SCAN: Cross Domain Object Detection with Semantic Conditioned Adaptation

Wuyang Li[1], **Xinyu Liu**[1], **Xiwen Yao**[2], **Yixuan Yuan**[1*]
[1] City University of Hong Kong
[2] Northwestern Polytechnical University
{wuyangli2, xliu423}-c@my.cityu.edu.hk, yaoxiwen517@gmail.com, yxyuan.ee@cityu.edu.hk

Abstract

The domain gap severely limits the transferability and scalability of object detectors trained in a specific domain when applied to a novel one. Most existing works bridge the domain gap by minimizing the domain discrepancy in the category space and aligning category-agnostic global features. Though great success, these methods model domain discrepancy with prototypes within a batch, yielding a biased estimation of domain-level distribution. Besides, the category-agnostic alignment leads to the disagreement of class-specific distributions in the two domains, further causing inevitable classification errors. To overcome these two challenges, we propose a novel Semantic Conditioned AdaptatioN (SCAN) framework such that well-modeled unbiased semantics can support semantic conditioned adaptation for precise domain adaptive object detection. Specifically, class-specific semantics crossing different images in the source domain are graphically aggregated as the input to learn an unbiased semantic paradigm incrementally. The paradigm is then sent to a lightweight manifestation module to obtain conditional kernels to serve as the role of extracting semantics from the target domain for better adaptation. Subsequently, conditional kernels are integrated into global alignment to support the class-specific adaptation in a well-designed Conditional Kernel guided Alignment (CKA) module. Meanwhile, rich knowledge of the unbiased paradigm is transferred to the target domain with a novel Graph-based Semantic Transfer (GST) mechanism, yielding the adaptation in the category-based feature space. Comprehensive experiments conducted on three adaptation benchmarks demonstrate that SCAN outperforms existing works by a large margin.

Introduction

Object detection (Ren et al. 2015; Lin et al. 2017; Tian et al. 2019) aims to recognize and localize object instances of predefined categories, which plays a critical part in several applications like self-driving and video analysis, etc. While these approaches have achieved remarkable performance when trained and tested in a specific domain, they will suffer severe performance degradation if evaluated in a novel domain due to the domain gap (Chen et al. 2018).

To address this challenge, a variety of studies (Chen et al. 2018; Xu et al. 2020; Hoffman et al. 2016) have been conducted that introduce Unsupervised Domain Adaptation (UDA) to adapt object detectors trained in the annotated source domain to an unlabeled target domain. A natural idea is to adapt features with adversarial learning (Chen et al. 2018; Saito et al. 2019; Li et al. 2020), giving a pixel-to-pixel adaptation on feature pyramids. Instead of aligning the overall feature maps, some works (Kim et al. 2019; Xu et al. 2020) focus on RPN-based region proposals and adopt a foreground adaptation to align those regions of interest. Recently, some works (Xu et al. 2020; Zhang, Wang, and Mao 2021) explore the category-level adaptation and model semantic knowledge with category centers, referred to as prototypes within inner-batch. They measure the distance between prototypes as domain discrepancy and minimize this discrepancy to bridge the domain gap.

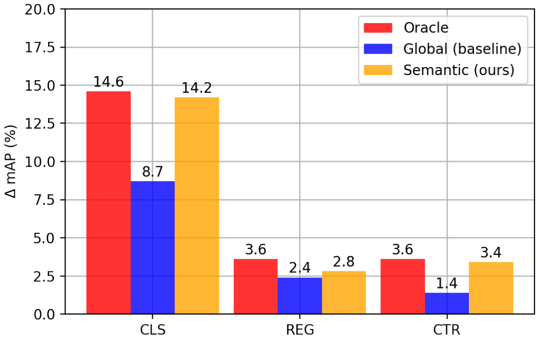

Figure 1: Illustration of the mAP (%) improvement after replacing the output maps, including classification (CLS), regression (REG), and centerness (CTR) output maps, compared with source only model.

Though great success, there are two challenges in existing category-level approaches. Firstly, these works model category prototypes within a batch, which inevitably bring about a biased estimation of domain-level distribution due to the limited and noisy information within the batch-wise observation. Since object occlusion and imbalanced object categories always appear in image batches, prototypes within a batch are hard to provide critical cues to establish an implicit probabilistic model fitting class-specific distributions well, leading to the difficulty of the alignment. Furthermore,

[*] Yixuan Yuan is the corresponding author.

directly transferring this biased semantic batch-to-batch (Xu et al. 2020; Zhang, Wang, and Mao 2021) from source to target domain is risky, as the unsatisfied categorical representations harm the adaptation. Besides, existing prototypes only capture current iteration semantics, lacking robustness to the dynamic model optimization. Hence, we aim to model unbiased semantics with category knowledge in the dynamic training procedure with a robust representation.

The second challenge lies in that these methods highly rely on global feature alignment (Chen et al. 2018), which ignores introducing semantic knowledge in the adversarial feature alignment. As shown in Figure 1, we conduct an experiment to delve into the importance of semantic knowledge in feature alignment. Given an object detector trained in the source domain (18.4% mAP) (Hsu et al. 2020a), we replace the output maps in the target domain with the ones from the source domain[1], which eliminates the influence of domain gap on each task, as shown in the *Red bars*. It can be observed a remarkable 14.6% mAP gain in classification, indicating that *bridging the domain gap in classification gives the most significant benefit in domain adaptive object detection*. After that, output maps are replaced by the ones from our baseline model (Hsu et al. 2020a), which deploys a class-agnostic alignment (*Blue bars*), showing the consistent importance in classification with an 8.7 % mAP gain. However, *we find a noticeable mAP gap in classification between global alignment and oracle (8.7% vs.14.6%)*, demonstrating that numerous classification errors still cannot be solved in class-agnostic alignment. This observation motivates us to introduce semantic knowledge in global alignment to bridge the domain gap in a category-to-category manner.

To overcome these two challenges, we propose a **S**emantic **C**onditioned **A**daptatio**N** (SCAN) framework, which achieves semantic conditioned adaptation with well-adapted category distributions. SCAN aims to overcome the challenge of biased semantics and introduce category knowledge in the global alignment. In the source domain, cross-image semantics are aggregated with the graph structure and then modeled with a *Time-Category-Distribution* three-dimensional paradigm. Based on the modeled semantics, we further introduce the conditional kernel to manifest semantics with activation maps. In the target domain, we model unbiased semantics with a novel conditional graph established in the pixel-level and category-level space. To achieve the semantic conditioned adaptation, we propose a Conditional Kernel guided Alignment (CKA) module to guide class-specific alignment and introduce a Graph-based Semantic Transfer (GST) mechanism to transfer unbiased semantics from the source to the target domain. After adopting our method, we observe a significant mAP improvement in the classification branch with an oracle-neared result, as shown in Figure 1 *Yellow bars*. To summarize, our main contributions are as follows.

- We propose a **S**emantic **C**onditioned **A**daptatio**N** (SCAN) framework[2] for cross-domain object detection. This work represents the first attempt at modeling unbiased semantics in cross-domain object detection to the best of our knowledge.
- SCAN utilizes a Conditional Kernel guided Alignment module and a Graph-based Semantic Transfer mechanism to achieve category-level adaptation, representing the first work introducing unbiased semantic knowledge in global adaptation instead of conventional class-agnostic alignment.
- Comprehensive experiments on three benchmarks demonstrate that SCAN achieves state-of-the-art results and outperforms existing works by a large margin.

Related Work

Cross Domain Object Detection

Cross-domain object detection aims to reduce the performance deterioration caused by the domain gap between training and inference datasets. Extensive works have been conducted to overcome this challenge, including image-level style translation (Inoue et al. 2018; Kim et al. 2019; Hsu et al. 2020b), pixel-level feature alignment (Chen et al. 2018; Saito et al. 2019; Li et al. 2020), region-level proposal adaptation (Kim et al. 2019) and pseudo label self-training (Inoue et al. 2018). Recently, some works (Xu et al. 2020; Zheng et al. 2020; Zhang, Wang, and Mao 2021; VS et al. 2021) introduce category-level adaptation to bridge the domain gap in the semantic space. GPA (Xu et al. 2020) introduces a graph to model prototype among region proposals in each image and minimizes prototype distance between the source and target domain. The authors in (Zhang, Wang, and Mao 2021) utilize RPN prototypes to model feature distributions of RPN-based region proposals. Another idea is to design a memory module (VS et al. 2021) to save large amounts of proposal features in the source domain and load them into the target domain. However, these works model prototypes within inner-batch, resulting in biased estimation of domain-level distribution. In this work, we aggregate semantics with the graph structure and propose a three-dimensional paradigm to model unbiased semantics.

Conditional Convolution

Different from conventional convolutions with fixed parameters, conditional convolutions learn dynamic kernels depending on the conditioned input. This idea is first explored in (Jia et al. 2016) to improve object recognition with a dynamic feature network. Some works have recently introduced conditional convolutions into the detection community, fully utilizing its dynamic property to model instance-level representations. CondInst (Yang et al. 2019) learns kernels conditioned on the regression output maps to extract semantic masks for each instance. Similarly, Sparse RCNN (Sun et al. 2021) and Implicit PointRend (Cheng,

[1] We replace output maps (CLS: classification; REG: regression; CTR: the centerness denoted as foregrounds) obtained from Foggy Cityscapes *validation* set by Cityscapes *validation* set. Note that they have the same annotations, and the only difference between them is the weather-based domain gap.

[2] Open source: https://github.com/CityU-AIM-Group/SCAN.

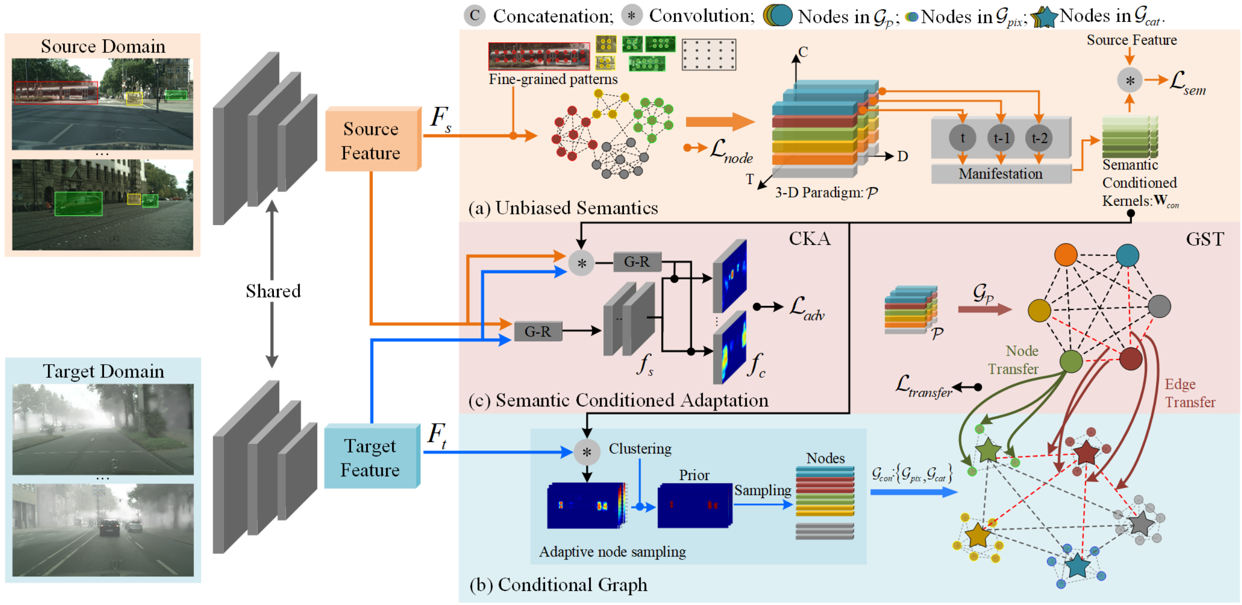

Figure 2: Overall of the proposed SCAN framework. G-R denotes the Gradient Reversal layer for adversarial training. (a) We aggregate cross-image semantics with graph structure, and then model unbiased semantics with a 3-D paradigm $\mathcal{P}$, which acts as the conditioned input to learn semantic conditioned kernels $\mathbf{W}_{con}$. (b) Based on $\mathbf{W}_{con}$, we proposed a conditional graph $\mathcal{G}_{con}$ to model semantics in the target domain, including a pixel-level sub-graph (marked as circles) and a category-level sub-graph (marked as stars). (c) CKA (left) and GST (right) utilize $\mathbf{W}_{con}$ and $\mathcal{P}$ to achieve semantic conditioned adaptation.

Parkhi, and Kirillov 2021) utilize region proposals conditioned kernels to extract more discriminative region representations. In this work, we extend this idea to cross-domain object detection and propose a semantic conditioned adaptation to bridge the domain gap. Different from existing works adopting pixel-level and proposal-level conditions, we model conditions at the domain level, yielding unbiased semantic representations, to overcome the challenge of cross-domain object detection.

Proposed Method

The overall workflow of SCAN framework is shown in Figure 2. Given a batch of annotated source images $\{I_{s,i}, Y_{s,i}\}_{i=1}^{B_s}$ and unlabeled target images $\{I_{t,i}\}_{i=1}^{B_t}$, we first adopt backbone network to extract image features $F_{s/t}$, and then separate the workflow into three branches. (a) Source domain: We perform fine-grained sampling to obtain semantic patterns and establish a graph to aggregate cross-image semantics. Then, a three-dimensional semantic paradigm $\mathcal{P}$ is proposed to model unbiased semantics, with which we further introduce the semantic conditioned kernel $\mathbf{W}_{con}$. (b) Target domain: Based on the conditional kernel $\mathbf{W}_{con}$, a conditional graph $\mathcal{G}_{con}$ is established to model unbiased semantics, including pixel-level (marked as circles) and category-level (marked as stars) sub-graphs $\{\mathcal{G}_{pix}, \mathcal{G}_{cat}\}$. (c) Semantic Conditioned Adaptation: After modeling semantics in both domains, we use $\mathbf{W}_{con}$ to guide the global alignment in the CKA module (middle-left) and transfer the semantics from the unbiased semantic graph $\mathcal{G}_{\mathcal{P}}$ to the conditional graph $\mathcal{G}_{con}$ using graph-based message propagation (middle-right).

Source Domain: Unbiased Semantics (US)

Given a labeled image batch $\{I_{s,i}, Y_{s,i}\}_{i=1}^{B_s}$ in the source domain, we first adopt domain-shared backbone to extract features F_s, based on which we perform fine-grained sampling to collect semantic patterns according to the ground-truth label, as shown in Figure 2(a). We sample those pixels inside object instances as category-specific foreground samples and perform spatial-uniformed sampling to obtain an equal number of background samples.

Semantic Modeling. Considering the critical role of long-distance semantic dependency (Chen et al. 2019), we propose a cross-image graph to aggregate fine-grained semantic patterns within a batch. Specifically, we first adopt a nonlinear projection on sampled semantic patterns to obtain node embedding $X = \{(x_i, y_i)\}_{i=1}^{N_s^{node}}, x_i \in \mathrm{R}^D$, and then establish a scalable graph covering all nodes followed by the graph reasoning as (Zhu et al. 2021) to obtain enhanced node representations:

$$\tilde{X} = Norm(softmax(W_f X, W_g X^T) W_h X + X), \quad (1)$$

where $\tilde{X} = \{\tilde{x}_i\}_{i=1}^{N_s^{node}}$ represents enhanced nodes, $W_{(\cdot)}$ represents the learnable weight, and $Norm$ is the layer normalization. To train the parameters in the graph and enhance the semantics of nodes, we perform an auxiliary node classification task with Cross Entropy loss as follows,

$$\mathcal{L}_{node} = -\sum_{i=1}^{N_s^{node}} y_i log(softmax(f_{cls}(\tilde{x}_i))), \quad (2)$$

where f_{cls} is a nonlinear classifier. In addition to establishing the cross-image semantic dependence, graph reasoning can relieve some adverse effects of noisy pixels inside bounding boxes belonging to the background with only box-level annotations.

To avoid the biased estimation within inner-batch, we innovatively design a *Time-Category-Distribution* three-dimensional paradigm $\mathcal{P} \in \mathrm{R}^{T\times C\times D}$ to model unbiased semantic knowledge at the domain level. With this time-included definition, $\mathcal{P}$ not only captures the category knowledge in the current iteration, but also saves $T-1$ historical representations, which provide critical robustness during model optimization. To update this paradigm $\mathcal{P}$, an incremental update strategy is proposed, fully utilizing aggregated graph nodes $\tilde{X} = \{\tilde{x}_i\}_{i=1}^{N_s^{node}}$. Specifically, we first define the semantics within batch for category c at the current iteration as $p = \frac{1}{N_{s,c}^{node}} \sum_{y_i=c} \tilde{x}_i$. Then, we adopt a time-axis translation to preserve $T-1$ historical semantics, and update the the current state with p:

$$\begin{aligned}
&\mathcal{P}_k^c \leftarrow \mathcal{P}_{k+1}^c, 1 \leq k < T, \\
&\mathcal{P}_k^c \leftarrow cos(p, \mathcal{P}_{k-1}^c)p + (1 - cos(p, \mathcal{P}_{k-1}^c))\mathcal{P}_k^c, k = T,
\end{aligned} \tag{3}$$

where $cos(x, y) = \frac{x\cdot y}{\|x\|_2 \cdot \|y\|_2}$, $\mathcal{P}_k^c$ indicates unbiased semantics for *Category* c at *Time* k. The cosine-based update (VS et al. 2021) accelerates the learning process in the early training stage when the local and global semantics are inconsistent, and gradually slows down to find a global-optima of unbiased semantics. Besides, the memory queue introduces historical semantic dependency and relives the noise caused by unstable adversarial learning. With this three-dimensional format, $\mathcal{P}$ utilizes a contiguous sequence with length T to model category knowledge during training, which preserves semantic evolution across iterations.

Semantic Conditioned Manifestation. Based on the well-modeled semantic paradigm $\mathcal{P}$, we further manifest this implicit semantics with semantic conditioned convolution. As the conditioned input, $\mathcal{P}$ is sent to a lightweight manifestation module to learn the parameters of conditional kernels. Since $\mathcal{P}$ covers a time-included dynamic procedure, this manifestation module is designed with a Recurrent Neural Network (RNN) based unit to capture cross-iteration semantic relationship and model inherent semantic evolution:

$$\mathbf{W}_{con} = Conv(tanh(RNN(\mathcal{P}))), \tag{4}$$

where $\mathbf{W}_{con}$ is the learned semantic conditioned kernels, $Conv$ is a $n \times 1$ convolution layer, and RNN is a two-layer recurrent network. Then, we perform conditional convolution on source features to obtain class-specific semantic activation $S_s = softmax(\mathbf{W}_{con} * (F_s))$, where $*$ denotes the convolution operation. Focal Loss is adopted on S_s to train this dynamic branch to manifest implicit semantics:

$$\mathcal{L}_{sem} = -\sum_{u,v} S_s^{(u,v)} \alpha (1 - S_s^{(u,v)})^{\gamma} log(Y_s^{(u,v)}) \tag{5}$$

where (u, v) denotes the location in the semantic maps, and α and γ are two common parameters in Focal Loss (Lin et al. 2017). The semantic conditioned kernel manifests unbiased semantic knowledge with class-specific activation, serving for modeling semantics in the target domain and migrating domain gap in a category-to-category manner.

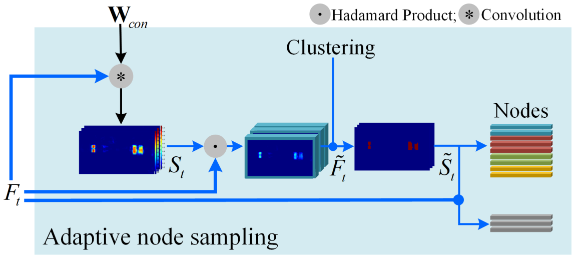

Figure 3: Illustration of the adaptive node sampling. $\mathbf{W}_{con}$ represents the semantic conditioned kernels.

Target Domain: Conditional Graph (CG)

Given an image batch $\{I_{t,i}\}_{i=1}^{B_t}$ in the target domain, we adopt domain-shared backbone to extract features F_t. To model semantics in the target domain, we establish a conditional graph $\mathcal{G}_{con}$ using the learned conditional kernels $\mathbf{W}_{con}$, which breaks the barrier of inaccessible category annotations, as shown in Figure 2(b). We first establish a pixel-level sub-graph $\mathcal{G}_{pix}$ to model fine-grained semantic patterns, which is further extended to category space, formatted as a category-level sub-graph $\mathcal{G}_{cat}$ to model the quadratic relationships between two categories.

For the pixel-level sub-graph $\mathcal{G}_{pix}$, we propose an adaptive node sampling strategy, as shown in Figure 3, to obtain fine-grained graph nodes. Given image features $F_t \in \mathrm{R}^{B_t\times D\times H\times W}$ and the learned conditional kernels $\mathbf{W}_{con} \in \mathrm{R}^{C\times D}$, we first perform conditional convolution to obtain explicit semantic activation maps $S_t \in \mathrm{R}^{C\times B_t\times H\times W}$. Then, class-specific feature representations are generated with semantic activation $\tilde{F}_{t,c} = S_{t,c} \bigodot F_t, \tilde{F}_t \in \mathrm{R}^{C\times B_t\times D\times H\times W}$, where $\bigodot$ is broadcasted hadamard product. After that, we adopt Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Schubert et al. 2017) algorithm to obtain a noiseless sampling prior defined with clusters, yielding clean semantic maps $\tilde{S}_t \in \{0, 1\}^{C\times B_t\times H\times W}$. Based on this prior $\tilde{S}_t$, i.e., class-specific clusters, we perform fine-grained node sampling on the target feature F_t and adopt spatial-uniformed sampling to obtain equal number of background node samples: $\{(v_i, \tilde{S}_{t,i})\}_{i=1}^{N_t^{node}}, v_i \in \mathrm{R}^D$. The nonlinear projection in the source domain is also adopted to obtain node embedding. Finally, we establish $\mathcal{G}_{pix}$ with nodes $\mathcal{V}_{pix} = \{v_i\}_{i=1}^{N_t^{node}}$, following the graph reasoning procedure as Eq.1 to obtain enhanced node representations $\{\tilde{v}_i\}_{i=1}^{N_t^{node}}$ (marked as circles in Figure 2(b)). Since both domains share the same category space (Chen et al. 2018), we share the parameters of graph reasoning in the source and target domain.

To obtain robust semantics in category space, we further extend $\mathcal{G}_{pix}$ to implicit semantic space with a category-level sub-graph $\mathcal{G}_{cat} =< \mathcal{V}_{cat}, \mathcal{E}_{cat} >$, where $\mathcal{V}_{cat}^c = \frac{1}{N_{t,c}^{node}} \Sigma_{\tilde{S}_{t,i}=c} \tilde{v}_i$ is the node embedding of class c, and $\mathcal{E}_{cat}^{(i,j)} = cos(\mathcal{V}_{cat}^i, \mathcal{V}_{cat}^j)$ represents the quadratic relationship between class i and class j (marked as stars in Figure 2(b)). With the proposed conditional graph $\mathcal{G}_{con}$, the semantics are modeled comprehensively in the target domain across fine-grained pixel-level space and implicit category-level space.

Semantic Conditioned Adaptation

After modeling unbiased semantic knowledge in both domains, we present semantic conditioned adaptation to achieve domain adaptation, as shown in Figure 2(c).

Conditional Kernel Guided Alignment (CKA). We propose a CKA module to integrate unbiased semantic knowledge in the global alignment, which is a semantic-aware discriminator with a multi-head architecture. Given extracted image features $F_{s/t}$ and the conditional kernels $\mathbf{W}_{con}$, we perform conditional convolution to obtain semantic activation maps $S_{s/t}$ for both domains. Then, $F_{s/t}$ are sent to the CKA module, including shared convolutions f_s followed by C class-specific branches, to obtain semantic-aware features: $\hat{F}_{s/t,c} = [f_s(F_{s/t}) : S_{s/t,c}]$, where [:] represents concatenation operation. After that, independent class-specific domain classifiers f_c are adopted the C separated branches with Binary Cross-Entropy (BCE) loss:

$$\begin{aligned}\mathcal{L}_{adv} = -\frac{1}{C}\sum_{u,v}\sum_{c}\{&\frac{DS_{s,c}^{(u,v)}}{\sum_{u,v} S_{s,c}^{(u,v)}} log(f_c(\hat{F}_{s,c}^{(u,v)})) \\ &+ \frac{(1-D)S_{t,c}^{(u,v)}}{\sum_{u,v} S_{t,c}^{(u,v)}} log(f_c(\hat{F}_{t,c}^{(u,v)}))\},\end{aligned} \tag{6}$$

where (u, v) denotes the location in feature/semantic maps, C is the category number, and D is the binary domain labels. With the proposed CKA module, the feature belonging to the same category will be aligned, yielding the mitigation of the domain gap in a category-to-category manner.

Graph Based Semantic Transfer (GST). To narrow the domain gap in category-based feature space, we further propose a GST mechanism, which utilizes unbiased semantics $\mathcal{P}$ to guide the optimization in the target domain through a graph-based message propagation. Specifically, CKA first converts $\mathcal{P}$ to a unbiased semantic graph $\mathcal{G}_{\mathcal{P}} =< \mathcal{V}_{\mathcal{P}}, \mathcal{E}_{\mathcal{P}} >$, where $\mathcal{V}_{\mathcal{P}}^c = \frac{1}{T}\sum_T \mathcal{P}_T^c$ indicates the graph node and $\mathcal{E}_{\mathcal{P}}^{(i,j)} = cos(\mathcal{V}_{\mathcal{P}}^i, \mathcal{V}_{\mathcal{P}}^j)$ represents the edge. After that, we establish a graph-based semantic transfer from the unbiased semantic graph $\mathcal{G}_{\mathcal{P}}$ modeled in the source domain to the conditional graph $\mathcal{G}_{con}$ in the target domain, including the node transfer and edge transfer. For the node transfer, GST utilizes $\mathcal{V}_{\mathcal{P}}$ to guide the node embedding of pixel-level sub-graph $\mathcal{G}_{pix}$ ($\mathcal{G}_{pix} \subset \mathcal{G}_{con}$), providing unbiased semantic consistency for each instance sample. For the edge part, quadratic relationships between two categories $\mathcal{E}_{\mathcal{P}}^{(i,j)}$ should be consistent in both domains, and be used to guide the edge of category-level graph $\mathcal{G}_{cat}$ ($\mathcal{G}_{cat} \subset \mathcal{G}_{con}$). Therefore, our transfer loss, including node and edge terms are as follows,

$$\begin{aligned}\mathcal{L}_{transfer} =& \frac{1}{N_t^{node}C}\sum_i\sum_c\{\mathcal{L}_{kl}(\tilde{v}_{t,i}^c, \mathcal{V}_{\mathcal{P}}^c)\} \\ &+ \frac{1}{C}\sum_c\{\mathcal{L}_{cos}(\mathcal{E}_{cat}^{(i,j)}, \mathcal{E}_{\mathcal{P}}^{(i,j)})\},\end{aligned} \tag{7}$$

where the first item represents node transfer, the second item defines edge transfer, C is the category number, $\mathcal{L}_{kl}$ and $\mathcal{L}_{cos}$ represent the Kullback-Leibler divergence and Cosine Embedding loss, respectively, which generate explicit gradient flows in the target domain to guide the model optimization. The proposed GST mechanism transfers the class-specific knowledge (node) and the inter-class quadratic relationship (edge) to the target domain jointly, bridging the domain gap in implicit category space comprehensively.

Model Optimization

In the training stage of the proposed SCAN framework, the whole loss function $\mathcal{L}$ consists of five main components denoted as,

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{node} + \mathcal{L}_{sem} + \alpha\mathcal{L}_{adv} + \beta\mathcal{L}_{transfer} \tag{8}$$

where $\mathcal{L}_{det}$ is the detection loss in (Tian et al. 2019), $\mathcal{L}_{node}$ and $\mathcal{L}_{sem}$ are two auxiliary loss terms to model unbiased semantic knowledge, $\mathcal{L}_{adv}$ is used in CKA module for semantic-aware global alignment, $\mathcal{L}_{transfer}$ is defined in GST for semantic transfer. Besides, α and β are two parameters controlling the adaptation weight.

Experiments

Experimental Settings

We conduct extensive experiments on three domain adaptation scenarios, following the standard settings in literature (Hsu et al. 2020a), i.e. training with labeled source data and unlabeled target data, and testing on the target data. Our baseline model is the FCOS (Tian et al. 2019) object detector combining with the Global Alignment module (Hsu et al. 2020a). Detection results are evaluated with mean Average Precision (mAP) using different IoU thresholds (Lin et al. 2014), denoted as mAP$_{IoU}$. Superscripts represent mAP$_{0.5}$ gains compared with corresponding source only results.

Cityscapes→Foggy Cityscapes: Cityscapes (Cordts et al. 2016) is a city landscape dataset under dry weather condition with eight annotated categories, which consists *train* set with 2975 images and *validation* set with 500 images. Foggy Cityscapes (Sakaridis, Dai, and Van Gool 2018) is a synthesized dataset from Cityscapes as foggy weather. Domain gap caused by the weather condition is explored in adaptation.

Sim10k→Cityscapes: Sim10k (Johnson-Roberson et al. 2017) is a simulated dataset with 10,000 images with the labels of annotated car bounding boxes. We give an exploration of the domain gap from synthesized to real-world images following existing literature.

KITTI→Cityscapes: KITTI (Geiger, Lenz, and Urtasun 2012) is a real-world scene dataset collected with different camera setups. KITTI consists of 7,481 images with car categories. We conduct the evaluation of the capability of cross-camera adaptation following literature.

Implementation Details

We use the ImageNet (Deng et al. 2009) pre-trained VGG-16 (Simonyan and Zisserman 2014) as the backbone network. We adopt the Stochastic Gradient Descent (SGD) optimizer with a 0.0025 learning rate and an 8 batch-size. α and β are set 0.1 and 1, respectively. We set an extremely low threshold (0.05) to accelerate node sampling to establish

Method	person	rider	car	truck	bus	train	motor	bike	mAP$_{0.5}$
F-RCNN (Chen et al. 2018)	17.8	23.6	27.1	11.9	23.8	9.1	14.4	22.8	18.8
EPM (Hsu et al. 2020a)	41.9	38.7	56.7	22.6	41.5	26.8	24.6	35.5	$36.0^{+17.6}$
IIOD (Wu et al. 2021a)	33.1	43.4	49.6	21.9	45.7	32.0	29.5	37.0	$36.6^{+17.8}$
RPNPA (Zhang, Wang, and Mao 2021)	33.6	43.8	49.6	32.9	45.5	46.0	35.7	36.8	$40.5^{+21.7}$
DSS (Wang et al. 2021)	42.9	51.2	53.6	33.6	49.2	18.9	36.2	41.8	$40.9^{+22.1}$
UMT (Deng et al. 2021)	33.0	46.7	48.6	34.1	56.5	46.8	30.4	37.4	$41.7^{+22.9}$
MeGA-CDA (VS et al. 2021)	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	$41.8^{+23.0}$
ICCR-VDD (Wu et al. 2021b)	33.4	44.0	51.7	33.9	52.0	34.7	34.2	36.8	$40.0^{+21.2}$
Source Only (Hsu et al. 2020a)	30.5	23.9	34.2	5.8	11.1	5.1	10.6	26.1	18.4
Baseline (Hsu et al. 2020a)	38.7	36.1	53.1	21.9	35.4	25.7	20.6	33.9	$33.2^{+14.8}$
SCAN w/o. CKA&US (ours)	40.4	40.7	54.5	28.1	43.1	38.9	27.6	38.5	$39.0^{+20.6}$
SCAN w/o. GST&CG (ours)	42.0	44.0	56.9	28.9	47.1	41.9	27.9	38.6	$40.9^{+22.5}$
SCAN (ours)	41.7	43.9	57.3	28.7	48.6	48.7	31.0	37.3	**42.1**$^{+23.7}$

Table 1: Comparison results in the Cityscapes→Foggy Cityscapes (%) domain adaptation scenario.

Method	mAP$_{0.5}$
F-RCNN (Chen et al. 2018)	34.3
EPM (Hsu et al. 2020a)	$49.0^{+9.2}$
DSS (Wang et al. 2021)	$44.5^{+9.8}$
MEGA-CDA (VS et al. 2021)	$44.8^{+10.5}$
RPNPA (Zhang, Wang, and Mao 2021)	$45.7^{+11.1}$
UMT (Deng et al. 2021)	$43.1^{+8.8}$
Source Only (Hsu et al. 2020a)	39.8
Baseline (Hsu et al. 2020a)	$45.9^{+6.1}$
SCAN w/o. CKA&US (ours)	$49.6^{+9.8}$
SCAN w/o. GST&CG (ours)	$51.8^{+12.0}$
SCAN (ours)	**52.6**$^{+12.8}$

Table 2: Comparison Results on Sim10K→Cityscapes (%).

Method	mAP$_{0.5}$
F-RCNN (Chen et al. 2018)	30.2
EPM (Hsu et al. 2020a)	$43.2^{+8.8}$
DSS (Wang et al. 2021)	$42.7^{+9.8}$
MEGA-CDA (VS et al. 2021)	$43.0^{+12.8}$
RPNPA (Zhang, Wang, and Mao 2021)	$44.8^{+10.8}$
Source Only (Hsu et al. 2020a)	34.4
Baseline (Hsu et al. 2020a)	$39.1^{+4.7}$
SCAN w/o. CKA&US (ours)	$43.7^{+9.3}$
SCAN w/o. GST&CG (ours)	$45.4^{+11.0}$
SCAN (ours)	**45.8**$^{+12.4}$

Table 3: Comparison results on KITTI→Cityscapes (%).

conditional graph $\mathcal{G}_{con}$, ignoring many background samples. The nonlinear projection is deployed with the Conv-ReLU-Conv structure, and the nonlinear classifier f_s is in the Fc-ReLU-Fc format. The three-dimensional paradigm $\mathcal{P}$ records T=3 iterations and uses D=256 channels to model class-specific distributions.

Benchmark Comparison

Cityscapes→Foggy Cityscapes: As shown in Table 1, SCAN can surpass existing approaches with a promising 42.1% mAP$_{0.5}$, outperforming RPNPA (Zhang, Wang, and Mao 2021) and ICCR-VDD (Wu et al. 2021b) with 1.6%, and 2.1% mAP$_{0.5}$ gain. Besides, SCAN also achieves the best absolute gain (+24.5%) compared with source only model, breaking through existing record (+23.0%) made by MeGA-CDA (VS et al. 2021) in literature. Moreover, removing CKA&US and GST&CG can also achieve consistent performance gains with 39.0% and 40.9% mAP$_{0.5}$, compared with baseline mode (33.2% mAP$_{0.5}$), which demonstrates the strength of semantic condition adaptation.
Sim10k→Cityscapes: Adaptation results are recorded in Table 2. SCAN can achieve 52.6% mAP$_{0.5}$ and gives 49.6% and 51.8% mAP$_{0.5}$ without deploying CKA&US and GST&CG, respectively, outperforming existing works greatly. Besides, SCAN surpasses EPM (Hsu et al. 2020a) (49.0% mAP$_{0.5}$) with a 3.6% gain using the same single-stage pipeline. Note that multi-category semantics will degrade into foregrounds under single category scenarios, demonstrating that SCAN could also adapt foreground well with semantic conditional adaptation.
KITTI→Cityscapes: The comparison results are shown in Table 3. SCAN gives a 45.8 mAP$_{0.5}$, surpassing other methods by a large margin. Compared with our baseline model, eliminating CKA&US achieves 43.7% mAP$_{0.5}$ and removing GST&CG gives a 45.4% mAP$_{0.5}$, both of which outperform the baseline model with remarkable gains.

Ablation Study

We present detailed ablation studies, as shown in Table 4.
Conditional Kernel Guided Alignment (CKA). Since CKA consists of the semantic conditioned kernels, which are optimized with the unbiased semantics (US), we introduce a CKA† that using classification results to guide global alignment. Compared with the baseline model, introducing semantic knowledge in feature alignment could give a 2.4% mAP$_{0.5}$ gain (CKA†), demonstrating that semantic knowledge plays a critical role in bridging domain gap.
Unbiased Semantics (US). Introducing unbiased semantics achieves a significant mAP$_{0.5}$ improvement (35.6% → 40.9%), which yields a 40.9 mAP$_{0.5}$ and outperforms baseline model (33.2% mAP$_{0.5}$) with a remarkable 7.7% mAP$_{0.5}$. Besides, it can be observed that peeling semantic modeling (SM: 40.9%→38.7%) and semantic conditioned manifestation (SCM: 40.9%→36.9%) both makes noticeable adverse influences, showing that each component is necessary to model unbiased semantics.
Conditional Graph (CG). Introducing conditional graph

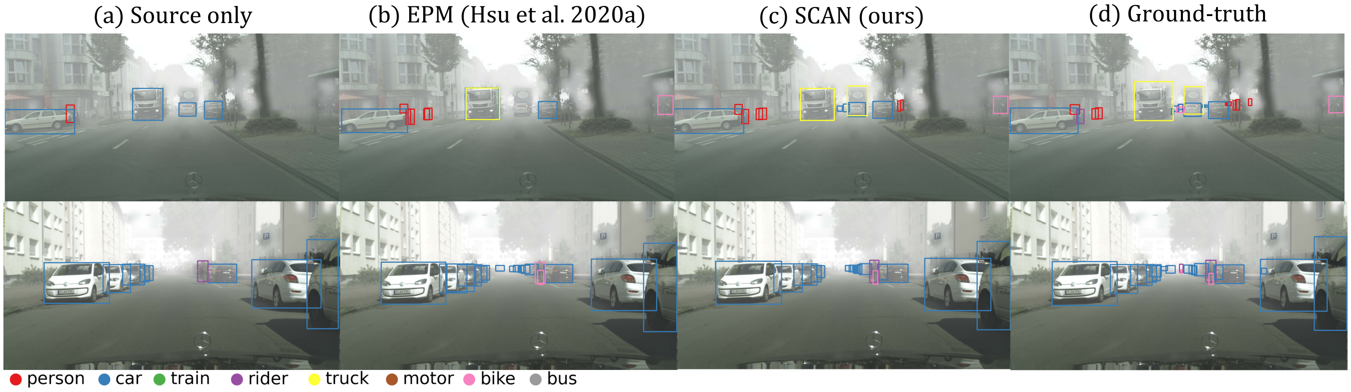
(a) Source only
(b) EPM (Hsu et al. 2020a)
(c) SCAN (ours)
(d) Ground-truth
● person ● car ● train ● rider ● truck ● motor ● bike ● bus

Figure 4: Qualitative results on the Cityscapes→Foggy Cityscapes adaptation scenario of (a) the source only model, (b) EPM (Hsu et al. 2020a), (c) SCAN, and (d) ground truth. (Zooming in for best view.)

Setting	w/o	mAP	$\text{mAP}_{0.5}$	$\text{mAP}_{0.75}$
Baseline	-	16.9	33.2	15.4
CKA[†]	-	18.5	$35.6^{+2.4}$	18.1
CKA+**US**	-	22.1	$40.9^{+7.7}$	21.0
CKA+**US**	SM	19.2	$38.7^{+5.5}$	19.0
CKA+**US**	SCM	18.8	$36.9^{+3.7}$	18.4
CKA+US+**CG**	-	22.5	$41.4^{+8.2}$	21.9
CKA+US+**CG**	ANS	22.2	$41.0^{+7.8}$	21.3
CKA+US+CG+**GST**	-	23.1	$42.1^{+8.9}$	21.2
CKA+US+CG+**GST**	Node	22.2	$41.5^{+8.3}$	21.0
CKA+US+CG+**GST**	Edge	22.8	$41.9^{+8.7}$	20.9

Table 4: Ablation studies on SCAN framework. SM represents semantic modeling, SCM indicates semantic conditioned manifestation, and ANS is adaptive node sampling.

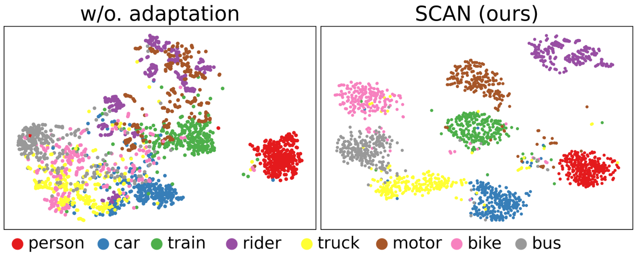

Figure 5: T-SNE feature visualization of object features in different categories. For each category, we randomly sample object features (marked as circles) inside bounding boxes in the source domain and target domain equally.

achieves 41.4% $\text{mAP}_{0.5}$ and outperforms baseline with 8.2% $\text{mAP}_{0.5}$, due to the well-modeled semantics in the target domain. Replacing our adaptive node sampling (ANS) strategy with handcraft threshold gives a 41.0% $\text{mAP}_{0.5}$ with a 0.4% $\text{mAP}_{0.5}$ reduction (41.4%→41.0%) compared with ANS, verifying its better semantic sampling capacity.

Graph-based Semantic Transfer (GST). Adopting node and edge transfer can both improve the performance with a 41.9% and 41.5% $\text{mAP}_{0.5}$ while adopting them together could give the best performance with 42.1% $\text{mAP}_{0.5}$, outperforming our baseline model with a significant 8.9% $\text{mAP}_{0.5}$.

Qualitative Results

T-SNE Visualization. As shown in Figure 5, we present category-specific feature distributions (distinguished in different colors) to demonstrate the effectiveness of the SCAN framework. We randomly sample the same number of object features (marked as circles) in the source domain and target domain for each category, respectively. After adopting SCAN, category-specific distributions of the source and target domain can be aligned well, demonstrating the effectiveness of semantic conditioned adaptation.

Detection Result Visualization. Comparisons of detection results are shown in Figure 4. SCAN can reduce false-negative cases (top row) caused by class-agnostic adaptation, such as the missed truck. Most importantly, SCAN also eliminates some classification errors, like the rider and bike in the bottom row, demonstrating the advantages of introducing semantic knowledge to bridge the domain gap.

Conclusion

We propose a **S**emantic **C**onditioned **A**daptatio**N** (SCAN) framework for cross-domain object detection. In the source domain, unbiased semantics are aggregated with a cross-image graph, modeled with the unbiased semantic paradigm, and manifested with semantic conditioned kernels. A conditional graph is established in the target domain to model unbiased semantic knowledge at the category level. To achieve adaptation, we propose a Conditional Kernel guided Alignment (CKA) module to align category distributions globally and a Graph-based Semantic Transfer (GST) mechanism to adapt semantics in implicit feature space. Comprehensive experiments on three adaptation scenarios demonstrate the superior performance of SCAN.

Acknowledgments

This work is supported by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA).

References

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 3339–3348.

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019. Multi-label image recognition with graph convolutional networks. In *CVPR*, 5177–5186.

Cheng, B.; Parkhi, O.; and Kirillov, A. 2021. Pointly-Supervised Instance Segmentation. *arXiv preprint arXiv:2104.06404*.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 3213–3223.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Deng, J.; Li, W.; Chen, Y.; and Duan, L. 2021. Unbiased Mean Teacher for Cross-Domain Object Detection. In *CVPR*, 4091–4101.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 3354–3361.

Hoffman, J.; Wang, D.; Yu, F.; and Darrell, T. 2016. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*.

Hsu, C.-C.; Tsai, Y.-H.; Lin, Y.-Y.; and Yang, M.-H. 2020a. Every Pixel Matters: Center-aware Feature Alignment for Domain Adaptive Object Detector. In *ECCV*, 733–748.

Hsu, H.-K.; Yao, C.-H.; Tsai, Y.-H.; Hung, W.-C.; Tseng, H.-Y.; Singh, M.; and Yang, M.-H. 2020b. Progressive domain adaptation for object detection. In *WACV*, 749–757.

Inoue, N.; Furuta, R.; Yamasaki, T.; and Aizawa, K. 2018. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, 5001–5009.

Jia, X.; De Brabandere, B.; Tuytelaars, T.; and Gool, L. V. 2016. Dynamic Filter Networks. In *Advances in Neural Information Processing Systems*, volume 29.

Johnson-Roberson, M.; Barto, C.; Mehta, R.; Sridhar, S. N.; Rosaen, K.; and Vasudevan, R. 2017. Driving in the Matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *ICRA*, 746–753.

Kim, T.; Jeong, M.; Kim, S.; Choi, S.; and Kim, C. 2019. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *CVPR*, 12456–12465.

Li, C.; Du, D.; Zhang, L.; Wen, L.; Luo, T.; Wu, Y.; and Zhu, P. 2020. Spatial attention pyramid network for unsupervised domain adaptation. In *ECCV*, 481–497. Springer.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *ICCV*, 2980–2988.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, 91–99.

Saito, K.; Ushiku, Y.; Harada, T.; and Saenko, K. 2019. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, 6956–6965.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9): 973–992.

Schubert, E.; Sander, J.; Ester, M.; Kriegel, H. P.; and Xu, X. 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *TODS*, 42(3): 1–21.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. 2021. Sparse r-cnn: End-to-end object detection with learnable proposals. In *CVPR*, 14454–14463.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 9627–9636.

VS, V.; Gupta, V.; Oza, P.; Sindagi, V. A.; and Patel, V. M. 2021. MeGA-CDA: Memory Guided Attention for Category-Aware Unsupervised Domain Adaptive Object Detection. In *CVPR*, 4516–4526.

Wang, Y.; Zhang, R.; Zhang, S.; Li, M.; Xia, Y.; Zhang, X.; and Liu, S. 2021. Domain-Specific Suppression for Adaptive Object Detection. In *CVPR*, 9603–9612.

Wu, A.; Han, Y.; Zhu, L.; and Yang, Y. 2021a. Instance-Invariant Domain Adaptive Object Detection via Progressive Disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wu, A.; Liu, R.; Han, Y.; Zhu, L.; and Yang, Y. 2021b. Vector-Decomposed Disentanglement for Domain-Invariant Object Detection. *ICCV*.

Xu, M.; Wang, H.; Ni, B.; Tian, Q.; and Zhang, W. 2020. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, 12355–12364.

Yang, B.; Bender, G.; Le, Q. V.; and Ngiam, J. 2019. CondConv: Conditionally Parameterized Convolutions for Efficient Inference. In *Advances in Neural inf. Process. Syst.*, volume 32.

Zhang, Y.; Wang, Z.; and Mao, Y. 2021. RPN Prototype Alignment for Domain Adaptive Object Detector. In *CVPR*, 12425–12434.

Zheng, Y.; Huang, D.; Liu, S.; and Wang, Y. 2020. Cross-domain object detection through coarse-to-fine feature adaptation. In *CVPR*, 13766–13775.

Zhu, C.; Chen, F.; Ahmed, U.; Shen, Z.; and Savvides, M. 2021. Semantic relation reasoning for shot-stable few-shot object detection. In *CVPR*, 8782–8791.