Catching the Unknown with Limited Data: Bi-Directional Prompt Tuning in CLIP for Few-Shot Open-Set Adaptation

Moloud Abdar[1]
m.abdar1987@gmail.com
Md Mehedi Hasan[2]
mmhasan@deakin.edu.au
Biplab Banerjee[3]
getbiplab@gmail.com
Abbas Khosravi[2]
abbas.khosravi@deakin.edu.au
Pietro Lio[4]
pl219@cam.ac.uk

[1] CHIRP, Child Health Research Centre,
The University of Queensland,
Brisbane, Australia

[2] Institute for Intelligent Systems,
Deakin University,
Geelong, Australia

[3] Centre of Studies in Resources Engineering,
Indian Institute of Technology Bombay,
Mumbai, India

[4] Department of Computer Science and Technology,
University of Cambridge,
Cambridge, UK

Abstract

Domain-adaptive few-shot open-set (DAFSOS) learning aims to transfer knowledge from a label-rich source domain to a target domain where only a few labeled samples are available. Both domains contain disjoint sets of classes, and at inference time, models must accurately classify unlabeled samples from novel target-domain classes using limited support examples, while also recognizing outliers. Existing approaches often rely on meta-learning with sophisticated loss functions to obtain domain-invariant, discriminative features, yet overlook semantic information. We propose a novel prompt-learning framework for CLIP, named BIMAP-CLIP [1], tailored to the DAFSOS setting. Its core innovation is a bi-directional, multi-modal prompt-learning mechanism that shares learnable tokens across both image and text encoders, enhancing visual–semantic alignment. We also introduce a hybrid prompt initialization strategy that integrates manual and learnable prompts to make the prompts resilient to domains and incorporates fine-grained class attributes to tackle the scarcity of training samples. In parallel, a thresholding-based method on prompt-image similarity scores is employed for outlier detection. Across multiple benchmarks, BIMAP-CLIP consistently outperforms existing counterparts.

[1] Bi-directional Multi-modal Attribute-enriched Prompting for CLIP

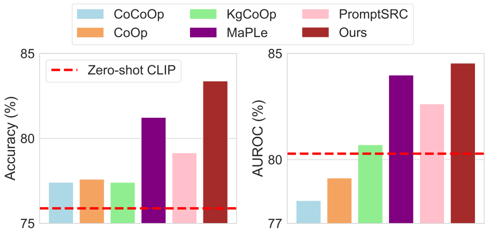


Figure 1: Performance comparison of our method with other prompt learning methods (Zero-shot CLIP, CoOp, CoCoOp, KgCoOp, MaPLe, PrompSRC). The scores are obtained by averaging over all datasets. Our method offers better class detection performance while still maintaining the ability to identify open-set samples in both domains.

1 Introduction

Domain Adaptive Few-Shot Open-Set Learning (DAFSOS) addresses the demanding scenario of adapting a trained model to a new domain where limited labeled data is available for a few known classes, while additional classes remain unknown at test time [14]. This problem combines two core challenges: (1) adjusting to domain shifts when labeled data from the source and target domains may have fundamentally different distributions, and (2) recognizing open-set unknowns, not seen during training. In practice, such scenarios arise in medical diagnostics (adapting to new patient demographics and detecting novel health conditions) or autonomous navigation (handling unseen hazards in unfamiliar environments). Due to the need for accurate classification under minimal supervision while managing unseen classes, DAFSOS must surmount significant obstacles such as domain alignment, few-shot learning constraints, and unbounded open space.

Existing paradigms like Domain Adaptive Few-Shot Learning (DAFSL) and Few-Shot Open-Set Learning (FSOSL) partially address these issues but fail to handle both simultaneously. DAFSL methods optimize for domain alignment with minimal data but lack the capacity to detect novel, unlabelled classes [17]. FSOSL approaches focus on unknown detection but are restricted to a single domain and overlook domain shifts [10]. DAFSOS-Net [14] bridges some of these gaps via adversarial alignment and prototype-based open-set recognition. However, its purely feature-based design can struggle with fine-grained open-set detection when unknown classes closely resemble known ones [2], and it relies on pseudo-open samples that may be challenging to obtain in real-world cases.

However, in DAFSOS settings, the learned prompts may overfit to the source domains while failing to generalize to novel target domains with unknown classes. Furthermore, most current multi-modal prompting relies on unidirectional text-to-vision coupling, limiting the model's capacity to dynamically reinforce representations across modalities—especially critical when unknown classes resemble known ones at a fine-grained level. Figure 1 highlights these challenges by comparing several prompt learning methods, showing that existing solutions either sacrifice detection accuracy or open-set recognition capabilities.

To overcome these drawbacks, we propose a bi-directional prompt coupling strategy enriched with large-language-model-generated class-discriminative attributes, which ensures both vision and text encoders continually inform each other. We also guard against overfitting by retaining zero-shot CLIP features through diversified manual prompts, and use CLIP's logit scores for open-set detection, as our empirical results confirm the effectiveness of maximum logit thresholds in discriminating unknown samples.

In summary, our main contributions are listed as follows: (1) We introduce bidirectional prompt coupling in multi-modal prompt learning with CLIP-based models. (2) We effectively incorporate class discriminative attributes in the text prompt in multi-modal prompt learning. (3) We also modify and adapt the challenging DAFSOS problem in prompt learning scenarios.

2 Related Work

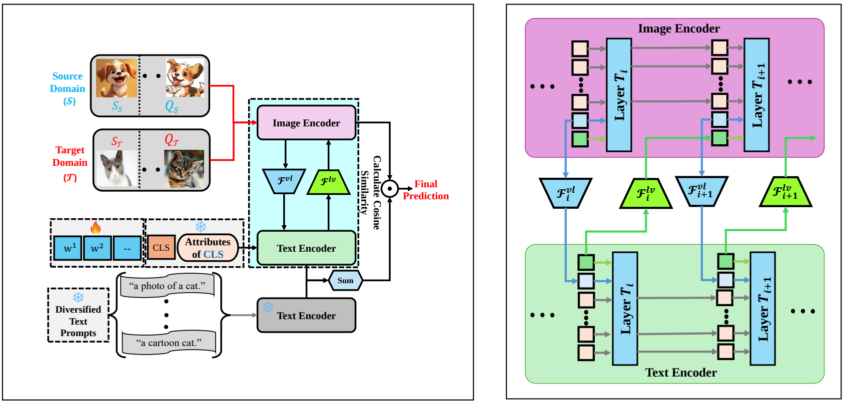
Domain Adaptive Few-Shot Open Set Recognition: Handling unknown class samples in few-shot learning (FSL) remains a major challenge. Recent methods often rely on learning explicit outlier representations. For instance, Gaussian cluster formation has been employed to model known class distributions using limited samples [9], while [5] introduced confidence optimization for outlier detection. Other notable efforts include episodic optimization of a binary closed-open discriminator [12] and generative augmentation [13].

Cross-domain FSL introduces the additional complexity of adapting to significant domain shifts between disjoint base and novel classes. Although the combination of feature extraction with domain alignment objectives has been explored [4], direct domain alignment can fail due to label shifts [1]. To address this, [18, 23] proposed methods that align domains while preserving global class discrimination, and [3] employed cross-domain mixup to produce more generalized features. Building on these advancements, [14] introduced the Domain Adaptive Few-Shot OpenSet Learning (DAFSOS) framework, which specifically tackles the issue of unknown classes in few-shot scenarios across disjoint domains.

Vision-Language Model and Prompt Learning: With the advent of vision-language models such as CLIP [16] and ALIGN [6, 19], prompt engineering has emerged as a compelling alternative to full fine-tuning. One of the earliest works in automated prompt learning is Context Optimization (CoOp) [25], which replaces handcrafted prompts (e.g., "a photo of a <class>") with learnable context vectors. Building on this, CoCoOp (Conditional CoOp) [24] conditions prompt learning on input images, dynamically adjusting prompts based on image features to enhance generalization across unseen domains and novel classes. Kg-CoOp [22] enriches CoOp with knowledge-grounded information, while PromptSRC [8] employs self-regularization to mitigate overfitting to base classes. ProMetaR [15] uses meta-regularization for improved cross-task generalization. Cascade Prompt Learning [21] leverages a cascade structure for progressive adaptation, and STAR-Prompt [11] applies semantic residual prompts for continual learning.

MCPL [20] introduces a collaborative strategy that jointly optimizes text and image prompts for medical vision-language models. MaPLe [7] integrates both visual and textual prompts within a unified optimization framework, demonstrating its effectiveness in tasks such as domain-adaptive FSL and open-set recognition. However, within this unified approach, only the text encoder's tokens are projected to the vision encoder, potentially limiting cross-modal synergy. To address this limitation, we propose a two-way projection of tokens, termed bi-directional coupling, to facilitate richer interactions between modalities.

Additionally, prior multi-modal prompting methods often overlook the inclusion of attributes due to their inherent complexity and the challenges of integrating them effectively. However, incorporating these attributes can enrich feature representations and improve the model's ability to differentiate between classes. Hence, we improve multi-modal prompting by integrating attributes, substantially boosting the model's capacity to tackle diverse tasks



(a) (b)

Figure 2: Overall overview of the proposed BIMAP-CLIP model (as indicated by (a) at the left) with Bi-directional prompt coupling (as indicated by (b) at the right). Our method not only conditions the vision prompts on language prompts but also language prompts on the vision prompts to enhance the mutual synergy between the two modalities. It enables the flow of information in both directions, reducing information leaks by utilizing zig-zag connections.

and domains with minimal labeled data.

3 Problem Definition

We consider two distinct domains: a *source domain* $\mathcal{S}$ and a *target domain* $\mathcal{T}$. The source domain $\mathcal{S}$ is associated with a labeled dataset $\mathcal{D}_s$, containing classes $\mathcal{C}_s$, while the target domain $\mathcal{T}$ has a dataset $\mathcal{D}_t$, with classes $\mathcal{C}_t$. We assume the following:

$$P(\mathcal{S}) \neq P(\mathcal{T}), \quad \mathcal{C}_s \cap \mathcal{C}_t = \emptyset, \tag{1}$$

where $P(\mathcal{S})$ and $P(\mathcal{T})$ denote the underlying distributions of the source and target domains, respectively. This assumption implies both distinct class labels and possible domain shifts in data distribution (e.g., image style, resolution, or sensor type). Furthermore, we assume that the source domain has abundant labeled data, whereas the target domain has only a few training samples per class.

During testing, the trained model may encounter new samples from a new set of target classes $\mathcal{C}_{t'} \subset \mathcal{T}$, where $\mathcal{C}_{t'} \cap \mathcal{C}_t = \emptyset$. Throughout training and testing, we follow an N-way K-shot setting in the target domain, meaning that each of the N classes has K training samples. These target training samples form the **Support** set **S**, while the **Query** set **Q** consists of unlabeled samples from $\mathcal{C}_{t'}$, as well as potential outliers.

We propose formulation of DAFSOS as a prompt-learning problem for large vision-language models (VLMs). The primary goal is to *fine-tune learnable prompts* such that the resulting classifier (built on a large VLM) not only accurately classifies samples in **Q** either in classes from $\mathcal{C}_{t'}$ and distinguishes open-set samples from the known classes.

4 Our Method

Figure 2 demonstrates the architectural overview of our proposed model. As we expect a domain shift and very few samples in the target domain, we want the text features to depend on the features of the vision encoder. To achieve this dependency, we propose bi-directional coupling where both language and vision prompts are projected to vision and text encoders, respectively. To enhance the discriminative capability of text features, we append class discriminative attributes to text prompts. CLIP [16] shows strong performance in a dynamic environment, including open-set recognition tasks. To retain the generalization capability of CLIP, we utilize fixed prompts with a pre-trained CLIP model. Finally, we use maximum logit scores directly generated by the model to calculate AUROC scores. We argue that maximum logit scores should be lower for open-set samples compared to known samples under well-learned prompts.

4.1 Bi-directional Prompt Coupling

We argue that uni-directional projection of prompts from language to vision encoder might limit the scope of information sharing between the encoders. To further enhance coupling between the encoders, we propose bi-directional coupling where vision prompts are projected to the language encoder, and language prompts are projected to the vision encoders. In uni-directional coupling, the flow of information is inherently asymmetric, favoring one modality over the other. This asymmetry could lead to suboptimal adaptation, especially in open-set scenarios, where both modalities need to evolve simultaneously to handle unknown data effectively. Bidirectional coupling allows the vision and language encoders to share information symmetrically, fostering a more cohesive understanding of incoming data of both modalities, thus achieving better open-set recognition capability.

Following [7], we incorporate learnable tokens into the language branch to dynamically adjust the model's representations. Specifically, a set of learnable tokens, denoted as $W_0 = [w_0^1, w_0^2, .., w_0^a] \in \mathbb{R}^{a \times d_t}$ is introduced into the language branch. These tokens are concatenated with the fixed input tokens $E_0 = [e_0^1, e_0^2, .., e_0^N]$. The resulting embedding is then processed by a series of transformer layers. To refine the model's contextual understanding further, additional learnable tokens are introduced at deeper layers (up to J^{th} layer) of the transformer enabling it to capture progressively complex language patterns.

$$[\_, E_i] = S_i([\tilde{W}_{i-1}, E_{i-1}]), i = 1, 2, ..J. \tag{2}$$

Here, S_i denotes i^{th} layer's function through which the embeddings get updated, $[\cdot, \cdot]$ indicates the concatenation operation, and $\tilde{W}$ consists of tokens from both text and vision encoders.

$$\tilde{W}_i = [W_i, \mathcal{F}_i^{vl}(Q_i)], i = 0, 1, ..J-1, \tag{3}$$

where $\mathcal{F}^{vl}$ denotes the coupling function that essentially maps d_p-dimensional vision prompts to d_t-dimensional language prompts and Q stands for the learnable tokens in vision encoder. After the J^{th} layer, no new tokens are included, and each layer processes the previous layer's prompts.

$$[\tilde{W}_i, E_i] = S_i([\tilde{W}_{i-1}, E_{i-1}]), i = J+1, ..K. \tag{4}$$

The final text representation U is produced as:

$$U = \text{TextProj}(e_K^N). \tag{5}$$

Similarly, a set of b learnable tokens, denoted as $Q_0 = [q_0^1, q_0^2, .., q_0^b] \in \mathbb{R}^{b \times d_p}$ are incorporated into the visual encoder to enhance the model's ability to understand and represent visual content. These learnable tokens are introduced to the image encoder at various depths up to a specified depth $J+1$. At each layer, the tokens are processed along with the input image features to generate refined representations.

Unlike [7], we do not use the same layer coupling for both vision and text prompts. Though, it seems to be an obvious way to achieve bi-directional coupling. However, the same layer coupling might inhibit the information flowing through the encoder layers. Furthermore, it might cause information leaks between vision and text prompts within the same layer. To prevent these phenomena, we couple the text prompt from i^{th} layer in text encoder with the vision prompt from $(i+1)^{th}$ layer in the vision encoder. For each layer i, the input to transformer block T_i consists of patch embeddings P_{i-1} augmented with learnable class tokens c_{i-1}. Thus, the deep vision prompting can be expressed as follows:

$$[c_i, P_i, \_] = T_i([c_{i-1}, P_{i-1}, \tilde{Q}_{i-1}]), i = 1, 2, .., J+1. \tag{6}$$

$$[c_i, P_i, \tilde{Q}_i] = T_i([c_{i-1}, P_{i-1}, \tilde{Q}_{i-1}]), i = J+2, .., K. \tag{7}$$

Here, $\tilde{Q}$ is estimated as:

$$\tilde{Q}_i = \begin{cases} Q_i & for \quad i = 0 \\ [Q_i, \mathcal{F}_i^{lv}(W_{i-1})] & for \quad 1 \leq i \leq J-1 \\ \mathcal{F}_i^{lv}(W_{i-1}) & for \quad i = J \end{cases} \tag{8}$$

where the coupling function $\mathcal{F}^{lv}$ maps d_t-dimensional language prompts to d_p-dimensional vision prompts.

This two-way flow of learnable tokens through learnable projection layers helps the mutual propagation of the gradients in both directions. It should be noted that both $\mathcal{F}^{lv}$ and $\mathcal{F}^{vl}$ consist of a single linear layer.

4.2 Incorporating Fixed Prompts

The pre-trained CLIP models are good at capturing generalizable zero-shot knowledge, but they lack task-specific information. Conversely, learned prompts excel at task adaptation, but may lose generalizability. To balance this, the authors of [8] encourage the learned prompts to align with the representations generated by the frozen CLIP model. We also want text features to revolve around the pre-trained clip text features. Instead of maximizing the similarity between the learned and fixed text features, we adopt a weighted aggregation method. To be precise, our approach is more like a better initialization rather than a regularization approach in [8]. For a learnable prompt, we obtain textual feature U and pre-trained textual feature $\hat{U}$, then the final textual feature:

$$U_f = \alpha U + (1-\alpha)\hat{U}, \tag{9}$$

where α is a learnable variable that controls the relative significance of U and $\hat{U}$ in the textual feature.

Furthermore, we aggregate the textual tokens of the fixed prompts of the pre-trained CLIP to the learnable tokens before the first vision encoder layer. We again follow weighted aggregation here with the weight being a learnable variable. Features from pre-trained CLIP using fixed prompts helps to reduce overfitting even with limited labeled data and leverage CLIP's pre-learned ability to generalize to new, unseen data.

Architecture	Method	Office-home Real World-Clipart ACC	Office-home Real World-Clipart AUROC	Domainnet Clipart-Painting ACC	Domainnet Clipart-Painting AUROC	Domainnet Real-Clipart ACC	Domainnet Real-Clipart AUROC	Domainnet Real-Painting ACC	Domainnet Real-Painting AUROC	Mini-imagenet to CUB Mini-imagenet-CUB ACC	Mini-imagenet to CUB Mini-imagenet-CUB AUROC
ViT-B/32	DAFOS-NET	51.79±0.67	50.54±0.54	56.21±0.21	57.26±0.49	52.42±0.37	54.84±0.36	55.06±0.34	54.16±0.33	49.25±0.62	50.17±0.28
	CLIP	75.29±0.2	78.63±0.1	77.15±0.13	81.44±0.09	84.92±0.06	85.54±0.04	81.45±0.13	83.8±0.07	60.61±0.54	72.0±0.21
	CoCoOp	73.5±0.28	77.0±0.08	80.27±0.28	78.82±0.19	86.94±0.04	85.21±0.12	83.66±0.15	80.1±0.15	62.68±0.12	69.22±0.27
	CoOp	70.34±0.41	74.29±0.08	79.1±0.14	80.74±0.09	86.27±0.06	85.78±0.09	83.56±0.15	83.97±0.06	68.7±0.24	70.88±0.18
	KgCoOp	71.57±0.36	78.84±0.1	78.85±0.29	81.15±0.08	85.91±0.09	87.87±0.06	82.99±0.29	83.97±0.04	67.72±0.39	71.67±0.21
	MaPLe	77.71±0.23	**81.45±0.07**	81.24±0.18	84.3±0.09	88.88±0.06	89.36±0.13	85.31±0.1	**87.53±0.1**	73.02±0.3	77.25±0.39
	PromptSRC	72.95±0.37	80.17±0.03	81.46±0.15	84.19±0.1	88.2±0.07	88.8±0.06	85.37±0.07	86.22±0.04	67.77±0.44	73.74±0.28
	Ours	**79.26±0.23**	81.34±0.08	**82.22±0.08**	**84.67±0.13**	**90.65±0.04**	**90.77±0.05**	**86.55±0.11**	87.38±0.1	**78.19±0.33**	**78.54±0.53**

Table 1: Main result. ACC and AUROC comparison on multiple datasets demonstrate that our method achieves consistent improvements compared to previous methods. The **bold** and green colored numbers indicate the highest and the second highest scores, respectively.

Architecture	Method	Domainnet Clipart-Painting ACC	Domainnet Clipart-Painting AUROC	Domainnet Real-Painting ACC	Domainnet Real-Painting AUROC
ViT-B/32	with prompt sharing	81.72±0.12	85.19±0.12	85.1±0.12	87.78±0.14
	w/o prompt sharing	79.1±0.14	80.74±0.09	83.56±0.15	83.97±0.06

Table 2: Ablation study on the deep prompt sharing. Both ACC and AUROC experience significant boosts with deep prompt coupling.

4.3 Class Discriminating Attributes

To further enhance the discriminative capability of our model, we introduce multiple class attributes to the text prompts of the textual encoder. The intuition is that these attributes encompass more intricate information than a simple class name, which will help distinguish between different classes more effectively. For example, instead of using only the class name 'goose', we append the following characterizing information 'long neck, webbed feet, broad body' to the class name. Furthermore, attributes like "shape" or "structure" can provide robust anchors for cross-domain alignment, as they are often invariant to domain shifts, ensuring that the model's performance remains consistent even under significant domain shifts. It should be noted that we resort to using a publicly available large-language model (GPT-3.5) to generate these class attributes. For more clarity, we provide a list of the generated attributes in the Supplementary Material.

5 Experiments

In this section, we present a quantitative evaluation of our proposed method across 3 widely used image classification tasks with 5 different domain adaptive setups. We also conducted an ablation study on the different components of our method. Details about experimental setup and datasets are provided in the supplementary.

5.1 Main Results

Table 1 presents the classification accuracy (ACC and AUROC) achieved by various methods across different datasets and tasks. Each method employs a ViT-B/32 backbone for a consistent basis of comparison. First, we notice that zero-shot CLIP offers a strong baseline in terms of AUROC though the accuracy is poor due to fixed text prompts being unable to generalize well in different domains. As expected, all prompt learning approaches outperform zero-shot CLIP in terms of ACC. We notice that CoOp and zero-shot CLIP outperform

Architecture	Method	Office-Home Real World-Clipart ACC	Office-Home Real World-Clipart AUROC	Mini-imagenet to CUB Mini-imagenet-CUB ACC	Mini-imagenet to CUB Mini-imagenet-CUB AUROC
ViT-B/32	with attributes	78.37±0.29	80.01±0.04	75.16±0.44	77.84±0.29
	w/o attributes	78.23±0.21	81.04±0.08	74.82±0.35	77.55±0.35

Table 3: Ablation study on the attributes of the text prompts.

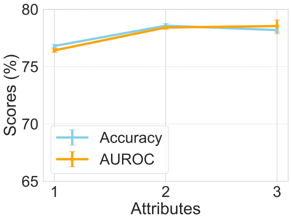


Figure 3: Performance varying the number of the attributes. The experiment is done with the MiniImageNet/CUB dataset.

CoCoOp in AUROC. We assume the decrease in performance is due to the fact that the output of the vision encoder is added to the learnable text prompt. The meta-net used to project the visual features as the token might not generalize well in both domains as we have very limited samples in the target domain. Finally, it is noticeable that MaPLe proves to be the best baseline in both ACC and AUROC confirming the strength of the multi-modal prompt learning. In our method, we further enhance the multi-modal prompt learning with bidirectional deep prompt sharing and textual class attributes.

In summary, our method consistently surpasses state-of-the-art baselines, including PromptSRC and MaPLe, across all metrics (Accuracy and AUROC) and tasks. This indicates the efficacy of our approach in addressing the challenges of domain adaptive FSL with open-set recognition. The significant gains in AUROC further demonstrate the model's robustness in detecting and managing open-set samples, a critical aspect of real-world applications.

5.2 Ablation Study

In this section, we show the ablation results for each component of our method. The pivotal component of our method is the bi-directional deep prompt coupling. Table 2 demonstrates the effectiveness of bi-directional coupling. The drastic boosts in the AUROC scores tell that bi-directional prompt sharing increases the model's ability to identify the open set, even with very low samples per class during fine-tuning. It should be noted that ACC also increases significantly.

We present the ablation report on the attributes of the text prompts in Table 3. The increase in ACC scores confirms that class discriminative attributes help to identify class samples even under domain shift. We also try to figure out how many attributes can effectively contribute to the increase in performance. For example, a class 'frying pan' has three attributes 'flat circular base', 'short handle', and 'raised edges'. We determine whether all three attributes are required to increase the performance of the model. Figure 3 shows the ablation study on the number of attributes in the text prompts. It is noticeable that the scores do not change significantly beyond the second attribute. Finally, the last component of our method is the diverse fixed prompts. It is evident from Table 4 that the addition of fixed prompts upgrades both ACC and AUROC scores by a good margin. In summary, we conclude that each component in our method contributes to learning a better domain agnostic

Architecture	Method	Domainnet			
		Real-Clipart		Real-Painting	
		ACC	AUROC	ACC	AUROC
ViT-B/32	with fixed prompts	89.68±0.06	91.55±0.06	85.75±0.13	88.51±0.11
	w/o fixed prompts	88.78±0.09	90.82±0.08	85.1±0.12	87.78±0.14

Table 4: Ablation study on the fixed prompts. Both ACC and AUROC decreases significantly with the absence of the fixed prompts.

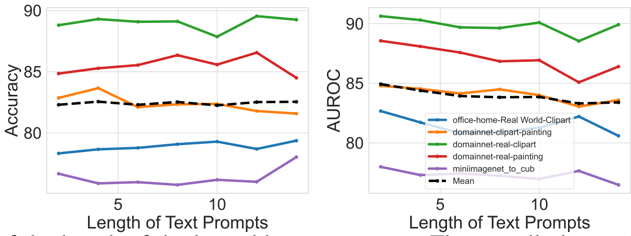


Figure 4: Effect of the length of the learnable text prompts. The overall observation is that smaller prompt lengths offer better AUROC scores.

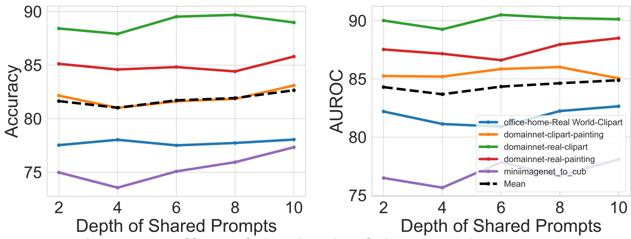


Figure 5: Effect of the depth of the shared prompts.

prompt while being able to detect the open set samples in both domains.

5.3 Prompt Length

We demonstrate the effect of different prompt lengths on the performance of our method in Figure 4. The average scores over all datasets and domain setups are indicated by the dashed black lines. We notice that overall the accuracy is hardly affected by the learnable prompt lengths. However, a slight but consistent decrease in the AUROC score is vivid with increasing prompt length, signifying a gradual inability to identify the open set samples. We argue that the scarcity of samples in the target domain makes it difficult to learn a higher number of generalizable prompts. Following [7], we set the length of the text prompt as 4.

5.4 Depth of the Shared Prompts

Finally, in Figure 5, we study the effect of the depth of the bi-directional prompt coupling. We observe an almost linear relationship between the depth and both evaluation metric (ACC and AUROC). Although at lower depth, overall performance seems to degrade, but it then increases consistently with the increase in prompt depth. However, to fairly compare the results with [7], we set the depth to 9 for all of our experiments.

6 Takeaways and Future Scope

In this paper, we introduce BIMAP-CLIP, a new model that uses a bi-directional approach along with a multi-modal prompt-learning mechanism. Through this approach, it allows for the sharing of learnable tokens between image and text encoders, strengthening visual–semantic alignment. Additionally, we propose a novel hybrid prompt initialization strategy that combines manual and learnable prompts, enhancing their resilience across domains and incorporating fine-grained class attributes to address the challenge of limited training samples. Simultaneously, we use a thresholding-based technique on prompt-image similarity scores to detect outliers. The proposed BIMAP-CLIP achieves superior performance across multiple benchmarks compared to existing methodologies. For a possible future goal, we will develop a novel uncertainty quantification technique integrated with our proposed BIMAP-CLIP model to further enhance its robustness and reliability.

References

[1] David Berthelot, Rebecca Roelofs, Kihyuk Sohn, Nicholas Carlini, and Alex Kurakin. Adamatch: A unified approach to semi-supervised learning and domain adaptation. *arXiv preprint arXiv:2106.04732*, 2021.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[3] Yuqian Fu, Yanwei Fu, and Yu-Gang Jiang. Meta-fdmixup: Cross-domain few-shot learning guided by labeled target data. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5326–5334, 2021.

[4] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[5] Minki Jeong, Seokeon Choi, and Changick Kim. Few-shot open-set recognition by transformation consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12566–12575, 2021.

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

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

[8] 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.

[9] Bo Liu, Hao Kang, Haoxiang Li, Gang Hua, and Nuno Vasconcelos. Few-shot open-set recognition using meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[10] Bo Liu, Hao Kang, Haoxiang Li, Gang Hua, and Nuno Vasconcelos. Few-shot open-set recognition using meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2020.

[11] Martin Menabue, Emanuele Frascaroli, Matteo Boschini, Enver Sangineto, Lorenzo Bonicelli, Angelo Porrello, and Simone Calderara. Semantic residual prompts for continual learning. In *European Conference on Computer Vision*, pages 1–18. Springer, 2025.

[12] Debabrata Pal, Valay Bundele, Renuka Sharma, Biplab Banerjee, and Yogananda Jeppu. Few-shot open-set recognition of hyperspectral images with outlier calibration network. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2091–2100, 2022. doi: 10.1109/WACV51458.2022.00215.

[13] Debabrata Pal, Shirsha Bose, Biplab Banerjee, and Yogananda Jeppu. Morgan: Meta-learning-based few-shot open-set recognition via generative adversarial network. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6284–6293, 2023. doi: 10.1109/WACV56688.2023.00623.

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

[15] Jinyoung Park, Juyeon Ko, and Hyunwoo J Kim. Prompt learning via meta-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26940–26950, 2024.

[16] 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.

[17] Doyen Sahoo, Hung Le, Chenghao Liu, and Steven C. H. Hoi. Meta-learning with domain adaptation for few-shot learning under domain shift, 2019. URL https://openreview.net/forum?id=ByGOuo0cYm.

[18] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. *arXiv preprint arXiv:2001.08735*, 2020.

[19] Dongsheng Wang, Miaoge Li, Xinyang Liu, MingSheng Xu, Bo Chen, and Hanwang Zhang. Tuning multi-mode token-level prompt alignment across modalities. *Advances in Neural Information Processing Systems*, 36, 2024.

[20] Pengyu Wang, Huaqi Zhang, and Yixuan Yuan. Mcpl: Multi-modal collaborative prompt learning for medical vision-language model. *IEEE Transactions on Medical Imaging*, 43(12):4224–4235, 2024. doi: 10.1109/TMI.2024.3418408.

[21] Ge Wu, Xin Zhang, Zheng Li, Zhaowei Chen, Jiajun Liang, Jian Yang, and Xiang Li. Cascade prompt learning for vision-language model adaptation. In *European Conference on Computer Vision*, pages 304–321. Springer, 2025.

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

[23] An Zhao, Mingyu Ding, Zhiwu Lu, Tao Xiang, Yulei Niu, Jiechao Guan, and Ji-Rong Wen. Domain-adaptive few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1390–1399, 2021.

[24] 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.

[25] 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.