Unknown Prompt, the only Lacuna: Unveiling CLIP's Potential for Open Domain Generalization

Mainak Singha[1†] Ankit Jha[2] Shirsha Bose[3] Ashwin Nair[4]
Moloud Abdar[5] Biplab Banerjee[2]
[1] Aisin Corporation, Japan [2] IIT Bombay, India
[3] TU Munich, Germany [4] IISER Thiruvananthapuram, India [5] Deakin University, Australia
{mainaksingha.iitb, ankitjha16, shirshabosecs, ashwin9084yt, m.abdar1987, getbiplab}@gmail.com

Abstract

We delve into Open Domain Generalization (ODG), marked by domain and category shifts between training's labeled source and testing's unlabeled target domains. Existing solutions to ODG face limitations due to constrained generalizations of traditional CNN backbones and errors in detecting target open samples in the absence of prior knowledge. Addressing these pitfalls, we introduce ODG-CLIP, harnessing the semantic prowess of the vision-language model, CLIP. Our framework brings forth three primary innovations: Firstly, distinct from prevailing paradigms, we conceptualize ODG as a multi-class classification challenge encompassing both known and novel categories. Central to our approach is modeling a unique prompt tailored for detecting unknown class samples, and to train this, we employ a readily accessible stable diffusion model, elegantly generating proxy images for the open class. Secondly, aiming for domain-tailored classification (prompt) weights while ensuring a balance of precision and simplicity, we devise a novel visual style-centric prompt learning mechanism. Finally, we infuse images with class-discriminative knowledge derived from the prompt space to augment the fidelity of CLIP's visual embeddings. We introduce a novel objective to safeguard the continuity of this infused semantic intel across domains, especially for the shared classes. Through rigorous testing on diverse datasets, covering closed and open-set DG contexts, ODG-CLIP demonstrates clear supremacy, consistently outpacing peers with performance boosts between 8%-16%. Code will be available at https://github.com/mainaksingha01/ODG-CLIP.

1. Introduction

Domain Generalization (DG) [56] outlines an inductive learning strategy wherein a classifier is trained across multiple distinct source domains with diverse data distributions and is then applied to unknown target domains. Unlike closed-set DG, which presupposes uniform semantic categories across domains [26, 67], Open Domain Generalization (ODG) [48] is designed to handle both shared and unique classes within its training ambit. Consequently, during inference, the unlabeled target domain may include both familiar and exclusively new categories. Despite its relevance to various real-world scenarios, such as autonomous driving and remote sensing, research specifically focused on ODG remains scarce. Prominent ODG techniques, including DAML [48] and MEDIC [57], incorporate meta-learning to enhance the reliability of classifiers for training classes, thereby facilitating outlier detection alongside recognition of known classes at inference. However, these approaches face obstacles related to the heterogeneous nature of open-domain data. Their generalization capabilities are also compromised, in part because they rely on conventional CNN frameworks, which exhibit limited adaptability.

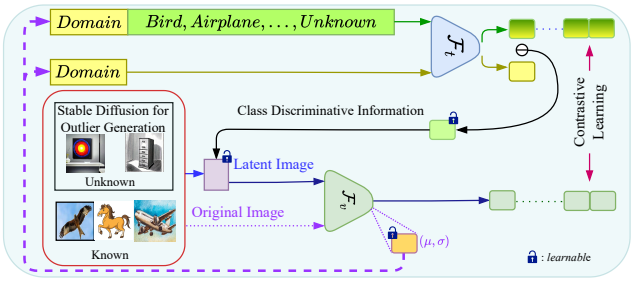

Figure 1. **ODG-CLIP** operates as a multi-class classifier leveraging prompt learning for effective management of known categories and outliers in an ODG context. Central to its methodology is a novel `unknown`-class prompt, designed for open-set samples and integrated with CLIP's unaltered image and text encoders, $\mathcal{F}_v$ and $\mathcal{F}_t$. For the training of `unknown`-class prompt weights, ODG-CLIP employs pseudo-unknown image generation via stable diffusion (SD) [44]. Diverging from existing methods [3,41,65], ODG-CLIP focuses on creating a refined latent visual space to improve visual embeddings and address domain disparities efficiently.

[†]This work is partially done while studying at IIT Bombay, India

Foundation models like CLIP [41] excel by generating rich embedding spaces through multi-modal contrastive training, benefiting especially from prompt learning mechanisms [65, 66] that align classification weights with textual class labels. STYLIP [3] advances this by integrating visual style and content in prompt learning, outperforming other counterparts in closed-set DG, yet it has limitations. Initializing prompts from visual properties and refining them through separate projectors adds complexity and may limit the relevance of the downstream target domains. Moreover, STYLIP and similar CLIP-based prompt learning strategies struggle with outlier recognition, reducing their suitability for ODG tasks. Efforts to tackle open-set recognition by enhancing closed-set knowledge [34] or developing specialized prompts within the CLIP framework [55] have shown promise but remain unoptimized for cross-domain scenarios. The challenge of visual ambiguities also poses significant hurdles for contrastive learning in CLIP, especially across pronounced domain differences.

We identify three key research gaps in using CLIP for ODG: **Prompt design**: Emphasizing the superiority of domain conditional prompts in DG [3, 50], crafting concise, domain-adaptable prompts merging domain-specific and generic tokens is crucial for effective novel domain adaptation. **Multi-class classification over one-against-all recourse for ODG [55]**: Inspired by the efficacy of a unified multi-class classifier for open-set recognition [35, 37], applying this in a CLIP framework promises enhanced performance but poses challenges, particularly in prompt representation for the open space and sourcing training images. **Domain-agnostic visual embeddings**: Boosting the generalizability and discriminative power of visual embeddings could significantly enhance CLIP's multi-domain efficacy. Investigating the extraction of this information from learned prompts is worth exploring.

Our proposed ODG-CLIP: We aim to solve the issues in our proposal, ODG-CLIP (Fig. 1), as described below.

- **Rethinking CLIP for open sample detection:** To unify the classification of known classes and outliers using CLIP, we propose a unique `unknown`-class prompt tailored for detecting the open-set samples. To gather training data for this prompt, our strategy involves generating pseudo-open images that are semantically distinct from existing categories. Moving away from conventional generative or mix-up methods [35, 54], which often yield semantically inferior images, we opt to leverage a pre-trained conditional diffusion model [44], renowned for its superior and diverse image generation capabilities. To ensure the semantic distinctiveness of these synthesized images, we propose to consider *negative prompts* encompassing the known class names, complementing the *positive prompt* of the form `Generate [domain] of an unknown class.`

- Interplay between novel prompt learning for DG and enhancing visual embeddings from CLIP: Our contributions in this regard encompass two main aspects. Firstly, we introduce an innovative approach to prompt learning for DG that incorporates visual style information from CLIP's vision encoder into a specialized *domain token* while also incorporating semantics through a distinct set of learnable *generic tokens*, demonstrating a better balance between complexity and performance than others counterparts [3, 20]. Secondly, we delineate a technique to augment the caliber of image embeddings. This involves the fusion of a learnable, class-centric channel with the images to create latents that are more discriminative than the raw image data, as is conventionally practised [41], for visual feature extraction from CLIP. Precisely, we propose to deploy a dual-prompt strategy per image: one influenced by both style and class information, as aforementioned, and its counterpart driven purely by style. We theorize that the disparities in embeddings of these paired prompts retain class-specific discerning information while capturing the visual distributions of the domain. We introduce a novel loss objective to ensure these differential vectors resonate consistently for communal classes over varied domain pairs. Our salient contributions are therefore:

[-] We propose a CLIP-based method, ODG-CLIP, to solve the challenging ODG problem. To our knowledge, ours is the first approach to utilizing vision-language models (VLMs) for solving ODG.

[-] In ODG-CLIP, we introduce a novel prompt learning for DG, with a specialized prompt to tackle open-class samples and propose a way to use a pre-trained diffusion model to obtain the pseudo-open training samples. Also, we show how the prompt information can be leveraged to enhance the quality of CLIP's visual embeddings.

[-] We perform extensive validations on open and closed-set DG tasks. ODG-CLIP is found to produce the new state-of-the-art results on six benchmarks for both settings.

2. Related Works

Open-set Recognition (OSR) and Open-set Domain Adaptation (OSDA): The OSR challenge [2, 23, 37, 52] centers on proficiently discerning novel-class samples during evaluation using training exemplars from closed-set classes. The generative OSR techniques [12, 15, 64] augment the training set with artificially synthesized categories outside the training set, typically using a GAN-based model or mixing input images randomly to create pseudo-open data [23, 33]. However, these images are mostly restricted in semantics and confined to a lower-dimensional manifold in the open space. The discriminative models, on the other hand, rely on the confidence of the closed-set classifier, reconstruction loss for the samples, or metric learning to detect the open data [8, 62]. Recently, CLIPN [55] has introduced a negative prompt learning approach for OSR using

CLIP, outperforming other few existing VLM-based counterparts [14, 34] significantly. Notably, these models cannot handle distribution shifts between training and test domains. Similarly, OSDA [4, 24, 38, 46] follows an OSR-like scenario within a *transductive* cross-domain setting.

ODG, with its *inductive* nature and absence of target domain knowledge, presents more formidable challenges than OSR and OSDA. Also, diverging from approaches in OSR, we harness a pre-trained conditional diffusion model, yielding images that represent the open space comprehensively.

(Open) DG: DG refers to the problem of constructing a supervised learning model that is generalizable across target distributions without the availability of any prior. The initial studies in closed-set DG focused on DA models [28, 29, 59] due to the disparity in domain distributions. Several DG methods have since been developed, such as self-supervised learning [6], ensemble learning [60], and meta-learning [18, 25, 27, 32, 39, 58]. To address the domain disparity, the concept of domain augmentation [19, 31, 63, 67, 68] was introduced, which involves generating pseudo-domains and adding them to the available pool of domains. Subsequently, the notion of ODG was introduced in [48], which is based on domain-augmented meta-learning. MEDIC [57] recently tackled some of the problems of [48] and proposed to consider both domain-wise and class-wise gradient matching to learn a balanced decision boundary for the closed-set classes. Moving forward, [69] and [61] further extended the idea of multi-source ODG to accommodate a single source domain. Unlike these models, we focus on exploring VLMs to solve the ODG task from the perspective of prompt processing and tackle the associated challenges.

VLMs and prompt learning: The advent of multi-modal foundation models significantly enhances textual and visual integration for image recognition, leveraging BERT [9] and GPT [42] alongside CNN and ViT for content analysis. Key VLMs, such as CLIP [41] and VisualBERT [30], initially relied on intricate manual prompts. Prompt learning, emerging to customize these prompts for specific tasks, involves methods like [5, 20, 21, 49, 65, 66] to make token embeddings learnable, use projector networks for their evolution, or consider the notion of token sharing between the visual and textual modalities for multi-modal prompting. StyLIP [3] uniquely adapts prompt learning for DG, focusing on deriving prompt tokens from visual properties. Further discussions on prompt processing are mentioned in the `Supplementary`.

Our prompts, blending domain-specific and generic tokens, show enhanced domain adaptability compared to [3]. Additionally, we explore the novel approach of improving CLIP's visual embeddings via prompt utilization, without the need for fine-tuning CLIP.

3. Proposed Methodology

In the context of ODG, we utilize multiple source domains, denoted as $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_\mathcal{S}\}$. Each domain exhibits distinct data distributions and comprises a combination of categories specific to the respective domain and categories shared among them. In the training phase, we employ labeled data samples from each domain $\mathcal{D}_s = \{(x_s^i, y_s^i)\}_{i=1}^{n_s}$, where $y_s \in \mathcal{Y}_s$ represents the label associated with $x_s \in \mathcal{X}_s$. The total number of unique classes present in $\mathcal{D}$ is denoted as $\mathcal{C}$ for the joint label space $\mathcal{Y}$.

The target domain, denoted as $\mathcal{D}_\mathcal{T} = \{x_t^j\}_{j=1}^{n_t}$, differs in its data distribution compared to $\mathcal{D}$. This target domain consists of unlabeled data samples, which can either belong to one of the known classes in $\mathcal{Y}$ or to novel classes not encountered during training. We aim to develop a unified classifier considering the notion of prompt learning in CLIP given $\mathcal{D}$ for effectively discerning outliers while accurately recognizing samples from the known classes in $\mathcal{D}_\mathcal{T}$.

3.1. Discussing the principles of ODG-CLIP

Learning objectives: Our proposed ODG-CLIP is founded on three key principles: **i)** Our classification strategy encompasses $\mathcal{C}+1$ classes, where the $\mathcal{C}+1^{th}$ index is designated for the novel `unknown`-class. This class utilizes specific prompts, the weights of which are shaped by synthetic images generated using a diffusion model [44]. **ii)** We advocate for adaptive prompt learning across all classes, enabling the capture of domain-specific distributions and overarching semantic contents through distinct token sets. **iii)** To further refine visual-textual contrastive learning for ODG-CLIP, we focus on enhancing the discriminability of visual embeddings. This is achieved by establishing a latent visual space, guided by the prompts we've developed. Our goal is to address these aspects cohesively in ODG-CLIP.

Walking through ODG-CLIP: ODG-CLIP (Fig. 2) is built upon the frozen vision and text encoders, $\mathcal{F}_v$ and $\mathcal{F}_t$, from CLIP. The pre-trained stable diffusion model $\mathcal{F}_{diff}$ takes a pair of positive and negative prompts (`PP` and `NP`) as input, and the generated images are represented by $\mathcal{D}_{open}$. The new training set is $\mathcal{D}_{aug} = \mathcal{D} \cup \mathcal{D}_{open}$, and the label space is $\mathcal{Y}_{aug} = \mathcal{Y} \cup$ `unknown`. In the following, we outline the methodologies implemented to achieve our objectives.

- Pseudo-open image synthesis using stable diffusion: As mentioned, to train the `unknown`-class prompts to recognize the outlier samples, we seek to generate pseudo-open samples with improved quality and semantic versatility, which the traditional interpolation/extrapolation or adversarial approaches fail to deliver. As a remedy, we propose to utilize the pre-trained Stable Diffusion v1-5 model [44]. While numerous text-to-image generation techniques exist [36, 43, 45], our preference stems from the impressive inference speed, the latent diffusion models provide. For generating images, we use direct `PP` like `a [domain]`

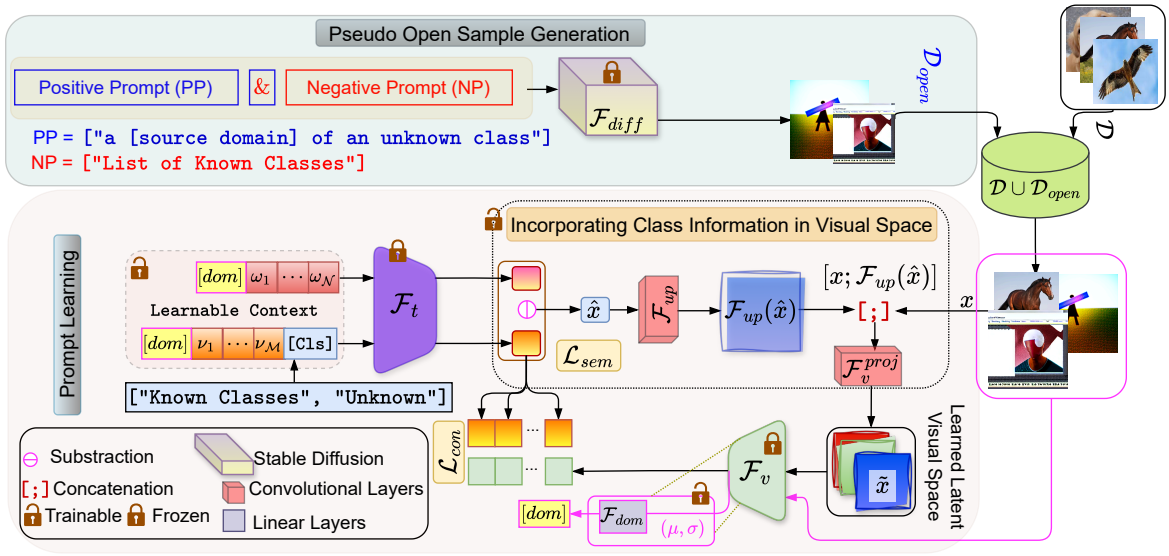

Figure 2. **Model architecture of ODG-CLIP**, which consists of three main components for designing a multi-class closed-open class classifier using prompt learning with a novel `unknown`-class prompt for the outliers. Firstly, we propose to generate pseudo-open samples using a pre-trained diffusion model by employing specialized positive and negative textual instructions. The combined images $\mathcal{D} \cup \mathcal{D}_{open}$ go through the prompt learning stage with specialized projectors ($\mathcal{F}_{dom}$), where two types of prompts are learned per image, one using domain+class information, and the other using only domain information. Their difference is used to obtain the latent visual representation $\tilde{x}$ for a given image x conditioned on the class labels from $\mathcal{Y}_{aug}$, through $\mathcal{F}_{up}$ and $\mathcal{F}_v^{proj}$. The model is trained using $\mathcal{L}_{con} + \mathcal{L}_{sem}$ given all the source domains in $\mathcal{D} \cup \mathcal{D}_{open}$. During inference, we create the latent representations for a target image with respect to all the class labels, and the class maximizing Eq. 2 is selected.

`of an unknown class` along with *category* `NP` that include class names from $\mathcal{Y}$, where `[domain]` refers to the names of the source domains. Our aim is to create potential open-class images that conform to the style characteristics of these known source domains, but diverge from the known semantics of $\mathcal{Y}$ simultaneously. Positive prompts typically guide the structural composition of images within the latent space. However, by adjusting unconditional sampling to incorporate negative prompts and purposefully excluding negative latents from the conditioned set, we strive to generate images with distinctive characteristics from $\mathcal{D}$. Please refer to [44] for further details.

However, upon observation, it became evident that some of the generated images were devoid of significant semantics, leading to images mainly marked by uniform regions. To address this issue, we decided against incorporating such subpar images. Instead, we introduced a simple *filtering* method: applying a threshold to the entropy of the grayscale renderings of the generated images, proficiently weeding out the low-entropy images from $\mathcal{D}_{open}$.

- Domain-aware prompt learning: Our method offers a dynamic approach to prompt learning, seamlessly weaving style elements from the visual domain into a set of task-specific, learnable tokens. This form of domain conditioning proves highly effective for DG, allowing the model to adapt fluidly to new target domains as needed. We adopt a straightforward technique to capture domain information, utilizing visual feature statistics from $\mathcal{F}_v$, akin to the method used in [3]. However, our approach diverges from [3]'s more complex setup, which initializes all tokens with multi-scale visual style or content features. Instead, we integrate a singular domain token, with the remaining tokens positioned as adaptable variables, tailored to suit the task-centric requirements. This configuration achieves a more harmonious balance between task specificity and domain adaptability compared to [3].

Specifically, we propose defining two types of prompts for serving classification and upgrading the visual embeddings simultaneously. One is conditioned on domain and class information, denoted as $\mathcal{P}_{dom_x,cls}(x) = \Big[[dom_x], [\nu_1], \cdots, [\nu_\mathcal{M}], [cls]\Big]$. The other is conditioned solely on the domain, denoted as $\mathcal{P}_{dom_x}(x) = \Big[[dom_x], [\omega_1], \cdots, [\omega_\mathcal{N}]\Big]$. The domain token, $[dom_x]$, is based on the mean and standard deviation, calculated from the final visual feature embedding $\mathcal{F}_v(x)$ for x, represented as $[\mu^x; \sigma^x]$. It is mapped to the textual space through the projector $\mathcal{F}_{dom}$: $dom_x = \mathcal{F}_{dom}([\mu^x; \sigma^x])$. $[cls]$ denotes the embeddings of the class names in $\mathcal{Y}_{aug}$. In contrast, the generic tokens ν and ω are directly learned. $\mathcal{F}_t(\mathcal{P}_{dom,cls})$ and $\mathcal{F}_t(\mathcal{P}_{dom})$ denote the prompt embeddings.

- Generating latent visual space for obtaining improved visual embeddings: To enhance visual-textual contrastive learning with $\mathcal{L}_{con}$ (Eq. 1) amidst domain diversity, we aim to create a semantically rich and class-discriminative latent space from images. We intend to ensure uniformity in features $\mathcal{F}_v$ extracted from these representations for similar-class samples across domains, guided

by our consistency loss $\mathcal{L}_{sem}$ (Eq. 3). Our method leverages prompt embeddings $\mathcal{F}_t(\mathcal{P}_{dom,cls})$ and $\mathcal{F}_t(\mathcal{P}_{dom})$ to imbue images with discriminative qualities, facilitating latent space formation. These prompts inherently capture domain nuances, making the latent space robust against overfitting. Precisely, our proposed approach involves generating a discriminative knowledge per class, denoted by $\hat{x}^y$, for a given x, via element-wise subtraction ($\ominus$) of $\mathcal{F}_t(\mathcal{P}_{dom_x,y}(x))$ from $\mathcal{F}_t(\mathcal{P}_{dom_x}(x))$. An upsampling convolution block, $\mathcal{F}_{up}$, then resizes $\hat{x}^y$ to align with the spatial dimensions of x. Following this, x and $\mathcal{F}_{up}(\hat{x}^y)$ are concatenated and processed through the vision projector $\mathcal{F}_v^{proj}$, yielding the latent representation $\tilde{x}^y = \mathcal{F}_v^{proj}([x; \mathcal{F}_{up}(\hat{x}^y)])$ and which is then inputted into $\mathcal{F}_v$ to extract the visual embeddings $\mathcal{F}_v(\tilde{x}^y)$. This is followed for all $y \in \mathcal{Y}_{aug}$ to aid in the calculation of the loss functions.

3.2. Loss functions, training, and inference

$\mathcal{L}_{con}$: Visual-textual contrastive learning using proposed embeddings: We train the prompt tokens using $\mathcal{D}_{aug}$ for both known and unknown class names in $\mathcal{Y}_{aug}$. In opposition to all the CLIP-based models [3, 41, 66] that take $\mathcal{F}_v(x)$ as the visual embeddings for contrastive learning, our proposal considers $\mathcal{F}_v(\tilde{x})$ instead. Precisely, given $\mathcal{F}_v(\tilde{x}^y)$ and $\mathcal{F}_t(\mathcal{P}_{dom_x,y}(x))$, $\mathcal{L}_{con}$ is defined as follows,

$$\mathcal{L}_{con} = \min_{\mathcal{P}_{dom,class},\mathcal{P}_{dom},\mathcal{F}_{dom},\mathcal{F}_{up},\mathcal{F}_v^{proj}} \mathop{\mathbb{E}}_{(x,y)\in\mathcal{P}(\mathcal{D}_{aug})} -\log p(y|\tilde{x}^y, x) \tag{1}$$

where we calculate $p(y|\tilde{x}^y, x)$ as follows, given δ as the `cosine` similarity and τ as a hyper-parameter.

$$p(y|\tilde{x}^y, x) = \frac{\exp(\delta(\mathcal{F}_t(\mathcal{P}_{dom_x,y}(x)), \mathcal{F}_v(\tilde{x}^y))/\tau)}{\sum_{y'\in\mathcal{Y}_{aug}} \exp(\delta(\mathcal{F}_t(\mathcal{P}_{dom_x,y'}(x)), \mathcal{F}_v(\tilde{x}^{y'}))/\tau)} \tag{2}$$

An intriguing aspect of $\mathcal{L}_{con}$ is that while the generated latent space is utilized for visual feature extraction, the style information is derived from the original image in the prompt embeddings. This intuitive strategy forms the cornerstone of our enhanced contrastive mapping.

$\mathcal{L}_{sem}$: Proposed cross-domain semantic consistency loss: Our objective is to ensure uniformity in the information derived from $|\mathcal{F}_t(\mathcal{P}_{dom,cls}) \ominus \mathcal{F}_t(\mathcal{P}_{dom})|$ through $\mathcal{L}_{sem}$, aiming to cultivate a robust class-wise correlation in the derived latent visual representations across images from different domains but sharing identical class labels. Additionally, this helps implicitly disentangle style and semantics within the prompt space, leading to a highly generic embedding space. The other prompt learning approaches, including [3], do not offer such insights.

To illustrate this mathematically, consider two images, x_i and x_j, both belonging to class y but originating from different source domains, designated as $(x_k^i, y) \in \mathcal{D}_k$ and $(x_l^j, y) \in \mathcal{D}_l$, respectively, and $\mathcal{D}_k, \mathcal{D}_l \in \mathcal{D}_{aug}$. $\mathcal{L}_{sem}$ is then conceptualized as the `cosine` distance between their respective prompt differential vectors, as follows,

$$\mathcal{L}_{sem} = \min_{\mathcal{P}_{dom,class},\mathcal{P}_{dom}} \mathop{\mathbb{E}}_{\mathcal{P}(\mathcal{D}_{aug})} (1 - \delta(|\mathcal{F}_t(\mathcal{P}_{dom_{x_k^i},y}(x_k^i)) \ominus \mathcal{F}_t(\mathcal{P}_{dom_{x_k^i}}(x_k^i))|, |\mathcal{F}_t(\mathcal{P}_{dom_{x_l^j},y}(x_l^j)) \ominus \mathcal{F}_t(\mathcal{P}_{dom_{x_l^j}}(x_l^j))|)) \tag{3}$$

Total loss: We train the prompts and the projectors $(\mathcal{F}_{dom}, \mathcal{F}_{up}, \mathcal{F}_v^{proj})$ with the combined loss $\mathcal{L}_{con} + \mathcal{L}_{sem}$.

Inference: During inference, we generate $\tilde{x}_t^{y'}$ for a given $x_t \in \mathcal{D}_{\mathcal{T}}$ for all the $y' \in \mathcal{Y}_{aug}$. The y' maximizing $p(y'|\tilde{x}_t^{y'}, x_t)$ (Eq. 2) is selected as the predicted label.

4. Experimental Evaluations

Datasets: We evaluate the efficacy of ODG-CLIP using six benchmark datasets: PACS [26], VLCS [13], Office-Home [53], Multi-Dataset [48], Digits-DG [67], and the large-scale Mini-DomainNet [40]. Details on the dataset splits are provided in the `supplementary`.

Architecture details: In our proposed architecture, $\mathcal{F}_{up}$ is composed of four transpose convolution layers complemented with ReLU activations. For the final layer, we employ bilinear interpolation to ensure a perfect alignment with the input of $\mathcal{F}_v$. Meanwhile, $\mathcal{F}_v^{proj}$ incorporates just one convolutional layer, tasked with reducing the input channel count from four to three. On the other hand, $\mathcal{F}_{dom}$ is designed with a single dense layer. We select ViT-B/32 [10] as the backbone for $\mathcal{F}_v$ and the Transformer [51] for $\mathcal{F}_t$ for all the CLIP-based experiments.

Training and evaluation protocols: We conduct training over 10 epochs, starting with a warm-up learning rate of 0.01 and using the Adam optimizer [22] in conjunction with a scheduler. The batch size is set to 32 for the PACS, VLCS, Office-Home, and Digits-DG, while for the Multi-dataset and Mini-DomainNet, we use a batch size of 8. To counteract the bias arising from an excess number of unknown labeled images, we limit our generation to only 25% of the batch size's amount of images for each source domain during training. While generating the synthetic images, a threshold of 0.2 was fixed for rejecting images with low-entropy values in $\mathcal{D}_{open}$. We consider a consistent context length of four for all the CLIP-based models, following [66]. To assess our model's performance, we employ two primary metrics in line with the `leave-one-domain-out` [3] protocol where the model is trained in all but one domain which is used during evaluation. We first use the top-1 accuracy (Acc) to gauge the model's effectiveness on closed-set classes. The harmonic mean (H-score) is also calculated to represent performance across closed-set and open-set samples. For closed-set DG, we showcase the top-1 accuracy. The reported results denote the average over three runs.

Table 1. Comparative analysis for PACS, VLCS, Office-Home, Digits-DG, Multi-Dataset and Mini-DomainNet in ODG setting on average ACC and H-score over all the domain combinations following leave-one-domain-out protocol. Here, SD [44] represents stable diffusion.

	Methods	PACS		VLCS		OfficeHome		Digits-DG		Multi-Dataset		Mini-DomainNet		Average	
		Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score
CNN-based	Cumix [33]	57.85	41.05	52.46	50.11	51.67	49.40	58.13	54.20	42.18	46.91	50.27	39.16	52.09	46.81
	MixStyle [68]	63.35	48.30	52.30	50.61	53.52	49.53	60.23	56.35	42.18	46.91	50.43	40.25	53.67	48.66
	DAML [48]	65.49	51.88	53.53	51.59	56.45	53.34	59.51	55.61	46.61	51.71	52.81	43.63	55.73	51.29
	MEDIC [57]	89.81	83.03	57.28	55.73	60.26	57.91	83.28	66.30	50.74	53.13	55.29	45.71	66.11	60.30
CLIP-based	CLIP [41]	95.16	76.77	91.84	72.94	81.43	63.62	77.08	61.95	77.88	72.19	84.50	68.94	84.65	69.40
	CLIP + OpenMax [2]	93.45	79.13	92.09	73.67	81.00	61.54	76.93	62.78	78.34	73.26	81.89	69.40	83.95	69.96
	CLIP + OSDA [38]	92.62	75.40	90.21	70.89	82.58	67.35	80.53	65.70	74.45	75.22	82.00	73.62	83.73	71.36
	CoOp [66]	78.77	26.87	92.02	39.26	73.85	36.26	58.54	34.81	66.03	44.34	61.13	68.34	71.72	41.65
	CoCoOp [65]	85.76	32.93	89.47	37.01	75.38	34.38	52.77	33.50	64.84	47.57	60.63	56.30	71.48	40.28
	MaPLe [20]	93.97	48.47	89.70	43.33	79.47	33.06	70.54	43.83	69.34	62.20	74.67	60.57	79.62	48.58
	LASP [5]	88.45	30.37	90.67	39.41	76.13	34.52	60.89	35.23	66.78	50.22	62.34	61.56	74.21	41.89
	PromptSRC [21]	94.53	43.32	90.13	42.78	80.21	36.40	75.34	44.25	65.51	59.45	73.60	62.56	79.89	48.13
	CLIPN [55]	96.24	45.00	84.82	50.72	84.55	42.83	81.70	45.56	77.16	62.60	77.38	66.92	83.64	52.27
	STYLIP [3]	95.36	50.74	90.75	65.66	84.73	60.97	80.59	58.15	79.88	71.99	80.22	69.11	85.26	62.77
	CLIPN + STYLIP	96.37	64.46	84.65	68.02	83.67	76.50	82.14	59.24	76.93	72.15	86.59	76.18	85.06	69.43
	MaPLe + SD	91.47	82.60	91.70	72.67	85.02	80.60	79.92	65.82	77.62	72.83	83.79	79.30	84.92	75.64
	PromptSRC + SD	93.21	87.95	90.34	72.62	84.60	83.31	80.92	65.37	78.44	77.89	83.87	82.95	85.23	78.35
	STYLIP + SD	91.78	87.42	92.11	73.34	85.51	81.22	81.45	68.10	79.05	78.52	84.12	83.21	85.67	78.64
	ODG-CLIP	**99.53**	**99.70**	**95.71**	**86.53**	**98.32**	**96.08**	**91.53**	**78.27**	**84.60**	**90.00**	**95.68**	**94.48**	**94.23±0.19**	**90.84±0.26**

4.1. Comparison to the literature

Competitors: We carried out in-depth comparisons of our ODG-CLIP model against traditional ODG methods, including [33, 48, 57, 68]. Notably, these methods are grounded on conventional CNN backbones like ResNet-18 [16] and use a confidence-driven classification for OSR, where a sample with low classification probability for the closed-set classes is marked as an outlier. Furthermore, we conducted exhaustive evaluations against CLIP-based models: **Baseline CLIP** [41]: This method evaluates the OSR prowess by gauging the prediction confidence of target images concerning manually defined prompts for the known classes, exemplified as `a photo of a [CLS]`. **CLIP features paired with OpenMax (OSR)** [2]: Here, we combined pre-trained CLIP features with the OpenMax technique to cultivate a joint closed-open set classifier using the amalgamated source domains. **CLIP features paired with OSDA** [38]: We combined the pre-trained CLIP with OSDA-BP [38] for open-set DA, considering a blended source domain and assuming that the target domain is known. **Prompt learning techniques**: We assessed a variety of prompt learning strategies, inclusive of [3, 20, 21, 65, 66], adopting a confidence-centric open-set prediction approach, similar to the baseline CLIP evaluation mentioned above. **Incorporation of models with an unknown-class prompt (Model + SD)**: For models such as [20], [21] and [3], we enriched them by adding the `unknown`-class prompt that relies on our considered diffusion-based pseudo-open sample synthesis for training this prompt. For LASP [5], an extra `unknown`-class was considered for text-to-text contrastive loss during training, and the open-set novel class samples were classified into this unknown class during evaluation. **CLIP-based open-set classification (CLIPN)** [55]: This method employs CLIP for OSR through the training of a sophisticated encoder for negative-class tokens. Additionally, we integrated the prompt learning of STYLIP [3] into the CLIPN architecture, replacing the hand-crafted tokens. When it comes to closed-set DG, we scrutinized leading non-CLIP methods [1, 7, 17], as well as the prompt learning-based approaches.

Discussions on ODG and closed-set DG performance: In Table 1, we compare ODG across six datasets. ODG-CLIP notably surpasses Cumix [33], MixStyle [68], DAML [48] and MEDIC [57] in H-score, posting gains of 44.03%, 42.18%, 39.55% and 30.54%, respectively. Remarkably, against the zero-shot CLIP approach, ODG-CLIP exhibits a marked superiority, registering a significant boost of 21.44% in the H-score. When juxtaposed with other prompt learning techniques, ODG-CLIP continues to impress. While integrating explicit `unknown`-class prompt learning in [3, 20, 21] does improve performance relative to their confidence-centric prediction counterparts, they remain eclipsed by ODG-CLIP. Furthermore, while CLIPN [55] manages to outscore several competitors in H-score, ODG-CLIP ultimately trumps CLIPN, benefiting from its embrace of cross-domain learning — a facet absent in CLIPN. In this regard, STYLIP+CLIPN improves the performance of CLIPN substantially but is still poor than ODG-CLIP by $\approx 21\%$ in H-score. Finally, STYLIP+SD provides the best result among the competitors, resulting in the average H-score of 78.64%. However, it lags ODG-CLIP, which outputs an average H-score of 90.84%.

This improvement in performance can be attributed to enhanced visual feature extraction and our innovative approach to generalizable prompt learning, which fosters a more integrated alignment of image and prompt attributes. Additionally, by simultaneously addressing open and closed-set classification tasks, we enhance the discriminative quality of the embedding space. This leads to a balanced and harmonious performance across both closed and open classes, a synergy often lacking in other models.

In line with the trends observed in ODG tasks, ODG-

Table 2. Mean leave-one-domain-out performance on PACS, VLCS, Office-Home, Digits-DG and Mini-DomainNet for DG.

	Methods	PACS	VLCS	O.H.	D-DG	M.DNet	Avg.
CNN	SWAD [7]	88.10	79.10	70.60	-	-	79.27
	EoA [1]	88.60	79.10	72.50	-	-	80.07
	DandelionNet [17]	89.20	81.60	70.40	-	-	80.40
CLIP-based	CLIP [41]	94.89	82.14	78.40	64.59	78.73	79.75
	CoOp [66]	97.11	83.34	81.33	77.11	72.30	82.23
	CoCoOp [65]	96.54	85.02	81.05	79.36	71.51	82.70
	MaPLe [20]	97.72	86.75	83.52	80.25	73.87	84.42
	LASP [5]	97.02	87.25	84.13	79.92	70.67	83.80
	PromptSRC [21]	98.02	86.34	83.89	82.40	76.10	85.35
	StyLIP [3]	98.17	87.21	85.94	81.62	80.43	86.67
	ODG-CLIP	**99.83**	**95.74**	**96.91**	**96.38**	**96.65**	**97.10**

CLIP consistently outperforms all competitors in closed-set DG tasks across all datasets. Closed-set DG represents a special case of ODG, which does not include any category shift among the domains. As illustrated in Table 2, ODG-CLIP showcases its superiority over traditional DG methods based on the conventional CNN backbones like SWAD [7], EoA [1], and DandelionNet [17]. Notably, ODG-CLIP outperforms SWAD, EoA, and DandelionNet by an average Acc of 17.83%, 17.03%, and 16.70%, respectively. Furthermore, when compared against prompt learning benchmarks, ODG-CLIP demonstrates significantly improved performance, outperforming CoOp [66] by 14.87%, CoCoOp [65] by 14.40%, MaPLe [20] by 12.68%, and StyLIP [3] by 10.43%.

4.2. Ablation analysis [1]

(i) How do $\mathcal{L}_{sem}$ and $\hat{x}$ help ODG? While $\tilde{x}$ introduces class-discriminative information to images, $\mathcal{L}_{sem}$ maintains the consistency of this information across various domains. Their roles are interconnected, influencing the ODG performance significantly. In this context, we begin by assessing the influence of $\mathcal{L}_{sem}$, as detailed in Table 3. Across all datasets, we notice a consistent enhancement in closed-set performance by approximately $3-4\%$ and nearly a 5% surge in H-score compared to the model without $\mathcal{L}_{sem}$.

Table 3. Ablation analysis for $\mathcal{L}_{sem}$ and $\hat{x}$ in our proposed ODG-CLIP. Manual $\hat{x}$ refers to the case where $\hat{x}$ is derived from the ready-made embeddings of the class names.

Methods	PACS		O.H.		M.Data		M.DNet	
	Acc	H	Acc	H	Acc	H	Acc	H
w/o $\hat{x}$ and $\mathcal{L}_{sem}$	90.47	88.34	92.21	87.00	73.56	75.73	87.24	83.51
w/o $\mathcal{L}_{sem}$, with $\hat{x}$	94.21	92.56	95.67	91.56	80.34	85.32	91.24	90.88
Manual $\hat{x}$	93.54	92.82	95.31	91.22	78.53	79.26	90.65	86.52
Full (ours)	**99.53**	**99.70**	**98.32**	**96.08**	**84.60**	**90.00**	**95.68**	**94.48**

We also examined the impact of $\hat{x}$ by considering a model configuration in which $\hat{x}$ is not combined with x. As indicated in Table 3, incorporating $\hat{x}$ leads to marked improvements in performance for both closed and open classes. Furthermore, we investigated an alternative scenario where $\hat{x}$ is derived from the embeddings of class names. In this setup, each class name yields a singular $\hat{x}$, regardless of the domain distinctions. This approach contrasts with our methodology, where $\hat{x}$ inherently captures the visual space distribution, reflecting domain-specific dynamics. In comparison, this static method of manually deriving $\hat{x}$ does not account for the variability across domains, leading to lesser adaptability. The advantages of our approach are evident in the results, demonstrating a consistent 5-11% increase in the H-score compared to the manual approach.

[1]More analysis on domain alignment, prompts management, qualitative visualizations, model complexity analysis, *etc*. are mentioned in the `supplementary`.

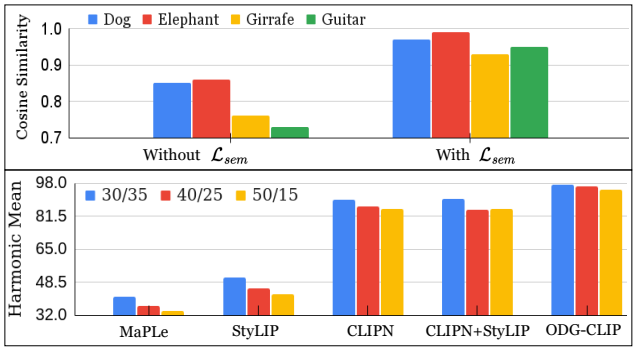

Figure 3. **Top**: Ablation on the average cosine similarity values of $\hat{x}$ on four shared classes across the domains in PACS, **Below**: Openness analysis of different methods on Office-Home.

Our hypothesis posits that employing $\mathcal{L}_{sem}$ ensures consistent representation in $\hat{x}$ for images from shared classes across different domains. To substantiate this, we conducted an in-depth analysis of the average pairwise `cosine` similarity of $\hat{x}$ for four shared classes within the PACS dataset, as depicted in Figure 3 (Top). We examined two specific scenarios: **(i)** where $\hat{x}$ is derived using our method but without the integration of $\mathcal{L}_{sem}$ in our training objective, and **(ii)** utilizing the complete ODG-CLIP model with $\mathcal{L}_{sem}$. The results reveal that the full ODG-CLIP model, incorporating $\mathcal{L}_{sem}$, exhibits a higher average `cosine` similarity compared to the version without $\mathcal{L}_{sem}$. This outcome supports our assertion that $\mathcal{L}_{sem}$ indeed promotes uniformity in the augmented features embedded into the images, reducing the domain divergence considerably. We further report the Frećhet [11] distance between the source and target domains to justify the better domain alignment offered by ODG-CLIP in the `Supplementary`.

(ii) Openness analysis of ODG-CLIP: To assess ODG-CLIP's effectiveness in scenarios with varying levels of 'openness', defined by the number of unknown classes compared to the known classes in $\mathcal{D}_{\mathcal{T}}$, we segmented the Office-Home dataset into subsets with different distributions of known and unknown classes: specifically, splits of $30/35$, $40/25$, and $50/15$. When compared with existing techniques like [3, 20, 55] and their combinations, as shown in Figure 3 (Below), ODG-CLIP demonstrates superior capability in differentiating known from unknown classes. This proficiency is highlighted by margins of 7.41%, 9.79%, and 9.50% in the respective dataset splits.

(iii) Comparison of the proposed diffusion-based

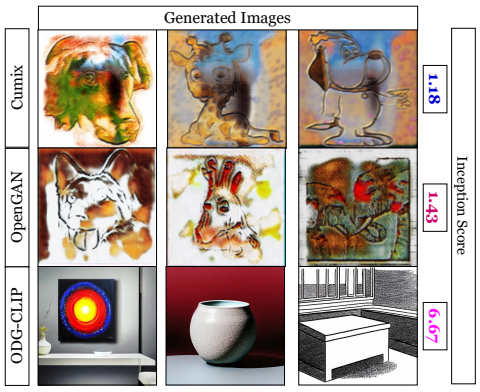

Figure 4. Comparison of ODG-CLIP with Cumix [33] and OpenGAN [23] on inception score for the generated open-set samples.

pseudo-open image generation to the literature: To establish the superiority of the diffusion-based pseudo-open image synthesis, we compared it against two established counterparts. The first, cumix [33], blends two images with a Dirichlet coefficient, classifying the resultant image as outside the known class set. The second, OpenGAN [23], employs adversarial learning to create outlier images distinct from the known classes. We integrated both these methods with ODG-CLIP for open sample synthesis, replacing our diffusion-based approach.

Table 4. Analysis of the effects of pseudo-open image generation by our diffusion model and methods from the OSR literature.

Methods	PACS		O.H.		M.Data		M.DNet	
	Acc	H	Acc	H	Acc	H	Acc	H
OpenGAN [23]	95.89	91.35	91.57	90.23	79.93	80.41	92.80	89.76
Cumix [33]	95.61	93.08	95.42	92.11	80.97	85.92	91.63	91.02
With SD [44] (**Ours**)	**99.53**	**99.70**	**98.32**	**96.08**	**84.60**	**90.00**	**95.68**	**94.48**

Our analyses, detailed in Table 4, highlight the diffusion method's enhanced performance, evidenced by consistently higher H-scores. This suggests the diffusion model's proficiency in effectively mapping the open space, leading to more accurate classification weights for the `unknown`-class prompts. The qualitative analysis in Fig. 4, along with the inception scores [47] of the generated images, further affirm the diffusion method's effectiveness. Images generated using cumix and OpenGAN exhibited lower quality, as indicated by their inception scores of 1.18 and 1.35, respectively. In contrast, the diffusion-based approach successfully synthesized high-quality open-set samples, achieving a significantly higher inception score of 6.67.

Table 5. Analysis of prompts in ODG-CLIP. B1-3 defines the different baseline cases as mentioned below.

Baselines	PACS		O.H.		M.Data		M.DNet	
	Acc	H	Acc	H	Acc	H	Acc	H
B1-manual	92.28	80.56	84.50	65.85	78.23	75.59	82.24	70.41
B2-manual	93.42	84.61	88.52	73.89	80.95	80.67	88.56	83.60
B3-Gaussian	93.72	93.84	94.60	88.34	78.92	76.34	90.51	88.78
ODG-CLIP	**99.53**	**99.70**	**98.32**	**96.08**	**84.60**	**90.00**	**95.68**	**94.48**

(iv) Analysis of the prompts in ODG-CLIP: In order to explore the nuances of prompt processing in ODG-CLIP, we conducted an ablation study, detailed in Table 5, focusing on three distinct configurations: **(B1-manual)**: This setup involves the use of manually defined prompts for both $\mathcal{P}_{dom,cls}$ and $\mathcal{P}_{dom}$, *e.g.* `[dom] of a [cls]` and `This is a [dom]`, where [dom] can sketch/painting *etc*. **(B2-manual)**: In this scenario, the $[dom]$ token is manually initialized, but the generic prompts (ω, ν) are set as learnable parameters. **(B3-Gaussian)**: This approach aligns with our proposed method for prompt processing, except we initialize ω and ν with noise sampled from a normal distribution $\mathcal{N}(0, \mathbb{I})$. This is in contrast to our full model, which begins the refinement process with these variables initialized from the embedding of the phrase '`Image of a`'. The results demonstrate that our strategies for prompt processing outperform the other baselines convincingly.

(v) Impact of `NP` on performance: In this regard, we seek to evaluate ODG-CLIP's performance with stable diffusion-driven pseudo-open samples generated using only `PP` and combined `PP` and `NP`. Positive prompts alone proved ambiguous for precise content generation, while negative prompts enhanced clarity on excluding elements for unknown classes. Without `NP`, the lack of specificity degraded visual-textual mapping in ODG-CLIP, affecting classification accuracy. Employing both positive `PP` and `NP` yielded pseudo-open images with varied granularities, improving classifier's robustness for the open samples (Table 6).

Table 6. Importance of negating prompts (`NP`) together with positive prompts (`PP`) for open image synthesis using diffusion.

Methods	PACS		O.H.		M.Data		M.DNet	
	Acc	H	Acc	H	Acc	H	Acc	H
Only `PP`	92.45	92.16	93.72	90.83	78.20	81.57	91.30	89.78
`PP` + `NP`	**99.53**	**99.70**	**98.32**	**96.08**	**84.60**	**90.00**	**95.68**	**94.48**

5. Takeaways & Future Scope

This paper presents ODG-CLIP, an innovative solution tailored to the complex and relatively unexplored domain of open-domain generalization, viewed through the lens of prompt learning in CLIP. Central to ODG-CLIP are three pivotal innovations: *Domain-aware prompt learning*, *Prompt-driven visual embedding enhancement*, and *Unified classification for known and novel categories*. We develop a unique unknown-class prompt to handle the outliers during testing, specifically trained with data generated by capitalizing on conditional diffusion models. In our extensive evaluations across both closed-set and open-set DG settings, ODG-CLIP has demonstrated superior performance over existing methodologies. A possible future direction may consider the dense prediction tasks in the ODG setting.

The authors gratefully **acknowledge** the immense support provided by Aisin Corporation, Japan.

References

[1] Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *Advances in Neural Information Processing Systems*, 35:8265–8277, 2022. 6, 7

[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. 2, 6

[3] 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. 1, 2, 3, 4, 5, 6, 7

[4] Silvia Bucci, Mohammad Reza Loghmani, and Tatiana Tommasi. On the effectiveness of image rotation for open set domain adaptation. In *European conference on computer vision*, pages 422–438. Springer, 2020. 3

[5] Adrian Bulat and Georgios Tzimiropoulos. Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23232–23241, 2023. 3, 6, 7

[6] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019. 3

[7] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021. 6, 7

[8] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081, 2021. 2

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[11] DC Dowson and BV666017 Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982. 7

[12] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022. 2

[13] 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)*, December 2013. 5

[14] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021. 3

[15] ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*, 2017. 2

[16] 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

[17] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Dandelionnet: Domain composition with instance adaptive classification for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19050–19059, 2023. 6, 7

[18] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pages 124–140. Springer, 2020. 3

[19] 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. 3

[20] 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, June 2023. 2, 3, 6, 7

[21] 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

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[23] 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. 2, 8

[24] Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, R Venkatesh Babu, et al. Towards inheritable models for open-set domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12376–12385, 2020. 3

[25] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 3

[26] 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. 1, 5

[27] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455, 2019. 3

[28] Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in Neural Information Processing Systems*, 33:3118–3129, 2020. 3

[29] Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. Maximum density divergence for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3918–3930, 2021. 3

[30] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[31] Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8886–8895, 2021. 3

[32] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *International Conference on Machine Learning*, pages 3915–3924. PMLR, 2019. 3

[33] 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. 2, 6, 8

[34] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in Neural Information Processing Systems*, 35:35087–35102, 2022. 2, 3

[35] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 613–628, 2018. 2

[36] Jonas Oppenlaender. The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference*, pages 192–202, 2022. 3

[37] 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, January 2023. 2

[38] 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, 6

[39] Novi Patricia and Barbara Caputo. Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1442–1449, 2014. 3

[40] 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. 5

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 5, 6, 7

[42] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *https://www.mikecaptain.com/resources/pdf/GPT-1.pdf*, 2018. 3

[43] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[44] 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. 1, 2, 3, 4, 6, 8

[45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[46] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 153–168, 2018. 3

[47] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 8

[48] 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, 5, 6

[49] Mainak Singha, Ankit Jha, and Biplab Banerjee. Gopro: Generate and optimize prompts in clip using self-supervised learning. 2023. 3

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

[51] 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. 5

[52] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? *arXiv preprint arXiv:2110.06207*, 2021. 2

[53] 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. 5

[54] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR, 2019. 2

[55] 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. 2, 6, 7

[56] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 1

[57] 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. 1, 3, 6

[58] Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3622–3626. IEEE, 2020. 3

[59] Ziqi Wang, Marco Loog, and Jan van Gemert. Respecting domain relations: Hypothesis invariance for domain generalization. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9756–9763. IEEE, 2021. 3

[60] Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014. 3

[61] Shiqi Yang, Yaxing Wang, Kai Wang, Shangling Jui, and Joost van de Weijer. One ring to bring them all: Towards open-set recognition under domain shift. *arXiv preprint arXiv:2206.03600*, 2022. 3

[62] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4016–4025, 2019. 2

[63] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. Towards principled disentanglement for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8024–8034, 2022. 3

[64] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2021. 2

[65] 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. 1, 2, 3, 6, 7

[66] 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, 5, 6, 7

[67] 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, 5

[68] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. 3, 6

[69] Ronghang Zhu and Sheng Li. Crossmatch: Cross-classifier consistency regularization for open-set single domain generalization. In *International Conference on Learning Representations*, 2021. 3