Sim10K→Cityscapes adaptation tasks with VGG-16 backbone.

IGT_v	IGT_l	CGT_{vl}	S→C	C→F
-	-	-	49.9	42.8
✔	-	-	51.4	43.0
-	✔	-	49.8	42.2
-	-	✔	51.3	43.3
✔	-	✔	53.6	44.2
-	✔	✔	54.1	44.5
✔	✔	-	56.6	44.4
✔	✔	✔	**58.5**	**44.7**

Table 6. Comparison results on the linguistic graph reasoning and linguistic graph loss. The baseline method only uses visual graphs. LG denotes linguistic graphs used in the method. E-dimension is the dimension of linguistic graph node embeddings. LG-loss denotes the linguistic graph matching loss used in the method.

method	LG	E-dimension	LG-loss	mAP
baseline	-	-	-	54.5
CIGAR	✔	100	-	57.6
CIGAR	✔	300	-	57.2
CIGAR	✔	100	✔	58.3
CIGAR(ours)	✔	300	✔	**58.5**

tic graph matching loss. By introducing linguistic graph reasoning to the baseline method, we can improve the detection performance. We also conduct experiments with different dimensions of graph node embeddings. The results show that the dimension has no significant influence on the detection performance. Introducing the matching loss over linguistic graphs improves the mAP by a margin of 1.3.

4.5. Qualitative Results

Result comparison. Fig. 4 presents some qualitative comparison results in the adaptation task from Cityscapes to Foggy Cityscapes. Our CIGAR can detect car and bicycle objects with fewer missing errors, showing its advantage of dense detection ability.

Figure 5. The selected discriminative features of *person*, *bicycle*, and *car* category objects with our proposed discriminative feature selector (DFS). The red circles denote the selected features.

Discriminative feature selection. Fig. 5 shows the selected discriminative features of three categories. These selected features, *i.e.,* red circles, are sparse and correspond to different characteristic components of a category of objects, and the graph built by these features is representative.

5. Conclusion

We propose a Cross-modality Graph Reasoning Adaptation (CIGAR) method for domain adaptive object detection. We use graphs to represent the visual modality knowledge extracted from image features and solve the domain adaptive object detection task as the graph matching problem. To select the most discriminative features for constructing visual graphs, we propose a Discriminative Feature Selector to estimate the information richness of each feature via subspace learning. Besides, we perform cross-modality graph reasoning between the linguistic graph and visual graphs to enhance their representations. In addition, we employ the linguistic graph matching loss to regulate the update of linguistic graphs and maintain their semantic representations during the training process. Comprehensive experiments indicate the effectiveness of our CIGAR.

Acknowledgement. The authors wish to acknowledge the financial support from: (i) Natural Science Foundation China (NSFC) under Grant no. 62172285; (ii) Guangdong Basic and Applied Basic Research Foundation under Grant no. 2023A1515012110; and (iii) Shenzhen Science and Technology Program under Grant no. JCYJ20220818103000001.

References

[1] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, pages 3722–3731, 2017. 2

[2] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual Bipartite Graph Learning: A General Approach for Domain Adaptive Object Detection. In *ICCV*, pages 2683–2692, 2021. 7

[3] Chaoqi Chen, Jiongcheng Li, Hong-Yu Zhou, Xiaoguang Han, Yue Huang, Xinghao Ding, and Yizhou Yu. Relation Matters: Foreground-aware Graph-based Relational Reasoning for Domain Adaptive Object Detection. *TPAMI*, pages 1–1, 2022. 1, 2, 3, 6, 7

[4] 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, 7

[5] Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, et al. Learning domain adaptive object detection with probabilistic teacher. *arXiv preprint arXiv:2206.06293*, 2022. 1

[6] 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

[7] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *ICCV*, pages 1992–2001, 2017. 2

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 6

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1

[10] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, 2021. 1, 6, 7

[11] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111:98–136, 2015. 6

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189. PMLR, 2015. 2

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, pages 3354–3361, June 2012. 5, 6

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[16] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In *ECCV*, pages 309–324. Springer, 2020. 7

[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*, pages 733–748. Springer, 2020. 1, 2, 6, 7, 8

[18] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *ECCV*, pages 59–75. Springer, 2020. 2

[19] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation. In *CVPR*. arXiv, Mar. 2018. 6

[20] Junguang Jiang, Baixu Chen, Jianmin Wang, and Mingsheng Long. Decoupled Adaptation for Cross-Domain Object Detection. In *ICLR*, May 2022. 6, 7

[21] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. 33:3569–3580, 2020. 2

[22] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *ICCV*, pages 480–490, 2019. 1

[23] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *ECCV*, pages 481–497. Springer, 2020. 7

[24] Wuyang Li, Xinyu Liu, Xiwen Yao, and Yixuan Yuan. SCAN: Cross Domain Object Detection with Semantic Conditioned Adaptation. In *AAAI*, volume 36, pages 1421–1428, June 2022. 2, 3, 6, 7

[25] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, pages 5291–5300, 2022. 1, 2, 3, 4, 5, 6, 7

[26] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *CVPR*, pages 7581–7590, 2022. 1

[27] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *ICCV*, pages 8771–8780, 2021. 6, 7

[28] Mingbao Lin, Rongrong Ji, Yan Wang, Yichen Zhang, Baochang Zhang, Yonghong Tian, and Ling Shao. Hrank: Filter pruning using high-rank feature map. In *CVPR*, pages 1529–1538, 2020. 4

[29] Chengliang Liu, Jie Wen, Xiaoling Luo, and Yong Xu. Incomplete multi-view multi-label learning via label-guided masked view- and category-aware transformers. 2023. 2

[30] Chengliang Liu, Zhihao Wu, Jie Wen, Yong Xu, and Chao Huang. Localized sparse incomplete multi-view clustering. *TMM*, 2022. 2

[31] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. In *ECCV*, volume 9905, pages 21–37, 2016. 1

[32] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional Adversarial Domain Adaptation. In *NeurIPS*. arXiv, Dec. 2018. 2

[33] Yadan Luo, Zi Huang, Zijian Wang, Zheng Zhang, and Mahsa Baktashmotlagh. Adversarial Bipartite Graph Learning for Video Domain Adaptation. In *ACMMM*, pages 19–27, Oct. 2020. 2

[34] Xinhong Ma, Tianzhu Zhang, and Changsheng Xu. Gcan: Graph convolutional adversarial network for unsupervised domain adaptation. In *CVPR*, pages 8266–8276, 2019. 2

[35] Muhammad Akhtar Munir, Muhammad Haris Khan, M Sarfraz, and Mohsen Ali. Ssal: Synergizing between self-training and adversarial learning for domain adaptive object detection. 34:22770–22782, 2021. 1, 6, 7

[36] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 5, 6

[37] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 4

[38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 1

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*, volume 28. Curran Associates, Inc., 2015. 1

[40] Farzaneh Rezaeianaran, Rakshith Shetty, Rahaf Aljundi, Daniel Olmeda Reino, Shanshan Zhang, and Bernt Schiele. Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection. In *ICCV*, pages 9204–9213, 2021. 6

[41] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 1, 3, 7

[42] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic Foggy Scene Understanding with Synthetic Data. *IJCV*, 126(9):973–992, Sept. 2018. 6

[43] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning Transferrable Representations for Unsupervised Domain Adaptation. In *NeurIPS*, volume 29. Curran Associates, Inc., 2016. 2

[44] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. SCL: Towards Accurate Domain Adaptive Object Detection via Gradient Detach Based Stacked Complementary Losses. *arXiv preprint arXiv:1911.02559*, Dec. 2019. 7

[45] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*. arXiv, Apr. 2015. 6

[46] Anurag Singh, Naren Doraiswamy, Sawa Takamuku, Megh Bhalerao, Titir Dutta, Soma Biswas, Aditya Chepuri, Balasubramanian Vengatesan, and Naotake Natori. Improving semi-supervised domain adaptation using effective target selection and semantics. In *CVPR*, pages 2709–2718, 2021. 2

[47] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *GlobalSIP*, pages 245–248. IEEE, 2013. 6

[48] Kun Tian, Chenghao Zhang, Ying Wang, Shiming Xiang, and Chunhong Pan. Knowledge mining and transferring for domain adaptive object detection. In *ICCV*, pages 9133–9142, 2021. 6, 7

[49] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019. 6

[50] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 7167–7176, 2017. 2

[51] Jinghua Wang, Ming-Ming Cheng, and Jianmin Jiang. Domain shift preservation for zero-shot domain adaptation. *TIP*, 30:5505–5517, 2021. 2

[52] Jinghua Wang and Jianmin Jiang. Learning across tasks for zero-shot domain adaptation from a single source domain. *TPAMI*, 44(10):6264–6279, 2022. 2

[53] Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, YangYang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection. In *CVPR*, pages 9603–9612, 2021. 6, 7

[54] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. Vector-decomposed disentanglement for domain-invariant object detection. In *ICCV*, pages 9342–9351, 2021. 3, 6

[55] Tianyi Wu, Yu Lu, Yu Zhu, Chuang Zhang, Ming Wu, Zhanyu Ma, and Guodong Guo. Ginet: Graph interaction network for scene parsing. In *ECCV*, pages 34–51. Springer, 2020. 5

[56] Zhihao Wu, Chengliang Liu, Chao Huang, Jie Wen, and Yong Xu. Deep object detection with example attribute based prediction modulation. In *ICASSP*, pages 2020–2024, 2022. 1

[57] Zhihao Wu, Chengliang Liu, Jie Wen, Yong Xu, Jian Yang, and Xuelong Li. Selecting high-quality proposals for weakly supervised object detection with bottom-up aggregated attention and phase-aware loss. *TIP*, 2022. 1

[58] 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, 7

[59] 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. 6

[60] Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *CVPR*, pages 12425–12434, 2021. 6, 7

[61] 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, 6

[62] Wenzhang Zhou, Dawei Du, Libo Zhang, Tiejian Luo, and Yanjun Wu. Multi-Granularity Alignment Domain Adaptation for Object Detection. In *CVPR*. arXiv, Mar. 2022. 6, 7

[63] Chenchen Zhu, Fangyi Chen, Uzair Ahmed, Zhiqiang Shen, and Marios Savvides. Semantic relation reasoning for shot-stable few-shot object detection. In *CVPR*, pages 8782–8791, 2021. 2