The Finer the Better: Towards Granular-aware Open-set Domain Generalization

Yunyun Wang[1], Zheng Duan[1], Xinyue Liao[1], Ke-Jia Chen[1], Songcan Chen[2]

[1]School of Computer Science, University of Posts and Telecommunications, Nanjing, China
[2]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
{wangyunyun, 1223045211, 1224045627, chenkj}@njupt.edu.cn, s.chen@nuaa.edu.cn

Abstract

Open-Set Domain Generalization (OSDG) tackles the realistic scenario where deployed models encounter both domain shifts and novel object categories. Despite impressive progress with vision-language models like CLIP, existing methods still fall into the dilemma between structural risk of known-classes and open-space risk from unknown-classes, and easily suffers from over-confidence, especially when distinguishing "hard unknowns" that share fine-grained visual similarities with known classes. To this end, we propose a Semantic-enhanced CLIP (SeeCLIP) framework that explicitly addresses this dilemma through fine-grained semantic enhancement. In SeeCLIP, we propose a semantic-aware prompt enhancement module to decompose images into discriminative semantic tokens, enabling nuanced vision-language alignment beyond coarse category labels. To position unknown prompts effectively, we introduce duplex contrastive learning with complementary objectives, that is, repulsion to maintain separability from known classes, and cohesion to preserve semantic proximity. Further, our semantic-guided diffusion module synthesizes pseudo-unknowns by perturbing extracted semantic tokens, generating challenging samples that are visually similar to known classes yet exhibit key local differences. These hard negatives force the model to learn finer decision boundaries. Extensive experiments across five benchmarks demonstrate consistent improvements of 3% accuracy and 5% H-score over state-of-the-art methods.

Appendix — https://github.com/Leagelab/seeclip

Introduction

Domain Generalization (DG) focuses on training models maintaining robustness when deployed in unseen target domains (Wang et al. 2022a). The target domains typically have data distributions that vary substantially from those of the source domains. Traditional DG assumes that all domains share the same label space (Li et al. 2017; Zhou et al. 2020), while practical applications often encounter scenarios where the target domain contains not only familiar classes but also entirely new ones. It gives rise to Open-Set Domain Generalization (OSDG) that needs to handle both known and unknown classes under distribution shift (Wang et al. 2023). It is common across various applications (Sun and Dong 2023; Yang et al. 2024), such as autonomous driving, remote sensing, and medical imaging, in which detecting unforeseen objects is often critical.

Recently, a few researches have started focusing on OSDG, and the core challenge lies in identifying unknown samples never encountered during training (Vaze et al. 2022). Existing methods commonly employ meta-learning to boost classifier learning on known classes (Shu et al. 2021; Wang et al. 2023), enabling the model to distinguish unknown classes from known ones during inference. Further, pseudo open-set samples are generated in learning, recasting the open-set recognition as a closed-set problem (Singha et al. 2024; Gupta et al. 2025). However, they often struggle to accommodate the variability and complexity of real unknown data. Recently, large-scale vision-language models (Radford et al. 2021; Jia et al. 2021; Cherti et al. 2023) that leverage semantic information in text prompts to better understand images have demonstrated impressive performance for OSDG.

Nevertheless, previous methods have two main limitations. First, they fail to sufficiently model the fine-grained semantics (Lang et al. 2024), leading to over-confidence in distinguishing semantically similar categories. That is, they fail to capture the subtle semantic differences between similar categories like "persian" and "siamese" cats. Consequently, they may misclassify unknown objects that have subtle semantic differences from known classes, sacrificing the open-space risk for structural risk. Second, in pseudo-open generation, capturing the real data distribution is often challenging due to its openness and diversity. Generating unknown samples too far from or too close to known categories both leads to model bias (Bele et al. 2024; Gupta et al. 2025), increasing the generalization risk. In fact, they commonly generate samples that diverge excessively from known classes, e.g., "desk" vs. known "cat", causing the model to overlook challenging unknown samples like "tiger". Consequently, the model becomes biased to known classes and falls into the dilemma between structural risk and open-space risk, leading to sub-optimal performance.

To this end, we propose a Semantic-enhanced CLIP (SeeCLIP) framework that leverages fine-grained semantics in both prompt learning (Zhou et al. 2022a,b) and pseudo-open generation, so as to enable precise discrimination among categories. Specifically, we introduce a semantic-

aware prompt enhancement module, which utilizes spatial attention analysis to extract key semantic features from CLIP, and dynamically integrates them into text prompts using learnable weights, establishing a fine-grained vision-language alignment. Duplex contrastive learning is proposed for prompt learning with two complementary constraints. The repulsion loss aims to push the unknown prompts away from known classes, while the cohesion loss restricts their relative distance for semantic proximity. In this way, the unknown prompt will be similar to known prompts, while exhibiting key semantic differences. Moreover, we present a semantic-guided diffusion module, in which the key semantic features extracted are deliberately perturbed and injected into a pre-trained diffusion model as control conditions. Consequently, pseudo-open samples that are globally similar yet locally different from known classes are generated. In fact, these generated samples can be regarded as hard unknowns, which help capture the core semantic features for individual categories, and construct a compact feature space for known classes. In this way, the model considers both structural and open-space risks, and the robustness of open-set recognition will be enhanced.

We also formulate a generalization bound for OSDG, and show that SeeCLIP achieves a lower generalization risk. There are also fine-grained DG researches (Yu et al. 2024; Bi et al. 2025) recently, aiming to identify fine-grained categories within a closed-set environment, and the training data commonly includes structured fine-grained labels. Unlike previous studies, we focus on hard open-set classes without supervised fine-grained knowledge, a significant challenge for OSDG in real-world scenarios. We aim to balance the structural risk of known-classes and open-space risk of unknown-classes, achieving more robust generalization capability. The main contributions are summarized as follows,

- We propose SeeCLIP for OSDG, which utilizes fine-grained semantics in both prompt learning and pseudo-open generation, so as to address the under-utilization of fine-grained semantic details, and improve model capability for recognizing indistinguishable unknowns.
- We propose a generalization bound for OSDG, under which we show that SeeCLIP accommodates both the structural risk and the open-space risk, achieving a lower generalization risk for OSDG.
- Extensive experiments on multiple benchmark datasets demonstrate that SeeCLIP significantly enhances the robustness of open-set recognition. It achieves an average improvement of 3% in accuracy and approximately 5% increase in H-score, over the state-of-the-art methods.

Related Works

Open-Set Domain Generalization

OSDG extends traditional DG by requiring models to handle both known and unknown classes under distribution shifts, distinct from both Open Set Recognition (OSR) (Bendale and Boult 2016; Kong and Ramanan 2021) and Open Set Domain Adaptation (OSDA) (Panareda Busto and Gall 2017). Pioneering work in OSDG was led by Shu et al., who propose a domain-augmented meta-learning framework to simulate open-set scenarios. Subsequent advancements have explored diverse technical paths. Methods like MEDIC (Wang et al. 2023) match gradients at both domain and class levels. Adversarial approaches (Bose et al. 2023; Rakshit et al. 2022) generate pseudo-unknowns for training, recasting open-set recognition as a closed-set problem. Recent efforts have shifted toward integrating prompt tuning. A representative example is ODG-CLIP (Singha et al. 2024), which leverages an explicit "unknown" prompt and a pre-trained diffusion model (Rombach et al. 2022) to generate pseudo-open samples, to mimic domain-specific styles while exhibiting semantic divergence. However, previous methods often fail to model fine-grained semantic distinctions or generate pseudo-samples with excessive divergence from known classes, leading to over-confidence boundaries that misclassify unknowns with localized variations.

Vision-Language Models and Prompt Learning

Vision-Language Models (VLMs) like CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) have revolutionized zero-shot recognition through contrastive learning of image-text alignment. Early VLMs relied on hand-crafted prompts, e.g., "a photo of a {class}", but static prompts lack adaptability to domain-specific nuances or fine-grained semantics. Subsequent prompt learning techniques dynamically optimized prompt tokens for task alignment. CoOp (Zhou et al. 2022b) learns continuous prompt embeddings via gradient descent. DualPrompt (Wang et al. 2022b) decouples domain-invariant and task-specific prompts to enable multi-domain adaptation. MaPLe (Khattak et al. 2023) further enhances visual-semantic alignment by joint cross-modal prompt tuning. However, they still face challenges in open-set scenarios: 1) They prioritize global feature alignment but overlook localized semantic cues critical for fine-grained discrimination. 2) Hand-crafted or rigidly learned prompts struggle to reject unknowns highly similar to known classes. Thereby, we propose SeeCLIP by incorporating fine-grained semantic cues to effectively reject hard unknown samples, thereby enhancing open-set recognition capability in OSDG.

Proposed Methodology

Task Definition and Model Architecture

In OSDG, we consider multiple source domains, each with its own distinct data distribution, yet sharing a common set of classes. Formally, given M source domains $\{\mathcal{D}_1, \ldots, \mathcal{D}_M\}$, where each $\mathcal{D}_k = \{(x_i^s, y_i^s)\}_{i=1}^{n_k}$ consists of n_k labeled samples from C categories. The target domain $\mathcal{D}_t = \{x_j^t\}_{j=1}^{n_t}$ exhibits a different distribution from source domain. It contains unlabeled samples that may belong to known classes or novel classes not encountered during training. The known label space is $\mathcal{Y}^s = \{1, \ldots, C\}$, and the unknown label space in the target domain is $\mathcal{Y}^u = \{C+1\}, \quad \mathcal{Y}^t = \mathcal{Y}^s \cup \mathcal{Y}^u$. The learning goal is to develop a robust classifier capable of distinguishing between known and unknown classes in the target domain, while classifying samples from known classes accurately.

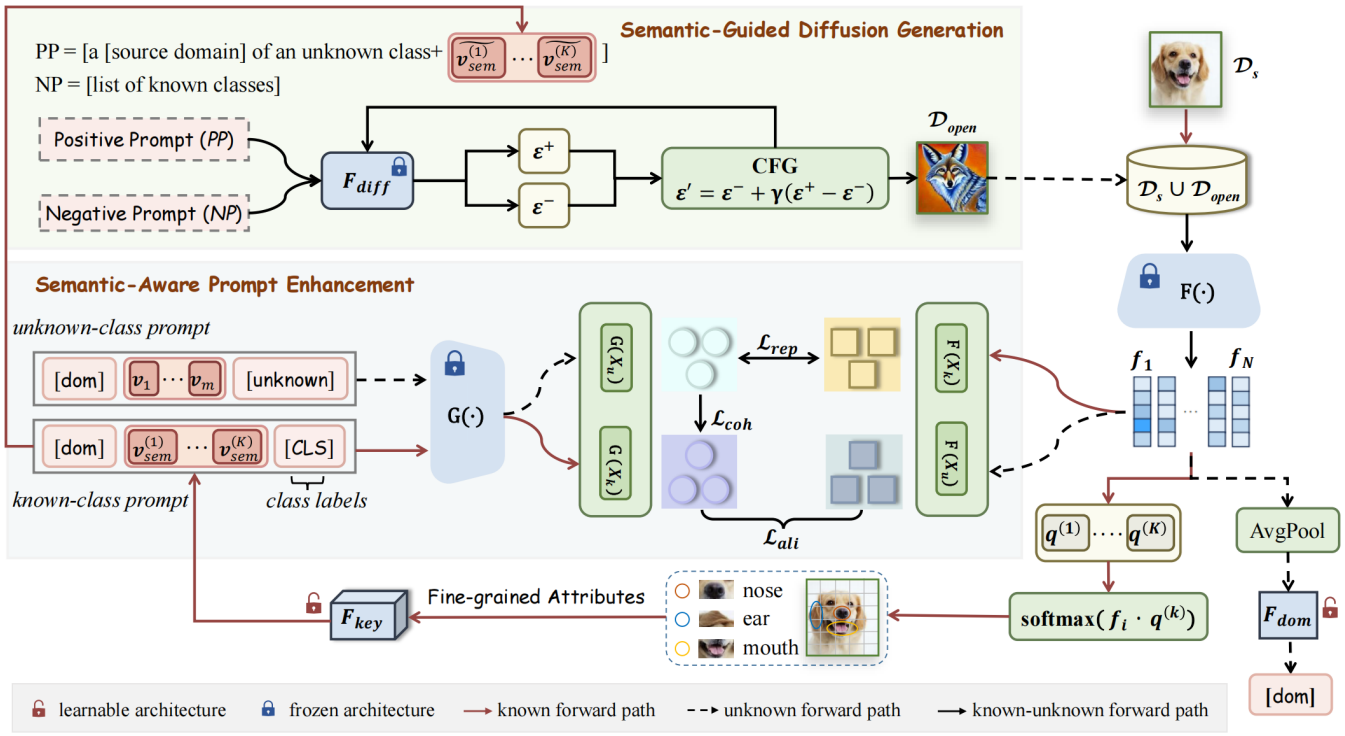

Figure 1: Overall architecture of SeeCLIP, which consists of a semantic-aware prompt enhancement module for boosting prompt learning with fine-grained semantics, and a semantic-guided diffusion module to generate pseudo-unknowns as hard unknowns.

We propose a SeeCLIP framework, which enhances OSDG by effectively leveraging fine-grained semantics, so as to boost the robustness of open-set recognition. SeeCLIP consists of a semantic-aware prompt enhancement module, and a semantic-guided diffusion module, as shown in Figure 1. In semantic-aware prompt enhancement learning, fine-grained semantic tokens $\left\{v_{\text{sem}}^{(1)}, \ldots, v_{\text{sem}}^{(K)}\right\}$ are extracted via K-head attention using learnable queries $\{q^{(1)}, \ldots, q^{(K)}\}$, and combined with domain tokens [dom] to construct enhanced prompts. For unknown prompts, learnable class-agnostic semantic tokens $\{v_1, \ldots, v_m\}$ and the token $[unknown]$ are used. In semantic-guided diffusion module, perturbed semantic tokens are fused with a positive prompt PP to form the positive condition, while a negative prompt NP listing all known classes serves as a counter-guidance to generate pseudo-unknowns. Classifier-Free Guidance (CFG)(Ho and Salimans 2022) is employed to strengthen the distinction between positive and negative conditions. Three losses, including alignment, repulsion, and cohesive losses, jointly optimize the prompts, enforcing a clear boundary between known and unknown categories.

Semantic-Aware Prompt Enhancement

To distinguish semantically similar unknowns in open domains, we aim to enhance prompt learning with both domain-level semantics and multi-token discriminative features, enabling fine-grained vision-language alignment under distribution shifts.

Representation of Prompts. The input data is encoded by the vision encoder of CLIP, then we build a domain token $v_{dom}^{(k)}$ for source domain $\mathcal{D}_k$ by a simple average over all samples. Each sample is also transformed into a sequence of patch embeddings $\{f_i\}_{i=1}^{N}$, where each $f_i \in \mathrm{R}^d$ represents the feature embedding of a localized region. To extract fine-grained semantics, we adopt a K-head attention pooling mechanism,considering that only a few visual heads in attention mechanisms dominate visual understanding by focusing on key image regions linked to fine-grained semantics(Zhang et al. 2025). K learnable query vectors $\{q^{(k)}\}_{k=1}^{K}$ are introduced for each class, each focusing on a specific semantic region, and are shared across all domains. An attention weight $w_i^{(k)}$ is assigned for the k-th attention head,

$$\omega_i^{(k)} = \frac{\exp\left(q^{(k)} \cdot f_i\right)}{\sum_{j=1}^{N} \exp\left(q^{(k)} \cdot f_j\right)} \tag{1}$$

The k-th semantic token $v_{sem}^{(k)}$ is then formulated as a weighted aggregation of all patch embeddings, i.e.,

$$v_{sem}^{(k)} = \sum_{i=1}^{N} \omega_i^{(k)} \cdot f_i \tag{2}$$

It generates K semantic tokens, each adept at capturing the discriminative features from a distinct part, e.g., tail, eyes, ears. Further, to construct an enhanced prompt for each known class, we integrate both domain and semantic tokens

using learnable projections $\Phi(\cdot)$ and $\Psi(\cdot)$, respectively,

$$p_c = [\Phi(v_{\text{dom}})], [\Psi_1(v_{\text{sem}}^{(1)}), \ldots, \Psi_K(v_{\text{sem}}^{(K)})], [\textit{classname}] \tag{3}$$

For unknown categories lacking semantic description, we construct the prompt as,

$$p_{unk} = [\Phi(v_{\text{dom}})], [v_1, \ldots, v_m], [\textit{unknown}] \tag{4}$$

where tokens $v_1, \ldots, v_m$ are a set of learnable semantic vectors encoding generalizable, class-agnostic patterns frequently observed in unknown classes. $[unknown]$ serves as a trainable label token, indicating that the prompt corresponds to an unknown category. It helps represent and distinguish samples that fall outside known-class distribution.

Learning for Class Prompts. We introduce a duplex contrastive formulation for prompt learning. To enforce a clear open-set boundary, a margin-based repulsion loss pushes the unknown prompt away from all known embeddings,

$$\mathcal{L}_{rep} = \sum_{c=1}^{C} \max(0, \delta - \text{sim}(F_t(p_{\text{unk}}), F_v(\mathbf{X}_c))) \tag{5}$$

where δ is the margin hyperparameter, F_t and F_v are the text and vision encoders of CLIP, respectively, $\mathbf{X}_c$ denotes data from the c-th class. Besides, to prevent unknown prompt from drifting too far from known categories, we encourage its proximity to the center of known prompt embeddings with a cohesive loss,

$$\mathcal{L}_{coh} = \left\| F_t(p_{\text{unk}}) - \frac{1}{C} \sum_{c=1}^{C} F_t(p_c) \right\|_2^2 \tag{6}$$

We also apply an L_1 regularization to each projected semantic token to suppress overfitting and encourage a sparse utilization of fine-grained semantics, i.e.,

$$\mathcal{L}_{reg} = \lambda \sum_{k=1}^{K} \left\| \Psi_k \left(v_{\text{sem}}^{(k)} \right) \right\|_1 \tag{7}$$

In this way, the model can automatically select the most discriminative semantic components while filtering out noisy or redundant ones, bringing more robust and transferable representations for open-set recognition across different domains.

Semantic-Guided Diffusion Generation

Building on the semantic token representations, we aim to synthesize pseudo-unknown samples that globally resemble known classes while exhibiting local deviations, focusing on the core semantics of known categories.

We first perturb each semantic token with Gaussian noise,

$$\widetilde{v_{\text{sem}}^{(k)}} = v_{sem}^{(k)} + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \sigma^2 I) \tag{8}$$

where σ regulates the perturbation intensity. A small σ ensures coherence semantic with known classes, yet discriminative local variations. The perturbed tokens serve as conditional inputs to the diffusion model.

We use a positive textual prompt describing the source domain with unknown-class context,

PP = "A [*source domain*] image of an unknown class"

where [*source domain*] is the specific domain name, such as "photo" or "art painting" in PACS dataset. Its embedding E_{text}^{+} is then concatenated with visual condition $E_{\text{visual}} = \left[\Psi_1 \left(\widetilde{v_{\text{sem}}^{(1)}} \right), \ldots, \Psi_K \left(\widetilde{v_{\text{sem}}^{(K)}} \right) \right]$ to form a joint condition,

$$E_{\text{joint}} = \left[E_{\text{text}}^{+}; E_{\text{visual}} \right] \tag{9}$$

It is subsequently fed into the cross-attention layers of the diffusion network, ensuring that the generated samples preserve global domain structure while introducing controlled visual divergence.

Further, to suppress the high belongingness to known categories, we introduce a negative textual prompt including all known-class names,

NP = "[*Known Class* 1], ..., [*Known Class C*] "

By integrating positive and negative prompt conditions, model will generate pseudo unknowns that balance semantic validity with distributional novelty. It ensures that the generated samples mimic real-world open-set categories, thereby enhancing generalization to unseen classes.

Loss Functions, Training, and Inference

SeeCLIP learns by simultaneously aligning image representations with their corresponding prompts, enforcing separation between known and unknown prompts, and regularizing semantic token projections to prevent overfitting as well.

A symmetric contrastive alignment loss is adopted to pull each sample and the corresponding prompt closer in the shared embedding space,

$$\mathcal{L}_{ali} = -\log \frac{\exp(\text{sim}(F_v(x_i), F_t(p_i))/\tau)}{\sum_{j=1}^{C} \exp(\text{sim}(F_v(x_i), F_t(p_j))/\tau} \tag{10}$$

where p_i is the class prompt of x_i, p_j is the prompt of classes that x_i does not belong to, τ is a temperature scaling factor.

Finally, the total loss for SeeCLIP is defined as,

$$\mathcal{L} = \mathcal{L}_{ali} + \alpha \mathcal{L}_{rep} + \beta \mathcal{L}_{coh} + \gamma \mathcal{L}_{reg} \tag{11}$$

where α, β and γ are hyper-parameters.

During training, the visual and textual encoders of CLIP are frozen to retain pre-trained semantic stability. Optimization is performed only on the learnable query vectors, the semantic/domain token projection layers, and the unknown prompt embeddings. After the diffusion generation, the end-to-end training procedure utilizes both known-class and synthesized pseudo-unknown samples, alternating between two phases: i) the alignment phase, which aligns known visual embeddings with enhanced prompts, and ii) the repulsion phase, which pushes the unknown prompt away from known features while preventing excessive semantic drift. In inference, samples are predicted by similarity with enhanced prompts, including both known and unknown class prompts, as shown in Algorithm 1.

Algorithm 1: Training of SeeCLIP

Require: Source domains $\{\mathcal{D}_k\}_{k=1}^{M}$, frozen CLIP encoders
Ensure: Semantic tokens and projections

Initialize $\{q^{(k)}\}_{k=1}^{K}$, $\Phi(\cdot)$, $\{\Psi_k(\cdot)\}_{k=1}^{K}$, and $\{v_i\}_{i=1}^{m}$
for each training iteration **do**
  for each domain $\mathcal{D}_k$ **do**
    $v_{\text{dom}}^{(k)} \leftarrow \frac{1}{|F_k|}\sum_{f_i \in F_k} f_i$
    for each attention head k **do**
      $\omega_i^{(k)} \leftarrow \frac{\exp(q^{(k)} \cdot f_i)}{\sum_{j=1}^{N} \exp(q^{(k)} \cdot f_j)}$ as in Eq. (1)
      $v_{\text{sem}}^{(k)} \leftarrow \sum_{i=1}^{N} \omega_i^{(k)} \cdot f_i$ as in Eq. (2)
    end for
  end for
  $p_c \leftarrow [\Phi(v_{\text{dom}})], [\Psi_1(v_{\text{sem}}^{(1)}), \ldots, \Psi_K(v_{\text{sem}}^{(K)})], [\textit{classname}]$
  $p_{\text{unk}} \leftarrow [\Phi(v_{\text{dom}})], [v_1, \ldots, v_m], [\textit{unknown}]$
  $\widetilde{v_{\text{sem}}^{(k)}} \leftarrow v_{\text{sem}}^{(k)} + \epsilon_k, \epsilon_k \sim \mathcal{N}(0, \sigma^2 I)$ as in Eq. (8)
  Generate pseudo-unknowns via diffusion with $[\widetilde{v_{\text{sem}}^{(k)}}]$
  $\mathcal{L} \leftarrow \mathcal{L}_{ali} + \alpha\mathcal{L}_{rep} + \beta\mathcal{L}_{coh} + \gamma\mathcal{L}_{reg}$ as in Eq. (11)
  Update parameters via gradient descent
end for

Theoretical Analysis

Theorem 1 (OSDG error bound) Let $\gamma := \min_{\pi \in \Delta_M} d_{\mathcal{H}}\left(\mathcal{P}_{\mathcal{X}}^{t}, \sum_{i=1}^{M} \pi_i \mathcal{P}_{\mathcal{X}}^{i}\right)$ with minimizer π^* be the distance of $\mathcal{P}_{\mathcal{X}}^{t}$ from the convex hull of sources Λ, and $\mathcal{P}_{\mathcal{X}}^{*} := \sum_{i=1}^{M} \pi_i^* \mathcal{P}_{\mathcal{X}}^{i}$ be the best approximator within Λ. $\rho := \sup_{\mathcal{P}_{\mathcal{X}'}, \mathcal{P}_{\mathcal{X}''} \in \Lambda} d_H(\mathcal{P}_{\mathcal{X}'}, \mathcal{P}_{\mathcal{X}''})$ is the diameter of Λ. $\pi^{unk} = Pr(\mathcal{Y}^t \in \mathcal{Y}^u)$ denotes the prior probabilities of unknown class in the target domain. Then for any hypothesis $h \in \mathcal{H}$, the target risk $\mathcal{R}^t(h)$ is,

$$\mathcal{R}^t(h) \leq \sum_{i=1}^{M} \pi_i^* \mathcal{R}^i(h) + \frac{\gamma + \rho}{2} + \lambda_{\mathcal{H}(P_{\mathcal{X}}^t, P_{\mathcal{X}}^*)} + \pi^{unk} \cdot \mathcal{R}^{\text{OS}}(h) \tag{12}$$

where $\mathcal{R}^i(h)$ is the risk of the i-th source domain. $\lambda_{\mathcal{H}(\mathcal{P}_{\mathcal{X}}^t, \mathcal{P}_{\mathcal{X}}^*)}$ is the ideal joint risk across the target domain and the domain with the best approximator distribution $\mathcal{P}_{\mathcal{X}}^*$. $\mathcal{R}^{\text{OS}}(h) = \mathbf{E}_{(x,y)\sim\mathcal{D}^t \cap \mathcal{Y}^u}\mathbf{I}(h(x) \in \mathcal{Y}^s)$ is the open-space risk, which represents the risk of misclassifying unknown-class samples in target domain to known classes.

Lemma 1 Let $dis(c, d)$ and $dis_{\text{sem}}(c, d)$ denote the discrepancy between original and enhanced prompts of the c-th and d-th classes, respectively, $\forall c, d \in \{1, ..., C\}$, $c \neq d$, then it holds that $dis_{\text{sem}}(c, d) > dis(c, d)$.

Remark 1. From Lemma 1, by introducing fine-grained semantics, discrepancy between class prompts will be enlarged. Consequently, the inter-class discrimination will be enhanced, which helps reduce the weighted source structural risks $\sum_{i=1}^{M} \pi_i^* \mathcal{R}^i(h)$ in $\mathcal{R}^t(h)$.

Remark 2. The high-similarity pseudo-unknowns force model to learn core features of known classes, reducing its over-generalization to unknown classes. They also help compress the feature space of known classes, thereby help lower the open-space risk $\mathcal{R}^{\text{OS}}(h)$. The proof can be found in the Appendix. SeeCLIP addresses both risks, thus effectively lowers the generalization risk of OSDG.

Experiments

Experimental Setups

Datasets. We evaluate SeeCLIP on five benchmark datasets: Office-Home, PACS, VLCS, Mini-DomainNet, and Multi-Dataset, following standard known-novel class splits. Office-Home contains 65 categories across four domains. PACS comprises 7 categories with four domains. VLCS includes 5 categories across four domains. Mini-DomainNet contains 126 categories with four domains. Multi-Dataset combines samples from Office-Home, PACS, and VLCS. For each dataset, we follow the leave-one-domain-out protocol, where one domain is selected as the target and the remaining domains are used as sources for training.

Evaluation Metrics. Following standard protocols in OSDG, we evaluate with two key metrics: 1) top-1 accuracy (ACC) for closed-set classes, and 2) H-score for unknown class detection, which represents the harmonic mean of known and unknown accuracies.

Compared Methods. We compare SeeCLIP with two broad families of baselines: 1) Traditional OSDG/OSR methods built on ResNet-18, with a confidence-threshold rule to reject unknowns: Cumix, MixStyle, DAML, and MEDIC. 2) CLIP-based models, evaluated under the same leave-one-domain-out protocol. We also cite three strong closed-set DG methods for context: SWAD, EoA, and DandelionNet.

Architecture Details. For all experiments, we implement SeeCLIP on the CLIP ViT-B/32 architecture, with Transformer serving as the textual encoder. During training, we freeze the parameters of both vision and text encoders, and optimize only the learnable components specific to SeeCLIP. The number of semantic token heads K is set to 4, and the prompt tokens number M for unknown class is set to 3. We employ the Stable Diffusion v1.5 with 50 denoising steps and a perturbation standard deviation σ of 0.2.

Training and Evaluation. We train for 10 epochs using the AdamW optimizer with a learning rate of 1e-4. Batch sizes are set per dataset, i.e., 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. The textual prompt context length is set to 4. We set $\alpha = 0.5, \beta = 0.3,$ and $\gamma = 0.1$. The margin δ is set to 0.2, and τ is set to 0.07.

Performance Comparison

OSDG Setting. Table 1 presents comparison results across five benchmark datasets under the OSDG setting.

Better Performance among Compared Methods. As shown in Table 1, CNN-based methods show limited effectiveness in OSDG. Standard CLIP-based approaches perform better, but prompt learning methods without explicit unknown handling yield poor H-scores despite reasonable accuracy. SeeCLIP outperforms all baselines across metrics, validating our proposal. It achieves consistent gains across datasets and exhibits robustness to diverse domain shifts. On

Methods	PACS		VLCS		Office Home		Multi-Dataset		Mini-DomainNet		Average	
	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score
CNN-based												
Cumix	57.85	41.05	52.46	50.11	51.67	49.40	42.18	46.91	50.27	39.16	52.09	46.81
MixStyle	63.35	48.30	52.30	50.61	53.52	49.53	42.18	46.91	50.43	40.25	53.67	48.66
DAML	65.49	51.88	53.53	51.59	56.45	53.34	46.61	51.71	52.81	43.63	55.73	51.29
MEDIC	89.81	83.03	57.28	55.73	60.26	57.91	50.74	53.13	55.29	45.71	66.11	60.30
CLIP-based												
CLIP	95.16	76.77	91.84	72.94	81.43	63.62	77.88	72.19	84.50	68.94	84.65	69.40
CLIP+OpenMax	93.45	79.13	92.09	73.67	81.00	61.54	78.34	73.26	81.89	69.40	83.95	69.96
CLIP+OSDA	92.62	75.40	90.21	70.89	82.58	67.35	74.45	75.22	82.00	73.62	83.73	71.36
CoOp	78.77	26.87	92.02	39.26	73.85	36.26	66.03	44.34	61.13	68.34	71.72	41.65
CoCoOp	85.76	32.93	89.47	37.01	75.38	34.38	64.84	47.57	60.63	56.30	71.48	40.28
MaPLe	93.97	48.47	89.70	43.33	79.47	33.06	69.34	62.20	74.67	60.57	79.62	48.58
LASP	88.45	30.37	90.67	39.41	76.13	34.52	66.78	50.22	62.34	61.56	74.21	41.89
PromptSRC	94.53	43.32	90.13	42.78	80.21	36.40	65.51	59.45	73.60	62.56	79.89	48.13
CLIPN	96.24	45.00	84.82	50.72	84.55	42.83	77.16	62.60	77.38	66.92	83.64	52.27
STyLIP	95.36	50.74	90.75	65.66	84.73	60.97	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	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	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	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	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	84.60	90.00	95.68	94.48	94.23	90.84
SeeCLIP	99.90	99.97	98.30	89.49	99.50	98.97	88.70	92.80	98.87	97.06	**97.05**	**95.66**

Table 1: Comparative results under the OSDG setting over leave-one-domain-out combinations. *SD* denotes *stable diffusion*.

PACS with artistic style variations, it delivers near-perfect results. Its 0.27% H-score improvement over ODG-CLIP seems modest yet meaningful given the high baseline. On VLCS, it outperforms ODG-CLIP by 2.59% and 2.96% respectively, highlighting superior cross-domain generalization. *Effectiveness in Fine-grained Recognition.* SeeCLIP exhibits its most significant gains on Office-Home, outperforming ODG-CLIP by 1.18% and 2.89%, respectively. Given the challenges posed by fine-grained object categories across diverse domains for unknown class detection, these improvements are particularly notable. The results validate the our effectiveness in capturing subtle semantic differences, critical for distinguishing similar classes.

DG Setting. We also evaluate SeeCLIP under the standard DG setting. As shown in Table 2, SeeCLIP shows exceptional performance across all benchmarks. It significantly outperforms traditional CNN-based methods and maintains substantial performance gains when compared against CLIP-based methods. Most notably, SeeCLIP achieves superior performance compared to the previous state-of-the-art method ODG-CLIP, with improvements of 0.50% in accuracy, demonstrating its effectiveness in closed-set setting.

Methods	**PACS**	**VLCS**	**O. H.**	**M. DNet**	**Avg.**
CNN-based					
SWAD	88.10	79.10	70.60	–	79.27
EoA	88.60	79.10	72.50	–	80.07
DandelionNet	89.20	81.60	70.40	–	80.40
CLIP-based					
CLIP	94.89	82.14	78.40	78.73	83.54
CoOp	97.11	83.34	81.33	72.30	83.52
CoCoOp	96.54	85.02	81.05	71.51	83.53
MaPLe	97.72	86.75	83.52	73.87	85.47
LASP	97.02	87.25	84.13	70.67	84.77
PromptSRC	98.02	86.34	83.89	76.10	86.09
StyLIP	98.17	87.21	85.94	80.43	87.94
ODG-CLIP	99.83	95.74	96.91	96.65	97.28
SeeCLIP	**99.89**	**96.52**	**97.43**	**97.28**	**97.78**

Table 2: Performance comparison in standard DG setting.

Ablation Analysis

Analysis for Components. The individual components in SeeCLIP reveal distinct contributions relative to the baseline, as shown in Table 3. Semantic-Aware Prompt Enhancement(SAPE) drives the largest improvements, boosting accuracy by 6.93% and H-score by 16.36%, validating that fine-grained semantic tokens enhance the discrimination of similar categories. Semantic-Guided Diffusion Generation(SGDG) also yields meaningful gains, increasing accuracy by 4.13% and H-score by 9.67%, by generating samples that help sharpen the decision boundaries. Meanwhile, Duplex Contrastive Learning(DCL) achieves notable improvements, with accuracy up 5.54% and H-score up 13.30% through effective feature space separation.

Analysis for Losses. Table 4 summarizes the ablation results for each loss component. The alignment loss contributes most significantly: its removal reduces accuracy by 2.89% and H-score by 3.15%, confirming its critical role. The repulsion loss is also essential, with its exclusion decreasing

Methods	PACS		O. H.		M. Data		M. DNet	
	Acc	H	Acc	H	Acc	H	Acc	H
Baseline	95.16	76.77	81.43	63.62	77.88	72.19	84.50	68.94
+ SAPE	97.34	88.42	91.67	84.73	84.23	86.51	93.45	87.29
+ SGDG	96.28	82.15	87.92	76.84	81.56	79.67	89.73	81.52
+ DCL	96.89	85.67	89.45	81.92	83.12	82.34	91.67	84.78
+ SAPE & SGDG	98.45	92.78	95.23	91.45	86.78	89.67	96.34	93.12
+ SAPE & DCL	98.89	94.56	96.78	93.82	87.45	90.89	97.23	94.67
+ SGDG & DCL	97.67	91.23	93.56	88.34	85.34	87.78	94.78	90.45
Full (SeeCLIP)	**99.90**	**99.97**	**99.50**	**98.97**	**88.70**	**92.80**	**98.87**	**97.06**

Table 3: Ablation study of individual modules in SeeCLIP.

Configuration	PACS		O. H.		M. Data		M. DNet	
	Acc	H	Acc	H	Acc	H	Acc	H
Full (SeeCLIP)	**99.9**	**99.7**	**99.5**	**99.0**	**88.7**	**92.8**	**98.9**	**97.1**
w/o $\mathcal{L}_{\text{ali}}$	97.8	97.3	97.5	96.3	85.0	89.0	95.1	93.2
w/o $\mathcal{L}_{\text{rep}}$	98.6	98.0	98.3	97.4	86.9	91.5	97.0	95.5
w/o $\mathcal{L}_{\text{coh}}$	99.1	98.9	99.0	98.7	88.0	92.0	98.3	96.6
w/o $\mathcal{L}_{\text{reg}}$	99.5	99.3	99.3	98.9	88.4	92.6	98.7	96.9

Table 4: Ablation analysis of loss functions in SeeCLIP.

accuracy by 1.54% and H-score by 1.60%. The cohesion loss, though individually modest, is crucial for semantic coherence. Its removal lowers accuracy by 0.64% and H-score by 0.65%. Besides, the regularization loss also provides consistent improvement.

Image Generation Analysis

We compare the unknown generation of SeeCLIP against three established methods in Table 5, from which SeeCLIP with SGDG achieves the best performance. In Figure 2, both Cumix and Open-GAN generate semantically ambiguous pseudo-open images, ODG-CLIP generates unknowns far from the known samples, thus may lead to overfitting to known classes. SeeCLIP adopts semantic token perturbation with Gaussian noise and dual-prompt conditioning to generate pseudo-unknowns. The pseudo-unknowns from SeeCLIP are similar to known classes, yet exhibit key semantic discrepancies. It helps generate a clear boundary that accounts for both risks of known and unknown categories.

SeeCLIP with	PACS		O.H.		M.Data		M.DNet	
	Acc	H	Acc	H	Acc	H	Acc	H
OpenGAN	96.25	92.18	92.34	91.05	81.15	81.23	93.42	90.58
Cumix	96.89	94.26	96.18	92.85	82.34	86.74	92.85	91.84
ODG-CLIP	99.64	99.78	98.85	96.74	86.23	91.15	96.42	95.31
SGDG	**99.90**	**99.97**	**99.50**	**98.97**	**88.70**	**92.80**	**98.87**	**97.06**

Table 5: Comparison of pseudo-open sample generation.

Semantic Attention Visualization

We visualize the learned attention distributions in Figure 3, each of the K=4 attention heads automatically focuses on distinct discriminative regions without part-level supervision. For the dog, attention heads concentrate on ears, nose, and mouth. For the horse, heads capture ears, face, front leg, and tail. For the elephant, attention focuses on ear, tusk, trunk, and hind leg. These visualizations demonstrate that SAPE successfully extracts category-specific fine-grained semantic features, enabling the model to distinguish semantically similar categories and handle hard unknowns in OSDG scenarios.

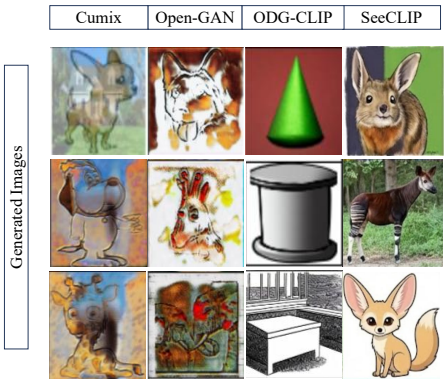

Figure 2: Comparison of unknown sample generation methods w.r.t. known classes over PACS. SeeCLIP generates semantically coherent pseudo-unknowns that are globally similar yet locally distinctive from known classes.

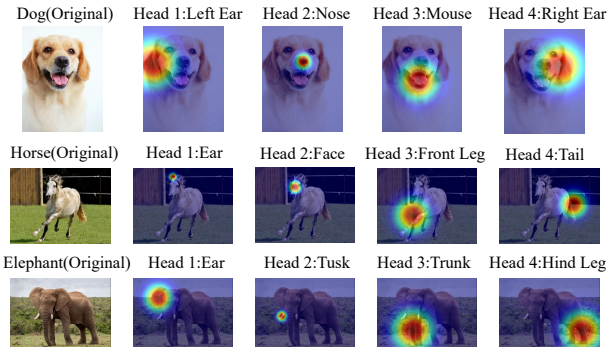

Figure 3: Visualization of multi-head semantic attention across different categories. Each row represents a distinct category (Dog, Horse, Elephant), with the leftmost column showing the original image. The four attention heads (columns 2-5) focus on discriminative semantic regions.

Conclusions

This paper presents SeeCLIP, a semantic-enhanced framework for fine-grained open-set domain generalization. Integrating semantic-aware prompt enhancement, semantic-guided diffusion generation, and duplex contrastive learning, it effectively distinguishes semantically similar unknowns. Extensive experiments validate consistent gains over SOTA methods, establishing a semantic-prioritized paradigm for open-set recognition under distribution drift.

Acknowledgments

This work was supported by the National Natural Science Foundation of China No. 62576174.

References

Bele, P.; Bundele, V.; Bhattacharya, A.; Jha, A.; Roig, G.; and Banerjee, B. 2024. Learning class and domain augmentations for single-source open-domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1816–1826.

Bendale, A.; and Boult, T. E. 2016. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1563–1572.

Bi, Q.; Yi, J.; Zhan, H.; Ji, W.; and Xia, G.-S. 2025. Learning fine-grained domain generalization via hyperbolic state space hallucination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1853–1861.

Bose, S.; Jha, A.; Kandala, H.; and Banerjee, B. 2023. Beyond boundaries: A novel data-augmentation discourse for open domain generalization. *Transactions on Machine Learning Research*.

Cherti, M.; Beaumont, R.; Wightman, R.; Kalantidis, Y.; Adeli, E.; Ranzato, M.; Caron, M.; and Alayrac, J.-B. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2818–2829.

Gupta, D.; Singha, M.; Rongali, S. B.; Jha, A.; Khan, M. H.; Banerjee, B.; et al. 2025. OSLoPrompt: Bridging Low-Supervision Challenges and Open-Set Domain Generalization in CLIP. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10110–10120.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PmLR.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19113–19122.

Kong, S.; and Ramanan, D. 2021. Opengan: Open-set recognition via open data generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 813–822.

Lang, N.; Snæbjarnarson, V.; Cole, E.; Mac Aodha, O.; Igel, C.; and Belongie, S. 2024. From coarse to fine-grained open-set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17804–17814.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, 5542–5550.

Panareda Busto, P.; and Gall, J. 2017. Open set domain adaptation. In *Proceedings of the IEEE international conference on computer vision*, 754–763.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rakshit, S.; Bandyopadhyay, H.; Bharambe, P.; and Patel, V. M. 2022. Open-set domain adaptation under few source-domain labeled samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4029–4038.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shu, Y.; Cao, Z.; Wang, C.; Wang, J.; and Long, M. 2021. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9624–9633.

Singha, M.; Jha, A.; Bose, S.; Nair, A.; Abdar, M.; and Banerjee, B. 2024. 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*, 13309–13319.

Sun, J.; and Dong, Q. 2023. A survey on open-set image recognition. *arXiv preprint arXiv:2312.15571*.

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2022. Generalized Category Discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7492–7501.

Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; and Yu, P. S. 2022a. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8): 8052–8072.

Wang, X.; Zhang, J.; Qi, L.; and Shi, Y. 2023. Generalizable decision boundaries: Dualistic meta-learning for open set domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11564–11573.

Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.-Y.; Ren, X.; Su, G.; Perot, V.; Dy, J.; et al. 2022b. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, 631–648. Springer.

Yang, Z.; Yue, J.; Ghamisi, P.; Zhang, S.; Ma, J.; and Fang, L. 2024. Open set recognition in real world. *International Journal of Computer Vision*, 132(8): 3208–3231.

Yu, W.; Chen, D.; Wang, Q.; and Hu, Q. 2024. Fine-Grained Domain Generalization with Feature Structuralization. *arXiv preprint arXiv:2406.09166*.

Zhang, W.; Zhang, B.; Teng, Z.; Luo, W.; Zou, J.; and Fan, J. 2025. Less Attention is More: Prompt Transformer for Generalized Category Discovery. In *Proceedings of the Com-*

puter Vision and Pattern Recognition Conference, 30322–30331.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, K.; Yang, Y.; Hospedales, T.; and Xiang, T. 2020. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, 561–578. Springer.