OSLoPrompt: Bridging Low-Supervision Challenges and Open-Set Domain Generalization in CLIP

Mohamad Hassan N C[1] Divyam Gupta[1] Mainak Singha[1] Sai Bhargav Rongali[1] Ankit Jha[2]
Muhammad Haris Khan[3] Biplab Banerjee[1]
[1]Indian Institute of Technology Bombay [2]The LNM Institute of Information Technology (LNMIIT)
[3]Mohamed Bin Zayed University of Artificial Intelligence

Abstract

We introduce Low-Shot Open-Set Domain Generalization (LSOSDG), a novel paradigm unifying low-shot learning with open-set domain generalization (ODG). While prompt-based methods using models like CLIP have advanced DG, they falter in low-data regimes (e.g., 1-shot) and lack precision in detecting open-set samples with fine-grained semantics related to training classes. To address these challenges, we propose OSLoPrompt*, an advanced prompt-learning framework for CLIP with two core innovations. First, to manage limited supervision across source domains and improve DG, we introduce a domain-agnostic prompt-learning mechanism that integrates adaptable domain-specific cues and visually guided semantic attributes through a novel cross-attention module, besides being supported by learnable domain- and class-generic visual prompts to enhance cross-modal adaptability. Second, to improve outlier rejection during inference, we classify unfamiliar samples as "unknown" and train specialized prompts with systematically synthesized pseudo-open samples that maintain fine-grained relationships to known classes, generated through a targeted query strategy with off-the-shelf foundation models. This strategy enhances feature learning, enabling our model to detect open samples with varied granularity more effectively. Extensive evaluations across five benchmarks demonstrate that* OSLoPrompt *establishes a new state-of-the-art in LSOSDG, significantly outperforming existing methods.*[1]

1. Introduction

Domain Generalization (DG) [3] enhances model robustness by training across multiple source domains to enable generalization to unseen targets. Traditional DG typically operates within a fully-supervised, closed-set framework, leveraging abundant labeled data and assuming semantic alignment across domains [18, 42, 60]. However, fields like healthcare often face data scarcity, motivating Few-Shot DG (FSDG) [36], which extends conventional few-shot learning [54] in the DG setting. Yet, the closed-set assumption falls short in real-world applications, where models encounter novel classes dynamically. ODG [4, 43] addresses this by incorporating mechanisms to reject outliers during inference but assumes a fully supervised training regimen, overlooking low-shot scenarios.

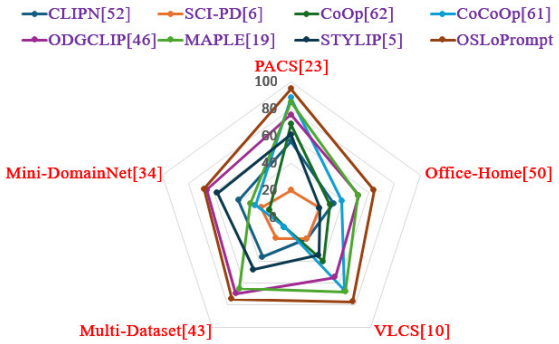

Figure 1. **Harmonic score (H-score) (between known and novel class performances) comparisons** of various CLIP-based DG/ODG/open-set recognition techniques versus our approach in LSOSDG setting with one-training example per known class, demonstrating the improved performances of OSLoPrompt.

To this end, we introduce LSOSDG, a novel paradigm extending conventional ODG by imposing data scarcity on training classes (1/5-shot). Unlike domain-adaptive few-shot open-set learning [31], which relies on predefined meta-training and meta-testing domains, LSOSDG requires learning a domain-generic classifier from limited source supervision while detecting outliers at inference without prior knowledge. This setup is ideal for dynamic, open-world scenarios—such as autonomous vehicles adapting to shifting conditions or medical diagnostics identifying emerging diseases—where data is sparse, domain shifts are unpredictable, and new classes appear spontaneously.

Existing CNN-based FSDG models [26, 36] struggle in LSOSDG due to conflicting supervision requirements. In

[1]https://github.com/has97/Osloprompt

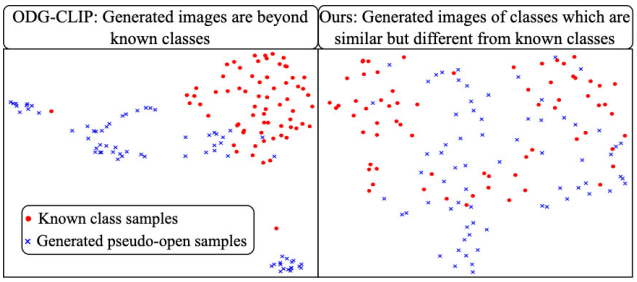

Figure 2. t-SNE [48] of **known-class** and **pseudo-open samples** generated by ODG-CLIP [46] (left) and our method (right). Our approach produces fine-grained pseudo-open samples, creating a sharper closed-open class boundary and enhancing feature learning, resulting in an improvement over [46] on Mini-DomainNet [34] (Table 3), significantly boosting open-set detection.

contrast, domain-agnostic prompt learning[2] in VLMs like CLIP [37] shows promise for DG by visual content and style conditioned prompt learning [5, 45]. However, these methods are challenged in LSOSDG, where undefined open spaces and limited supervision hinder visual style discernment and performance (Fig. 1).

ODG-CLIP [46] addresses ODG with a unified `Unknown` class prompt and diffusion-generated pseudo-open training samples for the outlier class. While effective in ODG, it falters in low-shot settings, akin to [5]. Additionally, these pseudo-open samples lack the nuanced distinctions needed to separate visually coherent known and open-set samples, a notable gap in ODG research (Fig. 2).

Besides, our analysis (Table 3) highlights three main limitations in current domain-agnostic prompt modeling for LSOSDG. **First**, existing methods over-rely on textual cues, missing the synergy between visual and textual prompts needed for robust, domain-agnostic features in multi-domain DG. Effective DG requires learnable visual prompts to complement textual ones—a frequently overlooked aspect. **Second**, DG prompting approaches like [5, 46] use learnable contexts [62] that lack the structured knowledge found in manually crafted prompts, as demonstrated by KG-CoOp [56]; we argue for a hybrid prompt strategy to reduce misguidance under limited supervision. **Finally**, while recent work [9, 21, 29, 35, 47, 55] shows that attribute-based prompting improves generalization through class-sharable attributes, these approaches often isolate visual and textual attributes. Instead, we advocate cross-referencing visual and textual attributes for richer semantic explanations of visual objects with a frozen CLIP backbone.

Our contributions: We propose a novel model, Open-Set LOw-shot PROMPT learning (OSLOPROMPT), to address these challenges, with two key innovations:

- Fine-grained pseudo-open sample synthesis: Building on [46], we introduce a novel prompt to classify open samples as `Unknown` and employ pseudo-open image synthesis to train this prompt. Unlike the generic pseudo-open sample generation in [46], which struggles with detecting fine-grained open samples, we propose a controlled synthesis approach where generated pseudo-open samples retain fine-grained semantic discrimination with respect to the training classes (Fig. 2). This is achieved by strategically prompting GPT-4o [1] and stable diffusion [40].

[2]Domain-agnostic prompts are designed in a way that can be applied to any unknown target domain without prior knowledge.

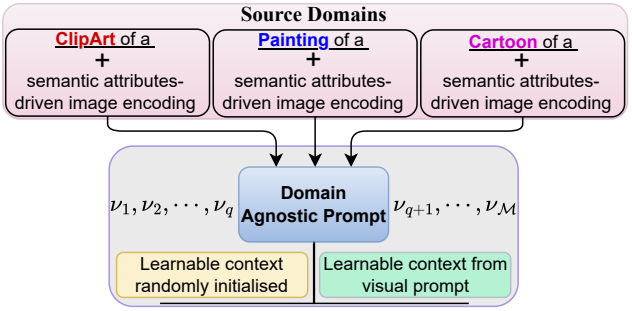

Figure 3. **Proposed prompt learning**: We develop a novel strategy for learning domain-agnostic prompts with tokens $\{\nu_{1:q:\mathcal{M}}\}$, inheriting context from source-specific prompts enriched with image-to attributes encodings. Some tokens also integrate knowledge from visual prompts spanning all training domains and classes. We differ considerably from the DG literature.

- Effective prompt learning in low-data regime: In OSLOPROMPT, we improve domain-agnostic semantic prompt learning by addressing two key challenges: enhancing prompt versatility across domains and categories, and bridging the visual-semantic gap effectively.

To make them universally applicable, we align the domain-agnostic prompt's context tokens with manually crafted source-domain-specific counterparts enriched with semantic attribute-based visual encodings. These attributes, generated by GPT-4o, capture class-specific details and are further refined through a novel image-to-attributes cross-attention module. This structured regularization enhances both domain independence and semantic richness.

To connect visual and semantic information further, we introduce a multi-modal prompting approach. Here, a learnable visual prompt within CLIP's image encoder captures diverse source-domain features, which are then used to initialize some of the tokens in the agnostic prompts. This setup aligns the prompts more closely with the visual characteristics in the data, supporting better generalization (Fig. 3). In summary, our major contributions are:

[-] We introduce the LSOSDG problem setting, highlighting the limitations of existing DG methods and presenting a comprehensive solution through OSLOPROMPT.

[-] We enhance prompt generalization using structured regularization, visually-driven semantic attribute guidance, and visual prompt learning. Additionally, we propose synthesizing superior pseudo-open samples to train a versatile outlier detection strategy through targeted prompting.

[-] We benchmark OSLOPROMPT in LSOSDG setting on five datasets, backed by thorough ablation studies.

2. Related Works

DG and FSDG: DG was introduced to address domain shifts by training models across diverse source domains, thereby enhancing performance on unseen targets [3, 18, 60]. Beyond traditional methods, recent advancements in prompt-based approaches within VLMs [5, 7, 15, 44, 46] have shown considerable improvements in DG. In contrast, FSDG [8, 25, 26, 36, 39] aims to generalize with limited supervision. It relies on transferrable knowledge from a distinct set of classes, often utilizing meta-learning and domain-specific adaptation. Nevertheless, we note that LSOSDG and FSDG follow distinct problem settings.

ODG: ODG extends DG to handle unknown classes during inference, a concept initially introduced by [43] using domain-augmented meta-learning. Unlike Open Set Recognition (OSR) [2, 22] and Open Set Domain Adaptation (OSDA) [32], ODG faces the unique challenge of operating in an inductive setting without target domain data during training. MEDIC [53] advanced this field by matching domain and class-wise gradients, while methods like [4] focus on disentangled feature learning and adversarial sample synthesis for open-set detection. Prompt tuning in foundation models, such as those based on CLIP, and generalizable prompting techniques [5, 52, 61, 62] have also been adapted for ODG. However, these methods often struggle with diverse domains and establishing optimal confidence thresholds for detecting open samples. ODG-CLIP [46] introduces an `unknown` class for the outliers and generates pseudo-open samples using a diffusion model [40] to train the respective prompt. Concurrent approaches like [6] explore style perturbation and knowledge distillation but generally speaking, all of them face difficulties with fine-grained differentiation between known and open classes.

VLMs and prompt learning: Vision-Language foundation models such as CLIP [37], ALIGN [16], LiT [59], FILIP [57], and Florence [58] have made significant strides in image recognition by leveraging large-scale image-text pairings to capture rich multi-modal representations. Despite their strength in open-vocabulary tasks, adapting these models to specific challenges while preserving generalization remains complex. Approaches like CoOp [62], CoCoOp [61], MaPLe [19], PromptSrc [20], and Kg-CoOp [56] enhance token embeddings for task-specific adaptation, while recent methods such as [9, 21, 24, 29, 35, 41, 47] incorporate textual or visual attributes into prompts to better generalize from base to novel classes (more discussions in **Sup Mat**). Furthermore, [14] explores prompt regularization for low-shot training but is limited to typical zero-shot inference. Contrary to the literature, we introduce a comprehensive domain-agnostic and enhanced prompting strategy to handle the nuanced structure of LSOSDG, showing a more balanced performance on both the known and open-set classes than other counterparts (Table 1).

3. Proposed Methodology

We consider a scenario with $\mathcal{N}$ source domains, each defined as $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ where $x^s \in \mathcal{X}_s$ are input images and $y^s \in \mathcal{Y}_s$ are the corresponding labels. These domains, denoted by $\mathcal{D} = \{\mathcal{D}_s\}_{s=1}^{\mathcal{N}}$, have unique distributions ($\mathcal{P}(\mathcal{D}_s) \neq \mathcal{P}(\mathcal{D}_{s'})$ for $s \neq s'$) and feature a mix of shared and domain-specific classes, often resulting in class imbalance. Classes across the source domains are collectively represented by $\mathcal{C} = \bigcup_{s=1}^{\mathcal{N}} \mathcal{Y}_s$, with each class typically having limited samples (e.g., 1-shot or 5-shot). During testing, the model encounters an unlabeled target domain $\mathcal{T}$ with its dataset $\mathcal{D}_t = \{x_j^t\}_{j=1}^{n_t}$, where $\mathcal{P}(\mathcal{D}_t)$ differs from $\mathcal{P}(\mathcal{D}_s)$ for all source domains. The target domain's label set, $\mathcal{Y}_t$, includes both known classes, $\mathcal{Y}_t^{\text{known}} = \mathcal{C}$, and novel classes or outliers, $\mathcal{Y}_t^{\text{novel}} = \mathcal{Y}_t \setminus \mathcal{Y}_t^{\text{known}}$.

To address the LSOSDG challenge, we frame it as a $|\mathcal{C}|+1$-class classification task within the CLIP framework, as inspired by [46]. We design prompts for each of the $|\mathcal{C}|$ known classes and an additional `Unknown`-class prompt for cohesively classifying outlier instances during inference, without utilizing open-set knowledge during training.

Our approach, OSLOPROMPT, focuses on two key strategies: **(a)** We synthesize robust pseudo-open samples that are proximal to known classes in the embedding space to train the `Unknown` class prompt. This ensures clear separation and reduces inference ambiguity. **(b)** We develop a domain-agnostic prompt learning strategy that distills contexts from structured, hand-crafted domain-specific prompts with semantic attribute encodings derived from the visual space, besides utilizing insights from learnable, domain- and class-generic visual prompts. The detailed architecture of OSLOPROMPT is illustrated in Fig. 4. Further specifics for both methods are provided below.

At the onset, we denote the frozen CLIP text and image encoders as $\mathcal{F}_t$ and $\mathcal{F}_v$, respectively. The textual encodings for the s^{th} source domain and class label y are represented as *[Domain$_s$]* and *[CLS$_y$]*.

3.1. Synthesis of fine-grained pseudo-open images

To generate fine-grained pseudo-open samples that are closely related to known classes, we leverage Stable Diffusion [40], which offers substantial improvements over previous methods like [4, 28, 60] that used adversarial examples or manually combined image pairs to craft pseudo-open samples. By utilizing Stable Diffusion, our approach maintains domain-specific stylistic coherence and carefully controls semantic drift from known classes in $\mathcal{C}$. This stylistic consistency is essential for accurately distinguishing between closed and open classes; unmanaged style variations can lead models to mistake domain shifts for class differences—a limitation seen in prior methods [4, 60].

Unlike [46], which uses generic prompts in Stable Diffusion to generate out-of-support images as pseudo-open

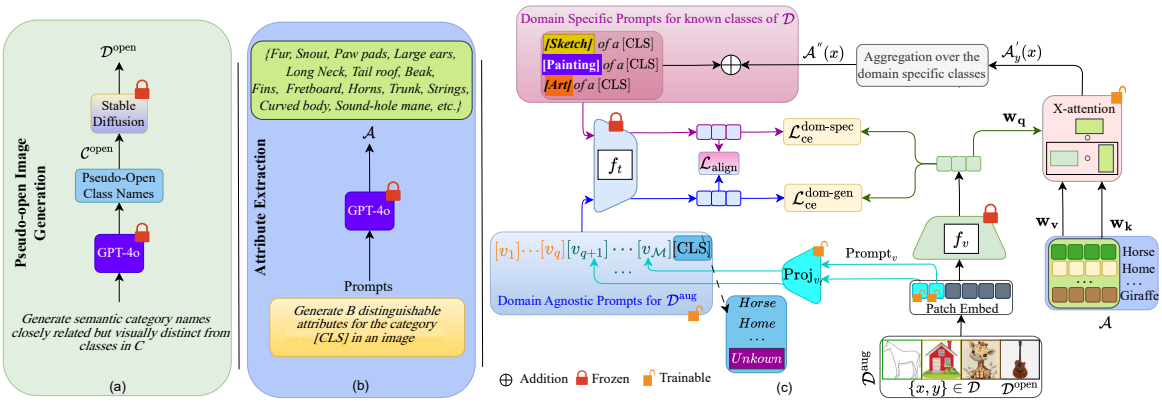

Figure 4. **Working principles of OSLoPrompt. (a)** Fine-grained pseudo-open samples $\mathcal{D}^{\text{open}}$ are generated using stable diffusion with pseudo-open class names $\mathcal{C}^{\text{open}}$ from GPT-4o. **(b)** GPT-4o generates attributes for each class in $\mathcal{C}$. **(c)** OSLoPrompt learns domain-agnostic prompts using tokens $\nu_{1:\mathcal{M}}$. The first q tokens follow [62], while tokens $q+1$ to $\mathcal{M}$ are initialized via learnable visual prompts, and transformed through a projector Proj_{vt}. Domain-agnostic prompts are regularized by domain-specific prompts enhanced with visually-guided semantic attributes, encoded through a cross-attention module with parameters $(\mathbf{w_k}, \mathbf{w_v}, \mathbf{w_q})$. The model is trained with a context alignment loss $\mathcal{L}_{\text{align}}$, along with visual-textual classification losses, handling known class samples for domain-specific prompts with $\mathcal{D}$ ($\mathcal{L}_{\text{ce}}^{\text{dom-spec}}$) and both known and pseudo-open class samples for domain-agnostic prompts with $\mathcal{D}^{\text{aug}}$ ($\mathcal{L}_{\text{ce}}^{\text{dom-gen}}$).

samples—often yielding semantically irrelevant images—our method enables controlled pseudo-open sample generation. The approach in [46] risks producing images that inadequately define decision boundaries, thus limiting the classifier's precision in segregating nuanced open-set samples from the known classes (Fig. 6).

To achieve this controlled generation, we first curate a set of pseudo-open class names, $\mathcal{C}^{\text{open}}$, derived from $\mathcal{C}$ (see **Sup Mat** for examples) to ensure similarity yet distinctiveness, guided by a targeted GPT-4o prompt:

> "*Generate semantic category names closely related but visually distinct from classes in $\mathcal{C}$.*"

For each class in $\mathcal{C}^{\text{open}}$, we use Stable Diffusion to synthesize images in the chosen source-domain-specific style (*Sketch, Clipart*, etc.), using prompts such as:

> "*Generate images in the style of [Domain] depicting [class from $\mathcal{C}^{\text{open}}$].*"

$\mathcal{D}^{\text{open}}$ denotes the synthesized images in this way. Furthermore, we introduce the augmented source dataset, $\mathcal{D}^{\text{aug}}$ by combining $\mathcal{D}$ with $\mathcal{D}^{\text{open}}$.

3.2. Proposed prompt learning strategy

Our domain-agnostic semantic prompt-learning framework is structured around three core concepts:

First, we implement a multi-modal prompting approach that enhances CLIP's capacity to learn visual abstractions invariant to domain and class distinctions, supporting robust DG. These learnable visual prompts also serve to initialize portions of the domain-agnostic prompts, improving feature encoding and reducing style biases. **Second**, we align the domain-agnostic prompt's contexts with those of the manually crafted, domain-specific prompts applied across all source domains in $\mathcal{D}$. While these domain-specific prompts capture unique domain features, their adaptability is limited by static design. **Third**, to address this limitation, we enrich the domain-specific prompts with visually guided class attribute encodings. Using a cross-attention mechanism, each image is represented as a combination of class attributes generated by GPT-4o, enabling the prompts to capture detailed visual semantics and improve versatility across domains. We detail these paradigms in the following.

(i) Proposed generic visual prompting: We incorporate the visual prompting strategy from [17] within the visual encoder $\mathcal{F}_v$. For images in the augmented dataset $\mathcal{D}^{\text{aug}}$, additional visual prompts, denoted as $\mathbf{Prompt_v} = \{p_v^{1:m}\}$ with m tokens, are introduced at the first ViT layer of $\mathcal{F}_v$ (denoted as $\mathcal{F}_v^1$) alongside the initial image patch embeddings (E_0). The forward pass through $\mathcal{F}_v$ is defined as:

$$[c_1, \_, E_1] = \mathcal{F}_v^1([c_0, \mathbf{Prompt_v}, E_0]) \tag{1}$$

$$[c_l, E_l] = \mathcal{F}_v^l([c_{l-1}, E_{l-1}]), \quad l = 2, 3, \ldots \tag{2}$$

Here, c_l denotes the conventional learnable token embedding of ViT after the l^{th} layer, and $[:,:]$ represents stacking and concatenation. The influence of the learned prompts $\mathbf{Prompt_v}$ is thereby integrated into any subsequent forward pass of $\mathcal{F}_v(x)$ during evaluation.

(ii) A novel image-to-semantic attributes connection, improving the domain-specific prompts: We begin by defining a general formulation for manually curated, domain-specific prompts for each source domain $s \in [1, \mathcal{N}]$ and its respective classes $y^s \in \mathcal{Y}_s$:

$$\mathbf{Prompt}_{\mathbf{s}}^{\mathbf{y^s}} = \text{“}[Domain_s] \text{ of a } [CLS_{y^s}]\text{”} \quad (3)$$

To enrich these prompts under limited supervision, we inject fine-grained, attribute-level knowledge into them; however, moving beyond static prompts with class descriptions [29], which often overlook the visual space, we seek to model the attribute's distributions within the images. This leads to our attribute-based image encoding scheme. Specifically, for each class y^s, we define a set of $\mathcal{B}$ domain-invariant attributes $\mathcal{A}_{y^s} = [a^1_{y^s}, a^2_{y^s}, \ldots, a^{\mathcal{B}}_{y^s}]$, generated via GPT-4o with a structured prompt:

> *"Generate $\mathcal{B}$ distinguishable attributes for the category [CLS] in an image."*

These attributes, consistent across domains, capture detailed semantic cues. For instance, generated attributes for `Dog` include *Fur, Snout, Tail, Paw Pads* (Fig. 4; additional details in **Sup Mat**).

To integrate these attributes into visual encoding, we use a cross-attention mechanism that enriches each image-label pair $(x^s, y^s) \in \mathcal{D}_s$ with class-specific attributes $\mathcal{A}_{y^s}$. Here, image features $\mathcal{F}_v(x^s)$ act as the query, while semantic attributes $\mathcal{F}_t(\mathcal{A}_{y^s})$ serve as keys and values, with learnable projections $\mathbf{w_q}$, $\mathbf{w_k}$, and $\mathbf{w_v}$ applied to obtain transformed representations $\mathcal{F}^q_v(), \mathcal{F}^k_t(), \mathcal{F}^v_t()$. The attribute-enhanced embedding for class y^s in image x^s is computed as:

$$\mathcal{A}'_{y^s}(x^s) = \text{Softmax}\left[\frac{\mathcal{F}^q_v(x^s)\mathcal{F}^k_t(\mathcal{A}_{y^s})^T}{\sqrt{d}}\right]\mathcal{F}^v_t(\mathcal{A}_{y^s}) \quad (4)$$

To obtain a class-agnostic encoding, we compute an average across classes:

$$\mathcal{A}''(x^s) = \frac{1}{|\mathcal{Y}_s|}\sum_{y^{s'}\in\mathcal{Y}_s}\mathcal{A}'_{y^{s'}}(x^s) \quad (5)$$

The final domain-specific prompt combines the base prompt $\mathbf{Prompt}_{\mathbf{s}}^{\mathbf{y^s}}$ with this averaged attribute embedding $\mathcal{A}''(x^s)$ through token-wise addition:

$$\overline{\mathbf{Prompt}_{\mathbf{s}}^{\mathbf{y^s}}}(x^s) = \mathbf{Prompt}_{\mathbf{s}}^{\mathbf{y^s}} + \mathcal{A}''(x^s) \quad (6)$$

The resulting prompt, $\overline{\mathbf{Prompt}_{\mathbf{s}}^{\mathbf{y^s}}}(x^s)$, is dynamic and discriminative, capturing nuanced visual variations beyond the static base prompt $\mathbf{Prompt}_{\mathbf{s}}^{\mathbf{y^s}}$ and enhancing domain-agnostic prompts substantially.

We train these prompts using a visual-textual contrastive objective:

$$\mathcal{L}_{\text{ce}}^{\text{dom-spec}} = \min_{\substack{\mathbf{w^q},\mathbf{w^k}\\ \mathbf{w^v},\mathbf{Prompt}_v}} \sum_{s=1}^{\mathcal{N}} \mathbb{E}_{(x^s,y^s)\sim\mathcal{P}(\mathcal{D}_s)}\left[-\log p(y^s|x^s)\right] \quad (7)$$

where the probability $p(y^s|x^s)$ is computed based on cosine similarity δ and a temperature parameter τ:

$$p(y^s|x^s) = \frac{\exp\left(\delta\left(\mathcal{F}_t\left(\overline{\mathbf{Prompt}_s^{y^s}}(x^s)\right), \mathcal{F}_v(x^s)\right)/\tau\right)}{\sum_{y^{s'}\in\mathcal{Y}_s}\exp\left(\delta\left(\mathcal{F}_t\left(\overline{\mathbf{Prompt}_s^{y^{s'}}}(x^s)\right), \mathcal{F}_v(x^s)\right)/\tau\right)} \quad (8)$$

(iii) Domain-generic prompt learning: Our objective is to learn domain-agnostic prompts, *e.g.* $\mathbf{Prompt}_{\text{gen}}^{\mathbf{y}}$ for class y, covering both known classes in $\mathcal{C}$ and the `Unknown` class in $\mathcal{D}^{\text{aug}}$. Full context learning for $\mathbf{Prompt}_{\text{gen}}$ as in [62] is insufficient, as it neglects visual cues. Additionally, since $\mathbf{Prompt}_{\text{gen}}$ addresses outliers without pre-defined class names or attributes, the attribute-based conditioning of Eq. 4 is ineffective. To address this, we propose a novel enrichment strategy for $\mathbf{Prompt}_{\text{gen}}$.

Precisely, a subset of context tokens within $\mathbf{Prompt}_{\text{gen}}$ is initialized by projecting knowledge from the learnable visual prompt, $\mathbf{Prompt}_{\mathbf{v}}$. Unlike existing methods [61] that rely on frozen per-image features extracted from $\mathcal{F}_v$ for conditioning the learnable prompts, risking the introduction of image-specific artifacts that can degrade prompt quality and affect performance (Table 3), our $\mathbf{Prompt}_{\mathbf{v}}$ is shared across all visual entities. This shared structure provides a more comprehensive and domain-independent source of knowledge. Formally, $\mathbf{Prompt}_{\text{gen}}^{\mathbf{y}}$ is defined as:

$$\mathbf{Prompt}_{\text{gen}}^{\mathbf{y}} = [\nu_{1:q}][\text{Proj}_{vt}(\mathbf{Prompt}_{\mathbf{v}})]_{q+1:\mathcal{M}}[\text{CLS}_y] \quad (9)$$

Where $\{\nu_{1:q}\}$ are the q directly learnable context tokens in $\mathbf{Prompt}_{\text{gen}}$, and $[\text{Proj}_{vt}(\mathbf{Prompt}_{\mathbf{v}})]_{q+1:\mathcal{M}}$ are the tokens derived from the learnable visual prompts through the projector function Proj_{vt}, and $\mathcal{M}$ denotes the total context length, which is similar to the that of $\mathbf{Prompt}_{\mathbf{s}}$.

Two main objectives drive the training of these domain-agnostic prompts:

- Proposed context alignment loss: We aim to align the context tokens of $\mathbf{Prompt}_{\text{gen}}$ with those of all domain-specific prompts $\{\overline{\mathbf{Prompt}_{\mathbf{s}}}\}_{s=1}^{\mathcal{N}}$ over the source domains. This alignment allows $\mathbf{Prompt}_{\text{gen}}$ to inherit domain-specific attributes and fine-grained semantic knowledge in the context tokens better, enabling it to generalize effectively. The alignment loss is defined as:

$$\mathcal{L}_{\text{align}} = \min_{\substack{\{\nu_{1:q}\},\mathbf{w^q},\mathbf{w^k},\mathbf{w^v},\\ \text{Proj}_{vt},\mathbf{Prompt}_{\mathbf{v}}}} \sum_{s=1}^{\mathcal{N}} \mathbb{E}_{(x^s,y^s)\in\mathcal{P}(\mathcal{D}_s)}\Big[1 - \text{cosine}\left(\mathcal{F}_t(\mathbf{Prompt}_{\text{gen}}^{\mathbf{y^s}}), \mathcal{F}_t(\overline{\mathbf{Prompt}_{\mathbf{s}}^{\mathbf{y^s}}}(x^s))\right)\Big] \quad (10)$$

- Supervised visual-textual contrastive loss: We further refine the domain-agnostic prompts using a supervised visual-textual contrastive framework, supplemented by a

Table 1. Comparative analysis across five datasets in the 1-shot(top) and 5-shot (bottom) LSOSDG setting, reporting both the average closed-set accuracy (Acc) and H-score over all domain combinations, following a leave-one-domain-out protocol. Best performance in **bold** and the second-best in **Red**. Results on the domain combinations are mentioned in **Sup Mat**.

Methods	CLIP-based	Venue	PACS		VLCS		OfficeHome		Multi-Dataset		Mini-DomainNet		Average	
			Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score
CLIP + OpenMax (OSR) [2]	✓	CVPR'16	20.24	31.97	20.59	31.83	20.00	32.64	11.74	20.87	16.92	28.05	17.90	29.07
CLIPN (OSR) [52]	✓	ICCV'23	64.03	55.79	25.34	19.49	44.18	32.83	39.84	36.28	47.63	40.91	44.20	37.06
MORGAN (FS-OSR) [30]	×	WACV'23	37.40	19.06	31.35	27.22	19.21	18.51	30.00	37.26	22.40	15.70	28.07	23.55
StyLIP (DG + OSR) [5]	✓	WACV'24	74.89	60.99	27.94	34.61	52.34	21.95	51.50	47.64	59.44	57.46	53.22	44.53
PromptSRC (DG + OSR) [20]	✓	ICCV'23	35.72	27.09	24.98	20.04	22.02	14.85	30.16	31.18	25.20	20.44	27.62	22.72
2LM (FSDG + OSR) [36]	×	CVPR'23	35.22	21.42	31.61	28.76	21.30	13.60	29.73	34.80	24.50	17.75	28.47	23.27
ODG-Net (ODG) [4]	×	TMLR'23	34.82	21.67	32.33	29.17	20.47	11.45	29.16	29.40	22.05	19.08	27.77	22.15
MEDIC (ODG) [53]	×	ICCV'23	33.91	21.40	32.94	26.28	21.31	11.75	30.35	33.11	23.73	19.05	28.45	22.32
SCI-PD (ODG) [6]	✓	CVPR'24	23.40	25.84	19.88	19.60	35.27	44.31	16.95	19.18	16.25	23.33	22.35	26.45
ODG-CLIP (ODG) [46]	✓	CVPR'24	68.89	75.56	52.43	54.70	48.69	52.93	63.74	69.53	61.05	65.50	58.96	63.64
OSLoPrompt (Ours)	✓	-	**92.71**	**94.86**	**78.89**	**76.89**	**69.73**	**64.04**	**76.30**	**74.49**	**69.00**	**67.57**	**77.32**	**75.57**

Methods	CLIP-based	Venue	PACS		VLCS		OfficeHome		Multi-Dataset		Mini-DomainNet		Average	
			Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score
CLIP + OpenMax (OSR) [2]	✓	CVPR'16	68.75	80.98	66.25	74.74	35.59	49.28	56.59	68.84	32.46	48.20	51.93	64.41
CLIPN (OSR) [52]	✓	ICCV'23	78.04	71.14	32.92	27.95	47.94	40.33	46.50	39.23	55.78	48.53	52.24	45.44
MORGAN (FS-OSR) [30]	×	WACV'23	46.27	24.06	42.16	38.70	36.20	18.63	35.47	42.80	37.81	27.06	39.58	30.25
StyLIP (DG + OSR)[5]	✓	WACV'24	80.10	70.01	45.78	48.93	61.87	42.46	54.58	49.76	64.03	60.68	61.27	54.37
PromptSRC (DG + OSR) [20]	✓	ICCV'23	46.86	30.23	36.16	32.36	31.10	20.35	35.68	38.12	36.37	31.32	37.24	30.28
2LM (FSDG + OSR) [36]	×	CVPR'23	46.70	24.06	41.67	37.36	29.38	18.95	35.04	35.38	38.43	28.70	38.24	28.89
ODG-Net (ODG) [4]	×	TMLR'23	46.66	25.92	43.05	37.71	34.52	15.96	34.20	36.93	39.95	23.72	39.68	28.05
MEDIC (ODG) [53]	×	ICCV'23	44.88	25.05	40.53	35.56	30.40	18.45	35.42	36.26	36.95	30.60	37.64	29.18
SCI-PD (ODG) [6]	✓	CVPR'24	35.16	34.53	30.11	30.48	32.98	42.50	32.20	28.89	21.25	30.57	30.34	33.39
ODG-CLIP (ODG) [46]	✓	CVPR'24	83.65	88.16	62.93	56.89	55.32	49.31	74.40	76.14	74.38	65.49	70.14	67.20
OSLoPrompt (Ours)	✓	-	**93.72**	**95.01**	**79.04**	**77.34**	**75.33**	**62.08**	**79.75**	**80.05**	**74.52**	**66.58**	**80.47**	**76.21**

cross-entropy loss $\mathcal{L}_{\text{ce}}^{\text{dom-gen}}$ over the augmented dataset $\mathcal{D}^{\text{aug}}$ similar to Eq. 7. This training aims to enhance the separation between known classes and pseudo-open samples.

The total loss to train OSLoPrompt is,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}}^{\text{dom-gen}} + \mathcal{L}_{\text{ce}}^{\text{dom-spec}} + \mathcal{L}_{\text{align}} \tag{11}$$

During inference, for an input x^t from the target domain $\mathcal{D}_t$, the model classifies x^t by maximizing the cosine similarity between $\mathcal{F}_v(x^t)$ and $\mathbf{Prompt}_{\text{gen}}^{\mathbf{y^t}}$:

$$\overline{y}^t = \underset{y^t \in \mathcal{C} \cup \texttt{Unknown}}{\text{argmax}} \; p(y^t | x^t, \mathcal{F}_v, \mathcal{F}_t, \mathbf{Prompt}_{\text{gen}}^{\mathbf{y^t}}) \tag{12}$$

Pseudo-code of the pipeline is mentioned in **Sup Mat**.

4. Experimental Evaluations

Datasets: We evaluate OSLoPrompt on five benchmark datasets: Office-Home [50], PACS [23], VLCS [10], Mini-DomainNet [34], and Multi-Dataset [43], following standard known-novel class splits [43, 46], but considering one and five training samples per class. Further details regarding the splits are mentioned in the **Sup Mat**. We also consider the ImageNet suite [61, 62] for evaluating the closed-set DG performance in the low-shot setting.

Architecture details: For all CLIP-based models including ours, ViT-B/32 is employed as the visual backbone $\mathcal{F}_v$, with a Transformer [49] serving as $\mathcal{F}_t$. We use the official implementations for the available methods, while implementing [36] on our own. For [30], which was originally designed with 3D-CNN, we re-implement with the ResNet-50 based backbone [11] to accommodate the RGB datasets.

Training and evaluation: Training is performed over 10 epochs with the AdamW optimizer [27]. Batch sizes are set per dataset: 6 for PACS/VLCS, 9 for Office-Home, Multi-Dataset, and Mini-DomainNet, with each batch incorporating three pseudo-open samples from each source domain. All CLIP-based methods use a textual prompt context length $\mathcal{M} = 4$ and a visual prompt context length $m = 2$. In $\mathbf{Prompt}_{\text{gen}}$, two tokens are initialized from $\mathbf{Prompt}_{\mathbf{v}}$, while the other two are initialized randomly. Proj_{vt} is implemented as a meta-net similar to the one in [61], with $\mathcal{B} = 4$ attributes per class added into $\mathcal{A}$. For domain-specific prompts, domain names are used directly where available. For VLCS and Multi-Dataset, prompts are formatted as `[Photo] of a [CLS]`. We filter out images with very low entropy of the grey-value distributions (≤ 0.2) as these images are irrelevant. We evaluate using two metrics under the leave-one-domain-out protocol [46]: **a)** top-1 accuracy (Acc) for closed-set classes, and **b)** the harmonic mean (H-score) for combined closed-set and open-set performance. Results are averaged over three runs.

Competing methods: Given the absence of existing LSOSDG methods in the literature, we design the following baselines: **Open-Set Recognition (OSR) and Few-Shot OSR methods:** We include CLIP + OpenMax [2], CLIPN [52], and MORGAN [30]. **FSDG methods:** We integrate the FSDG method [36] with OSR [2] capabilities. **CLIP-based closed-set DG + OSR methods:** This includes PromptSrc [20] and StyLIP [5] with an `Unknown`-

Table 2. Comparative analysis of **1-shot closed-set DG performance on the ImageNet benchmark** using ViT-B/32.

Methods	Source	Target				Average
	ImageNet	IN-V2	IN-Sketch	IN-A	IN-R	Average
CoOp [62]	63.8	56.5	41.4	31.4	66.8	49.02
CoCoOp [61]	**64.4**	56.4	41.0	**32.1**	66.7	49.05
KgCoOp [56]	64.2	56.2	41.7	31.4	67.7	49.25
MaPLe [19]	64.2	**56.7**	41.7	32.0	67.7	49.52
StyLIP [5]	64.0	56.6	40.9	31.7	66.7	48.98
ODG-CLIP [46]	61.3	53.1	38.7	28.9	63.5	46.05
OSLOPROMPT	64.2	56.5	**42.0**	**32.1**	**68.2**	**49.70**

class prompt like ours[3]. **Existing ODG techniques:** We evaluate both CLIP and non-CLIP based methods such as [4, 6, 46, 53][4]. In the `ImageNet` experiments, we benchmark against existing CLIP-based prompting techniques [5, 19, 46, 56, 61, 62] More details on the implementations are mentioned in **Sup Mat**.

Table 3. **Ablation analysis** for Office-Home (O.H.) and Mini-DomainNet (M.DNet) in a 1-shot setting on H-score. Details about these implementations are mentioned in **Sup Mat**.

Methods	O.H.	M.DNet
Analysis of domain-specific prompts		
✓ Manual prompting: `Domain of a CLS`	59.33	65.88
✓ Manual prompting with image conditioning	62.96	67.27
✓ Manual prompting expanded with ad-hoc attributes from $\mathcal{A}$ [29]	60.69	63.82
✓ Manual prompting with ad-hoc attributes and image conditioning	62.16	65.11
✓ Visual attributes learning [21]	58.35	60.57
✓ **Proposed cross-attention approach**	**64.04**	**67.57**
Analysis of domain-agnostic prompts		
✓ Full context learning [62]	60.81	53.43
✓ Image-cond. context learning [61]	63.10	59.61
✓ **Proposed multi-modal prompting**	**64.04**	**67.57**
Sensitivity to the number of attributes per class in $\mathcal{A}$		
✓ **4**	**64.04**	**67.57**
✓ 8	63.97	66.36
✓ 12	63.79	65.14
Importance of the loss terms		
✓ $\mathcal{L}_{\text{ce}}^{\text{dom-gen}}$ (no domain-specific guidance)	62.51	63.86
✓ $\mathcal{L}_{\text{ce}}^{\text{dom-gen}} + \mathcal{L}_{\text{ce}}^{\text{dom-spec}}$ (partial domain-specific guidance)	62.52	65.56
✓ $\mathcal{L}_{\text{ce}}^{\text{dom-gen}} + \mathcal{L}_{\text{ce}}^{\text{dom-spec}} + \mathcal{L}_{\text{align}}$	**64.04**	**67.57**
Pseudo-open image synthesis		
✓ Generic sample generation of [46]	41.09	49.07
✓ Mixup-based [28] pseudo-open images	57.26	64.85
✓ **Our fine-grained sample generation**	**64.04**	**67.57**

[3]The conventional prompting techniques like [19, 56, 61, 62] overfit the known classes, failing in general on the open-set detection. We do not report them in the tables.

[4]We did not compare with [33] as it requires training samples from all classes in all domains, unlike our setting that has no such constraint.

4.1. Discussions on the main results

Tables 1 shows that OSLOPROMPT consistently outperforms all competing methods in 1-shot and 5-shot settings. Non-CLIP-based models [4, 30, 36, 53] struggle with closed-set accuracy and open-sample classification, leading to lower H-scores. While CLIP-based methods [5, 6, 20, 52] improve on these metrics, they still face challenges in generalization and open-set detection, with STYLIP achieving the highest H-scores among them at 44.53% (1-shot) and 54.37% (5-shot). ODG-CLIP [46], the strongest competitor, achieves H-scores of 63.64% (1-shot) and 67.20% (5-shot), yet still lags behind OSLOPROMPT.

OSLOPROMPT attains average H-scores of 75.57% (1-shot) and 76.21% (5-shot) across five datasets, outperforming ODG-CLIP by 11.93% (1-shot) and 9.01% (5-shot), respectively. Although 5-shot performance generally improves, in Mini-DomainNet and Office-Home, open-set performance dips slightly, likely due to fine-grained distinctions causing mild overfitting, causing the H-score to decrease by $1-2\%$ than those of the 1-shot cases. Nonetheless, these consistent gains underscore OSLOPROMPT's superior generalization with limited data.

Additionally, we evaluated various prompting methods on the ImageNet suite [61, 62] for single-source multi-target DG, using a single training sample per class from `ImageNet` and testing on `ImageNet-v2` [38], `ImageNet-Sketch` [51], `ImageNet-A` [13], and `ImageNet-R` [12]. This closed-set evaluation used a generic domain-specific prompt `[Photo] of a [CLS]`, limiting the advantages of our context distillation objective $\mathcal{L}_{\text{align}}$. Notwithstanding this fact, as shown in Table 2, OSLOPROMPT outperforms competing methods in three of four target domains, with an average performance of 49.70%, 0.18% higher than the second-best method.

4.2. Main ablation analysis

Our comprehensive ablation analysis, outlined in Table 3, evaluates the impact of various components of our methodology across two datasets, Office-Home and Mini-DomainNet, in a 1-shot scenario. Further ablation analysis, specifically **analysis of the context lengths** and visualizations, are mentioned in the **Sup Mat**.

- Analysis of domain-specific prompts: To validate our attribute-enriched domain-specific prompts, we compared them to several alternatives: manual static prompts, expanded static prompts with class-wise attributes from $\mathcal{A}$ following [29], static prompts augmented with image features from $\mathcal{F}_v$, image conditioning within attribute-based static prompts, and the integration of visual attribute learning [21]. Our approach, which aligns visual embeddings with weighted semantic attributes, outperformed all others, exceeding the next best (image-conditioned static prompts) by 1.08% on Office-Home and 0.30% on Mini-DomainNet.

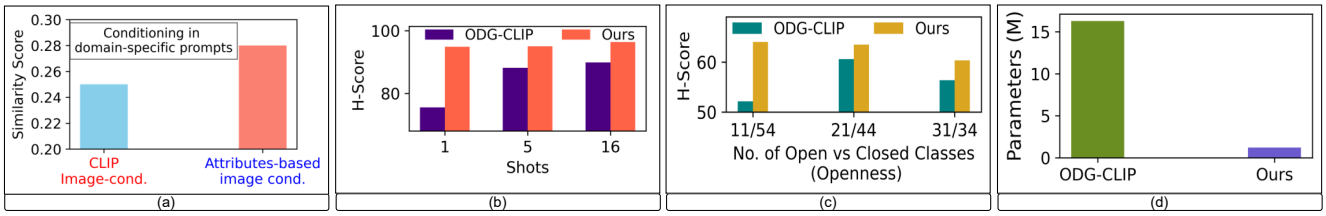

Figure 5. **(a)** Comparison of the **cosine similarity** between image and prompt embeddings from $\mathcal{F}_v$ and $\mathcal{F}_t$ under CLIP image feature conditioning and our proposed semantic attribute-driven encoding on the domain-specific prompts on PACS, showing improved image-prompt alignment with our approach. **(b) Sensitivity** of ODG-CLIP [46] and OSLOPROMPT **to the number of training samples per class** on PACS. **(c) Openness sensitivity** of OSLOPROMPT and ODG-CLIP in the 1-shot Office-Home case for different known and novel class ratios. **(d)** Comparison of **trainable** parameters between ODG-CLIP and our method.

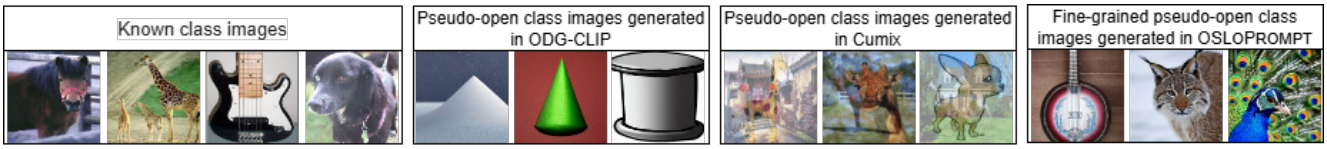

Figure 6. **Pseudo-open images** generated by ODG-CLIP [46] are highly **coarse-grained** in relation to the known classes. While CuMix [28] provides improved fine-grained details compared to ODG-CLIP, it still lacks proper semantic coherence. Our pseudo-open image generation achieves a **fine-grained** level of detail, maintaining both semantic relevance and class-specific granularity (for PACS).

Fig. 5 (a) illustrates the qualitative improvements of attribute guidance over full image-feature conditioning for image-prompt alignment.

- Analysis of domain-agnostic prompts: Among domain-agnostic prompt learning methods, our strategy—combining contexts from learnable visual prompts with directly learnable tokens—outperforms both full-context learning from initialization "`Photo of a`" [62] and image-conditioned full-context learning of [61]. This highlights the benefit of embedding generic visual insights into agnostic prompts, yielding approximately 8% improvement on Mini-DomainNet on H-score.

- Attribute count per class in $\mathcal{A}$: Optimizing the attribute count to four per class in $\mathcal{A}$ balances detail and model complexity, outperforming configurations with eight and twelve attributes. For Mini-DomainNet, four attributes yield 1.21% better performance than eight.

- Sensitivity to loss terms: Excluding $\mathcal{L}_{\text{ce}}^{\text{dom-spec}}$ and $\mathcal{L}_{\text{align}}$ from the total loss results in nearly an 3.71% H-score drop on Mini-DomainNet, underscoring domain-specific guidance's importance. Including $\mathcal{L}_{ce}^{\text{dom-spec}}$ alongside $\mathcal{L}_{\text{ce}}^{\text{dom-gen}}$, thus enabling domain-specific gradient updates for $\mathbf{Prompt_v}$, enhances performance by about 1.7% over the baseline. $\mathcal{L}_{\text{align}}$ further improves results by 2.01%.

- Sensitivity to varying shots and known-to-novel class ratios: As shown in Fig. 5 (b), our method sustains high H-scores as the number of training samples per class increases from 1 to 16, beating ODG-CLIP consistently. Additionally, Fig. 5 (c) illustrates the sensitivity of ODG-CLIP and OSLOPROMPT to varying ratios of open-to-closed classes, or *openness*. Our method consistently surpasses ODG-CLIP across a broad range of openness.

- Pseudo-open image generation (Fig. 6): Replacing our pseudo-open image generation with the coarse-level approach from [46] led to a 22.95% and 18.5% decrease in H-score across both datasets, highlighting the importance of a precise closed-open decision boundary in LSOSDG. While CuMix [28], which combines random image regions to form pseudo-open samples, produces finer results than [46] given the fact that the mixed images are not wildly divergent from the originally known set of classes, it still trails our approach by 6.78% and 2.72% in H-score. More visual results are shown in **Sup Mat**.

- Model complexity: Fig. 5 (d) illustrates the computational efficiency of OSLOPROMPT compared to ODG-CLIP. Overall, ODG-CLIP has 142M model parameters, contrary to 127M of ours. Amongst them, 16.3M parameters are trainable for ODG-CLIP while it is 1.21M for ours.

5. Conclusions

This paper introduces the LSOSDG problem setting and presents OSLOPROMPT as an effective solution. Our approach integrates a domain-agnostic prompt learning strategy with structured regularization through ad-hoc domain-specific prompts, enriched by image-to-semantic attribute encoding and generic visual prompts, enabling low-shot generalization and robust outlier detection. To enhance fine-grained outlier detection, we introduce a controlled mechanism for synthesizing pseudo-open samples by selectively querying diffusion models. OSLOPROMPT consistently outperforms other methods across benchmarks. Future work will extend OSLOPROMPT to broader open-world applications, including structured prediction tasks.

6. Acknowledgements

The authors sincerely acknowledge the support from AWL Japan and Adobe Research India.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016. 3, 6

[3] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011. 1, 3

[4] Shirsha Bose, Ankit Jha, Hitesh Kandala, and Biplab Banerjee. Beyond boundaries: A novel data-augmentation discourse for open domain generalization. *Transactions on Machine Learning Research*, 2023. 1, 3, 6, 7

[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, 3, 6, 7

[6] Zining Chen, Weiqiu Wang, Zhicheng Zhao, Fei Su, Aidong Men, and Hongying Meng. Practicaldg: Perturbation distillation on vision-language models for hybrid domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23501–23511, 2024. 3, 6, 7

[7] 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. 3

[8] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Selecting relevant features from a multi-domain representation for few-shot classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 769–786. Springer, 2020. 3

[9] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[10] Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 6

[11] 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

[12] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 7

[13] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 7

[14] Yuki Hirohashi, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Prompt learning with one-shot setting based feature space analysis in vision-and-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7761–7770, 2024. 3

[15] Zeyi Huang, Andy Zhou, Zijian Ling, Mu Cai, Haohan Wang, and Yong Jae Lee. A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11685–11695, 2023. 3

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 4

[18] Juwon Kang, Sohyun Lee, Namyup Kim, and Suha Kwak. Style neophile: Constantly seeking novel styles for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7130–7140, 2022. 1, 3

[19] 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 (CVPR)*, pages 19113–19122, 2023. 3, 7

[20] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023. 3, 6, 7

[21] Gahyeon Kim, Sohee Kim, and Seokju Lee. Aapl: Adding attributes to prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1572–1582, 2024. 2, 3, 7

[22] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2021. 3

[23] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 6

[24] Liunian Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang. Desco: Learning object recognition with rich language de-

scriptions. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[25] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Cross-domain few-shot learning with task-specific adapters. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7161–7170, 2022. 3

[26] Lu Liu, William Hamilton, Guodong Long, Jing Jiang, and Hugo Larochelle. A universal representation transformer layer for few-shot image classification. *arXiv preprint arXiv:2006.11702*, 2020. 1, 3

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[28] Massimiliano Mancini, Zeynep Akata, Elisa Ricci, and Barbara Caputo. Towards recognizing unseen categories in unseen domains. In *European Conference on Computer Vision*, pages 466–483. Springer, 2020. 3, 7, 8

[29] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022. 2, 3, 5, 7

[30] Debabrata Pal, Shirsha Bose, Biplab Banerjee, and Yogananda Jeppu. Morgan: Meta-learning-based few-shot open-set recognition via generative adversarial network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6295–6304, 2023. 6, 7

[31] Debabrata Pal, Deeptej More, Sai Bhargav, Dipesh Tamboli, Vaneet Aggarwal, and Biplab Banerjee. Domain adaptive few-shot open-set learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18831–18840, 2023. 1

[32] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 754–763, 2017. 3

[33] Kunyu Peng, Di Wen, Kailun Yang, Ao Luo, Yufan Chen, Jia Fu, M Saquib Sarfraz, Alina Roitberg, and Rainer Stiefelhagen. Advancing open-set domain generalization using evidential bi-level hardest domain scheduler. *arXiv preprint arXiv:2409.17555*, 2024. 7

[34] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 2, 6

[35] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 2, 3

[36] Xiaorong Qin, Xinhang Song, and Shuqiang Jiang. Bi-level meta-learning for few-shot domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15900–15910, 2023. 1, 3, 6, 7

[37] 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. 2, 3

[38] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 7

[39] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. *Advances in neural information processing systems*, 32, 2019. 3

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[41] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173, 2008. 3

[42] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018. 1

[43] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9624–9633, 2021. 1, 3, 6

[44] 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. PMLR, 2023. 3

[45] Mainak Singha, Ankit Jha, Bhupendra Solanki, Shirsha Bose, and Biplab Banerjee. Applenet: Visual attention parameterized prompt learning for few-shot remote sensing image generalization using clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2023–2033, 2023. 2

[46] Mainak Singha, Ankit Jha, Shirsha Bose, Ashwin Nair, Moloud Abdar, and Biplab Banerjee. Unknown prompt the only lacuna: Unveiling clip's potential for open domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13309–13319, 2024. 2, 3, 4, 6, 7, 8

[47] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Argue: Attribute-guided prompt tuning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28578–28587, 2024. 2, 3

[48] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 2

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 6

[50] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 6

[51] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 7

[52] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023. 3, 6, 7

[53] Xiran Wang, Jian Zhang, Lei Qi, and Yinghuan Shi. Generalizable decision boundaries: Dualistic meta-learning for open set domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11564–11573, 2023. 3, 6, 7

[54] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020. 1

[55] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023. 2

[56] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767, 2023. 2, 3, 7

[57] 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

[58] 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

[59] 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

[60] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pages 561–578. Springer, 2020. 1, 3

[61] 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, 5, 6, 7, 8

[62] 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. 2, 3, 4, 5, 6, 7, 8