This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# SEEN-DA: SEmantic ENtropy guided Domain-aware Attention for Domain Adaptive Object Detection

Haochen Li[1,4] Rui Zhang[2*] Hantao Yao[3] Xin Zhang[2] Yifan Hao[2] Xinkai Song[2]
Shaohui Peng[1] Yongwei Zhao[2] Chen Zhao[5] Yanjun Wu[5] Ling Li[1,4*]
[1]Intelligent Software Research Center, Institute of Software, CAS
[2]State Key Lab of Processors, Institute of Computing Technology, CAS
[3] School of Information Science and Technology, University of Science and Technology of China
[4] University of Chinese Academy of Sciences [5] Institute of Software, CAS
{haochen2021, pengshaohui, zhaochen, wuyanjun, liling}@iscas.ac.cn
{zhangrui, zhangxin, haoyifan, songxinkai, zhaoyongwei}@ict.ac.cn, yaohantao@ustc.edu.cn

## Abstract

*Domain adaptive object detection (DAOD) aims to generalize detectors trained on an annotated source domain to an unlabelled target domain. Traditional works focus on aligning visual features between domains to extract domain-invariant knowledge, and recent VLM-based DAOD methods leverage semantic information provided by the textual encoder to supplement domain-specific features for each domain. However, they overlook the role of semantic information in guiding the learning of visual features that are beneficial for adaptation. To solve the problem, we propose semantic entropy to quantify the semantic information contained in visual features, and design SEmantic ENtropy guided Domain-aware Attention (SEEN-DA) to adaptively refine visual features with the semantic information of two domains. Semantic entropy reflects the importance of features based on semantic information, which can serve as attention to select discriminative visual features and suppress semantically irrelevant redundant information. Guided by semantic entropy, we introduce domain-aware attention modules into the visual encoder in SEEN-DA. It utilizes an inter-domain attention branch to extract domain-invariant features and eliminate redundant information, and an intra-domain attention branch to supplement the domain-specific semantic information discriminative on each domain. Comprehensive experiments validate the effectiveness of SEEN-DA, demonstrating significant improvements in cross-domain object detection performance.*

## 1. Introduction

Object detection [8, 33, 44, 45] have been proven to attain encouraging performance assisted by high-quality images with accurate annotations. However, due to the noticeable discrepancy between domains, they suffer severe performance degradation when dealing with unseen data in distinct scenes. To alleviate this problem, domain adaptive object detection (DAOD) [3] aims to transfer a detector trained on the labelled source domain to a unlabelled target domain.

Traditional DAOD works [3, 30, 47] focus on reducing domain discrepancy by aligning visual features between domains to extract domain-invariant knowledge. Additionally, [24,32,49,58] decouple the unique features of each domain, *i.e.*domain-specific features, to prevent its interference during feature alignment. Although effective, they train the detector with domain-shared and semantically agnostic category labels (such as one-hot encoding), which ignores the linguistic properties of categories and inter-domain differences and commonality, *i.e.*semantic information. Recently, visual-language models (VLMs) [26,43,65] are proposed to ground visual with textual modalities. Through the textual encoder, they align visual features with the textual prompt, which is constructed from natural language and contains rich semantic information. Motivated by this, recent VLM-based DAOD methods, such as DA-Pro [23], utilize the textual encoder as detection head and tune the textual prompt to learn domain-invariant semantics while supplementing domain-specific semantics for each domain. However, they freeze the visual encoder and leverage semantic information solely in the detection head, limiting the discriminability of visual features.

Summarily, despite promising performance, existing methods ignore the incorporation of semantic information to guide the learning of visual features that are beneficial for adaptation. They typically optimize the visual encoder with semantically agnostic category labels and adversarial loss, or directly freeze it, resulting in two limitations. First, they fail to filter out *redundant information* from visual fea-

*Corresponding author.

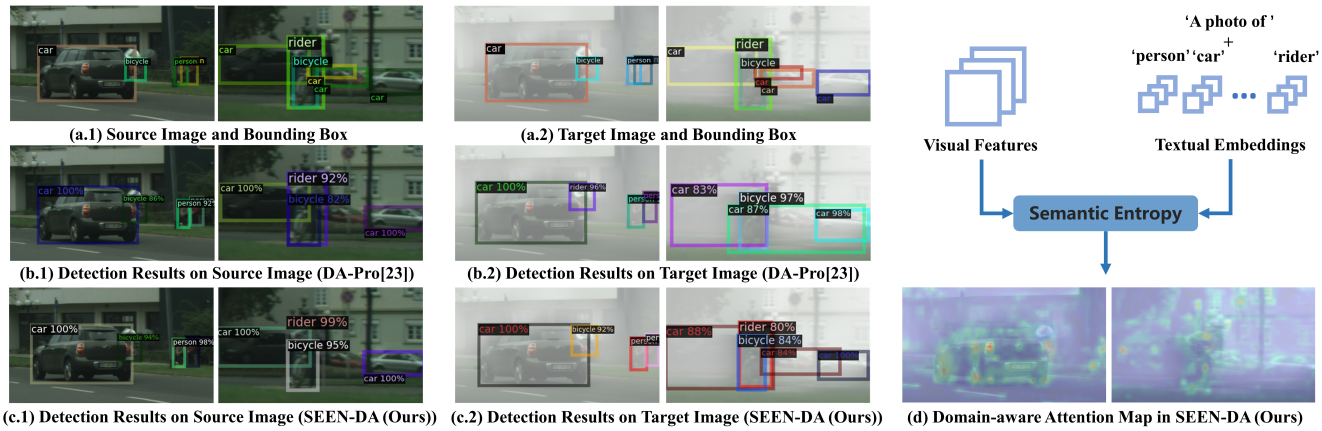



Figure 1. Visualization of the detection results and attention map. (a) Source and Target domain image and bounding boxes. (b) Existing methods ignore the incorporation of semantic information to guide the learning of visual features that are beneficial for adaptation. Infected by the redundant information and lack of domain-specific semantic information, some objects that are correctly detected on the source image are misclassified or missed on the target image. (c) We propose semantic entropy to evaluate the importance of visual features based on domain and category semantic information. By integrating it as domain-aware attention in SEEN-DA, we eliminate redundant information and enhance domain-specific features, thereby improving the detector's discriminability for DAOD.

tures, which is semantically irrelevant and can lead to incorrect feature alignment. Redundant information includes attributes unrelated to object identification, such as color, texture, etc. Although they do not contain domain and category semantic information, they are often extracted by adversarially trained visual encoder since they exist in both domains. Therefore, such semantically irrelevant information may interfere with the alignment of domain-invariant features. In Fig. 1(b.1), existing methods detect the bicycle covered with a white cloth on the source image (clear day). However, since target images (foggy day) tend to be white overall, they misclassify the white cloth as a rider in Fig. 1(b.2). Second, existing methods discard domain-specific features in the visual encoder, which contain rich domain semantic information and are important for the discriminability on the target domain. For example, they correctly detect the bicycle and rider on the source image Fig. 1(b.1), but miss them in Fig. 1(b.2) due to insufficient feature learning of domain-specific properties of target domain, *e.g.*fog.

To solve these problems, we propose semantic entropy which evaluates the importance of features based on domain and category semantic information, and accordingly design SEmantic ENtropy guided Domain-aware Attention (SEEN-DA) to eliminate redundant information and supplement domain-specific features for DAOD. The semantic entropy of visual features refers to the amount of semantic information related to a certain domain. As the textual encoder of VLM provides rich domain and category semantic information in textual embedding, the semantic entropy can be expressed as the information entropy of the classification probability between the visual and textual embedding. Therefore, semantic entropy reflects the importance of visual features in adaptation. Low semantic entropy indicates that visual features match the textual embedding of a certain category and are likely to be transferable foreground objects. On the contrary, high semantic entropy means that they are unpredictable and may be redundant or even harmful for detection.

Applying semantic entropy as attention, we design the SEEN-DA framework, introducing domain-aware attention modules into the visual encoder to select visual features with low semantic entropy and suppress regions with high ones. Specifically, the domain-aware attention module utilizes two parallel branches to learn inter- and intra-domain attention with semantic entropy. The inter-domain attention branch is shared between domains and evaluates the semantic entropy on domain-shared textual embedding, aiming to learn domain-invariant features while eliminating redundant information. The intra-domain attention branch is private in each domain to evaluate the semantic entropy on source/target textual embedding respectively, supplementing domain-specific features for each domain. The SEEN-DA is able to adaptively adjust visual features with the semantic commonalities and disparities between domains. As shown in Fig. 1(c), domain-aware attention map highlights features with rich semantic information, *i.e.*foreground regions, and hinders the negative impact of background regions. Thereby, the proposed SEEN-DA correctly detects objects misclassified by existing methods, improving the discriminability of the detector on the target domain.

We evaluate the proposed method on four benchmarks. The experimental results highlight the effectiveness of our SEEN-DA, *e.g.*, it reaches 57.5% mAP on Cityscapes→Foggy Cityscapes, surpassing the state-of-the-art DA-Pro [23] by 1.6%.

## 2. Related Work

**Visual-Languague Models** Large-scale visual-language models (VLMs) [25, 26, 43] embed visual and text modalities into a shared space, enabling cross-modal alignment. Pre-trained with an astonishing scale of image-text pairs, they demonstrate comprehensive visual understanding. As a pioneering work, CLIP [43] trains a visual encoder and a textual encoder with 400 million image-text pairs in a contrastive manner, showing promising classification ability for the seen and unseen classes. Following this spirit, [10, 16, 52, 65] freeze the textual encoder and adapt the visual encoder into Faster-RCNN [45] as a two-stage detector. In this work, we apply the textual encoder of CLIP [43] to infer textual embedding that contains semantic information, and the visual encoder of RegionCLIP [65] as the backbone.

**Domain Adaptive Object Detection (DAOD)** aims to adapt the object detector [45] trained on the labelled source domain to the unlabelled target domain. Previous approaches can be broadly divided into three categories: feature alignment, semi-supervised learning and VLM-based domain alignment. As a fundamental direction, feature alignment [12, 37, 39] aims to align the feature distributions of the two domains to generate domain-invariant features [3, 30, 47]. Typically, the domain discriminator is proposed to align feature distribution adversarially in image level [3], instance level [47, 63] and category level [29, 53]. To reduce the negative impact of aligning unique features of each domain, [24, 32, 34, 49, 54, 58] decouple the domain-invariant and domain-specific knowledge between two domains. In parallel, semi-supervised learning strives to generate unbiased training data with style transfer [6, 19, 35] and pseudo label [2, 7, 31, 50]. However, they train the detector with semantically agnostic category labels (such as one-hot encoding), ignoring the semantic information inherently contained in the category name. To overcome this, recent VLM-based domain alignment methods exploit the semantic information provided by the textual encoder of VLMs to learn semantic commonalities and disparities. Pioneeringly, DA-Pro [23] tuning the textual encoder into a domain-aware detection head, learning domain-invariant semantics while supplementing domain-specific semantics for each domain. Despite the promising performance, existing methods freeze the visual encoder and ignore using semantic information to guide the learning of visual features, which leads to two limiations. First, they fail to filter out *redundant information* from visual features, which is semantically irrelevant and can lead to incorrect feature alignment. Second, they discard domain-specific features in the visual encoder, which contain rich domain semantic information and are vital for DAOD.

**Entropy Optimization** Information entropy was first utilized as entropy minimization in [15] for semi-supervised learning to learn decision rules for unlabelled data. Considering its effectiveness, recent works introduce entropy optimization into DAOD. [40, 41, 46, 62] propose minimizing entropy on unlabelled data to make decisions consistent in the source and target domains. [55] apply entropy as metrics to measure the transferability of data samples. [38] employs entropy as attention to control the uncertainty of the classifier predictions, alleviating the effect of negative transfer. Instead of directly minimizing entropy, [60] further introduces diversity maximization to explore close-to-ideal domain adaptation solutions. However, the computation of entropy in existing methods is inherently *semantically agnostic*. Based on the output probabilities from a classifier, they treat categories merely as identifiers and convert the category labels into one-hot encoding. Therefore, they ignore the rich semantic information embedded within the category names, which is vital for DAOD. In this paper, we propose a novel semantic entropy, evaluating the transferability of visual features through the semantic commonalities and disparities between domains, improving the discrimination of the visual encoder.

## 3. Methodology

In this section, we present a novel SEmantic ENtropy guided Domain-aware Attention (SEEN-DA) for DAOD, which introduces domain-aware attention module to eliminate redundant information and supplement domain-specific features, guided by the proposed semantic entropy. Semantic entropy quantifies the amount of semantic information within visual features, providing a measure of their importance in domain adaptation. Utilizing semantic entropy as attention, the domain-aware attention module adopts an inter-domain attention branch to extract domain-invariant features and eliminate semantically irrelevant redundant information, and an intra-domain attention branch to supplement the domain-specific semantic information discriminative on each domain. As shown in Fig. 2(a), given input image $x$, the visual encoder of SEEN-DA is divided into $N$ frozen visual block $\{\mathcal{V}_i\}_{i=1}^{N}$, and each domain-aware attention module $\{\mathcal{A}_i\}_{i=1}^{N}$ is attached to $\mathcal{V}_i$ as a lightweight learnable module:

$$\mathbf{h}_0 = \mathcal{E}(x); \mathbf{h}_i = \mathcal{A}_i(\mathcal{V}_i(\mathbf{h}_{i-1})), \tag{1}$$

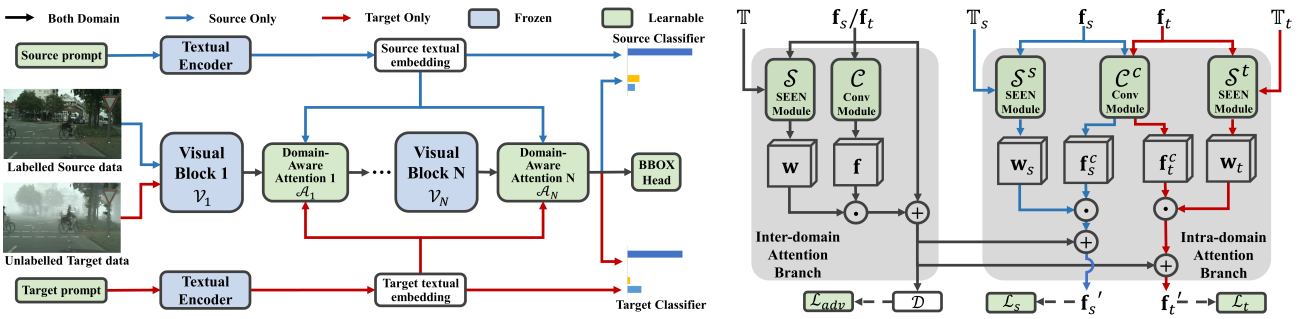
where $\mathcal{E}$ is the stem layer.

### 3.1. Semantic Entropy (SEEN)

Generally speaking, given feature $\mathbf{f}$, the probability that this feature belongs to a specific object category is:

$$p(c, \mathbf{f}) = \frac{\exp(v_c(\mathbf{f}))}{\sum_c \exp(v_c(\mathbf{f}))}, \tag{2}$$

where $c$, $v_c(\cdot)$ represents an object category and its corresponding classifier. $p(\cdot, \mathbf{f})$ is the probability distribution function, expected to output 1 on the corresponding category and 0 on other category, *i.e.*one-hot vector.




**(a) Overview of SEEN-DA** **(b) Domain-aware Attention Module**

Figure 2. (a) Overview of the proposed SEEN-DA for DAOD, where the semantic entropy is utilized as attention in domain-aware attention module. (b) The architecture of domain-aware attention module, consisting of an inter-domain and an intra-domain attention branch.

Tradition DAOD approaches focus on reducing domain discrepancy by aligning $\mathbf{f}$ between the source and target domain. Although effective, as shown in Eq. 2, they treat categories merely as one-hot vectors, which is inherently *semantically agnostic*. Therefore, they ignore the rich semantic information inherently embedded within the category names. As feature alignment relies heavily on a profound understanding of the semantic disparities between the source and target domains, this limitation hinders the learning of visual features for adaptation.

To address this, we propose semantic entropy, which quantifies the semantic information embedded in visual features by leveraging textual embeddings from Visual-language models (VLMs). VLMs utilize a textual encoder to encode textual prompts that describe category attributes into textual embeddings, aligning them with visual features. As a result, textual embeddings inherit rich semantic information from natural language, making them effective representations of the semantics of visual features. Motivated by this, we first infer textual embedding $\mathbb{T} = \{t_c\}_c^K$ for $K$ categories by feeding prompts like "A photo of $[Class]$ in $[Domain]$" into a textual encoder, where each $t_c$ contains semantic information of category $c$ and its domain. Thus, the probability distribution that $\mathbf{f}$ corresponds to certain semantics can be estimated by the similarity between the visual feature $\mathbf{f}$ and textual embedding $\mathbb{T}$:

$$p(t_c, \mathbf{f}) = \frac{\exp(\text{sim}(t_c, \mathbf{f})/\tau)}{\sum_c \exp(\text{sim}(t_c, \mathbf{f})/\tau)}, \tag{3}$$

where $\text{sim}(\cdot, \cdot)$ is the cosine similarity, and $\tau$ is a learnable temperature parameter. Though $p(t_c, \mathbf{f})$ provides a measure of semantic relevance, it does not directly quantify the amount of semantic information embedded in $\mathbf{f}$. Therefore, we draw inspiration from the concept of information entropy, which is widely used to measure the uncertainty and informativeness of a probability distribution. Formally, the proposed semantic entropy is defined as the information entropy of the probability distribution $p(t_c, \mathbf{f})$ to evaluate the semantic uncertainty of $\mathbf{f}$:

$$\text{SE}(\mathbb{T}, \mathbf{f}) = -\sum_c p(t_c, \mathbf{f}) \log(p(t_c, \mathbf{f})), \tag{4}$$

In the framework of DAOD, we apply semantic entropy as attention to select high discriminative visual features. According to Eq. 4, features with low semantic entropy exhibit clear and definite semantic information, expected to be foreground features suitable for transfer. Conversely, features with high semantic entropy indicate high semantic uncertainty, suggesting they may contain semantically irrelevant redundant information, which should be deemphasized during adaptation. To achieve this, we regulate Eq. 4 to propose semantic entropy-guided attention:

$$\text{SEAttention}(\mathbb{T}, \mathbf{f}) = \sum_c p(t_c, \mathbf{f}) \log(p(t_c, \mathbf{f})) + \log K, \tag{5}$$

where $K$ is the number of categories. Based on semantic entropy, Eq. 5 effectively adjusts the weight of visual features, focusing on features with rich semantic information while minimizing the influence of redundant information.

### 3.2. Domain-aware Attention Module

The domain-aware attention module is designed to adaptively refine visual features of each domain by utilizing the semantic entropy. As shown in Fig. 2(b), the module incorporates two complementary branches: an inter-domain attention branch to learn domain-invariant features and eliminate semantically irrelevant information, and an intra-domain attention branch to supplement the domain-specific features unique to each domain.

**Inter-domain Attention** The inter-domain attention is designed to modify domain-invariant features of the frozen visual encoder with semantic information. Specifically, for the output visual features $\mathbf{f}_d = \mathcal{V}_i(\mathbf{h}_{i-1})$ of the i-th visual block, we first forward it to a domain-shared convolution module $\mathcal{C}$ to capture task-related knowledge:

$$\mathbf{f} = \mathcal{C}(\mathbf{f}_d), \tag{6}$$

where $d \in \{\text{s}, \text{t}\}$ denotes whether $\mathbf{f}_d$ belongs to source domain s or target domain t.

Simultaneously, the visual feature $\mathbf{f}_d$ is forwarded to the SEEN module $\mathcal{S}$ to obtain attention map $\mathbf{w}$ guided by semantic entropy. With the help of VLM, we can represent the

domain-invariant semantic information through domain-shared prompts. Technically, we input textual prompts such as "A photo of $[Class]$" to the textual encoder to retrieve domain-invariant textual embedding $\mathbb{T}$. As described in Eq. 4, the similarity of visual features and textual embeddings is needed to calculate semantic entropy attention. To facilitate this, a projection layer $\mathcal{P}$ is firstly introduced to map $\mathbf{f}_d$ from the visual feature space to the textual embedding space:

$$\mathbf{f}_{\mathrm{p}} = \mathcal{P}(\mathbf{f}_d), \tag{7}$$

Then we calculate the semantic entropy-guided attention of the visual feature $\mathbf{f}_d$ over text embedding $\mathbb{T}$:

$$\mathbf{w} = \mathcal{S}(\mathbb{T}, \mathbf{f}_d) = \mathrm{SEAttention}(\mathbb{T}, \mathbf{f}_{\mathrm{p}}), \tag{8}$$

With the attention map $\mathbf{w}$, we can focus on semantically informative features and suppress redundant information in domain-invariant features. Typically, $\mathbf{w}$ is applied to $\mathbf{f}$ via an element-wise product, and the weighted feature is summed with the original visual feature:

$$\mathbf{f}' = \mathbf{w} \cdot \mathbf{f} + \mathbf{f}_d, \tag{9}$$

Expecting to be domain-invariant, the extracted $\mathbf{f}'$ is forwarded to a domain discriminator $\mathcal{D}$, which aligns the source and target domain adversarially.

**Intra-domain Attention** With the inter-domain attention branch, the extracted domain-invariant features contain sufficient semantic information shared between domains. However, it ignores the semantic discrepancy between the two domains, resulting in limited discriminability in the target domain. Therefore, we design the intra-domain attention branch to enrich visual features with domain-specific semantic information. Similar to the inter-domain attention branch, we first introduce a learnable layer $\mathcal{C}^{\mathrm{c}}$ to adapt visual features for to both domains:

$$\mathbf{f}_d^{\mathrm{c}} = \mathcal{C}^{\mathrm{c}}(\mathbf{f}_d), \tag{10}$$

Subsequently, we generate domain-aware textual embedding that captures the semantics of each domains. Motivated by DA-Pro [23], we tune learnable domain-adaptive prompts, *i.e.*"$[v^{\mathrm{c}}][v^{\mathrm{s}}][Class]$" for the source domain and "$[v^{\mathrm{c}}][v^{\mathrm{t}}][Class]$" for the target domain, to explicitly model the domain and category semantic commonalities and disparities. These prompts are forwarded to the textual encoder to derive domain-adaptive textual embedding $\mathbb{T}_{\mathrm{s}}, \mathbb{T}_{\mathrm{t}}$ for each domain.

Due to the visual and semantic differences between the source and target domain, a domain-shared SEEN module is inadequate. To calculate the semantic entropy for each domain, visual features must be independently mapped into their respective textual spaces. Therefore, we apply separate SEEN modules $\mathcal{S}^{\mathrm{s}}$ and $\mathcal{S}^{\mathrm{t}}$ with independent projection layers $\mathcal{P}^{\mathrm{s}}, \mathcal{P}^{\mathrm{t}}$ and textual embedding $\mathbb{T}_{\mathrm{s}}, \mathbb{T}_{\mathrm{t}}$:

$$\mathbf{w}_{\mathrm{s}} = \mathcal{S}^{\mathrm{s}}(\mathbb{T}_{\mathrm{s}}, \mathbf{f}_{\mathrm{s}}), \tag{11}$$

$$\mathbf{w}_{\mathrm{t}} = \mathcal{S}^{\mathrm{t}}(\mathbb{T}_{\mathrm{t}}, \mathbf{f}_{\mathrm{t}}), \tag{12}$$

where the $\mathbf{w}_{\mathrm{s}}$ and $\mathbf{w}_{\mathrm{t}}$ denotes the source- and target-specific attention maps.

The inter-domain attention branch extracts domain-invariant features while eliminating redundant information, whereas the intra-domain attention branch enriches visual features with domain-specific semantics. Overall, the output for the i-th domain-aware attention module $\mathcal{A}_i$ can be expressed as follows:

$$\mathbf{f}'_{\mathrm{s}} = \mathcal{A}_i(\mathbf{f}_{\mathrm{s}}) = \mathbf{f}' + \mathbf{w}_{\mathrm{s}} \cdot \mathbf{f}_{\mathrm{s}}^{\mathrm{c}}, \tag{13}$$

$$\mathbf{f}'_{\mathrm{t}} = \mathcal{A}_i(\mathbf{f}_{\mathrm{t}}) = \mathbf{f}' + \mathbf{w}_{\mathrm{t}} \cdot \mathbf{f}_{\mathrm{t}}^{\mathrm{c}}, \tag{14}$$

where $\mathbf{f}'_{\mathrm{s}}$ and $\mathbf{f}'_{\mathrm{t}}$ are output features of $\mathcal{A}_i$ for the source and target domain. Since $\mathbf{w}_{\mathrm{s}}$ and $\mathbf{w}_{\mathrm{t}}$ are involved in the computation of $\mathbf{f}'_{\mathrm{s}}$ and $\mathbf{f}'_{\mathrm{t}}$, SEEN modules $\mathcal{S}^{\mathrm{s}}$ and $\mathcal{S}^{\mathrm{t}}$ will be optimized by the classification losses from the two domains respectively, thereby learning domain-specific semantic information.

### 3.3. Optimization Objective

We introduce multiple domain-aware attention modules into the frozen visual encoder as learnable attachments. To this end, we expect them to capture visual features suitable for cross-domain adaptation and supplement domain-specific semantic information. First, the domain discriminator $\mathcal{D}$ is applied to the inter-domain attention branch to learn domain-invariant features. Formally, we minimize the adversarial loss over the output of the inter-domain attention branch $\mathbf{f}'$ for the source image $x_{\mathrm{s}}$ and the target image $x_{\mathrm{t}}$:

$$\mathcal{L}_{\mathrm{adv}} = -\sum_{i=1}^{N} [\mathbb{E}_{x_{\mathrm{s}}} ||\mathcal{D}_i(\mathbf{f}')||_2^2 + \mathbb{E}_{x_{\mathrm{t}}} ||\mathcal{D}_i(\mathbf{f}') - \mathbf{1}||_2^2], \tag{15}$$

Second, the detection loss is calculated over both domains to learn the semantic commonalities and disparities. For the source domain, we minimize the cross-entropy $\mathcal{L}_{\mathrm{s}}$ between the logits and ground truth. For the target domain, we use the prompt "A photo of $[Class]$ in foggy day" to generate high-confidence pseudo labels and minimize the cross-entropy $\mathcal{L}_{\mathrm{t}}$. Overall, the optimization objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{s}} + \lambda_{\mathrm{t}} \mathcal{L}_{\mathrm{t}} + \lambda_{\mathrm{adv}} \mathcal{L}_{\mathrm{adv}} + \mathcal{L}_{\mathrm{reg}}, \tag{16}$$

where $\mathcal{L}_{\mathrm{reg}}$ is the regression loss, and $\lambda_{\mathrm{t}}, \lambda_{\mathrm{adv}}$ are balance ratios.

Table 1. Comparison (%) with existing methods on Cross-Weather adaptation Cityscapes→Foggy Cityscapes (C→F), Cross-Fov adaptation KITTI→Cityscapes (K→C) and Sim-to-Real adaptation SIM10K→Cityscapes (S→C).

| Methods | Venue | C→F | | | | | | | | | K→C | S→ C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Person | Rider | Car | Truck | Bus | Train | Motor | Bicycle | mAP | mAP | mAP |
| DA-Faster [3] | CVPR 2018 | 29.2 | 40.4 | 43.4 | 19.7 | 38.3 | 28.5 | 23.7 | 32.7 | 32.0 | 41.9 | 38.2 |
| VDD [59] | ICCV 2021 | 33.4 | 44.0 | 51.7 | 33.9 | 52.0 | 34.7 | 34.2 | 36.8 | 40.0 | - | - |
| SCAN [28] | AAAI 2022 | 41.7 | 43.9 | 57.3 | 28.7 | 48.6 | 48.7 | 31.0 | 37.3 | 42.1 | 45.8 | 52.6 |
| TIA [64] | CVPR 2022 | 52.1 | 38.1 | 49.7 | 37.7 | 34.8 | 46.3 | 48.6 | 31.1 | 42.3 | 44.0 | - |
| LRA [42] | TNNLS 2024 | 45.6 | 47.1 | 59.7 | 31.2 | 52.4 | 44.6 | 28.1 | 39.5 | 43.5 | 49.4 | 55.7 |
| SIGMA++ [30] | TPAMI 2023 | 46.4 | 45.1 | 61.0 | 32.1 | 52.2 | 44.6 | 34.8 | 39.9 | 44.5 | 49.5 | 57.7 |
| CIGAR [36] | CVPR 2023 | 46.1 | 47.3 | 62.1 | 27.8 | 56.6 | 44.3 | 33.7 | 41.3 | 44.9 | 48.5 | 58.5 |
| OADA [61] | ECCV 2022 | 47.8 | 46.5 | 62.9 | 32.1 | 48.5 | 50.9 | 34.3 | 39.8 | 45.4 | 47.8 | 59.2 |
| MTM [57] | AAAI 2024 | 51.0 | 53.4 | 67.2 | 37.2 | 54.4 | 41.6 | 38.4 | 47.7 | 48.9 | - | 58.1 |
| AT [31] | CVPR 2022 | 56.3 | 51.9 | 64.2 | 38.5 | 45.5 | **55.1** | **54.3** | 35.0 | 50.9 | - | - |
| SOCCER [5] | ACMMM 2024 | 51.7 | 57.7 | 68.6 | 38.2 | 51.6 | 47.5 | 41.6 | 51.7 | 51.1 | - | 63.8 |
| DSD-DA [11] | ICML 2024 | 49.0 | 59.6 | 65.3 | 35.7 | 61.0 | 46.5 | 43.9 | 57.3 | 52.3 | 49.3 | 52.5 |
| CAT [22] | CVPR 2024 | 44.6 | 57.1 | 63.7 | 40.8 | **66.0** | 49.7 | 44.9 | 53.0 | 52.5 | - | - |
| NSA-UDA [66] | ICCV 2023 | 50.3 | 60.1 | 67.7 | 37.4 | 57.4 | 46.9 | 47.3 | 54.3 | 52.7 | 55.6 | 56.3 |
| REACT [24] | TIP 2024 | 51.4 | 57.9 | 67.4 | 37.7 | 58.4 | 52.8 | 44.6 | 54.6 | 53.1 | 59.5 | 58.6 |
| DA-Pro [23] | NeurIPS 2023 | 55.4 | 62.9 | 70.9 | 40.3 | 63.4 | 54.0 | 42.3 | 58.0 | 55.9 | 61.4 | 62.9 |
| RegionCLIP [65] | CVPR 2022 | 49.6 | 55.0 | 63.2 | 34.1 | 55.6 | 48.3 | 36.0 | 47.0 | 48.6 | 59.1 | 58.9 |
| SEEN-DA (Ours) | - | **58.5** | **64.5** | **71.7** | **42.0** | 61.2 | 54.8 | 47.1 | **59.9** | **57.5** | **67.1** | **66.8** |

Table 2. Comparison (%) with existing methods on Cross-Style adaptation task Pascal VOC→Clipart

| Methods | Aero | Bike | Bird | Boat | Bottle | Bus | Car | Cat | Chair | Cow | Table | Dog | Horse | Motor | Person | Plant | Sheep | Sofa | Train | Tv | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| UaDAN [17] | 35.0 | **73.7** | 41.0 | 24.4 | 21.3 | 69.8 | 53.5 | 2.3 | 34.2 | 61.2 | 31.0 | **29.5** | 47.9 | 63.6 | 62.2 | **61.3** | 13.9 | 7.6 | 48.6 | 23.9 | 40.2 |
| TFD [54] | 27.9 | 64.8 | 28.4 | 29.5 | 25.7 | 64.2 | 47.7 | 13.5 | 47.5 | 50.9 | **50.8** | 21.3 | 33.9 | 60.2 | 65.6 | 42.5 | 15.1 | 40.5 | 45.5 | 48.6 | 41.2 |
| FGRR [1] | 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 |
| UMT [6] | 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 |
| SIGMA [29] | 40.1 | 55.4 | 37.4 | 31.1 | 54.9 | 54.3 | 46.6 | 23.0 | 44.7 | 65.6 | 23.0 | 22.0 | 42.8 | 55.6 | 67.2 | 55.2 | 32.9 | **40.8** | 45.0 | 58.6 | 44.5 |
| ATMT [27] | 37.5 | 63.4 | 37.9 | 29.8 | 45.1 | 62.7 | 41.2 | 19.5 | 43.7 | 57.4 | 22.9 | 25.3 | 39.6 | 87.1 | 70.9 | 50.6 | 29.1 | 32.2 | 58.4 | 50.5 | 45.2 |
| CIGAR [36] | 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 |
| TIA [64] | 42.2 | 66.0 | 36.9 | 37.3 | 43.7 | 71.8 | 49.7 | 18.2 | 44.9 | 58.9 | 18.2 | 29.1 | 40.7 | **87.8** | 67.4 | 49.7 | 27.4 | 27.8 | 57.1 | 50.6 | 46.3 |
| SIGMA++ [30] | 36.3 | 54.6 | 40.1 | 31.6 | 58.0 | 60.4 | 46.2 | **33.6** | 44.4 | 66.2 | 25.7 | 25.3 | 44.4 | 58.8 | 64.8 | 55.4 | 36.2 | 38.6 | 54.1 | 59.3 | 46.7 |
| CMT [51] | 39.8 | 56.3 | 38.7 | 39.7 | **60.4** | 35.0 | **56.0** | 7.1 | **60.1** | 60.4 | 35.8 | 28.1 | **67.8** | 84.5 | **80.1** | 55.5 | 20.3 | 32.8 | 42.3 | 38.2 | 47.0 |
| RegionCLIP [65] | 38.1 | 70.4 | 48.8 | 37.3 | 44.8 | 55.8 | 43.5 | 14.4 | 48.2 | 47.8 | 14.3 | 18.3 | 58.3 | 78.4 | 67.9 | 22.2 | 30.1 | 16.9 | 48.4 | 50.2 | 42.7 |
| SEEN-DA(Ours) | **44.1** | 73.4 | **54.7** | **47.1** | 45.1 | **76.0** | 51.6 | 20.4 | 51.7 | 53.0 | 18.5 | 17.3 | 61.8 | 86.8 | 72.2 | 22.8 | **37.7** | 21.1 | **58.9** | 52.7 | **47.9** |

# 4. Experiment

We evaluate our method on mainstream DAOD scenarios, including the Cross-Weather, Cross-Fov, Sim-to-Real and Cross-Style. Further ablation studies are conducted to validate the effectiveness of the proposed domain-aware attention module.

## 4.1. Datasets

**Cross-Weather** Cityscapes [4] contains diverse street scenes in daylight, consisting of 2,975 training and 500 validation images annotated with eight classes. Foggy Cityscapes [48] simulates three distinct densities of fog on Cityscapes, containing 8,925 training images and 1,500 validation images. A standard configuration for cross-weather adaptation is to take the training set of Cityscapes as the source domain and the training set of foggy Cityscapes as the target domain, evaluating cross-weather adaptation performance on the 1500-sized validation set in all eight categories.

**Cross-FoV** KITTI [14] is a crucial dataset for self-driving that includes 7,481 photos annotated with cars. Collected by driving in rural areas and on highways, it provides data with a different Field of View (FoV). To fairly compare with other methods, we migrate KITTI to Cityscapes solely on the car category.

**Sim-to-Real** The synthetic dataset SIM10k [21] has 10,000 photos from the GTA V video game with bounding boxes of the class car. We follow existing works to perform sim-to-real adaptation and report the performance on class car.

**Cross-Style** Pascal VOC [9] dataset consists of 16551 images from the real world and has 20 categories. Clipart [20] contains 1000 comical images and has the same 20 categories as Pascal VOC. Following the mainstream splitting, we use Pascal VOC as the source domain and all Clipart images as the target domain.

Table 3. Comparison (%) of standard attention with domain-aware attention.

| Benchmark | Self-Attention | Cross-Attention | Domain-aware Attention |
|---|---|---|---|
| Cross-Weather | 54.8 | 55.5 | **57.5** |
| Cross-FoV | 61.9 | 63.5 | **67.1** |
| Sim-to-Real | 62.3 | 63.7 | **66.8** |
| Cross-Style | 43.1 | 45.2 | **47.9** |

Table 4. Ablation studies (%) of domain-aware attention on Cross-Weather adaptation.

| Inter-domain Attention | $\mathcal{L}_{adv}$ | Intra-domain Attention | mAP | Gains |
|---|---|---|---|---|
| | | | 52.6 | - |
| ✓ | | | 54.9 | +2.3 |
| ✓ | ✓ | | 55.8 | +0.9 |
| ✓ | ✓ | ✓ | 57.5 | +1.7 |

## 4.2. Implementation Details

Following [23], we adapt RegionCLIP(ResNet-50 [18]) with the Faster-RCNN [45] as the detector. In each iteration, one batch of source images with ground truth and one batch of target domain images with pseudo labels are forwarded to calculate the classification, adversarial and regression loss. The hyperparameter $\lambda_{adv}, \lambda_t$ is set to 0.1 and 1.0. We set the batch size of each domain to 8 and use the SGD optimizer with a warm-up learning rate. Mean Average Precision (mAP) with a threshold of 0.5 is taken as the evaluation metric. All experiments are deployed on 8 Tesla V100 GPUs.

## 4.3. Comparison to SOTA methods

We present representative state-of-the-art DAOD approaches for comparison, including feature alignment, semi-supervised learning and VLM-based domain alignment methods.

**Cross-Weather Adaptation Scenario** As shown in Table 1 (C→F), the proposed SEEN-DA outperforms all compared methods in terms of mAP and advances SOTA by 1.6%, from 55.9% to 57.5%. Specifically, our method improves performance over six categories (*i.e.* person, rider, car, truck, train, and bicycle) ranging from 0.8% to 3.1%.

**Cross-FOV Adaptation Scenario** Table 1(K→C) reports result for KITTI→Cityscapes. SEEN-DA achieves SOTA performance of 67.1% mAP, gaining an improvement of 5.7%.

**Sim-to-Real Adaptation Scenario** Table 1 (S→C) shows that the proposed method achieves the best results of 66.8% mAP, outperforming the previous best entry SOCCER [5] 63.8% with 3.0%.

**Cross-Style Adaptation Scenario** Additionally, we assess SEEN-DA on the more challenging Cross-Style adaptation, where the semantic hierarchy has broader discrepancies. In Table 2, SEEN-DA peaks with 47.9% mAP and improves six categories (aeroplane, bird, boat, bus, sheep and train).

Table 5. Comparison (%) of computational efficiency on Cross-Weather adaptation

| Method | Backbone Parm(M) | Learnable Parm(M) | mAP | Abs. Gains |
|---|---|---|---|---|
| DSS [56] | 26.834 | 26.834 | 40.9 | +4.2 |
| CSDA [13] | 33.645 | 33.645 | 45.3 | +6.9 |
| REACT [24] | 29.812 | 29.812 | 48.8 | **+7.9** |
| AT [31] | **39.225** | 18.723 | 50.9 | **+7.9** |
| DA-Pro [23] | 34.834 | **0.008** | 55.9 | +3.3 |
| SEEN-DA (Ours) | 36.701 | 1.875 | **57.5** | +4.9 |

## 4.4. Ablation Studies

We report detailed ablation studies (Table 3, 4, 5, Figure 4, 5) to validate the effectiveness of proposed SEEN.

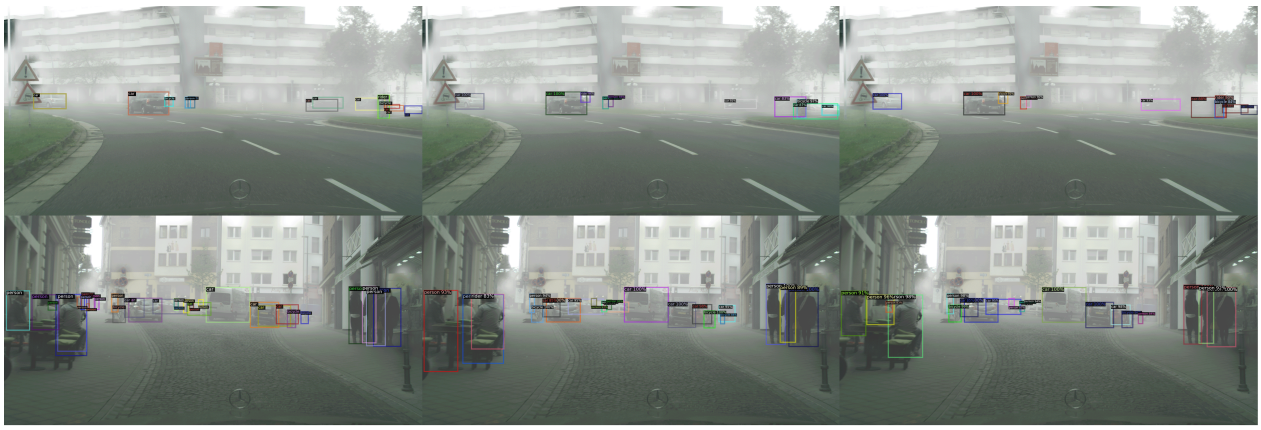
**Standard Attention *vs.* Domain-aware Attention** We compare the performance of domain-aware attention with existing attention methods, including self-attention and cross-attention, as shown in Table 3. With the guidance of textual information, the cross-attention improves the self-attention by 0.7% ∼ 2.1%. Furthermore, the proposed domain-aware attention gains improvements of 2.0% ∼ 33.6% over cross-attention, showing a superior ability to select features with high transferability.

**Ablation for Domain-aware Attention Module** We conduct ablation experiments to further analyze the significance of proposed domain-aware attention module in Table 4. Introducing the inter-domain attention branch attains an mAP of 54.9%, and optimizing with adversarial loss $\mathcal{L}_{adv}$ gains 0.9%. Meanwhile, applying the intra-domain attention branch improves 1.7% on the target domain. These results showing the effectiveness of the domain-aware attention module to eliminate redundant information and supplement domain-specific semantic information.

**Computational Efficiency** We discussed the computational efficiency of the proposed SEEN-DA in Table 5. Our method yields the same parameters scale as existing methods and achieves the highest mAP. In addition, SEEN-DA only needs to train very few parameters (1.875M), showing remarkable efficiency.

**Numbers of the Attention Module** We explore the impact of numbers of domain-aware attention modules in Figure 4. Starting from the last visual block, gradually introducing domain-aware attention modules to each visual block improves performance and peaks when using 3 modules. Considering efficiency, we apply 3 domain-aware attention modules into the visual encoder.

**Projection layer in SEEN module** We explore the projection layer in the SEEN module in Figure 5. Projecting textual embedding into visual space, denoted as T2V, only achieves 55.1%. Projecting textual and visual embedding into a latent space, denoted as T&V, improves performance by 2.1%. And projecting visual embedding into textual space achieves the best mAP of 57.5%. We analyze the reason for these results as follows. Visual embedding contains more sample noise and redundant information than textual



(a) Ground Truth (b) DA-Pro[23] (c) SEEN-DA (Ours)

Figure 3. Detection comparison on the Cross-Weather adaptation scenario. We visualize (a) the ground truth, the detection boxes of (b) SOTA DA-Pro [23] and (c) our methods SEEN-DA.

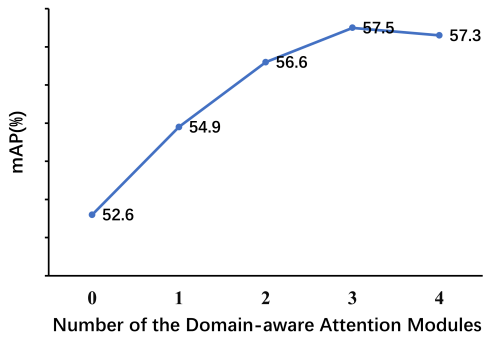



Figure 4. Ablation (%) on the number of attention modules on Cross-Weather adaptation.

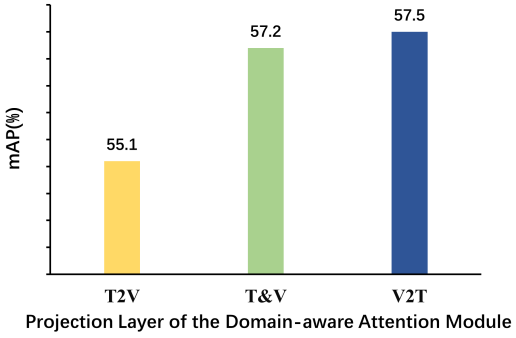



Figure 5. Ablation (%) on projection layer of the domain-aware attention module on Cross-Weather adaptation.

embedding. And learning to project visual embedding to textual space can reduce the interference of these factors. Therefore, we adopt the V2T projection in SEEN module.

### 4.5. Visualization

The detection results of ground truth (a), DA-Pro [23] (b) and the proposed SEEN-DA (c) on the target domain of Cross-Weather adaptation are displayed in Figure 3. The figure shows that SEEN-DA outperforms the SOTA method in terms of the ability to detect objects. Specifically, SEEN-DA detects the objects missed (*e.g.*the rider on the right of figures in 1st row and the person on the left of figures in 2nd row) and misclassified (*e.g.*the bicycle on the middle of figures in 1st row and the person on the left of figures in 2nd row), revealing the effectiveness of domain-aware attention. By utilizing semantic entropy to reduce redundant information and supplement domain-specific semantic information, the SEEN-DA improves the discriminability of the detector on the target domain.

## 5. Conclusion

In this paper, we propose a novel SEmantic ENtropy guided Domain-aware Attention (SEEN-DA) for DAOD. We design the semantic entropy to quantify the semantic information contained in visual features. Guided by semantic entropy, the domain-aware attention module is introduced to adaptively refine visual features with the semantic commonalities and disparities between domains. Precisely, it consists of an inter-domain attention branch for extracting domain-invariant features and eliminating redundant information between domains, and an intra-domain attention branch for supplementing the domain-specific semantic information for each domain. Extensive experiments over multiple benchmarks validate the effectiveness of SEEN-DA in inferring a discriminative detector.

**Acknowledgment** This work is partially supported by the National Key R&D Program of China(under Grant 2023YFB4502200), the NSF of China(under Grants 92364202, U22A2028, 62102398, 62102399, 62302478, 62302482, 62376268), Strategic Priority Research Program of the Chinese Academy of Sciences, (Grant No.XDB0660300, XDB0660301, XDB0660302), Major Program of ISCAS(Grant No. ISCAS-ZD-202402), Beijing Natural Science Foundation(4222039), CAS Project for Young Scientists in Basic Research(YSBR-029) and Youth Innovation Promotion Association CAS.

## References


[1] 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*, 45(03):3677–3694, 2023. 6

[2] 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. In *ICML*, pages 3040–3055, 2022. 3

[3] 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, 3, 6

[4] 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*, 2016. 6

[5] Yiming Cui, Liang Li, Jiehua Zhang, Chenggang Yan, Hongkui Wang, Shuai Wang, Heng Jin, and Li Wu. Stochastic context consistency reasoning for domain adaptive object detection. In *ACMMM*, pages 1331–1340, 2024. 6, 7

[6] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, 2021. 3, 6

[7] Jinhong Deng, Dongli Xu, Wen Li, and Lixin Duan. Harmonious teacher for cross-domain object detection. In *CVPR*, pages 23829–23838, 2023. 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[9] 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

[10] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul De Charette. Poda: Prompt-driven zero-shot domain adaptation. In *ICCV*, pages 18623–18633, 2023. 3

[11] Yongchao Feng, Shiwei Li, Yingjie Gao, Ziyue Huang, Yanan Zhang, Qingjie Liu, and Yunhong Wang. Dsd-da: Distillation-based source debiasing for domain adaptive object detection. In *ICML*, 2024. 6

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016. 3

[13] Changlong Gao, Chengxu Liu, Yujie Dun, and Xueming Qian. Csda: Learning category-scale joint feature for domain adaptive object detection. In *ICCV*, pages 11421–11430, 2023. 7

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 6

[15] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *NeurIPS*, 17, 2004. 3

[16] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 3

[17] Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, and Yanpeng Cao. Uncertainty-aware unsupervised domain adaptation in object detection. *TMM*, 24:2502–2514, 2021. 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7

[19] Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao. Cross domain object detection by target-perceived dual branch distillation. In *CVPR*, pages 9570–9580, 2022. 3

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

[21] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *ICRA 2017*, pages 746–753. IEEE, 2017. 6

[22] Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T Tan. Cat: Exploiting inter-class dynamics for domain adaptive object detection. In *CVPR*, pages 16541–16550, 2024. 6

[23] Haochen Li, Rui Zhang, Hantao Yao, Xinkai Song, Yifan Hao, Yongwei Zhao, Ling Li, and Yunji Chen. Learning domain-aware detection head with prompt tuning. *NeurIPS*, 36, 2024. 1, 2, 3, 5, 6, 7, 8

[24] Haochen Li, Rui Zhang, Hantao Yao, Xin Zhang, Yifan Hao, Xinkai Song, and Ling Li. React: Remainder adaptive compensation for domain adaptive object detection. *TIP*, 33:3735–3748, 2024. 1, 3, 6, 7

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 1, 3

[27] Kai Li, Curtis Wigington, Chris Tensmeyer, Vlad I. Morariu, Handong Zhao, Varun Manjunatha, Nikolaos Barmpalios, and Yun Fu. Improving cross-domain detection with self-supervised learning. In *CVPR*, pages 4745–4754, 2023. 6

[28] Wuyang Li, Xinyu Liu, Xiwen Yao, and Yixuan Yuan. Scan: Cross domain object detection with semantic conditioned adaptation. In *AAAI*, volume 6, page 7, 2022. 6

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

[30] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. *TPAMI*, 45(07):9022–9040, 2023. 1, 3, 6

[31] 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. 3, 6, 7

[32] 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. 1, 3

[33] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 1

[34] Dongnan Liu, Chaoyi Zhang, Yang Song, Heng Huang, Chenyu Wang, Michael Barnett, and Weidong Cai. Decompose to adapt: Cross-domain object detection via feature disentanglement. *TMM*, 25:1333–1344, 2022. 3

[35] Rui Liu, Yahong Han, Yaowei Wang, and Qi Tian. Frequency spectrum augmentation consistency for domain adaptive object detection. *arXiv preprint arXiv:2112.08605*, 2021. 3

[36] Yabo Liu, Jinghua Wang, Chao Huang, Yaowei Wang, and Yong Xu. Cigar: Cross-modality graph reasoning for domain adaptive object detection. In *CVPR*, pages 23776–23786, 2023. 6

[37] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105. PMLR, 2015. 3

[38] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *NeurIPS*, 31, 2018. 3

[39] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, pages 2208–2217, 2017. 3

[40] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulo. Autodial: Automatic domain alignment layers. In *ICCV*, pages 5067–5075, 2017. 3

[41] Pietro Morerio, Jacopo Cavazza, and Vittorio Murino. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. In *ICLR*, 2018. 3

[42] Zhengquan Piao, Linbo Tang, and Baojun Zhao. Unsupervised domain-adaptive object detection via localization regression alignment. *TNNLS*, 35(11):15170–15181, 2024. 6

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 3

[44] 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

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 1, 3, 7

[46] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *ICCV*, pages 8050–8058, 2019. 3

[47] 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

[48] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, Sep 2018. 6

[49] Sunandini Sanyal, Ashish Ramayee Asokan, Suvaansh Bhambri, Akshay Kulkarni, Jogendra Nath Kundu, and R Venkatesh Babu. Domain-specificity inducing transformers for source-free domain adaptation. In *ICCV*, pages 18928–18937, 2023. 1, 3

[50] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NPIS*, 30, 2017. 3

[51] VS Vibashan, Poojan Oza, and Vishal M Patel. Instance relation graph guided source-free domain adaptive object detection. In *CVPR*, pages 3520–3530, 2023. 6

[52] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *CVPR*, pages 3219–3229, 2023. 3

[53] Vibashan Vs, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *CVPR*, pages 4516–4526, 2021. 3

[54] Haoan Wang, Shilong Jia, Tieyong Zeng, Guixu Zhang, and Zhi Li. Triple feature disentanglement for one-stage adaptive object detection. In *AAAI*, pages 5401–5409, 2024. 3, 6

[55] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. In *AAAI*, volume 33, pages 5345–5352, 2019. 3

[56] 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. 7

[57] Weixi Weng and Chun Yuan. Mean teacher detr with masked feature alignment: A robust domain adaptive detection transformer framework. In *AAAI*, pages 5912–5920, 2024. 6

[58] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. *TPAMI*, 44(8):4178–4193, 2021. 1, 3

[59] 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. 6

[60] Xiaofu Wu, Suofei Zhang, Quan Zhou, Zhen Yang, Chunming Zhao, and Longin Jan Latecki. Entropy minimization versus diversity maximization for domain adaptation. *TNNLS*, 34(6):2896–2907, 2021. 3

[61] Jayeon Yoo, Inseop Chung, and Nojun Kwak. Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In *ECCV*, pages 691–708, 2022. 6

[62] Yifan Zhang, Hanbo Chen, Ying Wei, Peilin Zhao, Jiezhang Cao, Xinjuan Fan, Xiaoying Lou, Hailing Liu, Jinlong Hou,

Xiao Han, et al. From whole slide imaging to microscopy: Deep microscopy adaptation network for histopathology cancer image classification. In *MICCAI*, pages 360–368, 2019. 3

[63] Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *CVPR*, pages 12425–12434, 2021. 3

[64] Liang Zhao and Limin Wang. Task-specific inconsistency alignment for domain adaptive object detection. In *CVPR*, pages 14217–14226, 2022. 6

[65] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022. 1, 3, 6

[66] Wenzhang Zhou, Heng Fan, Tiejian Luo, and Libo Zhang. Unsupervised domain adaptive detection with network stability analysis. In *ICCV*, pages 6986–6995, 2023. 6