This ICCV 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.

# UPRE: Zero-Shot Domain Adaptation for Object Detection via Unified Prompt and Representation Enhancement

Xiao Zhang[1,2]*, Fei Wei[2], Yong Wang[2], Wenda Zhao[1]†, Feiyi Li[1], Xiangxiang Chu[2]
[1]Dalian University of Technology [2]AMAP, Alibaba Group
{1204855526@mail.dlut.edu.cn, zhaowenda@dlut.edu.cn, lifeiyi@mail.dlut.edu.cn}
{xixia.wf, wangyong.lz, chuxiangxiang.cxx}@alibaba-inc.com

## Abstract

*Zero-shot domain adaptation (ZSDA) presents substantial challenges due to the lack of images in the target domain. Previous approaches leverage Vision-Language Models (VLMs) to tackle this challenge, exploiting their zero-shot learning capabilities. However, these methods primarily address domain distribution shifts and overlook the misalignment between the detection task and VLMs, which rely on manually crafted prompts. To overcome these limitations, we propose the unified prompt and representation enhancement (UPRE) framework, which jointly optimizes both textual prompts and visual representations. Specifically, our approach introduces a multi-view domain prompt that combines linguistic domain priors with detection-specific knowledge, and a visual representation enhancement module that produces domain style variations. Furthermore, we introduce multi-level enhancement strategies, including relative domain distance and positive-negative separation, which align multi-modal representations at the image level and capture diverse visual representations at the instance level, respectively. Extensive experiments conducted on nine benchmark datasets demonstrate the superior performance of our framework in ZSDA detection scenarios. Code is available at https://github.com/AMAP-ML/UPRE.*




Figure 1. **(a)** Prompt-driven semantic augmentation generates pseudo target domain features with manually crafted prompts. **(b)** The proposed unified prompt and representation enhancement creates the synthetic target domain representations via learnable prompts. Learnable prompt and unified enhancement jointly mitigate detection and domain biases. IE represents the image encoder, and TE denotes the text encoder.

## 1. Introduction

Domain adaptation has gained widespread attention in recent years [4, 22, 31, 34, 39, 41, 42, 50]. However, obtaining image priors, even unlabeled, is not always feasible, thereby limiting their applicability in real-world scenarios. Zero-shot domain adaptation (ZSDA) [2, 14, 40, 54, 60, 62] aims to address this constraint by enabling adaptation to target domains without prior image exposure. The development of vision-language models (VLMs) [1, 7, 9, 25, 32, 33, 36, 56] has markedly advanced ZSDA, by utilizing textual prompts to describe unseen domains through their inherent zero-shot capabilities [14, 62]. Despite these advancements, our analysis reveals two primary limitations in existing VLM-based object detection approaches: 1) *domain bias*, where distribution shifts between source and target domains introduces task-agnostic noise, impairing model performance; 2) *detection bias*, where models such as CLIP [44] emphasize global image representations but overlook instance-level details crucial for precise object localization. This deficiency stems from manually constructed prompts, which inadequately capture the contextual attributes of foreground and

*Work partially done during the internship at AMAP, Alibaba Group.
†Corresponding author.

background objects.

Previous studies [14, 49, 62] have explored the mitigation of domain bias through prompt-driven semantic augmentation, as illustrated in Fig. 1(a). These methods utilize manually crafted prompts combined with image-level alignment to produce the representations of pseudo target domain. Despite their effectiveness in addressing domain bias, these approaches overlook and intensify detection bias during the fine-tuning process. Conversely, methods [12, 21, 30, 66] aimed at reducing detection bias tend to reinforce domain bias. They endeavors to minimize detection bias by learning prompt representations. However, in the absence of target domain images, the training of prompt representation is restricted to the source domain. Consequently, while these prompts effectively encapsulate representations for the detection task within the source domain, they unintentionally magnify domain bias. This gap highlights the pressing need for method that can simultaneously address both domain and detection biases.

In this paper, we propose the unified prompt and representation enhancement (UPRE) framework, a novel approach that jointly enhances textual prompt and visual representations. As illustrated in Fig. 1(b), our method establishes a cooperative relationship between prompt optimization and visual representation learning. Specifically, we design a multi-view domain prompt (MDP) that provides language-modal priors for the target domain while capturing diverse adaptation knowledge essential for cross-domain object detection. Furthermore, we introduce unified representation enhancement (URE), a shared module between both vision and language modalities designed to generate target domain representations from source domain data. The URE enhances the diversity of the domain styles under the guidance of prompt representations, effectively alleviating domain bias. Concurrently, the learnable prompt representations of MDP utilize the target domain representations generated by URE to acquire instance-level knowledge for object detection, thereby reducing detection bias.

To further enhance this interaction, we introduce two novel enhancement strategies: relative domain distance (RDD) and positive-negative separation (PNS), which together form a multi-level training framework. RDD aligns multi-modal representations at the image level to facilitate adaptation, while PNS leverages background context from negative proposals and object information from positive ones to capture diverse instance knowledge. Through the unified training process, our approach enhances both prompt and visual representations, thereby effectively mitigating detection and domain biases. This leads to notable improvements in adaptation and detection capabilities, empowering object detectors to accurately adapt to previously unseen domains. Our proposed method demonstrates superior performance across three domain adaptation tasks on nine datasets, including adverse weather conditions, cross-city scenarios, and virtual-to-real transitions.

In summary, our contributions are as follows:

- We identify and tackle the challenges of detection and domain biases in VLMs for ZSDA by introducing a unified prompt and representation enhancement (UPRE) approach tailored for object detection.
- We introduce the MDP mechanism to learn prompt representations tailored for cross-domain detection, and propose the URE to increase the diversity of domain styles.
- We develop the RDD technique to enable effective image-level adaptation, and the PNS to enhance the capability of precise localization at the instance level.

## 2. Related Work

### 2.1. Zero-shot Domain Adaptation

ZSDA [13, 15, 26, 28, 38, 43, 51–53] presents a challenging learning paradigm that transfers knowledge from a source domain to a target domain without accessing any target domain data. Min *et al.* [40] address generalized zero-shot learning by eliminating domain-specific visual biases. Wang *et al.* [54] introduce a domain shift preservation method, which relies on GAN [18] to maintain domain invariance during adaptation. Recent advances have explored VLMs to improve adaptation performance. Fahes *et al.* [14] propose a prompt-driven approach to guide domain adaptation. Yang *et al.* [62] also propose a unified language-driven method to bridge the gap between the source and target domains. In parallel, Single Domain Generalization [11, 27, 37, 45, 49, 58, 59], focuses on generalizing models to multiple unseen target domains simultaneously. Lee *et al.* [27] propose an object-aware domain generalization method to address misclassification and imprecise localization challenges. Subsequently, Liu *et al.* [37] introduce a structural causal model to analyze data and feature biases in the task. While recent works [14, 49, 62] leverage the zero-shot learning capabilities of VLMs to mitigate domain bias, they overlook downstream task biases, particularly detection bias, which remain underexplored for ZSDA.

### 2.2. Domain Prompt Learning

Recent advances in domain prompt learning [3, 5, 8, 16, 30, 48, 55, 61, 65] have demonstrated promising capabilities for cross-domain knowledge transfer through VLMs with minimal intervention. Existing approaches can be broadly categorized into two paradigms: domain-specific adaptation and doamin-invariant generalization. Domain-specific methods customizes prompts to capture domain related characteristic. For instance, Ge *et al.* [16] adapt domain-aware prompts to bridge domain gaps, while Cao *et al.* [6] dynamically adjust prompts based on domain features. Li *et al.* [30] enhance detection performance by incorporating prompt tuning in the

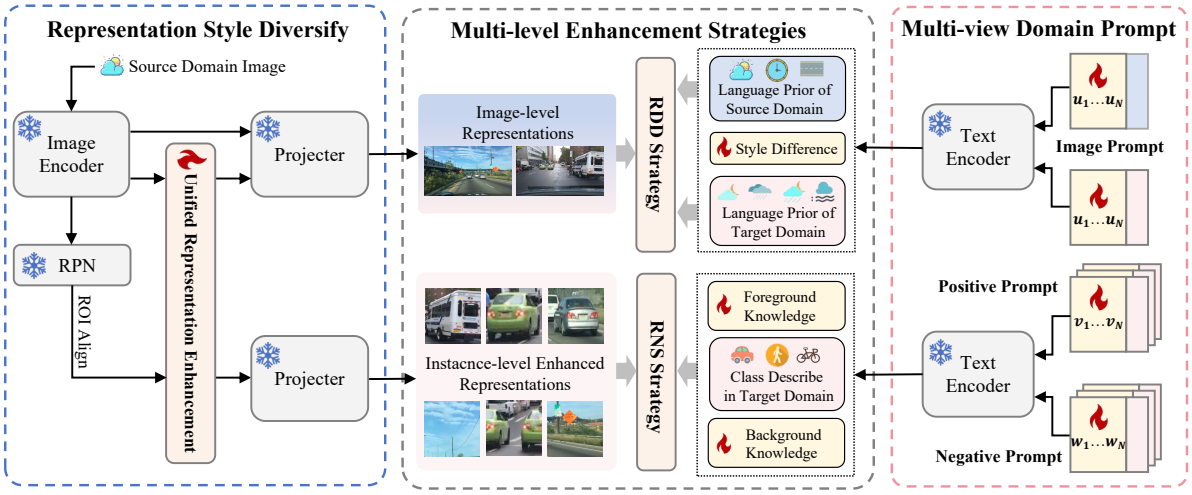



Figure 2. **Overview of unified prompt and representation enhancement**. Initially, the multi-view domain prompt provides language prior while also learns adaptation knowledge for object detection. Meanwhile, unified representation enhancement enriches source domain features with RPN, generating pseudo target domain representations. Finally, two enhancement strategies constrain language and visual representations, training the unified framework effectively. The symbols ⬅ represent losses.

detection head. Conversely, domain-invariant approaches focus on learning generalized prompts for unseen domains. Zhao *et al.* [65] focus on learning domain-invariant prompts for VLMs, utilizing domain-agnostic representations to enhance adaptability. Singha *et al.* [48] introduce AD-CLIP, adapting prompts within CLIP [44] representations to promote generalization. However, these methods face inherent limitations in ZSDA scenarios. Domain-specific approaches struggle to mitigate domain bias without image priors, while domain-invariant methods often decrease task-specific discriminability. The core challenge lies in effectively leveraging limited source-domain information to address both domain shift and downstream task misalignment. In this paper, we propose a novel approach that unified prompt learning and representation enhancement to address this challenge.

## 3. Method

To jointly address detection and domain biases in zero-shot domain adaptation, we propose a novel framework that cooperatively integrates prompt optimization with cross-domain representation synthesis. As depicted in Fig. 2, our approach establishes a unified learning system where domain-aware prompt representation and visual representation enhancement mutually reinforce each other. Our framework comprises three main components:

- Multi-view Domain Prompt (MDP): a hybrid prompts combining linguistic knowledge of target domain priors (image/instance-level) with learnable prompts to learn cross-domain knowledge for object detection.
- Unified Representation Enhancement (URE): a representation style diversify module through feature transformation and multi-modal projection.
- Multi-level Enhancement Strategies: incorporating Relative Domain Distance (RDD) and Positive-Negative Separation (PNS) for representation alignment between linguistic and visual modals.

**Problem Formulation.** Traditional object detection dataset is defined as $\mathcal{D} = \{(\mathcal{X}_n, \mathcal{O}_n)\}_{n=1}^{|\mathcal{D}|}$, where $\mathcal{X}_n$ is the $n$-th image, $\mathcal{O}_n = \{o_{nm}\}_{m=1}^{|\mathcal{O}_n|}$ is the corresponding set of annotated objects, and $m$ denotes the $m$-th objects. Each object $o$ is a pair of object bounding box $b \in \mathbb{R}^4$ and category label $y_c \in \mathcal{C}$, where $\mathcal{C}$ is the category space of the dataset. Zero-shot domain adaptation for object detection aims to detect objects in an unseen target domain using a model trained exclusively on the source domain. Specifically, the dataset space $\mathcal{D}$ is divided into source domain data $\mathcal{D}^s = \{(\mathcal{X}_n, \mathcal{O}_n)\}_{n=1}^{|\mathcal{D}^s|}$ and unseen target domain data $\mathcal{D}^t = \{\mathcal{X}_n\}_{n=1}^{|\mathcal{D}^t|}$, where $\mathcal{D} = \mathcal{D}^s \cup \mathcal{D}^t$ and $\mathcal{D}^s \cap \mathcal{D}^t = \varnothing$. For each category $c \in \mathcal{C}$, we use a pretrained text encoder $\mathcal{T}$ to encode its semantic embedding as: $\mathbf{t}_c = \mathcal{T}(\mathcal{R}(c))$, where $\mathcal{R}(c)$ is the textual prompt of class $c$, e.g. a photo of a $[class]$. Besides, we define a class of "$\mathcal{C}_{bg}$" to represent the background category. Given an image, the image encoder $\mathcal{I}$ of CLIP extracts a set of hierarchical feature maps, then the RPN generates a set of proposals $\mathcal{P}$. Subsequently, we perform ROI-Align to extract proposal features $\{F_p\}$ $p \in \mathcal{P}$, and feed it into the projector (the rest layer of CLIP) to extract proposal representations as: $\mathbf{e}_p = \mathcal{I}(F_p)$. The probability of a proposal belonging to category $c$ is computed as:

$$P_c = \frac{\exp(f(\mathbf{e}_p, \mathbf{t}_c))}{\sum_{c' \in \mathcal{C} \cup \{\mathcal{C}_{bg}\}} \exp(f(\mathbf{e}_p, \mathbf{t}_{c'}))} \tag{1}$$

where $f(\cdot, \cdot)$ denotes the cosine similarity function.

### 3.1. Multi-view Domain Prompt

Previous methods [6, 30, 65] rely on target domain images and assume that learnable prompts can fully replace manually designed prompts. In our approach, we retain the static human-defined prompt and prepend it with learnable prompts. By retaining the original human-defined context, our design allows the learnable prompt to concentrate on capturing multi-view knowledge. Specifically, we propose MDP to learn both domain adaptation knowledge for bridging domain gaps and object localization information for accurate dense predictions. As illustrated in Fig. 2 (right), MDP consists of three parts, including image, positive, and negative prompts. The image prompt provides language priors of target domain such as lighting and visibility, while further learning the style differences between domains at the image level. Initially, we define $\{u_l, v_l, w_l\}_{l=1}^{L}$ as the learnable context vectors with same dimension. The image prompt representation $\mathcal{R}_i^d$ for given domain $d \subset \{s, t\}$ is defined as:

$$\mathcal{R}_i^d = \left[u_1, u_2, \ldots, u_L, k_d\right], \tag{2}$$

where $k_d$ is the word representations of "a photo taken on a $[domain]$". Positive proposals encode the characteristics of foreground objects, whereas negative proposals focus on capturing the contextual information in background. To model the stylistic variations of foreground objects in the target domain, we introduce the positive prompt representation $\mathcal{R}_p$. Given target domain $t$ and class $c$, positive prompt representation $\mathcal{R}_p^t(c)$ is defined as:

$$\mathcal{R}_p^t(c) = \left[v_1, v_2 \ldots v_L, (k_t, k_c)\right], \tag{3}$$

where $(k_t, k_c)$ denotes the word representations of "a $[domain]$ photo of a $[class]$". Similarly, the negative prompt representation $\mathcal{R}_n^t(\mathcal{C}_{\text{bg}})$, which is designed to capture contextual background knowledge, is defined as:

$$\mathcal{R}_n^t(\mathcal{C}_{bg}) = \left[w_1, w_2 \ldots w_L, (k_t, k_{\mathcal{C}_{bg}})\right], \tag{4}$$

where $(k_t, k_{\mathcal{C}_{bg}})$ is the word representations of "a $[domain]$ photo of an $[unknown\ class]$".

### 3.2. Unified Representation Enhancement

Despite lacking prior knowledge of target domain images, VLMs [35, 44, 57, 64] can still describe unseen domains through prompts. CLIP-GAP [49] exploit it to train semantic augmentation for domain transformation but overlooks the deviation term. Later, PODA [14] addresses this by introducing AdaIN [23] to globally transfer feature styles. However, the style of an image often varies dynamically in different regions of the real world. For example, for a rainy and night image, nearby regions may reflect a style of "rainy" due to brighter lighting, while distant regions may reflect a "night" style due to insufficient lighting. Thus, basic style transform approach using the global style parameters is inadequate for complex domain adaptation detection task.

Subsequently, we propose the learnable mean enhancement $\mathcal{E}_\mu \in \mathbb{R}^{C\times M\times N}$ and learnable deviation enhancement $\mathcal{E}_\sigma \in \mathbb{R}^{C\times M\times N}$ to establish a fine-grained solution. Specifically, we align the pseudo target domain features $F_{s\to t}$, which are generated by enhancing the source image features $F_s$ through URE, with the linguistic counterparts of the target image features $F_t$. To generate pseudo target domain features $F_{s\to t}$, we first segment $F_s$ into $M \times N$ patches. After that, we apply $\mathcal{E}_\mu$ and $\mathcal{E}_\sigma$ to enhance the style-specific elements of $F_s$ as follows:

$$F_{s\to t} = \left\{\mathcal{E}_\sigma^j \cdot F_s^j + \mathcal{E}_\mu^j\right\}_{j=1}^{M\times N}, \tag{5}$$

where $j$ represents the $j$-th patch. The source image features $F_s$ are enhanced to pseudo target features $F_{s\to t}$, capturing the stylistic characteristics of the target domain in real-world scenarios.

### 3.3. Multi-level Enhancement Strategies

To further unlock the potential of VLMs, we propose RDD and PNS, which enable prompt and enhancement training at both the image and instance levels.

**Relative Domain Distance.** We introduce an image-level enhancement strategy to better describe the distribution of the target domain. Specifically, we feed $\mathcal{R}_i^d$ into text encoder $\mathcal{T}$ to generate text embedding $\mathbf{t}_i^s$ and $\mathbf{t}_i^t$. The image representations $\mathbf{e}_i^{s\to t}$ and $\mathbf{e}_i^s$ are computed by feeding $F_{s\to t}$ and $F_s$ to an image encoder $\mathcal{I}$. Then, the distance between image prompt representations and the enhanced representations is pushed closer by a loss $\mathcal{L}_a$:

$$\mathcal{L}_a = \mathbb{E}\left[1 - f(\mathbf{e}_i^{s\to t}, \mathbf{t}_i^t)\right], \tag{6}$$

It is crucial to regulate URE to ensure that it captures the style of target domain while maintaining the integrity of semantic information for objects. Thus, the constraint loss $\mathcal{L}_s$ is defined as:

$$\mathcal{L}_s = \mathbb{E}\left[\|\mathbf{e}_i^s - \mathbf{e}_i^{s\to t}\|_1\right], \tag{7}$$

where $\|\cdot\|_1$ represents $\mathcal{L}1$ distance metric. Besides, to efficiently learn pseudo target domain representations, we introduce a relative domain distance loss $\mathcal{L}_r$:

$$\mathcal{L}_r = \mathbb{E}\left[\|(\mathbf{e}_i^s - \mathbf{e}_i^{s\to t}) - (\mathbf{t}_i^s - \mathbf{t}_i^t)\|_1\right]. \tag{8}$$

The idea behind RDD is to guide the model in directly searching the generic embedding. This is achieved by refining the prompt representations through adjusting the relative positions of the image representations based on their language representations.

**Positive-Negative Separation.** DetPro [12] learns diverse contextual prompt representations by training on

cropped proposals across varying IoU thresholds. Similarly, we adopt an instance-level training strategy by separating positive and negative proposals to capture diverse contexts of target domains. Initially, instance prompt representations $\mathbf{t}_p^t(c)$ and $\mathbf{t}_p^t(\mathcal{C}_{bg})$ are generated by $\mathcal{T}$ from $\mathcal{R}_p^t(c)$ and $\mathcal{R}_p^t(\mathcal{C}_{bg})$. The representations of positive and negative proposals $\mathbf{e}_p^{s\to t}(c)$ and $\mathbf{e}_p^{s\to t}(\mathcal{C}_{bg})$ are then derived by ROI-Align. For the positive proposal, the probability belonging to a foreground category $c \in C$ is computed as:

$$P_{pc} = \frac{\exp(f(\mathbf{e}_p^{s\to t}(c), \mathbf{t}_p^t(c)))}{\sum_{c'\in\mathcal{C}} \exp(f(\mathbf{e}_p^{s\to t}(c), \mathbf{t}_p^t(c')))}. \tag{9}$$

Subsequently, the positive proposal loss is defined as:

$$\mathcal{L}_c = \mathbb{E}\big[-\sum_{c\in\mathcal{C}} y_c \log P_{pc}\big]. \tag{10}$$

For the negative proposal, the probability belonging to background category is computed as:

$$P_{nc} = \frac{\exp(f(\mathbf{e}_p^{s\to t}(\mathcal{C}_{bg}), \mathbf{t}_p^t(\mathcal{C}_{bg})))}{\sum_{c'\in\mathcal{C}\cup\mathcal{C}_{bg}} \exp(f(\mathbf{e}_p^{s\to t}(\mathcal{C}_{bg}), \mathbf{t}_p^t(c')))}. \tag{11}$$

Then, the loss of negative proposal is defined as:

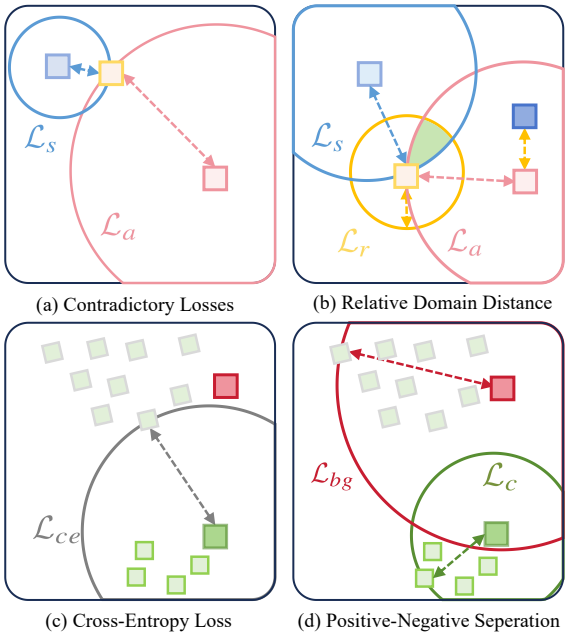
$$\mathcal{L}_{bg} = \mathbb{E}\big[-\sum_{c\in\mathcal{C}\cup\mathcal{C}_{bg}} y_{bg} \log P_{nc}\big], \tag{12}$$

where $y_{bg}$ is the reciprocal of the total number of classes, including the background class. By computing the difference from $y_{bg}$ and only considering positive differences, we limit the overconfidence of the background class.

### 3.4. Zero-shot Training and Testing

The training process consists of two stages, both conducted on the source domain. In the first stage, as illustrated in Fig. 2, we perform prompt and enhancement learning. Initializing MDP parameters randomly, we train URE to tranform source features into stylized pseudo-target representations. Simultaneously, these synthetic features drive MDP adaptation through alignment between visual and language representations. These processes are performed concurrently: the updated MDP refines URE, and the improved URE further enhances the pseudo target domain features, which in turn refine the MDP.

In the second stage, we freeze the optimized MDP, URE and text encoder $\mathcal{T}$ to preserve acquired domain knowledge, then fine-tune the CLIP backbone and RCNN detector. Following CLIP-GAP [49], we randomly apply enhancements to source features with a probability of 50%. The bounding boxes are predicted with a regression head, while classification scores of proposals are computed via CLIP. The




(a) Contradictory Losses (b) Relative Domain Distance

(c) Cross-Entropy Loss (d) Positive-Negative Seperation

Figure 3. **Illustration of Our Proposed Strategies.** Contradictory losses represent $\mathcal{L}_a$ and $\mathcal{L}_s$. The symbols ■ and ■ denote the representations of the source domain, pseudo target domain. ■ and ■ represent the representations of image prompts for the source and target domains. Additionally, the cross-entropy loss $\mathcal{L}_{ce}$ assigns proposals to all possible foregrounds. The symbols ■, ■, ■, and ■ indicate the representations of negative prompt, positive prompt, background objects, and foreground objects, respectively.

detector is trained with both regression and classification losses, following the framework of Faster R-CNN [46].

During inference on target domains, we disable URE transformations while maintaining identical regression and classification procedures, ensuring consistent feature processing between training and deployment phases.

### 3.5. Discussion

The transformation of source domain features requires a careful balance between preserving semantic integrity and enhancing feature style diversity. To achieve this, we need to align multi-modal representations while also address challenges arising from latent space variability and background contextual interference. Specifically, excessive enhancement must be controlled to avoid distorting the critical semantic information of the objects. However, the objectives of enhancement and constraint are inherently contradictory. As illustrated in Fig. 3(a), these two objectives interfere with each other during training, causing fluctuations that hinder the efficient learning of a generalized representation. To address this issue, we introduce representation regularization loss $\mathcal{L}_r$ to explore the multi-modal representation relationships between the source and target domains. It can constrain the search within the green overlap region, thereby

stabilizing the prompt and enhanement learning at image level, as illustrated in Fig. 3(b).

Proposals in object detection are categorized as positive or negative based on their IoU scores. These proposals, however, often contain not only the target object but also contextual information from the background and surrounding objects. As illustrated in Fig. 3(c), the vanilla cross-entropy loss is insufficient to model the varying latent space, thereby undermining the adaptation performance. To this end, we introduce $\mathcal{L}_c$ to shrink search space for foreground object and $\mathcal{L}_{bg}$ to assist former detecting background classes as shown in Fig. 3(d). Different from DetPro[12], which relies on pre-cropped proposals and only updates shared prompt representations with category keywords, our method cooperatively integrates prompt optimization with cross-domain representation synthesis. Most importantly, both the multi-view prompt representation and the enhancement representation are jointly trained in a unified framework.

# 4. Experiments

## 4.1. Datasets

We adopt nine benchmark object detection datasets across three cross-domain settings.

**Diverse Weather Conditions.** This setting incorporates five datasets representing various weather conditions from BDD100K [63], FoggyCityscapes [47], and Adverse-Weather [19]. Daytime Clear dataset comprises 19,395 training images. Night Clear dataset includes 26,158 images. Dusk Rainy and Night-Rainy datasets contain 3,501 and 2,494 images, respectively. Daytime Foggy dataset consists of 3,775 images. Training is exclusively conducted using the Daytime Clear dataset, while testing is performed on other adverse weather datasets.

**Cross-City Scenarios.** This setting involves datasets collected from different cities, highlighting geographical domain shifts. Specifically, Cityscapes dataset [10], comprising 2,975 images, is gathered from Germany. BDD100K dataset, collected across the US, contains 100k images, where 47,060 images captured under clear daytime are utilized. KITTI dataset [17] from Germany provides 7,481 labeled images. While Cityscapes and BDD100K share the same seven categories, we exclusively report results for the *car* category in KITTI. In this scenario, Cityscapes serves as the source domain, and BDD100K and KITTI serve as the target domains.

**Virtual-to-Real World Transitions.** Sim10K dataset [24], consisting of 10k synthetic images rendered from the Grand Theft Auto V game, is used as the source domain. The target domains in this setting are Cityscapes, BDD100K, and KITTI, and the evaluations are focused on the *car* category.

| Method | Venues | Daytime Foggy | Night Clear | Night Rainy | Dusk Rainy |
|---|---|---|---|---|---|
| Faster-RCNN[46] | NeurIPS'15 | 32.0 | 34.4 | 12.4 | 26.0 |
| S-DGOD [58] | CVPR'22 | 33.5 | 36.6 | 16.6 | 28.2 |
| CLIP-GAP [49] | CVPR'23 | 38.5 | 36.9 | 18.7 | 32.3 |
| PODA* [14] | ICCV'23 | 39.2 | 38.7 | 19.0 | 33.4 |
| OA-DG [27] | AAAI'24 | 38.3 | 38.0 | 16.8 | <u>33.9</u> |
| DAI-Net*[13] | CVPR'24 | 36.7 | <u>41.0</u> | 18.9 | 33.0 |
| PDD[29] | CVPR'24 | 39.1 | 38.5 | <u>19.2</u> | 33.7 |
| UFR[37] | CVPR'24 | <u>39.6</u> | 40.8 | <u>19.2</u> | 33.2 |
| UPRE(Ours) | - | **40.0** | **41.5** | **19.8** | **34.5** |

Table 1. Quantitative results on Diverse Weather Conditions. Results marked with (*) indicate those obtained through replication based on Faster-RCNN.

## 4.2. Implementation

We employ Faster R-CNN [46] with backbone initialized from the pre-trained CLIP ResNet-101 [20]. During training, input images are resized to 600×1067. The text encoder and the first two blocks of the image encoder remain frozen across all phases. Our model is optimized using SGD with momentum of 0.1, a weight decay of $10^{-4}$. All experiments are conducted on the Detectron2 platform using an NVIDIA 4090 GPU with a batch size of 4.

**Prompt and enhancement Learning.** Initially, we perform a warm-up, training the detector for 1.3k iterations on the source dataset with a learning rate of 0.001. We then proceed to prompt and enhancement training, where source features from the first layer are integrated with enhancements. This phase spans 5k iterations with a learning rate of 0.001, reduced by a factor of 10 at iteration 4.5k.

**Detector Fine-tuning.** During the fine-tuning phase, we freeze the parameters of MDP and URE. Training is conducted for 100k iterations, starting with a learning rate of 0.001, which decays by a factor of 0.1 at 40k iterations.

## 4.3. Main Results

We compare our method with the S-DGOD in object detection methods OA-DG [27], PDD [29], FDA [11] and UFR [37]. Besides, we also compare our method with the ZSDA in object detection methods DAI-Net [13] and PODA [14]. For each dataset, we follow the same protocol of existing works, and report the average precision ($AP_{50}$) of each class and the mean average precision (mAP) over all classes.

**Diverse Weather Conditions.** The qualitative and quantitative evaluation results are presented in Tab. 1 and Fig. 4. On average, our method improves mAP by 7.8% over Faster R-CNN across all conditions. In the Daytime Foggy scenario (Tab. 2), UPRE achieves an impressive 40.0% mAP, surpassing Faster R-CNN by 8.0%. Notably, we observe significant gains for challenging classes such as bus (+6.8%), motor (+9.9%), and rider (+9.1%). These improvements highlight our method's ability to capture fine-grained contextual details under low-visibility foggy conditions.

For the Night Rainy condition (Tab. 3), UPRE achieves



Figure 4. **Qualitative Results on Diverse Weather Conditions**. The top and bottom rows donate the results of OA-DG [27] and our proposed UPRE, respectively.

| Method | Bus | Bike | Car | Motor | Person | Rider | Truck | mAP |
|---|---|---|---|---|---|---|---|---|
| Faster-RCNN[46] | 30.7 | 26.7 | 49.7 | 26.2 | 30.9 | 35.5 | 23.2 | 32.0 |
| S-DGOD[58] | 32.9 | 28.0 | 48.8 | 29.8 | 32.5 | 38.2 | 24.1 | 33.5 |
| CLIP-GAP[49] | 36.1 | 34.3 | 58.0 | 33.1 | 39.0 | 43.9 | 25.1 | 38.5 |
| PODA* [14] | 34.6 | 34.1 | 50.7 | 31.2 | 38.5 | 44.0 | 25.9 | 39.2 |
| OA-DG[27] | - | - | - | - | - | - | - | 38.3 |
| DAI-Net*[13] | 33.7 | 32.8 | 57.9 | 33.1 | 37.9 | 41.0 | 24.7 | 36.7 |
| PDD[29] | 36.1 | 34.5 | 58.4 | 33.3 | **40.5** | 44.2 | 26.2 | 39.1 |
| UFR[37] | 36.9 | **35.8** | **61.7** | 33.7 | 39.5 | 42.2 | **27.5** | 39.6 |
| UPRE(Ours) | **37.5** | 35.6 | 58.4 | **36.1** | 39.0 | **44.6** | 27.0 | **40.0** |

Table 2. Per-class results on Daytime Clear to Daytime Foggy.

| Method | Bus | Bike | Car | Motor | Person | Rider | Truck | mAP |
|---|---|---|---|---|---|---|---|---|
| Faster-RCNN[46] | 22.6 | 11.5 | 27.7 | 0.4 | 10 | 10.5 | 19.0 | 12.4 |
| S-DGOD[58] | 24.4 | 11.6 | 29.5 | 9.8 | 10.5 | 11.4 | 19.2 | 16.6 |
| CLIP-GAP[49] | 28.6 | 12.1 | 36.1 | 9.2 | 12.3 | 9.6 | 22.9 | 18.7 |
| PODA* [14] | 26.8 | 11.7 | 36.4 | 9.0 | 12.7 | 11.5 | 28.8 | 19.0 |
| OA-DG[27] | - | - | - | - | - | - | - | 16.8 |
| DAI-Net*[13] | 24.6 | 11.7 | **37.5** | 9.0 | 12.9 | 11.2 | 21.4 | 18.9 |
| PDD[29] | 25.6 | 12.1 | 35.8 | **10.1** | **14.2** | 12.9 | 22.9 | 19.2 |
| UFR[37] | **29.2** | 11.8 | 36.1 | 9.4 | 13.1 | 10.5 | 23.3 | 19.2 |
| UPRE(Ours) | 27.7 | **15.7** | 36.7 | 9.4 | 13.5 | **14.7** | **25.7** | **19.8** |

Table 3. Per-class results on Daytime Clear to Night Rainy.

| Method | Bus | Bike | Car | Motor | Person | Rider | Truck | mAP |
|---|---|---|---|---|---|---|---|---|
| Faster-RCNN[46] | 36.8 | 15.8 | 50.1 | 12.8 | 18.9 | 12.4 | 39.5 | 26.0 |
| S-DGOD[58] | 37.1 | 16.9 | 50.9 | 13.4 | 19.7 | 16.3 | 40.7 | 28.2 |
| CLIP-GAP[49] | 37.8 | 22.8 | 60.7 | 16.8 | 26.8 | 18.7 | 42.4 | 32.3 |
| PODA* [14] | 38.2 | 24.9 | 58.1 | 18.5 | 27.5 | 17.7 | 43.6 | 33.4 |
| OA-DG[27] | - | - | - | - | - | - | - | 33.9 |
| DAI-Net*[13] | 36.9 | 26.1 | 59.0 | 17.5 | 28.2 | 16.5 | 42.7 | 33.0 |
| PDD[29] | **39.4** | 25.2 | 60.9 | **20.4** | **29.9** | 16.5 | 43.9 | 33.7 |
| UFR[37] | 37.1 | 21.8 | **67.9** | 16.4 | 27.4 | 17.9 | 43.9 | 33.2 |
| UPRE(Ours) | 39.1 | **29.2** | 65.5 | 19.7 | 28.6 | **19.2** | **45.7** | **34.5** |

Table 4. Per-class results on Daytime Clear to Dusk Rainy.

| Method | Bus | Bike | Car | Motor | Person | Rider | Truck | mAP |
|---|---|---|---|---|---|---|---|---|
| Faster-RCNN[46] | 37.7 | 30.6 | 49.5 | 15.4 | 31.5 | 28.6 | 40.8 | 34.4 |
| S-DGOD[58] | 40.6 | 35.1 | 50.7 | 19.7 | 34.7 | 32.1 | 43.4 | 36.6 |
| CLIP-GAP[49] | 37.8 | 22.8 | 60.7 | 16.8 | 26.8 | 18.7 | 42.4 | 36.9 |
| PODA* [14] | 39.6 | 34.7 | 58.5 | 19.5 | 37.4 | 27.2 | 42.6 | 38.7 |
| OA-DG[27] | - | - | - | - | - | - | - | 38.0 |
| DAI-Net*[13] | 41.6 | 38.6 | 59.2 | 20.5 | 45.2 | 31.1 | 43.7 | 41.0 |
| PDD[29] | 40.9 | 35.0 | 59.0 | **21.3** | 40.4 | 29.9 | 42.9 | 38.5 |
| UFR[37] | **43.6** | 38.1 | **66.1** | 14.7 | **49.1** | 26.4 | **47.5** | 40.8 |
| UPRE(Ours) | 42.4 | **38.7** | 61.4 | 20.6 | 45.7 | **31.6** | 44.3 | **41.5** |

Table 5. Per-class results on Daytime Clear to Night Clear.

a top mAP of 19.8%, surpassing Faster R-CNN by an average margin of 7.4%. Besides, we observe improvements in detecting occluded objects, such as bike (+4.2%) and rider (+4.2%). This demonstrates the effectiveness of our approach in handling nighttime rain effects, where poor illumination and reflective surfaces often degrade detection performance. In the Dusk Rainy setting (Tab. 4), UPRE achieves 34.5% mAP, surpassing Faster R-CNN and CLIP-GAP by 8.5% and 2.2%, respectively. Particularly, we achieve remarkable gains of 13.4% and 6.4% mAP for bike, underscoring the adaptability to transitional lighting and rainy conditions. Finally, in the Night Clear scenario (Tab. 5), our method achieves the best performance with an mAP of 41.5%, further confirming its adaptability across diverse scenarios.

**Cross-City Scenarios.** Geographic domain shifts between cities pose a significant challenge due to varying urban layouts and scene characteristics. However, previous methods [11, 27, 37, 49, 59] largely overlook cross-city adaptation scenarios. To address this limitation, we propose a novel approach that emphasizes more expressive prompt representations for handling cross-city variations.

Cityscapes primarily focuses on street scenes, while BDD100K and KITTI encompass diverse road environments, including highways and parking lots. To bridge these differences, we introduce a novel prompt representation $k_d$, which integrates weather and city scene descriptors. This representation enables the model to simulate knowledge transfer across different weather conditions and urban scenes, significantly enhancing its ability to handle geographic domain shifts. Quantitative results for Cityscapes to BDD100K and KITTI are presented in Tab. 6. Compared to state-of-the-art methods such as CLIP-GAP, PODA*, OA-DG, and DAI-Net*, our method (UPRE) achieves average mAP improvements of 1.9%, 1.4%, 1.4%, and 4.5%, respectively.

**Virtual to Real World Transitions.** The virtual world of GTAV completely differs from the real world in style. To address this gap, we harness easily accessible virtual data to facilitate detector adaptation to unseen real-world domains. Specifically, we integrate real-world weather conditions, such as rain, fog, and night, into the sunny virtual world, transforming it into a more realistic and diverse representation. This approach ensures consistent and transfer-

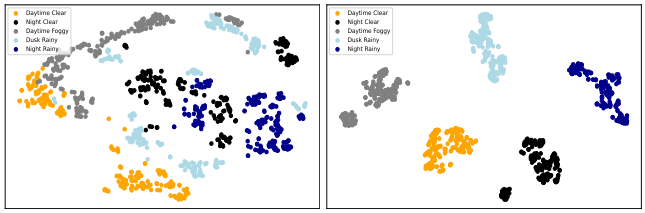



Figure 5. The t-SNE visualization of image embeddings across five domains using identical images and experimental settings. CLIP (Left) represents coarse generalization capabilities, whereas UPRE (Right) exhibits superior adaptation to each target domain.

| Method | BDD100K | | | | | | | | KITTI |
|---|---|---|---|---|---|---|---|---|---|
| | Bus | Bike | Car | Motor | Person | Rider | Truck | mAP | AP of Car |
| Faster-RCNN[46] | 21.9 | 22.7 | 36.9 | 17.8 | 24.1 | 25.4 | 22.6 | 24.5 | 72.5 |
| CLIP-GAP[49] | 22.4 | 23.2 | 47.2 | 20.2 | 20.1 | 30.2 | 15.0 | 26.3 | 72.9 |
| PODA* [14] | 23.1 | 22.7 | 46.5 | **20.4** | 20.9 | 31.0 | 18.5 | 26.7 | 73.6 |
| OA-DG[27] | 17.3 | 24.0 | 42.8 | 19.6 | 33.7 | 28.2 | 20.3 | 27.2 | 73.0 |
| DAI-Net*[13] | 18.1 | 22.0 | 36.9 | 17.3 | 30.1 | 29.2 | 18.4 | 22.8 | 71.3 |
| UPRE(Ours) | **23.8** | **24.5** | **48.7** | 19.7 | **34.7** | **31.9** | **21.5** | **28.7** | **74.3** |

Table 6. Quantitative results on Cross-City Scenarios.

able prompt representations across domains under varying weather conditions.

By simulating real world weather effects in the virtual domain, our method bridges the gap between synthetic and real world data. Quantitative results for virtual-to-real world transitions are presented in Tab. 7. Our method outperforms existing approaches across three real-world datasets, achieving mAP improvements of 1.6%, 1.5%, 1.8%, and 7.3% compared to CLIP-GAP, PODA*, OA-DG, and DAI-Net*, respectively. The superior performance of our method can be attributed to its ability to effectively simulate real-world conditions within the virtual domain.

**The visualization of the domain embeddings.** As depicted in Fig. 5 (left), although CLIP possesses generalization abilities, it lacks discrimination in embeddings for different weather conditions, such as the entanglement of Night Clear and Night Rainy embeddings. In Fig. 5 (right), UPRE enhances CLIP's performance and achieves superior domain adaptation capability.

### 4.4. Ablation Study

**Effect of Prompt Setting.** Tab. 8 compares different prompt settings, where either complete descriptions or only keywords are utilized. The learnable prompt with a complete prompt structure improves mAP by an average of 3.0%, demonstrating its effectiveness in capturing cross-domain knowledge. By maintaining the complete prompt, the learnable context can focus on acquiring domain-specific knowledge critical for cross-domain object detection, as described in Sec. 3.1. However, we find that using a learnable context with shared parameters hurts performance, e.g., reducing mAP by 2.4% on Night Rainy. This underscores the importance of prompt design, as mere learnability without proper constraints or separation is suboptimal.

**Analysis of Image-level Enhancement.** We study various loss combinations in the RDD in Tab. 9. Our default setting, which includes $\mathcal{L}_a$, $\mathcal{L}_s$ and $\mathcal{L}_r$, achieves the best and most stable performance compared to other configurations. Using only $\mathcal{L}_a$ leads to performance degradation, while adding $\mathcal{L}_s$ can result in instability due to conflicting optimization objectives. Notably, incorporating $\mathcal{L}_r$ into the enhancement further averagely improves mAP by 0.8% and MAD by 1.3%, demonstrating its effectiveness in aligning multi-modal representations and stabilizing training. This highlights the importance of balancing enhancement and constraint losses to achieve robust domain adaptation.

| Method | Cityscape | BDD100K | KITTI |
|---|---|---|---|
| Faster-RCNN[46] | 34.3 | 29.8 | 47.0 |
| CLIP-GAP[49] | 46.8 | 35.3 | 60.7 |
| PODA* [14] | 46.5 | 36.1 | 60.3 |
| OA-DG[27] | 47.0 | 35.6 | 59.5 |
| DAI-Net*[13] | 41.8 | 30.7 | 53.1 |
| UPRE(Ours) | **47.9** | **37.8** | **61.9** |

Table 7. Quantitative results on Virtual-to-Real World Transitions.

| Learnable | Static | Daytime Foggy | Night Clear | Night Rainy | Dusk Rainy |
|---|---|---|---|---|---|
| w/o | K | 37.2 | 37.5 | 17.0 | 31.7 |
| w/o | C | 38.2 | 39.3 | 17.9 | 32.2 |
| w | K | 38.7 | 40.1 | 18.6 | 33.0 |
| w + S | C | 38.0 | 39.7 | 17.4 | 32.2 |
| w | C | **40.0** | **41.5** | **19.8** | **34.5** |

Table 8. Effect study of prompt setting. w/o and w denote without and with learnable prompt, respectively. S represents using one shared learnable parameters. K defines that prompt only contains key word, while C represents using complete prompt in Sec. 3.1.

| $\mathcal{L}_a$ | $\mathcal{L}_s$ | $\mathcal{L}_r$ | Daytime Foggy | Night Clear | Night Rainy | Dusk Rainy | MAD |
|---|---|---|---|---|---|---|---|
| ✓ | - | - | 35.7 | 33.9 | 16.5 | 29.7 | 0.8 |
| ✓ | ✓ | - | 38.0 | 38.0 | 17.7 | 31.3 | 1.7 |
| - | - | ✓ | 37.2 | 37.5 | 17.2 | 30.8 | **0.3** |
| ✓ | ✓ | ✓ | **38.7** | **38.8** | **18.3** | **32.5** | 0.4 |

Table 9. Analysis of the RDD. The check mark denotes the addition of the loss. MAD is Mean Absolute Deviation, representing discrete fluctuation range of data.

## 5. Conclusion

In this paper, we propose a unified prompt and representation enhancement framework to mitigates both detection and domain biases for ZSDA in object detection. By jointly optimizing textual prompts and visual enhancements within object detection framework, our method effectively mitigates detection and domain biases. Specifically, MDP provides language priors for the target domain while also capturing diverse adaptation knowledge for cross-domain object detection. Meanwhile, URE diversifies domain styles under the guidance of the prompt representations. Additionally, our multi-level training framework, which incorporates the RDD and PNS enhancement strategies, effectively unlocks the potential of VLMs. Extensive experiments across multiple settings demonstrate the effectiveness of our approach in breaking the biases mitigation trade-off.

## References


[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1

[2] Hiroki Azuma, Yusuke Matsui, and Atsuto Maki. Zodi: Zero-shot domain adaptation with diffusion-based image transfer. *arXiv preprint arXiv:2403.13652*, 2024. 1

[3] Sikai Bai, Jie Zhang, Song Guo, Shuaicheng Li, Jingcai Guo, Jun Hou, Tao Han, and Xiaocheng Lu. Diprompt: Disentangled prompt tuning for multiple latent domain generalization in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27284–27293, 2024. 2

[4] Atif Belal, Akhil Meethal, Francisco Perdigon Romero, Marco Pedersoli, and Eric Granger. Multi-source domain adaptation for object detection with prototype-based mean teacher. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1277–1286, 2024. 1

[5] Shirsha Bose, Ankit Jha, Enrico Fini, Mainak Singha, Elisa Ricci, and Biplab Banerjee. Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5542–5552, 2024. 2

[6] Qinglong Cao, Zhengqin Xu, Yuntian Chen, Chao Ma, and Xiaokang Yang. Domain-controlled prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 936–944, 2024. 2, 4

[7] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 1

[8] De Cheng, Zhipeng Xu, Xinyang Jiang, Nannan Wang, Dongsheng Li, and Xinbo Gao. Disentangled prompt representation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23595–23604, 2024. 2

[9] Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546*, 2025. 1

[10] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[11] Muhammad Sohail Danish, Muhammad Haris Khan, Muhammad Akhtar Munir, M Saquib Sarfraz, and Mohsen Ali. Improving single domain-generalized object detection: A focus on diversification and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17732–17742, 2024. 2, 6, 7

[12] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 2, 4, 6

[13] Zhipeng Du, Miaojing Shi, and Jiankang Deng. Boosting object detection with zero-shot day-night domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12666–12676, 2024. 2, 6, 7, 8

[14] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul De Charette. Poda: Prompt-driven zero-shot domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18623–18633, 2023. 1, 2, 4, 6, 7, 8

[15] Guodong Fan, Shizhan Chen, Cuiyun Gao, Jianmao Xiao, Tao Zhang, and Zhiyong Feng. Rapid: Zero-shot domain adaptation for code search with pre-trained models. *ACM Transactions on Software Engineering and Methodology*, 33 (5):1–35, 2024. 2

[16] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 6

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[19] Mahmoud Hassaballah, Mourad A Kenk, Khan Muhammad, and Shervin Minaee. Vehicle detection and tracking in adverse weather using a deep learning framework. *IEEE transactions on intelligent transportation systems*, 22(7):4230–4242, 2020. 6

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[21] Weizhen He, Weijie Chen, Binbin Chen, Shicai Yang, Di Xie, Luojun Lin, Donglian Qi, and Yueting Zhuang. Unsupervised prompt tuning for text-driven object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2661, 2023. 2

[22] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 749–757, 2020. 1

[23] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 4

[24] 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? *arXiv preprint arXiv:1610.01983*, 2016. 6

[25] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. 1

[26] Mohammed Kutbi, Kuan-Chuan Peng, and Ziyan Wu. Zero-shot deep domain adaptation with common representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3909–3924, 2021. 2

[27] Wooju Lee, Dasol Hong, Hyungtae Lim, and Hyun Myung. Object-aware domain generalization for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2947–2955, 2024. 2, 6, 7, 8

[28] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4399–4409, 2021. 2

[29] Deng Li, Aming Wu, Yaowei Wang, and Yahong Han. Prompt-driven dynamic object-centric learning for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17606–17615, 2024. 6, 7

[30] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4

[31] Haochen Li, Rui Zhang, Hantao Yao, Xin Zhang, Yifan Hao, Xinkai Song, Xiaqing Li, Yongwei Zhao, Yunji Chen, and Ling Li. Da-ada: Learning domain-aware adapter for domain adaptive object detection. *Advances in Neural Information Processing Systems*, 37:103574–103598, 2024. 1

[32] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[34] Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[35] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 4

[36] Mingxing Li, Rui Wang, Lei Sun, Yancheng Bai, and Xiangxiang Chu. Next token is enough: Realistic image quality and aesthetic scoring with multimodal large language model. *arXiv preprint arXiv:2503.06141*, 2025. 1

[37] Yajing Liu, Shijun Zhou, Xiyao Liu, Chunhui Hao, Baojie Fan, and Jiandong Tian. Unbiased faster r-cnn for single-source domain generalized object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28838–28847, 2024. 2, 6, 7

[38] Rundong Luo, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Similarity min-max: Zero-shot day-night domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8104–8114, 2023. 2

[39] Giulio Mattolin, Luca Zanella, Elisa Ricci, and Yiming Wang. Confmix: Unsupervised domain adaptation for object detection via confidence-based mixing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 423–433, 2023. 1

[40] Shaobo Min, Hantao Yao, Hongtao Xie, Chaoqun Wang, Zheng-Jun Zha, and Yongdong Zhang. Domain-aware visual bias eliminating for generalized zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12664–12673, 2020. 1, 2

[41] Ba Hung Ngo, Nhat-Tuong Do-Tran, Tuan-Ngoc Nguyen, Hae-Gon Jeon, and Tae Jong Choi. Learning cnn on vit: A hybrid model to explicitly class-specific boundaries for domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28545–28554, 2024. 1

[42] 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. 1

[43] Kuan-Chuan Peng, Ziyan Wu, and Jan Ernst. Zero-shot deep domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 764–781, 2018. 2

[44] 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. 1, 3, 4

[45] Zhijie Rao, Jingcai Guo, Luyao Tang, Yue Huang, Xinghao Ding, and Song Guo. Srcd: Semantic reasoning with compound domains for single-domain generalized object detection. *arXiv preprint arXiv:2307.01750*, 2023. 2

[46] 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. 5, 6, 7, 8

[47] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 6

[48] Mainak Singha, Harsh Pal, Ankit Jha, and Biplab Banerjee. Ad-clip: Adapting domains in prompt space using clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4355–4364, 2023. 2, 3

[49] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3219–3229, 2023. 2, 4, 5, 6, 7, 8

[50] Vibashan VS, Poojan Oza, and Vishal M Patel. Instance relation graph guided source-free domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3520–3530, 2023. 1

[51] Jinghua Wang and Jianmin Jiang. Conditional coupled generative adversarial networks for zero-shot domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3375–3384, 2019. 2

[52] Jinghua Wang and Jianmin Jiang. Adversarial learning for zero-shot domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 329–344. Springer, 2020.

[53] Jinghua Wang and Jianmin Jiang. Learning across tasks for zero-shot domain adaptation from a single source domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6264–6279, 2021. 2

[54] Jinghua Wang, Ming-Ming Cheng, and Jianmin Jiang. Domain shift preservation for zero-shot domain adaptation. *IEEE Transactions on Image Processing*, 30:5505–5517, 2021. 1, 2

[55] Junxiang Wang, Guangji Bai, Wei Cheng, Zhengzhang Chen, Liang Zhao, and Haifeng Chen. Pond: Multi-source time series domain adaptation with information-aware prompt tuning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3140–3151, 2024. 2

[56] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR, 2022. 1

[57] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2022. 4

[58] Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 847–856, 2022. 2, 6, 7

[59] Fan Wu, Jinling Gao, Lanqing Hong, Xinbing Wang, Chenghu Zhou, and Nanyang Ye. G-nas: Generalizable neural architecture search for single domain generalization object detection. *arXiv preprint arXiv:2402.04672*, 2024. 2, 7

[60] Haifeng Xia and Zhengming Ding. Hgnet: Hybrid generative network for zero-shot domain adaptation. In *European Conference on Computer Vision*, pages 55–70. Springer, 2020. 1

[61] Yi Xin, Junlong Du, Qiang Wang, Ke Yan, and Shouhong Ding. Mmap: Multi-modal alignment prompt for cross-domain multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16076–16084, 2024. 2

[62] Senqiao Yang, Zhuotao Tian, Li Jiang, and Jiaya Jia. Unified language-driven zero-shot domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23407–23415, 2024. 1, 2

[63] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 6

[64] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 4

[65] Cairong Zhao, Yubin Wang, Xinyang Jiang, Yifei Shen, Kaitao Song, Dongsheng Li, and Duoqian Miao. Learning domain invariant prompt for vision-language models. *IEEE Transactions on Image Processing*, 2024. 2, 3, 4

[66] Xiaowei Zhao, Xianglong Liu, Duorui Wang, Yajun Gao, and Zhide Liu. Scene-adaptive and region-aware multi-modal prompt for open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16741–16750, 2024. 2