Semi-Supervised Domain Generalization for Object Detection via Language-Guided Feature Alignment

Sina Malakouti
sem238@pitt.edu
Adriana Kovashka
kovashka@cs.pitt.edu
University of Pittsburgh,
Pittsburgh, PA, USA

Abstract

Existing domain adaptation (DA) and generalization (DG) methods in object detection enforce feature alignment in the visual space but face challenges like object appearance variability and scene complexity, which make it difficult to distinguish between objects and achieve accurate detection. In this paper, we are the first to address the problem of semi-supervised domain generalization by exploring vision-language pre-training and enforcing feature alignment through the language space. We employ a novel Cross-Domain Descriptive Multi-Scale Learning (CDDMSL) aiming to maximize the agreement between descriptions of an image presented with different domain-specific characteristics in the embedding space. CDDMSL significantly outperforms existing methods, achieving 11.7% and 7.5% improvement in DG and DA settings, respectively. Comprehensive analysis and ablation studies confirm the effectiveness of our method, positioning CDDMSL as a promising approach for domain generalization in object detection tasks. Our code is available at https://github.com/sinamalakouti/CDDMSL.

1 Introduction

The recent success of fully-supervised object detectors (FSOD) heavily relies on the assumption that training data follows the same distribution as test data, which often fails in real-world applications. Domain adaptation (DA) addresses distribution alignment between a source and a particular target domain but struggles to generalize to domains unseen during training, limiting its real-world applicability [26, 49, 58]. On the other hand, domain generalization (DG) aims to learn a universal model that can generalize to any unseen domain. While DG has been studied in object recognition, DG in object detection (DGOD) is still heavily understudied. Pioneering work [26] utilizes multiple fully annotated source domains to learn domain-invariant features. However, obtaining multiple fully annotated labeled datasets is a formidable challenge in object detection.

Fortunately, obtaining unlabeled data is less intricate, and a large volume is available. To tackle the issue above, we formulate DGOD as a semi-supervised learning (SSL) paradigm, using one fully annotated and an additional unlabeled source domain. While semi-supervised

(a) Motivation (b) Consistency Objective

Figure 1: (a) Demonstrates that an image-captioning model, when trained with RegionCLIP, fails to produce descriptions across different styles. In contrast, CDDMSL effectively maintains overall image semantics across domains. (b) Our consistency objective is designed to preserve the overarching image semantics regardless of differing styles (green) while discouraging unrelated descriptions (red).

domain generalization (SSDG) [28, 42, 59, 60] has been studied in image recognition, to the best of our knowledge, this is the first work addressing SSDG in object detection.

Generally, existing DA and DG for object detection works learn invariant representations or disentangle domain-specific representations by enforcing their objectives in the visual space [9, 19, 25, 33, 53, 41, 49, 56]. However, disentangling domain-specific and domain-invariant features in object detection is extremely difficult due to inherent challenges such as variability in object appearance (e.g., size, shape), and scene complexity, which can make it difficult to distinguish between objects and accurately detect them. We believe high-level information, such as object relations, visible object characteristics, or actions being performed, can help perform better under extreme domain shifts. Extracting such contextual information is infeasible using visual features solely, but textual descriptions offer a semantically denser signal [10] that can help the model to focus on the underlying semantic information of the data, resulting in better localization and recognition of objects in complex scenes. Also, natural language descriptions provide more information than simple object labels; the latter may be ambiguous or insufficient. Thus, we enforce semantic consistency in the textual descriptions of images to ensure that useful information is preserved across domains.

We believe that by incorporating natural language captions in object detection models, the models can learn to associate object categories with their names and other semantic attributes, such as actions, relations, and context. This can help the models better generalize to new or unseen domains where objects' visual appearance may differ from what was observed in the training data. Note that previous works have shown the benefit of using image captions in object detection tasks, such as improving the performance of weakly supervised object detectors and generalizing over novel categories in open-vocabulary object detection tasks [7, 15, 44, 52, 62, 64, 67]. However, to the best of our knowledge, this is the first work to explore the benefits of incorporating image descriptions and vision-language pre-training in the context of generalizability to unseen domains for object detection.

In this paper, we first make three observations: (i) vision-language models, such as CLIP [36], have shown promise in improving generalizability in image recognition tasks, (ii) captions contain rich semantic information that can be helpful in both object localization and detection, and (iii) as illustrated in Fig. 1a, domain shift impacts the ability of image-captioning models to produce semantically consistent descriptions for the same image with different domain-specific characteristics (i.e., style). Based on these observations, we present

a simple yet effective novel approach called Cross-Domain Descriptive Multi-Scale Learning (CDDMSL) for object detection, which enforces learning domain-invariant features on the visual encoder through the language space. In particular, we first propose to initialize the object detector's backbone with RegionCLIP [57] weights, a state-of-the-art vision-language model in object detection, and adopt its text classifier to facilitate supervised object detection on the labeled data. Second, we exploit unlabeled data to learn domain-invariant features while preserving the semantics. To this end, we utilize the vision-to-language module of an image-captioning model to obtain descriptive features by projecting image features to the language space. To ensure semantically consistent descriptions, we apply a contrastive-based consistency objective over generated descriptive features of an image and its stylized version synthesized by a style transfer model (shown in Fig. 1b). We enforce the contrastive objective in the embedding space, which is crucial for learning semantically consistent features, as demonstrated in Sec. 4.2. We extend the proposed method to remedy existing domain bias at both the image level and instance level [26, 57]. Note our proposed approach only impacts the training and does not impose any computational overhead on the inference time.

Our experiments demonstrate that the proposed CDDMSL approach significantly outperforms existing DG and DA methods on two object detection benchmarks. In particular, CDDMSL achieves 11.7% (28%) and 7.5% (20.8%) improvement over the RegionCLIP (Faster-RCNN) in DG and DA settings, respectively, and significantly outperforms the state-of-the-art methods. Moreover, we provide comprehensive analysis and ablation studies, highlighting the benefit of enforcing generalizability through language space and the effectiveness of each component of the proposed method. Our findings indicate that CDDMSL is a promising approach for addressing the domain generalization challenge in object detection, making it a valuable contribution to the field.

2 Related Works

Domain Adaptation for Object Detection. A common DA approach is to align the feature space by adversarial learning or directly minimizing the distance between the feature distributions in the visual space [6, 19, 33, 39, 56]. Many works utilize self-training and highly confident predictions on the target (i.e., pseudo-labels) to progressively improve the model's performance on the target [9, 25, 33, 40]. For instance, [9, 25] leverage a Mean Teacher framework [43] to produce the pseudo-labels by the teacher model and guide the student to perform well on the target. MTOR [2] performs Mean Teacher to explore object relation in region-level, inter-graph, and intragraph consistency. Another work utilizes an unpaired image-to-image translation model to map the source data to target-like images to reduce distribution shift [8, 20, 33, 39, 55]. However, these methods require the pre-collection of target data and fail when the target domain is unknown.

Domain Generalization in Object Detection. Unlike DA, there are not many works addressing domain generalization for object detection (DGOD). Initially, DIDN [26] relied on fully-annotated source domains and adversarial learning to learn domain-invariant features and disentangle domain-specific ones with the help of domain-specific encoders. However, collecting multiple fully annotated source domains is cumbersome and time-consuming. To alleviate this issue, Single-DGOD [49] extended single-domain generalization (SDG) for object detection and employed self-distillation to disentangle domain-specific features. However, SDG methods rely on the quality of the labeled data and data augmentation techniques, reducing their ability to generalize over a wider range of domains. Unlike Single-DGOD,

we use unlabeled data along with only one labeled domain, i.e., semi-supervised domain generalization (SSDG). Pseudo-labeling and data augmentation are some common SSDG approaches. StyleMatch [60] propagates pseudo labels from the labeled domain to the unlabeled domain, and the style transfer model is used for data augmentation. MixStyle [59] inserted a plug-and-play augmentation module in convolutional layers to mix feature statistics between labeled and pseudo-labeled instances. However, the quality of pseudo-labels substantially degrades on out-of-domain data [9, 25] due to the existing distribution shift and challenges induced by producing pseudo-labeled bounding boxes [9, 25]. Unlike existing DA and DG methods, which learn domain-invariant features using objectives in the visual space, we propose to use text/descriptions.

Vision-Language Models. Various works have studied the benefit of learning image-text joint representations in various computer vision tasks [11, 15, 21, 24, 29, 36]. VirTex [11] pre-trained a visual encoder using image-caption pairs, which are then transferred for different downstream tasks. CLIP [36] and ALIGN [21] learn robust and rich visual features by performing cross-modal contrastive loss on large amounts of image-text pairs for open-vocabulary image classification. Recently, several studies used image-text pairs for open-vocabulary object detection. [1, 52] learned joint embedding by matching region features to word/text embeddings. ViLD [14] learned visual features by distilling from pre-trained CLIP encoders. RegionCLIP [57] and GLIP [23] use region-text matching and phrase-grounding, respectively, to pre-train the visual encoder for OVD. In contrast to these works, we study the effectiveness of language for generalizing over domain shift.

3 Approach

Let $\mathcal{X}$ and $\mathcal{Y}$ represent input space and label space, respectively. A domain can be formally described as a joint distribution P_{XY} sampled from $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \times \mathcal{Y}$ denotes the set of all probability distributions. Unlike conventional DG settings, in SSDG, we only have access to one fully labeled and one or more unlabeled source domains, defined as $\mathcal{S} = \{S_k\}_{k=1}^{K}$, where $S_1 = S_{\mathcal{L}} = (x_i^{\mathcal{L}}, y_i^{\mathcal{L}})_{i=1}^{N_{\mathcal{L}}}$ is a labeled source of size $N_{\mathcal{L}}$, $\{S_k\}_{k=2}^{K} = \{S_{\mathcal{U}_k}\}_{k=2}^{K}$ is the k-th unlabeled source of size $N_{\mathcal{U}_k}$, K is the number of source domains, and $\mathbf{y}_i^{\mathcal{L}} = (\mathbf{b}_i^{\mathcal{L}}, c_i^{\mathcal{L}})$ denotes the corresponding labels with bounding-box coordinates $\mathbf{b}$ and their associated categories c. Each source domain S_k is associated with a joint distribution P_{XY}^k and $P_{XY}^k \neq P_{XY}^{k'}$ for $k \neq k'$. In this work, we assume that the $\mathbf{y}_j \in \mathcal{Y}$ share the same set of classes across all domains. The ultimate goal is to utilize the information from the source domains to develop an object detection model capable of generalizing effectively to an unseen target domain $\mathcal{T} = \{\mathbf{x}_j^T\}_{j=1}^{N_T}$ with N_T examples, which is drawn from an unknown distribution P_{XY}^T.

3.1 Cross-Domain Descriptive Multi-Scale Learning

The primary motivation is to leverage the semantically rich information in image descriptions generated by image-captioning models to learn a robust object detector model. We enforce the backbone to learn domain-invariant features, enabling a caption generator to produce consistent descriptions for images with similar semantics but different domain-specific characteristics. This ensures that the model preserves the essential information across different styles and enhances the model's generalization capabilities across various domains. To this end, we propose a novel contrastive-based consistency to maximize the agreement between descriptions of an image across domains in the embedding space.

As illustrated in Fig. 2, we use a vision-to-language (*v2l*) module to obtain descriptive features by projecting image features to the language space. We propose to employ the mapping network of a pre-trained image-captioning model, which stays frozen throughout training. Specifically, we pre-train the ClipCap model [30] with the object detector backbone **B** as its vision encoder and extract its mapping network. We apply *v2l* to the image representation to obtain the desired descriptive features in the language space: $z^{\ell} = v2l(\mathbf{B}(x))$. We define the proposed consistency learning approach in the following text.

Instance-level Descriptive Consistency Learning. Consider a large labeled dataset $S_{\mathcal{L}}$ and a set of smaller unlabeled domains $S_{\mathcal{U}_k}$ (i.e., $N_{\mathcal{L}} >> N_{\mathcal{U}}$). We use the smaller domains to create a larger auxiliary domain $S_{\tilde{\mathcal{U}}}$ of size $N_{\tilde{\mathcal{U}}} = (K-1)\cdot N_{\mathcal{L}} + \sum_{k=2}^{K} N_{\mathcal{U}_k}$ by synthesizing $S_{\mathcal{L}}$ to K-1 unlabeled source domains. Particularly, we use CycleGAN [61] to stylize the images between the labeled and unlabeled source domains. For simplicity, in the rest of the paper, we assume that images are translated from a labeled domain to an unlabeled (i.e., $N_{\tilde{\mathcal{U}}} = N_{\mathcal{L}}$ and $K = 2$), but the unlabeled domain(s) can also be transferred to labeled in order to enforce consistency. Let us define $\{(x_i, \tilde{x}_i) | x_i \in S_{\mathcal{L}}, \tilde{x}_i \in S_{\tilde{\mathcal{U}}}\}$ where x_i and $\tilde{x}_i$ represent the same scene but belong to $S_{\mathcal{L}}$ and $S_{\tilde{\mathcal{U}}}$, respectively. RPN [37] generates region proposal $r_{i,j}$ for image x_i, which is then used to extract region features $v_{i,j}$ and $\tilde{v}_{i,j}$ by applying RoIAlign [17] on $z_i^v = \mathbf{B}(x_i)$ and $\tilde{z}_i^v = \mathbf{B}(\tilde{x}_i)$, respectively. Finally, descriptive region features can be computed by applying the *v2l* on each region feature, denoted as $z_{i,j}^{\ell} = v2l(v_{i,j})$ and $\tilde{z}_{i,j}^{\ell} = v2l(\tilde{v}_{i,j})$. For brevity, let us define all region descriptive feature pairs of all images in the batch as $Z^{\ell} = \{(z_i^{\ell}, \tilde{z}_j^{\ell})\}$, where $i \in \{1,\ldots,N\}$, $j \in \{1,\ldots,N\}$, $(z_i^{\ell}, \tilde{z}_i^{\ell})$ is a positive pair, $(z_i^{\ell}, \tilde{z}_j^{\ell})_{i \neq j}$ is a negative pair, and N is the total number of the proposals or region descriptive features in the batch. The instance-level contrastive loss can be defined:

$$\mathcal{L}_{inst-cont} = \frac{1}{N} \sum_{i} -\log \left(\frac{\exp(\mathrm{s}_{i,i}/\tau)}{exp(\mathrm{s}_{i,i}/\tau) + \sum_k \exp(\mathrm{s}_{i,k}/\tau)} \right);\ k \neq i,\ h_i = g(z_i^{\ell}) \tag{1}$$

$$\mathrm{s}_{i,j} = \mathrm{s}(h_i, \tilde{h}_j) = \frac{h_i^{\top} \cdot \tilde{h}_j}{||h_i|| \cdot ||\tilde{h}_j||} \tag{2}$$

where $g(\cdot)$ is a projection network to project region descriptive features to a lower dimension, $s_{i,j}$ = is a cosine similarity score, $\mathrm{s}_{i,i} = \mathrm{s}(h_i, \tilde{h}_i)$ is a positive pair, $\mathrm{s}_{i,k} = \mathrm{s}(h_i, \tilde{h}_k)$ is a negative pair, and τ is temperature parameter. One key aspect of our method is the choice of applying the contrastive loss in the embedding space (i.e., the output of the *v2l* layer) rather than directly in the language space (i.e., on the tokens). This choice has several advantages. The embedding space provides a continuous and compact representation of semantic content, enabling meaningful similarity metrics between images. By focusing on learning semantically consistent features in the embedding space, the model is less sensitive to specific phrasing or word choice. Additionally, optimizing in the embedding space leads to a more efficient optimization process without needing to backpropagate gradients through the entire language generation process, which is non-differentiable.

Image-level Descriptive Consistency Learning. The proposed instance-level contrastive loss can naturally extend to an image-level contrastive loss $\mathcal{L}_{img-cont}$ by simply replacing the region features with image features in Eq. 1. We thus encourage the generation of semantically consistent descriptive features in different domains by minimizing the contrastive loss at both image-level and instance-level. This removes the domain bias in the backbone **B** and results in a model capable of learning domain-invariant visual features at multiple scales, leading to better generalization performance on unseen target domains.

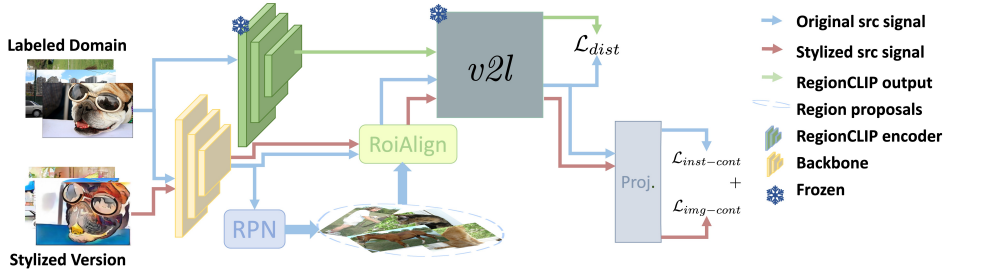


Figure 2: **Overview of CDDMSL**. The backbone processes the image and its stylized version to produce image features. Region proposals are created for the labeled domain, and RoiAlign extracts region features from the original and stylized versions. The *v2l* module produces region descriptive features, projected to a lower dimension by $g(\cdot)$. The backbone and projection heads are trained to maximize agreement between descriptive features across domains using a contrastive loss $\mathcal{L}_{img-inst}$. Image-level training $\mathcal{L}_{img-cont}$ follows a similar process, with RoiAlign directly passing input features to *v2l*. Finally, a knowledge distillation regularization loss $\mathcal{L}_{dist}$ is applied to the descriptive features of the labeled domain based on the backbone and RegionClip encoder output.

3.2 Regularization via Knowledge Distillation

Merely applying the contrastive loss may lead to converging to a trivial solution, generating consistent but meaningless descriptions. We propose a knowledge distillation-based (KD) regularization by exploiting that the image-captioning model produces meaningful descriptions, utilizing pre-trained RegionCLIP as its image-encoder. Concretely, for a given image $x_i \in \mathcal{S}_L$, the regularization loss is defined as:

$$\mathcal{L}_{dist} = \frac{1}{N_{\mathcal{L}}} \sum_{i=1}^{N_{\mathcal{L}}} \mathbf{d}(v2l(z_i), v2l(z_i^R)); \quad z_i^R = F_{R-CLIP}(x_i) \tag{3}$$

z_i and z_i^R represent the model's backbone (yellow encoder in Fig. 2) and the RegionCLIP (F_{R-CLIP}) output, respectively. $\mathbf{d}(\cdot,\cdot)$ is a distance function. In this work, the Manhattan distance metric is used. It is crucial to note that the RegionCLIP features remain frozen throughout the training.

3.3 Object Detector Training

We use the labeled source domain $S_{\mathcal{L}} = (x_i^{\mathcal{L}}, y_i^{\mathcal{L}})_{i=1}^{N_{\mathcal{L}}}$ to train a Faster-RCNN [37]. Unlike Faster-RCNN, we adopt a text-based classifier in the detection head. Following RegionCLIP, we create prompts for each object category by filling it into prompt templates (e.g., "A photo of a car," "A painting of a bus," etc.) and encode the prompts using the pre-trained CLIP language encoder. Finally, cross-entropy with focal scaling [35] is employed as the classification loss $\mathcal{L}_{cls}$. We use fixed all-zero embedding for the background category and apply a predefined weight to background regions following [24, 57]. The detector is then trained with a $\mathcal{L}_{det}$ loss which combines $\mathcal{L}_{cls}$ with standard RPN and regression loss as in [37]: $\mathcal{L}_{det} = \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{RPN}$. Lastly, training consists of two major stages. During the burn-up stage, the object detector is warmed-up in a supervised man-

ner, whereas in joint training, it is jointly learned by supervised and contrastive objectives: $\mathcal{L}_{tot} = \mathcal{L}_{det} + \mathcal{L}_{inst-contr} + \mathcal{L}_{img-contr} + \omega \cdot \mathcal{L}_{dist}$, where ω is a regularization parameter.

4 Experiments

Real-to-artistic generalization. We conduct our experiments based on 4 datasets: Pascal-VOC (VOC) [12, 13], Clipart1k [20], Watercolor2k [20], Comic2k [20] to assess generalizability on artistic and style domain shifts. VOC contains 20 categories of common objects from 16,551 natural images, Clipart consists of 20 categories of 3,165 objects from 1000 instances, and Watercolor and Comic contain 6 categories of 3,315 and 6,389 objects from 2,000 images, respectively. Experiments are conducted in three settings: (1) VOC and Clipart, (2) VOC and Watercolor, and (3) VOC and Comic are used as the source domains, in which only VOC is considered a labeled domain. In each setting, the remaining datasets are considered unseen targets to assess the model's generalizability across multiple domains (Table 1). We extend our model to DA and show the results in Table 4a.

Adverse weather adaptation. We experiment on Cityscapes [6] (City), Foggy-Cityscapes [6] (Foggy), and BDD100k [53] (Bdd) to evaluate generalizability across different time and weather conditions. City features 3,025 urban images from 50 cities, Foggy simulates fog on the City, and Bdd comprises 10,000 images of various weather conditions and times. All datasets contain 8 categories, but we ignore the "train" category on Bdd due to its low instance count in the validation set [26]. As Foggy is a simulated based on City, we use them as the source domains without the need for style transfer (City labeled and Foggy unlabeled), while Bdd is an unseen domain in the domain generalization task summarized in Table 2. Note we do not use the labels on the Foggy data throughout training. Following DIDN [26], since most previous works are on weather adaptation, we perform adaptation from City to Foggy and extensively compare against baselines and related works as shown in Table 3.

Baselines. Faster-RCNN [37] (F-RCNN) with ImageNet [8] and RegionCLIP (R-CLIP) are labeled source-only baselines, trained on VOC for real-to-artistic generalization/adaptation (Table 1 / Table 4a) and City for adverse weather generalization/adaptation (Table 2 / Table 3). For a fair comparison, we only compare against models using ResNet50 backbone as RegionCLIP [57] does not have ResNet101 pre-trained weights available. For real-to-artistic generalization, we further train Adaptive-MT [25], a state-of-the-art (sota) DA with ResNet50 backbone. We define *Direct Visual Alignment* (DVA) and *Caption Pseudo Labeling* (Caption-PL) as additional baselines in the ablation (Table 1 and Table 4a for DG and DA in real-to-artistic, respectively). DVA applies the descriptive consistency loss in the visual space without using *v2l*. Caption-PL uses pseudo labels obtained by applying ClipCap on the original image to compute image-captioning loss over tokens for the stylized image. IRG [45] is only reported for the (VOC and Clipart) setting as performance/pre-trained weights were reported only in this setting.

Implementation details. PyTorch [34] and Detectron2 [50] are used for development. We adopt RegionCLIP's ResNet50 [16] as the object detector backbone. Following ClipCAP[1], *v2l* comprises 8 multi-head self-attention layers, each with 8 heads. ClipCAP is pre-trained on COCO-captions [4, 27] using frozen RegionCLIP as the image encoder and GPT-2 as the language decoder, with only the mapper (*v2l*) being updated. Projection layer g (after *v2l* in Fig. 2) has output dimension of 256. For all experiments, we use SGD with 0.002

[1] https://github.com/rmokady/CLIP_prefix_caption

Table 1: **Real-to-artistic generalizations**. Numbers in parentheses show the difference from R-CLIP. Max ↑ shows maximum improvement over two target domains compared to F-RCNN. We report mAP (%). †/‡ denote DA methods/labeled source-only methods.

Method	VOC&Clip →Water,Com		VOC&Water→Clip,Com		VOC&Com→Clip,Water		Max ↑
	Watercolor	Comic	Clipart	Comic	Clipart	Watercolor	
F-RCNN‡ [67]	41.2	17.9	24.1	17.9	24.1	41.2	-
R-CLIP‡ [67]	44.7	34.2	33.9	34.2	33.9	44.7	16.3/16.3/9.8
Adaptive MT† [25]	40.6 (-4.1)	22.2 (-12.0)	29.0 (-4.9)	24.3 (-9.9)	25.7 (-8.2)	42.3 (-2.4)	4.3/6.4/1.6
IRG† [45]	48.1 (+3.4)	25.9 (-8.3)	-	-	-	-	8.0/-/-
DVA	45.6 (+0.9)	38.1 (+3.9)	32.6 (-1.3)	34.2 (+0.0)	35.9 (+2.0)	45.9 (+1.2)	20.2/16.3/11.8
Caption-PL	45.0 (+0.3)	36.4 (+2.2)	30.1 (-3.8)	30.3 (-3.9)	34.7 (+0.8)	42.1 (-2.6)	18.5/12.4/10.6
Ours	**49.8** (+5.1)	**45.9** (+11.7)	**38.7** (+4.8)	**43.5** (+9.3)	**39.8** (+5.9)	**49.4** (+4.7)	**28.0/25.6/15.7**

Table 2: **Adverse weather generalization.** *City, Foggy → Bdd* in %

Method	prsn	rider	car	truck	bus	motor	bike	mAP
F-RCNN [67]	27.9	27.5	43.1	16.6	15.1	5.6	21.0	19.6
R-CLIP [67]	40.6 (+12.7)	31.3 (+3.8)	47.9 (+4.8)	16.8 (+0.2)	12.0 (-3.1)	11.2 (+5.6)	23.2 (+2.2)	26.1 (+6.5)
DIDN [26] (ICCV'21)	34.5 (+6.6)	30.4 (+2.9)	44.2 (+1.1)	**21.2** (+4.6)	**19.0** (+3.9)	9.2 (+3.6)	22.8 (+1.8)	22.7 (+3.1)
Ours	**41.4** (+13.5)	**31.7** (+4.2)	**49.8** (+6.7)	18.1 (+1.5)	11.4 (-3.7)	**12.4** (+6.8)	**25.6** (+4.6)	**27.1** (+7.5)

initial lr, a linear scheduler, and a batch size of 8, except 12 for the RegionCLIP baseline. Experiments use 4 NVIDIA Quadro RTX 5000 with 10k burn-up and 20k joint-training iterations. The burn-up stage follows RegionCLIP [67], and SSDG experiments continue from this checkpoint. ω in $\mathcal{L}_{tot}$ is 1. We evaluate mAP with a 0.5 threshold. For realistic evaluation, following [31, 32], we avoided intensive hyperparameter search and manually chose reasonable parameters that resulted in stable training. However, a comprehensive search is expected to improve the performance.

4.1 Results

Domain generalization. Table 1 showcases the generalizability of our proposed method on real-to-artistic tasks. The table is divided into three sections according to the unlabeled domain (i.e., Clipart, Watercolor, and Comic). Our proposed method consistently enhances R-CLIP generalizability and achieves the best result compared to the DA baselines by a large margin. For example, it achieves an 11.7% and 9.3% improvement on Comic when Clipart and Watercolor is the unlabeled source domain, respectively. We also observe that baselines perform better on Watercolor, which is intuitive as Watercolor is the closest domain to VOC. Likewise, our proposed approach outperforms DIDN on *City, Foggy → Bdd* by 4.4% (Table 2). These results demonstrate the effectiveness of language-guided feature alignment and the proposed consistency learning method for developing a generalizable object detector.

Extension to domain adaptation. Our method can naturally extend to domain adaptation as it requires only one labeled source domain. The labeled and unlabeled source domains in the domain generalization task serve as source and target domains, respectively. Table 3 presents the result of *City → Foggy* adaptation against baselines and sota DA methods, while real-to-artistic adaptation can be found in the Table 4a. As illustrated in Table 3, our method improves F-RCNN and R-CLIP by 20.8% and 7.5%. It not only substantially outperforms DIDN (sota DGOD on this setting) by 8.6% but also achieves comparable results to the existing DA works. While CDDMSL is slightly under-performed compared to TTD by only 0.1%, contrary to DA methods, it is not designed to adapt to a particular domain.

[2] Source-free domain adaptation

Table 3: **Adverse weather adaptation.** *City→Foggy* in mAP (%), † from [18].

	Method	prsn	rider	car	truck	bus	train	motor	bike	mAP
-	F-RCNN† [34]	36.9	36.1	44.5.6	21.7	32.3	9.2	21.5	32.4	28.3 (-20.8)
	R-CLIP [34]	46.5	51.8	57.6	27.3	45.1	19.7	34.8	50.2	**41.6** (-7.5)
DA	SW-DA [40] (CVPR'19)	31.8	44.3	48.9	21.0	43.8	28.0	28.9	35.8	35.3
	D&Match† [24] (CVPR'19)	31.8	40.5	51.0	20.9	41.8	34.3	26.6	32.4	34.9
	SC-DA† [51] (CVPR'19)	33.8	42.1	52.1	26.8	42.5	26.5	29.2	34.4	35.9
	MTOR [1] (CVPR'19)	30.6	41.4	44.0	21.9	38.6	40.6	28.3	35.6	35.1
	AFAN† [46] (TIP'21)	42.5	44.6	57.0	26.4	48.0	28.3	33.2	37.1	39.6
	GPA† [50] (CVPR'20)	32.9	46.7	54.1	24.7	45.7	41.1	32.4	38.7	39.5
	ViSGA† [53] (ICCV'21)	38.8	45.9	57.2	29.9	50.2	**51.9**	31.9	40.9	43.3
	SFA† [41] (acmmm'21)	46.5	48.6	62.6	25.1	46.2	29.4	28.3	44.0	41.3
	DSS†[43] (CVPR'21)	**50.9**	**57.6**	61.1	**35.4**	50.9	36.6	38.4	51.1	47.8
	TTD+FPN† [19] (CVPR'22)	50.7	53.7	**68.2**	35.1	53.0	45.1	38.9	49.1	**49.2**
	IRG^2 [45] (CVPR'23)	37.4	45.2	51.9	24.4	39.6	25.2	31.5	41.6	37.1
DG	DIDN [26] (ICCV'21)	38.3	44.4	51.8	28.7	53.3	34.7	32.4	40.4	40.5 (-8.6)
	Ours	50.5	55.1	66.9	35.0	**56.2**	33.5	**41.0**	**54.3**	**49.1**

Table 4: (a) **Real-to-artistic adaptation**. VOC is labeled source domain. (b) **Ablation study**. *VOC, Clipart → Watercolor, Comic (DG), Clipart (DA)*. Results in mAP (%).

Method	Target Domain		
	Clipart	Watercolor	Comic
F-RCNN [34]	24.1	41.2	17.9
R-CLIP [34]	33.3	44.7	34.2
Adaptive MT [43]	30.5	43.7	23.4
IRG [45]	31.5	**53.0**	-
DVA	36.6	43.9	35.9
Caption-PL	35.2	44.2	34.2
Ours	**40.4**	49.7	**46.3**

(a) Domain adaptation on real-to-artistic

R-CLIP init.	$\mathcal{L}_{img-cont}$	$\mathcal{L}_{inst-cont}$	$\mathcal{L}_{dist}$	DA Clipart	DG Watercolor	DG Comic
				24.1	41.2	17.9
✓				32.3	44.7	34.2
✓	✓			32.3	41.7	35.1
✓		✓		34.6	45.0	35.4
✓	✓	✓		35.1	44.2	35.7
✓	✓	✓	✓	**40.4**	**49.8**	**45.9**

(b) Effectiveness of each component

4.2 Ablation & Discussion

We conduct an extensive ablation study by examining five aspects of CDDMSL: (i) Impact of vision-language pre-training (Sec. 3.3), (ii) Effectiveness of language space alignment compared to visual-space alignment (Sec. 3.1), (iii) Benefit of feature-space objective compared to token-space objective, (iv) Effectiveness of multi-scale learning (Sec 3.1), (v) Contribution of KD regularization (Sec. 3.2). Table 4b exhibits the effectiveness of each component.

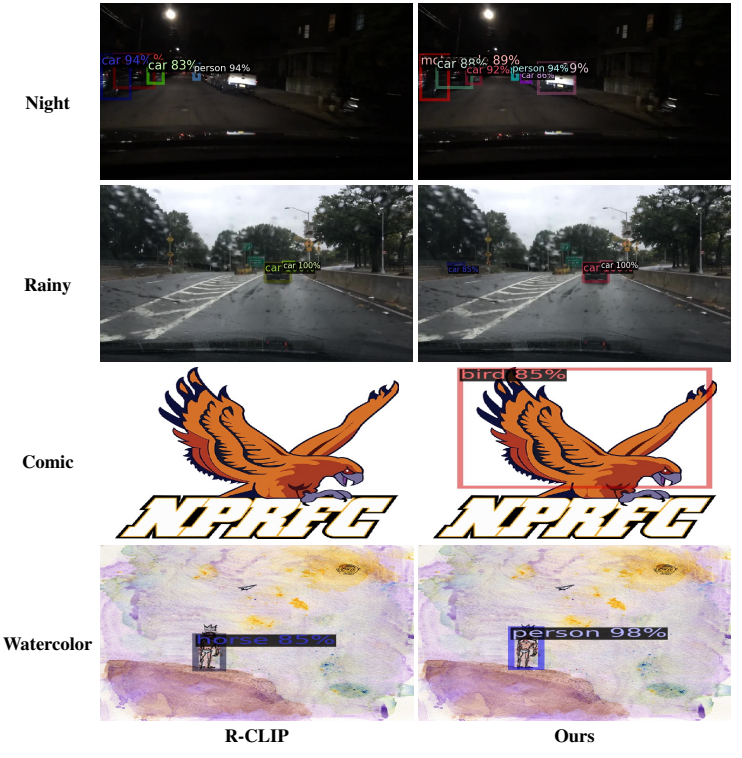
(i) **Vision-language pre-training.** According to Table 1, R-CLIP solely notably improves the generalizability of F-RCNN (pre-trained on Imagenet) by 9.8%, 3.5%, and 16.3% on Clipart, Watercolor, and Comic, respectively. Additionally, we observe 6.5% gain on *City,Foggy → Bdd* (Table 2), supporting the benefit of vision-language pre-training. Despite the improvement, vision-language pre-training does not guarantee optimal generalizability as it lacks explicit optimization for adaptation to a new domain, and the model may still overfit to domain-specific features during fine-tuning. Hence, we further improve R-CLIP initialization by incorporating descriptive consistency learning as described in Sec. 3.1.

(ii) **Language space vs. visual space alignment.** To better understand the effectiveness of enforcing domain-invariant learning through language space, we compare our model with the Direct Visual Alignment baseline (row 5 in Table 1 and Table 4a) under an identical setting. In all settings, our (last row) substantially exceeds DVA in DA and DG tasks.

(iii) **Feature-space vs. token-space.** Comparing Caption-PL (row 4) with our approach in Table 1 and Table 4a indicates the superiority of enforcing consistency in the feature space. Ours significantly improves Caption-PL in all settings and tasks. Token-space approaches may degrade the performance by focusing on unnecessary non-informative tokens.

(iv) **Effectiveness of Multi-Scale Learning.** Table 4b shows instance-level is slightly

Table 5: **Visualization.** *City, Foggy→ Bdd & VOC, Clipart→ Comic, Watercolor.*



more effective than image-level, but combined delivers the best result. This underscores the importance of mitigating bias and distribution shifts at both levels in object detection.

(v) **KD regularization's contribution.** Finally, adding a KD regularizer results in the best performance (compare last two rows in Table 4b), implying that KD regularization helps the model avoid trivial solutions and produce semantically consistent descriptions.

Visualization. Table 5 shows qualitative detection results from the R-CLIP baseline and CDDMSL. R-CLIP yields more false negatives, particularly in challenging conditions like rain and detecting distant small objects. Our multi-scale feature alignment improves object-background differentiation, leading to superior detection on the target domain. Additionally, CDDMSL ensures more accurate recognition by retaining semantically crucial information.

5 Conclusion

This paper introduces a novel approach, termed Cross-Domain Descriptive Multi-Scale Learning (CDDMSL), for semi-supervised domain generalization in object detection. Initially, we explore the effectiveness of vision-language pre-training in achieving a robust object detector. Subsequently, we present a unique method that promotes domain-invariant feature learning on the visual encoder through the language space. This is achieved using a contrastive learning-based approach at both image and instance levels. CDDMSL delivers state-of-the-art results in DA and DG tasks across various object detection benchmarks.

Acknowledgement: The work is in part supported by NSF Grant No. 2006885.

References

[1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018.

[2] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11457–11466, 2019.

[3] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[5] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *In IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4091–4101, 2021.

[10] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11162–11173, 2021.

[11] Zi-Yi* Dou, Aishwarya* Kamath, Zhe* Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, Jianfeng Gao, and Lijuan Wang. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *Advances in Neural Information Processing Systems*, 2022.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, .

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, .

[14] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=lL3lnMbR4WU.

[15] Tanmay Gupta, Kevin Shih, Saurabh Singh, and Derek Hoiem. Aligned image-word representations improve inductive transfer across vision-language tasks. In *In IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4213–4222, 2017.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *In IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[18] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9570–9580, 2022.

[19] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 733–748. Springer, 2020.

[20] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5001–5009, 2018.

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[22] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12456–12465, 2019.

[23] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10975, 2022.

[24] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

[25] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7581–7590, 2022.

[26] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *In IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8771–8780, 2021.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[28] Xiao Liu, Spyridon Thermos, Alison O'Neil, and Sotirios A Tsaftaris. Semi-supervised meta-learning with disentanglement for domain-generalised medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 307–317. Springer, 2021.

[29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32, 2019.

[30] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[31] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

[32] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12674–12684, 2020.

[33] Poojan Oza, Vishwanath A Sindagi, Vibashan Vishnukumar Sharmini, and Vishal M Patel. Unsupervised domain adaptation of object detectors: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library.

In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learni pdf.

[35] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HyhbYrGYe.

[36] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[38] Farzaneh Rezaeianaran, Rakshith Shetty, Rahaf Aljundi, Daniel Olmeda Reino, Shanshan Zhang, and Bernt Schiele. Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection. In *In IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9204–9213, October 2021.

[39] Adrian Lopez Rodriguez and Krystian Mikolajczyk. Domain adaptation for object detection via style consistency. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 232. BMVA Press, 2019. URL https://bmvc2019.org/wp-content/uploads/papers/1038-paper.pdf.

[40] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[41] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6956–6965, 2019.

[42] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9624–9633, June 2021.

[43] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf.

[44] Mesut Erhan Unal, Keren Ye, Mingda Zhang, Christopher Thomas, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Learning to overcome noise in weak caption supervision for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4897–4914, 2023. doi: 10.1109/TPAMI.2022.3187350.

[45] Vibashan VS, Poojan Oza, and Vishal M. Patel. Instance relation graph guided source-free domain adaptive object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3520–3530, June 2023.

[46] Hongsong Wang, Shengcai Liao, and Ling Shao. Afan: Augmented feature alignment network for cross-domain object detection. *IEEE Transactions on Image Processing*, 30:4046–4056, 2021.

[47] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. Exploring sequence feature alignment for domain adaptive detection transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1730–1738, 2021.

[48] Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, YangYang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9603–9612, 2021.

[49] Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 847–856, 2022.

[50] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[51] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12355–12364, 2020.

[52] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection. In *In IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[53] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[54] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14393–14402, 2021.

[55] Dan Zhang, Jingjing Li, Lin Xiong, Lan Lin, Mao Ye, and Shangming Yang. Cycle-consistent domain adaptive faster rcnn. *IEEE Access*, 7:123903–123911, 2019. doi: 10.1109/ACCESS.2019.2938837.

[56] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13766–13775, 2020.

[57] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16793–16803, 2022.

[58] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization in vision: A survey. *arXiv preprint arXiv:2103.02503*, 2021.

[59] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *The Tenth International Conference on Learning Representations, ICLR*, 2021. URL https://openreview.net/forum?id=6xHJ37MVxxp.

[60] Kaiyang Zhou, Chen Change Loy, and Ziwei Liu. Semi-supervised domain generalization with stochastic stylematch. *Int. J. Comput. Vis.*, 131(9):2377–2387, 2023. doi: 10.1007/s11263-023-01821-x. URL https://doi.org/10.1007/s11263-023-01821-x.

[61] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *In IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.

[62] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 687–696, 2019.