ODOV: Benchmark the Open-Domain Open-Vocabulary Object Detection

Yupeng Zhang[1,2] Ruize Han[3*] Fangnan Zhou[1] Wei Feng[1,2] Liang Wan[1,2]
[1]College of Intelligence and Computing, Tianjin University.
[2]Key Research Center for Surface Monitoring and Analysis of Relics, State Administration of Cultural Heritage.
[3]Faculty of Computer Science and Artificial Intelligence, Shenzhen University of Advanced Technology.
{zhangyupeng, zhoufangnan, wfeng, lwan}@tju.edu.cn, hanruize@suat-sz.edu.cn

Abstract

*Existing studies typically investigate domain shift and category shift as independent problems, however, in real-world scenarios, the two types of shifts often occur simultaneously and interact, leading to significant degradation in detection performance. To address this, we propose and systematically study a novel problem—Open-Domain Open-Vocabulary (ODOV) object detection—which aims to evaluate a model's ability to adapt to the compound domain and category shifts in **real-world environments**. We construct a new benchmark, OD-LVIS, which contains 46,949 images spanning 15 diverse real-world scenarios and 1,203 categories, for assessing object detection performance. Furthermore, we propose a novel ODOV detection baseline that fully leverages VLM's powerful multi-modal alignment capabilities and introduces two key mechanisms to enhance both category and domain generalization. One is the Domain-Agnostic Category Prompt (DAPmt), which strengthens category semantics while attenuating domain representations, enabling pure category representation. The other is the Domain Projection and Grafting (DP&G) module, which incorporates domain-specific features from input images, allowing the model to dynamically generalize across diverse open domains. These two components enable the model to maintain effective detection performance under simultaneous category and domain variations in real-world scenarios. We provide extensive benchmark evaluations for the proposed ODOV detection task and report experimental results. These results validate the soundness of the ODOV task, the practicality of the OD-LVIS dataset, and the superiority of the method.*

1. Introduction

Object detection is a fundamental task in computer vision, which aims at locating and identifying objects within images. Recent years have witnessed the rapid development of object detection, which, however, is still with unsolved problems for real-world applications. On one hand, due to the labor-intensive and costly manual annotation process for bounding boxes, the annotated object categories in the object detection dataset (for training) are limited, which are certainly smaller than the category vocabulary in the real world (**category shift**). On the other hand, most existing works have primarily focused on handling clear natural images. However, in many dynamically evolving real-world scenarios, such as autonomous driving, video surveillance, and internet search, the acquired images are often neither high-quality nor high-resolution. Instead, they may be plagued by various types of degradation or style differences (**domain shift**). For instance, the images captured in rainy or foggy weather (from the scene), with low resolution or various noises (during the imaging), or with different artistic styles (Internet data), *etc*.

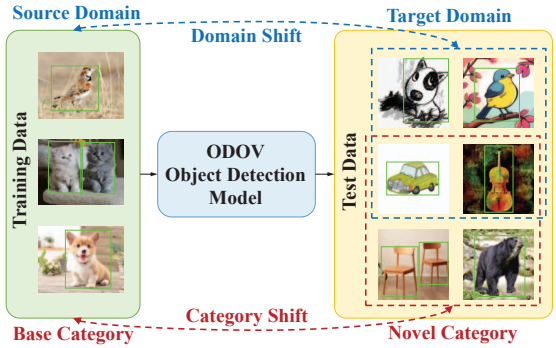

Figure 1. Illustration of the ODOV object detection model, showing the simultaneous occurrence of category and domain shifts. Blue dashed lines indicate domain shifts, primarily reflected in stylistic variations—for example, the training data comprise real-world photographs, while the test data may include cartoon or watercolor-style images. Red dashed lines indicate category shifts, characterized by changes in category distribution, such as new categories (*e.g.*, chairs, cars) appearing in the testing stage.

We find that current research on object detection (OD) in open scenes has noticed the above two problems, how-

*Corresponding author.

ever, they are often studied in isolation. Such as the one-shot OD [2, 15, 66, 76], open-world OD [36, 38, 59], and the most recent open-vocabulary OD [12, 62] are designed for the open-category detection problem. Besides, for the open-domain setting, domain adaption, domain generalization have been used for OD task [9, 33, 44, 52, 60, 61, 73], and some image restoration methods are also applied to the downstream OD task [54, 57, 58]. Actually, the *compound category and domain shifts are more common* in real-world scenes. For example, from indoor (training) to outdoor (testing) scenes, object categories often change significantly, while the visual domain of images also shifts, such as under different lighting, weather conditions, *etc*.

In this work, we propose a novel problem, namely open-domain open-vocabulary (ODOV) object detection to address the compound generalization of categories and domains, as shown in Fig. 1. Specifically, during training, we learn the detection model *only on the base categories with the natural domain*. In testing, we expect the model can detect and classify *the objects with unseen categories and domains*. Clearly, this problem is more practical yet challenging, it faces two main new difficulties.

- **Challenge distinction**: Category generalization and domain generalization usually rely on distinct mechanisms. Category generalization emphasizes enhancing a model's ability to recognize diverse objects, whereas domain generalization focuses on adapting to shifts in visual distributions, such as variations in artistic style or environmental conditions. Improper integration of these two mechanisms may undermine both category recognition and domain generalization, severely degrading overall detection performance. Thus, striking a balance between category and domain generalization constitutes the central challenge and research value of ODOV.
- **Benchmark shortage**: Existing large-scale detection datasets (*e.g*., LVIS) cover only a single domain (natural) images and fail to capture the complexity and diversity of real-world scenarios. In contrast, open-domain detection datasets (*e.g*., BDD100K [69]) include multiple visual domains but remain restricted to a few categories such as pedestrians and vehicles. As a result, current benchmarks either offer rich category diversity without multi-domain coverage or provide multi-domain variation with limited categories, and thus never simultaneously address the research needs of both domain and category shift.

Considering these two difficulties, we construct a new benchmark and establish the first unified framework for ODOV object detection, aiming to encourage attention to and exploration of solutions to this critical setting. *With respect to the benchmark*, we introduce a new evaluation dataset to systematically assess algorithms on the ODOV object detection task. Specifically, we design it based on the LVIS validation set and name it the Open-Domain LVIS benchmark (OD-LVIS). OD-LVIS contains 46,949 images, covering the same 1,203 categories as LVIS (including both base and novel categories), and is randomly distributed across 15 diverse and complex real-world domains, thereby more faithfully simulating continuously changing open environments. For models trained on LVIS base categories, OD-LVIS further incorporates rare (novel) categories and more challenging domain environments, thereby providing a more comprehensive and practical benchmark for ODOV object detection. *With respect to the method*, unlike previous approaches that leverage large vision-language models (VLMs) such as CLIP [43] and ALIGN [19] to address either domain generalization [49] or open-vocabulary detection [1, 12, 28, 32, 37, 55, 62, 65, 77], we propose a novel strategy, designed to fully exploit the cross-modal alignment capabilities of VLMs and uncover their generalization potential across both semantic and domain dimensions. By dynamically integrating domain representations with category descriptions, we construct adaptive prompts that align with the semantic and domain characteristics of the input image, effectively addressing the intertwined challenges of category shift and domain shift in real-world scenarios.

Specifically, we leverage CLIP, a model trained on massive categories and domains with powerful representational capacity. To enhance the robustness of text embeddings under category and domain shifts, we guide a large language model with instructions to generate descriptions that highlight distinctive category attributes while downplaying domain-specific information. These serve as Domain-Agnostic Category Prompts (DAPmt), mitigating the ambiguity (*e.g*., bat may refer either to a flying animal or a baseball bat) and domain constraints introduced by generic text prompts. We further design a Domain Projection and Grafting (DP&G) module that extracts domain-specific embeddings from input images and fuses them with DAPmt to generate **domain-customized category embeddings**. This mechanism enables the model to dynamically construct customized category text embeddings for each image, significantly enhancing its domain generalization capability in ODOV object detection. On OD-LVIS, we evaluate five backbones—CLIP RN50×16, CLIPSelf ViT-B/16 and ViT-L/14, and DeCLIP ViT-B/16 and ViT-L/14—and obtain gains of 1.8%, 2.1%, 1.5%, 1.8%, and 1.4% in AP, respectively. Moreover, our method consistently achieves the best overall performance compared with existing approaches.

2. Related Work

Open-vocabulary object detection (OVD) [71] aims to detect objects from novel categories unseen during training. Leveraging the zero-shot capabilities of vision-language models (VLMs), recent advancements in OVD have emerged. Works like [24–26, 50, 63] use region-aware training to integrate image-text pairs, improving clas-

sification, especially for novel categories. Studies such as [18, 22, 35, 77, 81] use large-scale image-text data, pre-trained VLMs, or pseudo-labels to predict novel categories and fine-tune the model with both pseudo and base labels. Other methods [1, 8, 12, 32, 41] employ knowledge distillation from VLMs to achieve OVD, transferring knowledge while enhancing localization. Some approaches build detectors on frozen VLMs [28, 37, 55, 56, 62], avoiding knowledge loss in fine-tuning and maximizing generalization. In this work, we use frozen VLMs as encoders, exploring their potential for category and domain generalization through prompt adjustment, revealing their strengths in open-domain and open-vocabulary object detection.

Domain Generalization (DG) based object detection aims to train a detector on multiple source domains to generalize to unseen target domains. Early work [33] used feature disentanglement for cross-domain generalization, followed by the Gated Disentangling Network [73], which activates feature channels for domain-invariant aspects. However, these methods rely on multiple domains and domain labels. Single Domain Generalized (SDG) object detection [60] addresses training with only one source domain. CDSD [60] separates domain-invariant from domain-specific representations, CLIP the Gap [52] uses pre-trained VLM, SRCD [44] reduces spurious correlations, and G-NAS [61] introduces a generalization loss to prevent Neural Architecture Search (NAS) overfitting. In this work, we adopt the SDG setting, using single-source domain data for training and evaluating multiple open domains. Our benchmark includes 15 diverse open domains, providing a more complex testbed than prior works.

Prompt learning for VLM adaptation. VLMs [19, 43, 51, 68, 70, 72] bridge image and text effectively. Pretrained on vast image-text pairs, models like CLIP [43] excel in open-scene recognition. However, adapting them to specific tasks with limited data is challenging. Text prompts guide VLMs, but even advanced prompt learning methods [23, 78, 79] still require training data. The recently proposed Test-time Prompt Tuning (TPT) [48] optimizes prompts by minimizing entropy through confidence-based selection, ensuring consistent predictions across different augmented views of each test sample. However, it remains inadequate when confronted with distribution shifts. Our work explores VLM adaptation for open-domain and open-vocabulary settings at test time. By fusing style features with category descriptors, we dynamically create task-relevant embeddings, enhancing generalization across new categories and domains. Unlike traditional test-time prompt tuning (*e.g.*, TPT), our approach does not require continuous adaptation on a single-style dataset. Instead, it directly addresses randomly varying distribution shifts, thereby substantially enhancing the model's cross-domain generalization across both categories and domains.

Table 1. Comparison of OD-LVIS and other detection datasets.

Name	#Year	#Image	#Category	#Domain	#Domain Type
MS COCO (val) [34]	2014	5,000	80	1	Normal
Cityscapes [5]	2016	3,475	8	1	Weather
Sim10K [21]	2016	10,000	1	1	Weather
WIDER FACE [67]	2016	32,000	1	1	Normal
Foggy Cityscapes [46]	2018	3,475	8	1	Weather
Clipart [17]	2018	1,000	20	1	Art
Watercolor [17]	2018	1,905	6	1	Art
Comic [17]	2018	1,905	6	1	Art
UFDD [39]	2018	884	1	1	Normal
RTTS [29]	2018	9,109	5	1	Weather
Objects365 [47]	2019	100,000	365	1	Normal
LVIS (val) [13]	2019	19,809	1,203	1	Normal
BDD100K [69]	2020	41,986	10	12	Weather
ODinW35 [30]	2022	20,000	314	1	Normal
MSOSB [75]	2024	76,146	80	5	Art
OD-LVIS	2025	**46,949**	**1,203**	**15**	**Art, Weather, Noise, Blur...**

3. ODOV Object Detection Benchmark

3.1. Motivation

As shown in Table 1, we summarize existing object detection datasets as follows. 1) *General-category detection*: Clipart [17], Watercolor [17], Comic [17], and MSOSB [75], with artistic styles; MS COCO [34], Objects365 [47], ODinW [30] and LVIS [13] consisting of common images with LVIS offering a more extensive set of categories often used for OV tasks. 2) *Specific-category detection*: WIDER FACE [67], a dataset representing common photographic scenes, focused solely on the face category. Besides, pedestrian detection and vehicle detection are also very popular with a series of datasets [6, 40, 74], *etc*. 3) *Traffic scene detection*: Cityscapes [5], BDD100K [69], Foggy Cityscapes [46], UFDD [39], RTTS [29], and Sim10K [21] contain unique weather conditions to test the domain generalization performance of models, but primarily focus on a limited set of objects common in traffic scenarios. Overall, the above **existing datasets do not meet the requirement of containing simultaneous open-domain and open-category scenes**.

This way, we build OD-LVIS, a dedicated evaluation benchmark designed specifically for ODOV object detection, encompassing a diverse range of categories and complex real-world scenarios. Specifically, we select the LVIS as our basic dataset. On the one hand, we retain all object categories from LVIS (including both base and novel categories) to ensure the benchmark's category diversity. On the other hand, to enhance the domain diversity, we extend the data by considering two aspects, *i.e.*, the *image styles* and the *imaging conditions*. Specifically, for the former, we collect nine distinct styles, *i.e.*, black-and-white pencil sketches, color pencil sketches, oil paintings, cartoons, watercolors, symbolism, impressionism, gothic art, and lyrical abstraction. For the latter, we also consider six imaging conditions, *i.e.*, rain, haze, illumination variations, low resolution, noise (Gaussian white noise and salt-and-pepper noise), and blur (Gaussian blur, motion blur, and out-of-

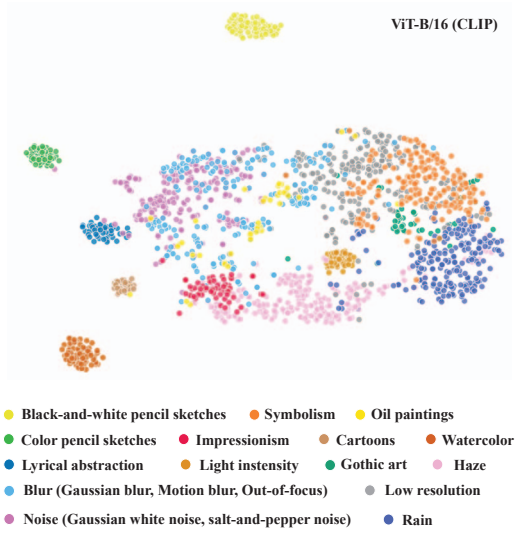

Figure 2. The t-SNE visualization of feature-level style statistics from the CLIP image encoder (ViT-B/16) outputs for 'apple' samples. The samples cluster by domain characteristics.

focus). As a result, OD-LVIS comprises 46,949 images across 15 different real-world domains, shares categories with LVIS, and adheres to its annotation guidelines, establishing a standardized benchmark for ODOV object detection evaluation. As shown in Fig. 2, we extract features of the apple category samples using the CLIP image encoder (ViT-B/16) and visualize them with t-SNE, clearly showing that the samples cluster according to their domain characteristics. Finally, since OD-LVIS shares categories with LVIS, it can be combined with LVIS as training data to further evaluate the domain generalization ability of object detection models across 15 different scenarios.

3.2. Preliminary Experiments on OD-LVIS

We investigate the impact of OD-LVIS on existing detector. Take the haze and noise for example, we evaluate the SOTA OVD method F-ViT (DeCLIP ViT-B/16) [56] in Table 2, where m and k control the levels of haze and noise, respectively (larger values indicate more severe shifts). Results show that as domain shifts increase, model performance declines significantly, demonstrating the necessity of the studying on ODOV detection and OD-LVIS benchmark.

Please refer to the supplementary material for details on Data Generation, Cleaning, and Annotation.

Table 2. Results on images with varying degrees of degradation.

Data	Degree / AP (%)	Degree / AP (%)
Source	- / 26.5	- / 26.5
Haze	m=0.05 / 24.5	m=0.08 / 20.7
Noise	k=0.04 / 13.7	k=0.06 / 12.3

4. The Proposed Baseline Method

4.1. Problem Formulation

We first provide the problem formulation of the proposed ODOV object detection problem. During the training stage, we use the data from the *single source domain* (*i.e.*, the natural image domain). Specifically, the training image dataset is notated as $\mathcal{D}^{\text{train}} = \{\mathbf{I}_i^{\text{train}}, L_i\}_{i=1}^{N}$, where N is the number of training images, $\mathbf{I}_i$ denotes an image from the source domain with the detection latel L_i for it. The label L_i is composed of $\{\mathbf{b}_i, \mathbf{c}_i\}$, in which $\mathbf{b}_i$ indicates all the annotated object bounding boxes in $\mathbf{I}_i$, and the corresponding object categories are stored in $\mathbf{c}_i$. Note that, all the (annotated) object categories contained in L_i of the training set $\mathcal{D}^{\text{train}}$ are from the *base category* set, *i.e.*, $\mathcal{C}^{\text{base}}$.

During the testing stage, the ODOV object detection task requires the model to be applied under open-domain conditions as well as open-vocabulary conditions. Specifically, the testing images are from hybrid open domains (*e.g.*, with various image styles), which are denoted as $\mathcal{D}^{\text{test}} = \{\mathbf{I}_j^{\text{test}}\}_{j=1}^{M}$ and M is the dataset scale. For each test image $\mathbf{I}_j^{\text{test}}$, the desired output is the predicted object bounding boxes with corresponding categories, *i.e.*, $\{\mathbf{b}_j, \mathbf{c}_j\}$. Note that, following the open-vocabulary detection setting, the predicted objects contain both the *base categories* $\mathcal{C}^{\text{base}}$ (appearing in training) and the *novel categories* $\mathcal{C}^{\text{novel}}$ (unseen during training), which are combined as the open-vocabulary category set, *i.e.*, $\mathcal{C}^{\text{open}} = \mathcal{C}^{\text{base}} + \mathcal{C}^{\text{novel}}$.

4.2. Overview of The Method

We aim to fully exploit the generalization capability of pre-trained VLMs (*e.g.*, CLIP) to handle the ODOV object detection, and propose a baseline method – O**DOV** Detec**tor** (**DVtor**). As illustrated in Fig. 3, we adopt a frozen CLIP encoder and propose a novel strategy that dynamically generates customized Domain-category embeddings based on test images, thereby accommodating both category diversity and domain variability. On the one hand, we introduce the *Domain-Agnostic Category Prompt (DAPmt)* to emphasize intrinsic category semantics while avoiding the limitations imposed by style information. On the other hand, we design a learnable *Domain Projection and Grafting (DP&G) network*, which integrates domain-specific features of the input image with DAPmt.

Since the training images originate only from a single source domain, we align prompts with image embeddings during training through implicit domain augmentation and contrastive learning.

4.3. Domain-Agnostic Category Prompt (DAPmt)

VLMs such as CLIP associate the visual images with text captions through large-scale pre-training. CLIP based open-scene detection has achieved high accuracy on multiple datasets, utilizing the manually crafted prompts (*e.g.*, 'a photo of {}') as the text prompts. Such simple prompts are easy to obtain, but can not make full use of the VLMs, which is especially highlighted in the ODOV setting since the same category of object from different domains shows

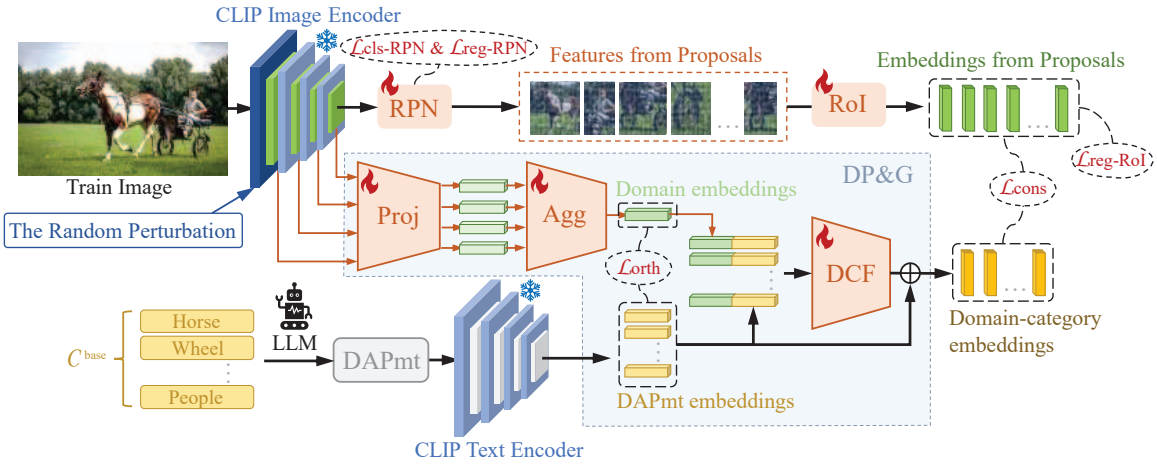

Figure 3. Overview of DVtor. We use the generalization ability of pre-trained VLM (CLIP) by decoupling domain-aware representations and combining them with domain-agnostic category embeddings to adapt to the ODOV object detection task.

various characteristics. Some recent works [1, 12, 28, 62, 77] manually design the prompt templates considering the dataset characteristics. For example, using templates like 'a black-and-white photo of {}' or 'a rainy photo of {}', which are time-consuming and also difficult to generalize to other datasets. Although several learnable and automated prompt generation methods [7, 10, 20, 27, 31] have been proposed, they often rely on generic category descriptions, overlooking category-specific attributes and the style information, as their focus remains on conventional OVD tasks.

To address the above challenge of generating low-ambiguity prompts applicable across diverse open domains, we advocate the use of large language models (LLMs) to generate such prompts. Specifically, the LLM produces category-specific yet domain-independent attribute descriptions for each category in the dataset, serving as Domain-Agnostic Category Prompt (DAPmt). This approach avoids the limitations caused by domain shifts and reduces ambiguities arising from category names through rich attribute descriptions. By generating text with fine-grained attributes, it not only enhances the discriminability of similar categories but also preserves generalizability across different domains.

Specifically, for each category, we provide the large language model (LLM, *e.g.*, ChatGPT-4o) with a unified instruction template: 'A/An {category} has/is [Appearance feature1], [Appearance feature2], ...'. During generation, we guide the model to ensure that attribute elements are non-redundant, emphasize each category's distinctive visual properties, and disregard domain information to preserve cross-domain consistency. For example, for the category {Airplane}, the generated description is: 'An airplane is a large vehicle with long wings and a streamlined body'. The resulting descriptions exhibit strong robustness to domain shifts, enabling them to generalize across diverse styles and contexts. At the same time, we deliberately exclude information such as color that is easily influenced by domain factors, reducing interference from illumination or artistic styles. Although the method still requires minor manual adjustments, it significantly reduces the workload compared with traditional approaches that rely heavily on handcrafted prompts. Overall, this strategy aims to produce domain-agnostic category prompts that enable the model to form a general and precise understanding of categories, thereby improving its robustness in ODOV object detection tasks.

4.4. Domain Projection and Grafting (DP&G)

To further address the ever-changing domain shift problem, we propose *Domain Projection and Grafting (DP&G)*. The core idea is to extract the **domain embeddings** (such as style or imaging conditions) of each input image and integrate them with the category text embeddings of DAPmt, producing customized embeddings that jointly capture both domain and category information. In the following, we present the detailed implementation of this strategy.

Domain-aware embedding extraction. First, we extract the multi-level domain-aware features. Specifically, during training on single source-domain data, we first apply random perturbations to the features based on the classic AdaIN [16] and NP [9] to simulate domain variations [1]. Then, for an input image $\mathbf{I}$, we extract the (disturbed) feature map $\mathbf{F}_l \in \mathbb{R}^{W\times H\times C}$ from the l^{th} layer of the image encoder $\mathcal{E}$, in which W, H, C denote the height, width, and channel, respectively. For the c^{th} channel, the mean μ_l^c and standard deviation σ_l^c are computed as $\mu_l^c{=}\text{avg}(\mathbf{F}_l^c(w,h)), \sigma_l^c{=}\sqrt{\text{avg}\big(\mathbf{F}_l^c(w,h){-}\mu_l^c\big)^2}$. We contact all C channels of μ_l^c and σ_l^c to obtain the mean/standard deviation vectors, as $\mu_l \in \mathbb{R}^C$ and $\sigma_l \in \mathbb{R}^C$. This way, we can obtain the initial domain-aware features by concatenating the mean and standard de-

[1]Following the rationale in [16], the mean and standard deviation in the feature map implicitly represent the domain-aware information.

viation vectors as $\mathcal{E}_l = [\mu_l; \sigma_l]$.

The domain feature of layer l namely $\mathcal{E}_l$ is then processed through adaptive average pooling and a fully connected layer to align the dimension with that of the DAPmt embeddings (Sec. 4.3). As shown in Fig. 3, we extract the domain features from multiple layers. These features are then input into an aggregation network to get the final domain embedding $\mathbf{E}_{\text{domain}}$, in which a set of learnable weights dynamically balances the contributions from each layer through weighted integration.

We then consider to project the domain embeddings $\mathbf{E}_{\text{domain}}$ into a regularized embedding space. Specifically, we hope the learned domain-aware embeddings are *irrelevant* (orthometric) to the category embeddings. For this purpose, we apply an orthogonality constraint loss, which minimizes the similarity between the domain embeddings and the category DAPmt embeddings, thereby effectively separating them, as

$$\mathcal{L}_{\text{orth}} = \sum_{g=1}^{\mathcal{C}^{\text{base}}} \text{cossim}\left(\mathbf{E}_{\text{domain}}, \mathbf{E}^g_{\text{category}}\right), \tag{1}$$

where $\mathbf{E}_{\text{domain}}$ represents the learnable domain embeddings, $\mathbf{E}^g_{\text{category}}$ denotes the DAPmt embeddings for the g^{th} category, $\mathcal{C}^{\text{base}}$ is the total number of (base) categories, and cossim denotes the cosine similarity.

Grafting the domain-specific embedding to DAPmt. Finally, we develop a **domain and category embedding fusion (DCF)** module. Specifically, we concatenate the domain and category embeddings, followed by a multi-layer perception (MLP) to project it into the fusion space as

$$\mathbf{E}^g_{\text{fusion}} = \text{MLP}([\mathbf{E}_{\text{domain}}; \mathbf{E}^g_{\text{category}}]), \tag{2}$$

which denotes the fused embedding for the g-th category.

After that, to preserve core semantic information from the category text descriptions, a residual connection of $\mathbf{E}^g_{\text{category}}$ is applied as

$$\mathbf{E}^g_{\text{DCF}} = \alpha \cdot \mathbf{E}^g_{\text{fusion}} + (1 - \alpha) \cdot \mathbf{E}^g_{\text{category}}, \tag{3}$$

where a learnable parameter α within the range $[0, 1]$ is used to dynamically balance the direct fusion embedding and the original category embedding $\mathbf{E}^g_{\text{category}}$.

After the DCF module, the proposed method grafts the domain-specific embedding (from the input image) on the domain-agnostic category embedding (from DAPmt), to generate the ***domain-customized category embeddings for each given image and category prompt***, thereby improving the ODOV detection ability.

Finally, during training, disturbed visual features extracted from the image encoder form $\mathbf{F}l$, while ROI Align [14] generates the object-level features $\mathbf{F}\text{RoI}$. The domain–category embedding $\mathbf{E}\text{DCF}^g$ is aligned with $\mathbf{F}\text{RoI}$ using a contrastive loss as

$$\mathcal{L}_{\text{cons}} = 1 - \frac{\mathbf{E}^g_{\text{DCF}} \cdot \mathbf{F}_{\text{RoI}}}{\|\mathbf{E}^g_{\text{DCF}}\| \|\mathbf{F}_{\text{RoI}}\|}, \tag{4}$$

where $\|\cdot\|$ is the L_2 norm of a vector.

4.5. Implementation Details

Training stage. During training, we apply random perturbations to the mean and standard deviation of the first and second layer features output by the image encoder, and simultaneously, the multi-layer domain features in DP&G are drawn from layers [3, 5, 7, 11] of ViT-B/16, layers [6, 10, 14, 23] of ViT-L/14, and all layers of ResNet. Our method is trained using 16 3090 GPUs, with a batch size of 10 per GPU. We use AdamW configured with a learning rate of 10^{-4} and a weight decay of 0.1. Training is conducted on the LVIS training set for 50 epochs.

Detailed illustrations and descriptions of testing stage are provided in the supplementary material.

5. Experimental Results

5.1. Setup

ODOV settings. For the open domain, we train the models on single-domain natural images (LVIS training set) and test on all 15 open-domains in OD-LVIS, similar to the single-domain generalization setting [42]. For the open vocabulary, we follow the OV-LVIS in ViLD [12] for dataset splitting. Among all categories in OD-LVIS, *i.e.*, in OV-LVIS, 405 'frequent' and 461 'common' categories are assigned as base categories for training, while 337 'rare' categories are novel categories only for testing. Note that, OD-LVIS is the *only benchmark meeting the ODOV setting*, so all main experiments are conducted on it.

Evaluation methods. To build the benchmark evaluation of the ODOV detection, we selected several mainstream VLM-based OVD methods for comparison on OD-LVIS. Specifically, we include the transfer learning approaches, *i.e.*, F-VLM [28], OWL-ViT [37], CLIPSelf [62], DeCLIP [56], MM-OVOD [65], and OV-DQUO [55], and several knowledge distillation methods, *i.e.*, RKDWTF [1], DK-DETR [32], RegionCLIP [77], and region-aware training method YOLO-World [4], YOLOE [53]. We also include two recent domain generalization (DG) methods, ALT [11], ABA [3], NP [9], MixStyle [80], and PhysAug [64] for comparison.

Evaluation metrics. For the evaluation metrics, the average precision on 'frequent' and 'common' categories, denoted as AP_f and AP_c, respectively, serves as the metric for base categories, while the average precision on 'rare' categories, denoted as AP_r, is used to evaluate novel categories. The average precision for all categories is denoted as AP.

5.2. Main Results on OD-LVIS

Table 3 shows the results of all comparative methods and our DVtor on OD-LVIS. We can first see that, the proposed DVtor (DeCLIP ViT-L/14) with the DAPmt and DP&G network, achieves the best performance among all competitors. Specifically, when using CLIPSelf ViT-L/14

Table 3. Comparison with SOTA on OD-LVIS (%).

Method	Backbone	Training Data	AP_f	AP_c	AP_r	AP
RegionCLIP [77]	RN50*	CC3M	16.6	13.0	9.7	13.9
	RN50x4*		19.5	15.8	12.4	16.7
OWL-ViT [37]	ViT-B/16	O365 + VG	13.1	13.9	13.2	13.5
	ViT-L/14		22.1	21.6	19.9	21.5
RKDWTF [1]	RN50* Base	LVIS-base + IN-L	14.6	12.4	8.7	12.6
	RN50* RKDPIS		13.4	12.1	10.3	12.3
	RN50* WTF		14.0	12.5	11.3	12.9
	RN50* WTF8x		15.8	14.3	11.9	14.5
DK-DETR [32]	RN50	LVIS-all	21.1	19.4	15.3	19.4
MM-OVOD [65]	RN50*_Agg	LVIS-base	20.5	19.8	14.0	19.0
+ DAPmt			20.8	20.4	14.5	19.5
MM-OVOD		LVIS-base + IN-L	20.4	20.4	15.9	19.6
+ DAPmt			21.1	20.9	16.0	20.1
YOLO-World [4]	YOLOv8-L*	O365 + GoldG	21.9	19.1	19.3	20.2
+ DAPmt			22.0	19.6	19.7	20.6
YOLOE [53]	YOLOv11-L*		13.7	8.6	6.8	10.3
+ DAPmt			14.3	9.5	9.1	11.3
OV-DQUO [55]	ViT-B/16	LVIS-base	12.8	14.8	14.8	14.0
+ DAPmt			13.5	15.3	15.8	14.7
OV-DQUO	ViT-L/14		16.4	20.6	21.2	19.1
+ DAPmt			17.2	21.5	22.3	20.0
F-VLM (CLIP) [28]	RN50x16		16.7	14.4	13.7	15.2
+ DAPmt			17.2	15.3	14.7	16.0
+ DP&G			17.5	16.2	15.2	16.5
DVtor (CLIP)			**17.6**	**16.9**	**15.8**	**17.0**
F-ViT (CLIPSelf) [62]	ViT-B/16		17.1	12.0	12.2	14.0
+ DAPmt			17.5	12.7	13.2	14.7
+ DP&G			18.6	13.8	13.7	15.7
DVtor (CLIPSelf)			**19.0**	**14.3**	**14.0**	**16.1**
F-ViT (CLIPSelf)	ViT-L/14		22.5	21.3	20.2	21.6
+ DAPmt			22.8	21.7	20.8	22.0
+ DP&G			23.5	22.2	21.5	22.6
DVtor (CLIPSelf)			**23.9**	**22.9**	**21.6**	**23.1**
F-ViT (DeCLIP) [56]	ViT-B/16		17.8	12.9	13.2	14.9
+ DAPmt			18.3	13.6	14.6	15.6
+ DP&G			18.8	14.1	14.9	16.1
DVtor (DeCLIP)			**19.5**	**14.7**	**15.5**	**16.7**
F-ViT (DeCLIP)	ViT-L/14		23.0	21.7	21.4	22.2
+ DAPmt			23.6	22.0	22.6	22.7
+ DP&G			24.1	22.3	22.8	23.1
DVtor (DeCLIP)			**24.9**	**22.6**	**23.2**	**23.6**

Notes: IN-L denotes the inclusion of images corresponding to the 997 categories shared between ImageNet-21k-P [45] and LVIS, '*' indicates that the backbone is not initialized with CLIP, O365 is an abbreviation for Objects365, CC3M, GoldG, and VG are all publicly available datasets.

(304.43M) as the backbone, DVtor surpasses F-ViT with the same backbone by 1.4%, 1.6%, 1.4%, and 1.5% on the frequent, common, rare, and overall categories, respectively. Moreover, with the DeCLIP ViT-L/14 backbone, DVtor further achieves improvements of 1.9%, 0.9%, 1.8%, and 1.4% over F-ViT across the same four category groups. For smaller networks, such as CLIPSelf and DeCLIP with ViT-B/16 (86.26M), our method achieves overall AP improvements of 2.1% and 1.8% compared with F-ViT using the same backbone. In addition, DVtor with RN50×16 (167.33M) surpasses F-VLM with the same backbone by 1.8% in overall AP, and notably, our RN50×16 model (17.0% @AP) even outperforms F-VLM with a much larger RN50×64 backbone (16.9% @AP). These results demonstrate the remarkable advantages of the DAPmt and DP&G modules in enhancing generalization for open-domain detection. Moreover, the overall performance of all methods on OD-LVIS remains relatively low, highlighting the challenging nature of this benchmark and the substantial room for further improvement.

5.3. Ablation Study

Effectiveness of the proposed DAPmt. As shown in Table 3, when we integrate DAPmt ('+ DAPmt') into F-VLM (CLIP RN50×16), F-ViT (CLIPSelf ViT-B/16 and ViT-L/14), and F-ViT (DeCLIP ViT-B/16 and ViT-L/14), it achieves improvements of 1.0%, 1.0%, 0.6%, 0.6%, and

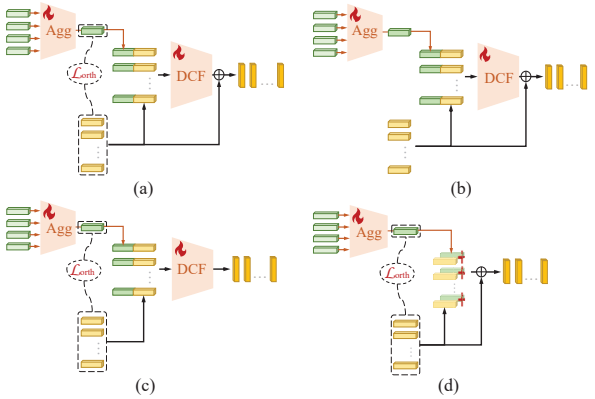

Figure 4. The different fusion structures.

Table 4. Comparison of the fusion structures on OD-LVIS (%).

Method	Struct.	AP_f	AP_c	AP_r	AP
DVtor (DeCLIP ViT-B/16)	(a)	19.5	14.7	15.5	16.7
	(b)	17.6	12.7	13.1	14.7
	(c)	18.1	13.1	9.5	14.4
	(d)	18.0	11.9	13.2	14.5

1.2% on the 'rare' categories, respectively. These results indicate that introducing DAPmt in the training process enables our model to effectively learn the category semantics from CLIP, thus enhancing the model's generalization capabilities. Moreover, we apply DAPmt to the inference of various OVD frameworks, including MM-OVOD, YOLO-World, YOLOE, and OV-DQUO, all of which achieve consistent and significant performance improvements, further demonstrating the generality and robustness of DAPmt.

Effectiveness of the proposed DP&G. To validate the DP&G solely, we train DP&G based on fixed prompt templates (*e.g.*, 'a photo of {}') instead of using DAPmt. As shown in Table 3, we integrate the DP&G network into both F-VLM and F-ViT frameworks ('+ DP&G'). The results show that, when using CLIPSelf ViT-B/16 and ViT-L/14 as backbones, DVtor surpasses F-ViT by 1.5%, 1.8%, 1.5%, and 1.7%, as well as by 1.0%, 0.9%, 1.3%, and 1.5% on the frequent, common, rare, and overall categories, respectively. Similar performance improvements are also observed when F-ViT adopts DeCLIP ViT-B/16 and ViT-L/14 as backbones, and when F-VLM uses RN50×16 as the backbone. We attribute these improvements to the DP&G module's ability to effectively extract and integrate visual domain information, enriching the semantic representation of customized prompts and further enhancing the model's generalization across diverse scenarios. Moreover, compared with models using only 'DAPmt' or 'DP&G', DVtor (*i.e.*, DAPmt + DP&G) achieves superior results across all metrics, indicating that removing either module leads to a decline in performance, thereby further verifying the effectiveness and complementarity of both components.

Ablation study of different structures. As shown in Fig. 4, we compare: (a) Ours; (b) Removing the orthogonal regularization loss; (c) Removing the residual connection; (d) Replacing the DCF module with a direct weighted

Table 5. Comparison with OVD methods on OV-LVIS (%).

Method	Backbone	Training Data	AP_f	AP_c	AP_r	AP
RegionCLIP [77]	RN50*	CC3M	34.0	27.4	17.1	28.2
	RN50x4*		36.9	32.1	22.0	32.3
OWL-ViT [37]	ViT-B/16	O365 + VG	-	-	20.6	27.2
	ViT-L/14		-	-	31.2	34.6
RKDWTF [1]	RN50* Base	LVIS-base + IN-L	26.4	19.4	12.2	20.9
	RN50* RKDPIS		25.5	20.9	17.3	22.1
	RN50* WTF		26.7	21.4	17.1	22.8
	RN50* WTF8x		29.1	25.0	21.1	25.9
MM-OVOD [65]	RN50*_Avg	LVIS-base	-	-	20.7	30.5
	RN50*_Agg		-	-	19.3	30.6
	RN50*_Avg	LVIS-base + IN-L	-	-	26.5	32.8
	RN50*_Agg		-	-	27.3	33.1
DK-DETR [32]	RN50	LVIS-all	40.2	32.0	22.2	33.5
YOLO-World [4]	YOLOv8-L*	O365 + GoldG	35.4	24.9	22.9	28.7
YOLOE [53]	YOLOv11-L*		36.5	35.0	29.1	35.2
OV-DQUO [55]	ViT-B/16	LVIS-base	23.8	27.7	29.4	26.5
	ViT-L/14		28.5	36.0	39.5	33.7
F-VLM (CLIP) [28]	RN50x16		-	-	30.4	32.1
F-ViT (CLIPSelf) [62]	ViT-B/16		29.1	21.8	25.3	25.2
	ViT-L/14		35.6	34.6	34.9	35.1
F-ViT (DeCLIP) [56]	ViT-B/16		29.8	22.4	26.8	26.0
	ViT-L/14		36.5	35.2	37.2	36.0
DVtor (CLIP)	RN50x16		33.0	34.1	33.1	33.5
DVtor (CLIPSelf)	ViT-B/16		30.4	23.2	26.3	26.6
	ViT-L/14		36.9	35.8	36.4	36.3
DVtor (DeCLIP)	ViT-B/16		31.0	23.7	28.1	27.3
	ViT-L/14		37.6	35.9	39.0	37.1

summation. As shown in Table 4, the structure in (b) results in an overall performance drop. This is because the domain embeddings extracted from the input images may still contain a small amount of object category information, and without maintaining orthogonality, the fusion text features contain reduced category discrimination than the original, which negatively impacts the model's recognition accuracy. Structure (c) leads to a significant drop in accuracy for novel categories, indicating that the fusion process may cause some semantic information of categories to be lost. This indicates that the residual connection helps retain original semantic information, thus enhancing the model's generalization capability for novel categories. Structure (d) also causes an accuracy drop. This direct fusion method is too simplistic to effectively integrate the two feature types, resulting in suboptimal performance.

5.4. More Results

Comparison with OVD methods on OV-LVIS [12]. We further evaluate our proposed method on the OV-LVIS validation set, as shown in Table 5. Our DVtor (based on DeCLIP ViT-L/14) achieves the highest overall AP among all methods. Moreover, when using the same backbone, the proposed method consistently outperforms F-ViT (CLIPSelf and DeCLIP) and F-VLM across all four settings. Overall, the performance improvement of our method on OV-LVIS is relatively modest compared to that on OD-LVIS, mainly because OV-LVIS exhibits lower domain diversity, whereas our method demonstrates stronger competitiveness on the more diverse OD-LVIS benchmark.

Comparison with DG Methods on OD-LVIS. In Table 6, we compare our proposed method with five domain generalization (DG) methods based on F-ViT using DeCLIP ViT-B/16 as the backbone on OD-LVIS. It can be observed that our method achieves the best performance.

Cross-Dataset Transfer Results. Table 7 presents the cross-dataset transfer results from OV-LVIS to Objects365 [47]. We compare our proposed method with Detic [81], MM-OVOD [22], F-VLM [28], and F-ViT (CLIPSelf [62] and DeCLIP [56]), using the standard bounding box AP metric on Objects365 for evaluation. In all experiments, Detic and MM-OVOD are trained on LVIS-all, where IN-L models use ImageNet-21k-P as additional weak supervision, while F-ViT (CLIPSelf and DeCLIP) and our method are trained on the base categories of OV-LVIS and evaluated on the Objects365 validation set. Following MM-OVOD, we define the bottom one-third of categories in Objects365, ranked by frequency, as rare categories. When using CLIP RN50×16 as the backbone, our method improves over F-VLM by 1.3% in AP_r and 2.3% in AP^{50}, surpasses F-ViT (CLIPSelf) by 0.7% and 0.5%, and outperforms F-ViT (DeCLIP) by 1.1% and 2.3%, respectively. With ViT-L/14 as the backbone, our method also achieves notable improvements over both F-ViT (CLIPSelf) and F-ViT (DeCLIP). Overall, these results demonstrate that our method exhibits significant advantages and strong generalization capabilities in cross-dataset transfer tasks.

Please refer to the supplementary material for visualized results and analysis, as well as the limitations.

Table 6. Comparison with DG methods (%).

Method	AP_f	AP_c	AP_r	AP
F-ViT (DeCLIP ViT-B/16)	17.8	12.9	13.2	14.9
+ ALT [11]	18.2	13.2	13.1	15.1
+ ABA [3]	18.0	13.4	13.8	15.3
+ NP [9]	18.6	14.0	14.2	15.8
+ MixStyle [80]	17.9	13.3	13.5	15.1
+ PhysAug [64]	18.3	13.6	13.7	15.5
DVtor (DeCLIP)	19.5	14.7	15.5	16.7

Table 7. Cross-dataset main results on Objects365 (%).

Method	Backbone	Training Data	AP_r	AP	AP^{50}
Detic [81]	RN50*	LVIS-all	9.5	13.9	19.7
		LVIS-all + IN-L	12.4	15.6	22.2
MM-OVOD [65]		LVIS-all	10.1	14.8	21.0
		LVIS-all + IN-L	13.1	16.6	23.1
F-VLM (CLIP) [28]	RN50x16	LVIS-base	14.9	16.2	25.3
F-ViT (CLIPSelf) [62]	ViT-B/16		16.8	19.0	32.3
	ViT-L/14		21.7	23.7	39.2
F-ViT (DeCLIP) [56]	ViT-B/16		17.6	20.2	33.1
	ViT-L/14		22.3	24.5	39.8
DVtor (CLIP)	RN50x16		16.2	17.9	27.6
DVtor (CLIPSelf)	ViT-B/16		17.5	19.2	32.8
	ViT-L/14		22.3	24.0	39.7
DVtor (DeCLIP)	ViT-B/16		18.7	21.1	35.4
	ViT-L/14		23.7	25.0	40.9

6. Conclusion

We have proposed to study a new yet practical problem of ODOV object detection, by considering both the domain and category shifts. For this purpose, we construct the benchmark OD-LVIS containing 15 domains and 1,203 categories. We have also developed a baseline method for ODOV detection, which can generate the domain-agnostic text prompts for category embedding, as well as the domain embeddings using a domain projection and grafting network. By combining both of them, we obtain the customized domain-specific category embedding for each test image, which well adapts ODOV detection. We provide the benchmark evaluation of a series of SOTA methods on OD-LVIS. Through these efforts, we hope to pave the way for the study of this new yet significant problem.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants U2574216 and 62402490; in part by the Emerging Frontiers Cultivation Program of Tianjin University Interdisciplinary Center; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515010101.

References

[1] Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35:33781–33794, 2022. 2, 3, 5, 6, 7, 8

[2] Ding-Jie Chen, He-Yen Hsieh, and Tyng-Luh Liu. Adaptive image transformer for one-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12247–12256, 2021. 2

[3] Sheng Cheng, Tejas Gokhale, and Yezhou Yang. Adversarial bayesian augmentation for single-source domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11400–11410, 2023. 6, 8

[4] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 6, 7, 8

[5] 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/CVF conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3

[6] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *2009 IEEE conference on computer vision and pattern recognition*, pages 304–311, 2009. 3

[7] Penghui Du, Yu Wang, Yifan Sun, Luting Wang, Yue Liao, Gang Zhang, Errui Ding, Yan Wang, Jingdong Wang, and Si Liu. Lami-detr: Open-vocabulary detection with language model instruction. In *European Conference on Computer Vision*, pages 312–328. Springer, 2024. 5

[8] 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. 3

[9] Qi Fan, Mattia Segu, Yu-Wing Tai, Fisher Yu, Chi-Keung Tang, Bernt Schiele, and Dengxin Dai. Towards robust object detection invariant to real-world domain shifts. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023. 2, 5, 6, 8

[10] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *European conference on computer vision*, pages 701–717. Springer, 2022. 5

[11] Tejas Gokhale, Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, Chitta Baral, and Yezhou Yang. Improving diversity with adversarially learned transformations for domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 434–443, 2023. 6, 8

[12] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2, 3, 5, 6, 8

[13] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 3

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[15] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation. *Advances in neural information processing systems*, 32:1–10, 2019. 2

[16] 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. 5

[17] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 3

[18] Joonhyun Jeong, Geondo Park, Jayeon Yoo, Hyungsik Jung, and Heesu Kim. Proxydet: Synthesizing proxy novel classes via classwise mixup for open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2462–2470, 2024. 3

[19] 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, 2021. 2, 3

[20] Sheng Jin, Xueying Jiang, Jiaxing Huang, Lewei Lu, and Shijian Lu. Llms meet vlms: Boost open vocabulary object detection with fine-grained descriptors. *arXiv preprint arXiv:2402.04630*, 2024. 5

[21] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016. 3

[22] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-modal classifiers for open-vocabulary object detection. In *International Conference on Machine Learning*, pages 15946–15969, 2023. 3, 8

[23] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 3

[24] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Contrastive feature masking open-vocabulary vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15602–15612, 2023. 2

[25] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Detection-oriented image-text pretraining for open-vocabulary detection. *arXiv preprint arXiv:2310.00161*, 2023.

[26] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11144–11154, 2023. 2

[27] Jooyeon Kim, Eulrang Cho, Sehyung Kim, and Hyunwoo J Kim. Retrieval-augmented open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17427–17436, 2024. 5

[28] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022. 2, 3, 5, 6, 7, 8

[29] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28:492–505, 2018. 3

[30] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35:9287–9301, 2022. 3

[31] Liunian Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang. Desco: Learning object recognition with rich language descriptions. *Advances in Neural Information Processing Systems*, 36:37511–37526, 2023. 5

[32] Liangqi Li, Jiaxu Miao, Dahu Shi, Wenming Tan, Ye Ren, Yi Yang, and Shiliang Pu. Distilling detr with visual-linguistic knowledge for open-vocabulary object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6501–6510, 2023. 2, 3, 6, 7, 8

[33] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8771–8780, 2021. 2, 3

[34] 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, 2014. 3

[35] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in neural information processing systems*, 36, 2024. 3

[36] Zeyu Ma, Yang Yang, Guoqing Wang, Xing Xu, Heng Tao Shen, and Mingxing Zhang. Rethinking open-world object detection in autonomous driving scenarios. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1279–1288, 2022. 2

[37] M Minderer, A Gritsenko, A Stone, M Neumann, D Weissenborn, A Dosovitskiy, A Mahendran, A Arnab, M Dehghani, Z Shen, et al. Simple open-vocabulary object detection with vision transformers. arxiv 2022. *arXiv preprint arXiv:2205.06230*, 2, 2022. 2, 3, 6, 7, 8

[38] Sahal Shaji Mullappilly, Abhishek Singh Gehlot, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Hisham Cholakkal. Semi-supervised open-world object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4305–4314, 2024. 2

[39] Hajime Nada, Vishwanath A Sindagi, He Zhang, and Vishal M Patel. Pushing the limits of unconstrained face detection: a challenge dataset and baseline results. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10, 2018. 3

[40] Yanwei Pang, Jiale Cao, Yazhao Li, Jin Xie, Hanqing Sun, and Jinfeng Gong. Tju-dhd: A diverse high-resolution dataset for object detection. *IEEE Transactions on Image Processing*, 30:207–219, 2021. 3

[41] Chau Pham, Truong Vu, and Khoi Nguyen. Lp-ovod: Open-vocabulary object detection by linear probing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 779–788, 2024. 3

[42] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12556–12565, 2020. 6

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021. 2, 3

[44] 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, 3

[45] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 7

[46] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 3

[47] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 3, 8

[48] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 3

[49] Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing clip to out-of-distributions. In *International Conference on Machine Learning*, pages 31716–31731, 2023. 2

[50] Hwanjun Song and Jihwan Bang. Prompt-guided transformers for end-to-end open-vocabulary object detection. *arXiv preprint arXiv:2303.14386*, 2023. 2

[51] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 3

[52] 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, 3

[53] Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yoloe: Real-time seeing anything. *arXiv preprint arXiv:2503.07465*, 2025. 6, 7, 8

[54] Jing Wang, Meimei Xu, Huazhu Xue, Zhanqiang Huo, and Fen Luo. Joint image restoration for object detection in snowy weather. *IET Computer Vision*, 2024. 2

[55] Junjie Wang, Bin Chen, Bin Kang, Yulin Li, Weizhi Xian, Yichi Chen, and Yong Xu. Ov-dquo: Open-vocabulary detr with denoising text query training and open-world unknown objects supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7762–7770, 2025. 2, 3, 6, 7, 8

[56] Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. Declip: Decoupled learning for open-vocabulary dense perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14824–14834, 2025. 3, 4, 6, 7, 8

[57] Xiaofeng Wang, Xiao Liu, Hong Yang, Zhengyong Wang, Xiaoyue Wen, Xiaohai He, Linbo Qing, and Honggang Chen. Degradation modeling for restoration-enhanced object detection in adverse weather scenes. *IEEE Transactions on Intelligent Vehicles*, 2024. 2

[58] Yongzhen Wang, Xuefeng Yan, Kaiwen Zhang, Lina Gong, Haoran Xie, Fu Lee Wang, and Mingqiang Wei. Togethernet: Bridging image restoration and object detection together via dynamic enhancement learning. In *Computer Graphics Forum*, pages 465–476, 2022. 2

[59] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting everything in the open world: Towards universal object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11433–11443, 2023. 2

[60] 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, 3

[61] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5958–5966, 2024. 2, 3

[62] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023. 2, 3, 5, 6, 7, 8

[63] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7031–7040, 2023. 2

[64] Xiaoran Xu, Jiangang Yang, Wenhui Shi, Siyuan Ding, Luqing Luo, and Jian Liu. Physaug: A physical-guided and frequency-based data augmentation for single-domain generalized object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21815–21823, 2025. 6, 8

[65] Yifan Xu, Mengdan Zhang, Xiaoshan Yang, and Changsheng Xu. Exploring multi-modal contextual knowledge for open-vocabulary object detection. *arXiv preprint arXiv:2308.15846*, 2023. 2, 6, 7, 8

[66] Hanqing Yang, Sijia Cai, Hualian Sheng, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Yong Tang, and Yu Zhang. Balanced and hierarchical relation learning for one-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7591–7600, 2022. 2

[67] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5525–5533, 2016. 3

[68] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 3

[69] 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. 2, 3

[70] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3

[71] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 2

[72] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 18123–18133, 2022. 3

[73] Haozhuo Zhang, Huimin Yu, Yuming Yan, and Runfa Wang. Gated domain-invariant feature disentanglement for domain generalizable object detection. *arXiv preprint arXiv:2203.11432*, 2022. 2, 3

[74] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3221, 2017. 3

[75] Yupeng Zhang, Shuqi Zheng, Ruize Han, Yuzhong Feng, Junhui Hou, Linqi Song, Wei Feng, and Liang Wan. Rethinking the one-shot object detection: Cross-domain object search. In *ACM Multimedia 2024*. 3

[76] Yizhou Zhao, Xun Guo, and Yan Lu. Semantic-aligned fusion transformer for one-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7601–7611, 2022. 2

[77] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022. 2, 3, 5, 6, 7, 8

[78] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 3

[79] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3

[80] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Mixstyle neural networks for domain generalization and adaptation. *International Journal of Computer Vision*, 132 (3):822–836, 2024. 6, 8

[81] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368, 2022. 3, 8