Domain-Controlled Prompt Learning

Qinglong Cao[1,2], Zhengqin Xu[1], Yuntian Chen[2*], Chao Ma[1], Xiaokang Yang[1]

[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo
{caoql2022, fate311}@sjtu.edu.cn, ychen@eitech.edu.cn, {chaoma, xkyang}@sjtu.edu.cn

Abstract

Large pre-trained vision-language models, such as CLIP, have shown remarkable generalization capabilities across various tasks when appropriate text prompts are provided. However, adapting these models to specific domains, like remote sensing images (RSIs), medical images, etc, remains unexplored and challenging. Existing prompt learning methods often lack domain-awareness or domain-transfer mechanisms, leading to suboptimal performance due to the misinterpretation of specific images in natural image patterns. To tackle this dilemma, we proposed a Domain-Controlled Prompt Learning for the specific domains. Specifically, the large-scale specific domain foundation model (LSDM) is first introduced to provide essential specific domain knowledge. Using lightweight neural networks, we transfer this knowledge into domain biases, which control both the visual and language branches to obtain domain-adaptive prompts in a directly incorporating manner. Simultaneously, to overcome the existing overfitting challenge, we propose a novel noisy-adding strategy, without extra trainable parameters, to help the model escape the suboptimal solution in a global domain oscillation manner. Experimental results show our method achieves state-of-the-art performance in specific domain image recognition datasets. Our code is available at https://github.com/caoql98/DCPL.

Introduction

With the emergence of deep learning technology, various visual understanding tasks, including classification (Simonyan and Zisserman 2014; He et al. 2016), semantic segmentation (Cao et al. 2023b,a), and object detection (Redmon et al. 2016; Girshick 2015), have witnessed remarkable progress. However, the success of these tasks heavily relies on access to large-scale, high-quality annotated datasets (Deng et al. 2009; Lin et al. 2014), which entail significant labor and expense for each specific visual task. To tackle this practical challenge, the Contrastive Language-Image Pretraining (CLIP) (Radford et al. 2021) has been introduced, aiming to provide transferable visual features that can be leveraged across a diverse range of downstream tasks. By employing contrastive learning with extensive image-text pairs, CLIP has demonstrated exceptional zero-shot generalization capabilities.

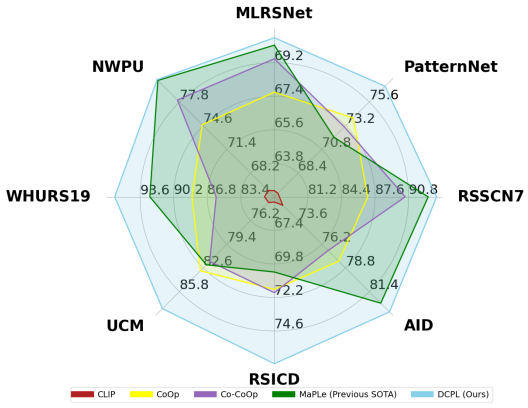

Figure 1: Using RSIs as examples. Our method achieves state-of-the-art performance on 8 different RSIs datasets.

In CLIP, visual categories are directly incorporated into carefully designed templates as prompts. Nonetheless, the creation of appropriate templates can be a time-consuming endeavor. To address this concern and drawing inspiration from prompt learning techniques, CoOp (Zhou et al. 2022b) proposes context optimization, employing learnable context vectors to enhance the zero-shot generalization performance. Following the prompt learning paradigm, numerous prompt learning algorithms have been developed for vision-language models, yielding notable advancements in zero-shot image recognition. For instance, CoCoOp (Zhou et al. 2022a) tackles the class shift problem by introducing input-conditional tokens, while MaPle adopts prompt learning for both vision and language branches to enhance the alignment between visual and linguistic representations.

Despite the progress made in prompt learning algorithms (Zhou et al. 2022b,a; Khattak et al. 2023; Wang et al. 2022b; Ge et al. 2022), they only consider the same-domain downstream task, while the adaptation problem from the natural image domain to specific domains like RSIs has rarely been considered. The domain-awareness or domain-transfer mechanisms are correspondingly been ignored. Naturally, existing prompt learning algorithms would approach

*Corresponding author

these domain-specific images with inappropriate natural image perception patterns, leading to suboptimal performance in specific domain recognition tasks.

To address this challenge and enable prompt learning to effectively model the necessary domain adaptations for specific domains like RSIs, medical images, etc, we propose a novel domain-controlled prompt learning approach. Our key idea is to generate domain-adaptive prompts for both the visual and language branches, instead of relying on existing domain-insensitive prompts, and experiments are implemented on RSIs and medical images to demonstrate the efficiency. Specifically, we introduce the newly open-sourced large-scale specific domain foundation model (LSDM) as the specific domain knowledge. By incorporating lightweight neural networks, the LSDM generates domain biases separately for the visual branch and the language branch. The domain bias for language is incorporated into the learnable context vector, while the domain bias for the visual branch is directly integrated into the image features. This approach controls the model to rightly understand the specific domain data, leading to more informed and contextually rich representations, ultimately enhancing the model's discriminative power and overall performance.

Meanwhile, CoCoOp has identified the overfitting problem caused by category shift and attempted to solve it in a conditional method. However, we tend to tackle it in a more explicit manner. To straightforwardly solve this, adopting dropout or mutation operations seems to be a plausible solution. However, these strategies only introduce randomness and variations to some extent, they are still constrained by their local-sampling nature (dropout) and point-based modifications (mutation), which means they are insufficient for escaping suboptimal solutions.

Inspired by the random sampling process in diffusion models (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021), which greatly facilitates exploration in complex spaces, we propose a novel noisy-adding strategy to handle it. This strategy induces global domain oscillation throughout the entire feature space by introducing adaptive random Gaussian noise. In contrast to local sampling and point jittering, this strategy allows for broader exploration across whole feature space, preventing model from being trapped in narrow solution regions. As shown in Figure 1, our proposed method outperforms existing prompt learning approaches across 8 diverse remote sensing image recognition datasets.

To sum up, the main contributions could be concluded as follows:

- To the best of our knowledge, we propose the first prompt learning paradigm for specific domains. By introducing the specific domain foundation model, the proposed domain-controlled prompt learning provides better domain-adaptive prompts.
- A novel noise-adding strategy is proposed to explicitly address the issue of overfitting in domain-controlled prompt learning, enabling a wider solution space.
- Our method is extensively evaluated on specific domain datasets. The experimental results demonstrate our method achieves state-of-the-art performance.

Related Work

Vision Language Models. Vision Language (V-L) models aim to build a cohesive alignment between images and languages to learn a shared embedding space that encompasses both modalities. Conventional V-L models typically comprise three key components: a visual encoder, a text encoder, and an alignment loss. Visual images were often processed using hand-crafted descriptors (Elhoseiny, Saleh, and Elgammal 2013; Socher et al. 2013) or neural networks (Frome et al. 2013; Lei Ba et al. 2015), while texts were typically encoded using pre-trained word vectors (Socher et al. 2013; Frome et al. 2013) or frequency-based descriptors (Schnabel et al. 2015; Gong et al. 2018). The visual and textual representations were then aligned using techniques like metric learning (Frome et al. 2013) or multi-label classification (Joulin et al. 2016).

However, recent advancements in V-L models (Radford et al. 2021; Jia et al. 2021; Yao et al. 2021; Yuan et al. 2021; Zhai et al. 2022) have revolutionized the field by seamlessly integrating the two modalities through joint learning of image and text encoders in an image-text pair alignment fashion. For example, models like CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) leverage an extensive corpus of approximately 400 million and 1 billion image-text pairs, respectively, to train their multi-modal networks. This approach enables the recent V-L models to generate highly informative cross-modality representations, leading to exceptional performance across various downstream tasks, including few-shot and zero-shot visual recognition (Gao et al. 2021; Zhang et al. 2021). Furthermore, by carefully tailoring V-L models and effectively utilizing the cross-modality representations, traditional image recognition (Conde and Turgutlu 2021; Fu et al. 2022), object detection (Feng et al. 2022; Bangalath et al. 2022), and semantic segmentation (Li et al. 2022; Lüddecke and Ecker 2022; Rao et al. 2022) tasks have also achieved promising performance improvements.

Prompt Learning in Vision Language models. V-L models can be adapted to downstream tasks using either full fine-tuning or linear probing approaches. However, full fine-tuning is computationally intensive and may degrade the previously learned cross-modality representations. On the other hand, linear probing limits the zero-shot capability of models like CLIP. To address these challenges, inspired by prompt learning in natural language processing, many algorithms (Zhou et al. 2022b,a; Khattak et al. 2023; Wang et al. 2022b; Ge et al. 2022) have been proposed to efficiently adapt V-L models in the prompt tokens learning manner. For instance, CoOp (Zhou et al. 2022b) introduces context optimization to adapt CLIP by using learnable context vectors while keeping the pre-trained parameters fixed. However, CoOp's learned context has limited generalizability and suffers from overfitting issues in base categories. To overcome these limitations, CoCoOp (Zhou et al. 2022a) proposes conditional context optimization, which provides instance-conditioned prompt tokens. While previous methods focus on efficient prompt learning in CLIP's language branch, the visual branch is few considered. Addressing this gap, MaPLe (Khattak et al. 2023) proposes multi-modal prompt learning to simultaneously adapt vi-

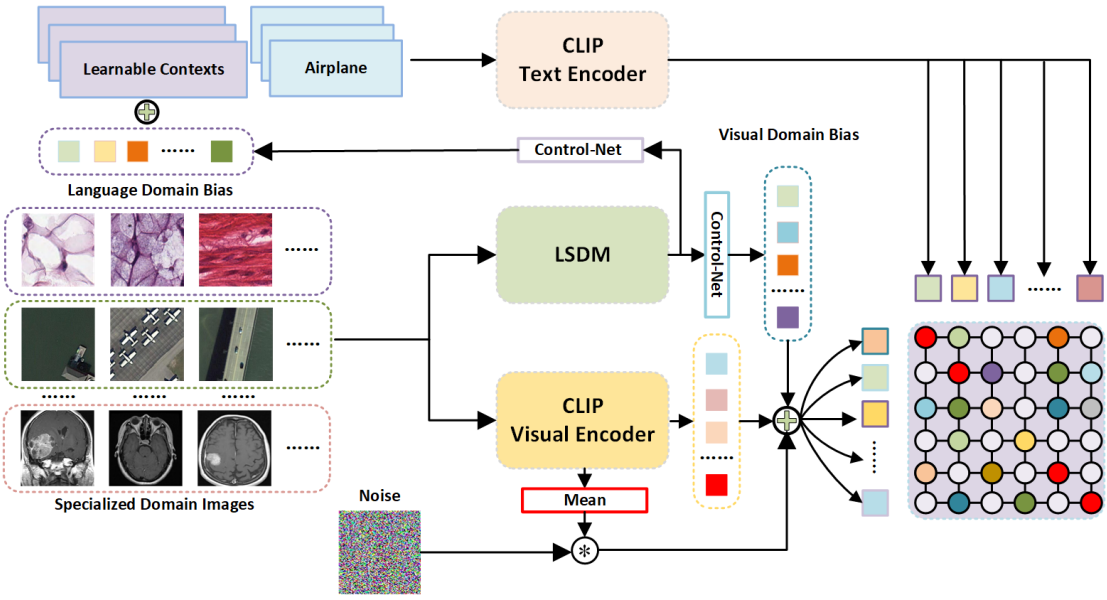

Figure 2: Overview of our proposed Domain-Controlled Prompt Learning (DCPL) framework. Introducing the large-scale specific domain foundation model (LSDM) to provide domain foundation knowledge, DCPL provides domain-adaptive prompts to respectively control the visual and language branch in a directly incorporating manner. Additionally, a noisy-adding strategy is further proposed to help the model escape the suboptimal solution in a global domain oscillation manner.

sion and language representations, which successfully improves the cross-alignment. However, existing algorithms rarely consider the adaptation problem when transitioning from the natural image domain to specific domains like RSIs, and medical images. This lack of domain awareness or domain-transfer mechanism leads to an inadequate perception of specific domain images and results in suboptimal performance. To handle this issue, we propose domain-controlled prompt learning for specific domain images to provide domain-adaptive prompts.

Method

To provide domain-adaptive prompts, our proposed method (DCPL) first introduces the large-scale specifc domain foundation model into CLIP to achieve domain-controlled prompt learning. Figure 2 shows the overall architecture. More specifically, to better transfer CLIP from the natural domain to specific domains, the large-scale specific domain foundation model (LSDM) is first introduced to provide specific domain features as the specific domain knowledge. Then, through the designed control nets, the specific domain features could be respectively transferred into language domain bias and visual domain bias. By adding the language domain bias into the learnable context vectors and incorporating the visual domain bias into the visual features, the networks are controlled to have domain-adaptive prompts. Simultaneously, to help the network search solution in a broader space, the noise is adaptively added to the visual features to perform a global domain oscillation. Below we first introduce the pre-trained CLIP (Radford et al. 2021) and the introduced LSDM (Wang et al. 2022a; Ma and Wang 2023). Then, we illustrate the proposed DCPL.

Review of CLIP

CLIP mainly contains a visual encoder and a text encoder, which could respectively generate image embeddings and corresponding text embeddings. We follow the setting in previous methods (Khattak et al. 2023; Zhou et al. 2022a) to adopt the vision transformer (ViT (Dosovitskiy et al. 2020)) based CLIP model. For the visual encoder, $I \in \mathbb{R}^{H\times W\times 3}$ would be firstly spilled into M fixed-size patches, which are further reshaped as patch embeddings $E_p \in \mathbb{R}^{M\times d_p}$. Then the patch embeddings would be propagated into the transformer layers with the learnable category tokens C_p. To obtain the final image embeddings $x \in \mathbb{R}^{d_t}$, the category tokens C_l from last layer would be projected into the common Visual-Language feature space:

$$x = VisProj(C_l) \tag{1}$$

In the test encoder, the text descriptions for images would be first tokenized into the words and further projecting them to word embedding W_t. Subsequently, the word embeddings would be inputted into transformer layers. Similarly, the text embeddings W_l from the last layer are projected into the common Visual-Language feature space to obtain the final text embeddings $\omega \in \mathbb{R}^{d_t}$:

$$\omega = WordProj(w_l) \tag{2}$$

With these image embeddings and the corresponding text embeddings, the CLIP would maximize the cosine similarity between the image and its matched text while minimize the cosine similarity between the image and its unmatched

text. After training, the CLIP would be directly leveraged to perform the zero-shot classification. Particularly, the ω_i is generated from the hand-craft prompt, such as "a photo of $category$", where $category$ is the i-th class name. Then, suppose there are C categories and the visual embedding of the image is x, the probability of the image belonging to i-th class name is produced by:

$$p(y|x) = \frac{\exp(sim(x, \omega)/\tau)}{\sum_{i=1}^{C} \exp(sim(x, \omega)/\tau)} \tag{3}$$

where sim denotes the cosine similarity and τ is the adjusting temperature parameter.

Large-Scale Specific Domain Foundation Model

The Large-Scale Specific Domain Foundation Model (LSDM) (Sun et al. 2022; Wang et al. 2022a; Ma and Wang 2023) is recently proposed to provide better representations for downstream image processing tasks like RSIs, medical images, etc. Inspired by this, the large-scale remote sensing foundation model (LRSM) (Wang et al. 2022a) and MedSAM (Ma and Wang 2023) are respectively utilized to provide basic domain knowledge for RSIs. The LRSM mainly adopt ViT (Dosovitskiy et al. 2020) and ViTAE (Xu et al. 2021) architectures, and the networks are trained in an MAE (He et al. 2022) manner with millions of RSIs. MAE aims to recover the masked images with the visible parts in an encoder-decoder architecture. The network is optimized by minimizing the loss between the recovered regions and the ground-truth masked regions. Harnessing the power of a meticulously curated dataset comprising over one million medical images, MedSAM (Ma and Wang 2023) are pre-trained for downstream medical image processing tasks. Since we need to control the visual and language branches in the cross-modality space, the pre-trained encoder of the LSDM network is leveraged to provide the specific domain embeddings R_b as the basic specific domain knowledge.

DCPL: Domain-Controlled Prompt Learning

Existing prompt learning methods all ignore the adaption problem from the natural domain to the specific domain like the remote sensing domain. This negligence would result in the specific domain images being handled with an inappropriate natural image processing pattern, further leading to suboptimal performance. To tackle this challenge, we introduce the LSDM to provide specific domain knowledge to control the visual and language to perceive the specific domain images with domain-adaptive prompts.

Give the input images $I \in \mathbb{R}^{H\times W\times 3}$, the input images would be propagated into the pre-trained encoder of LSDM to generate the specific domain embeddings $R_b \in \mathbb{R}^{d_r}$ as the basic specific domain knowledge:

$$R_b = \underset{LSDM}{Encoder}(I) \tag{4}$$

Control the Language Branch. To control the language branch, we first adopt the learnable context vectors in the CoOp (Zhou et al. 2022b) as the basic prompt. Assuming we have M context tokens $\{v_1^{ct}, v_2^{ct}, ..., v_M^{ct}\}$. The language domain bias $D_b^l \in \mathbb{R}^{d_t}$ for the language branch is generated by $D_b^l = f_{LC}(R_b)$, where $f_{LC}(\cdot)$ denotes the designed language control net. By directly incorporating the domain bias D_b into the context tokens, the basic context tokens are transferred into the specific domain:

$$v_m^{ct}(R_b) = v_m^{ct} + D_b^l, m \in \{1, 2, ...M\} \tag{5}$$

Then the final domain-adaptive prompt could be defined as $t_i(R_b) = \{v_1^{ct}(R_b), v_2^{ct}(R_b), ..., v_M^{ct}(R_b), C_i\}$, where i denotes i-th category, and C_i means the category name.

Control the Visual Branch. We first compute the visual domain bias $D_b^v \in \mathbb{R}^{d_t}$ through the designed visual control net$f_{LC}(\cdot)$: $D_b^v = f_{VC}(R_b)$. Since the specific domain embeddings have the same modality as the final image embeddings $x \in \mathbb{R}^{d_t}$. Thus, the generated domain bias could be directly fused with x to directly generate the domain-adaptive image features $x_d(R_b)$:

$$x_d(R_b) = x + D_b^v \tag{6}$$

In this manner, both prompts for the visual and language branches are directly controlled by the introduced specific domain knowledge R_b, which helps the model process the specific domain images in a correct specific domain perception manner.

Noisy-Adding Strategy. As discussed in CoCoOp (Zhou et al. 2022a), prompt learning methods tend to be overfitted in the base categories and not generalizable to wider unseen classes within the same task. CoCoOp tends to handle this problem with an instance-conditional network. However, we tend to solve it in a more explicit manner. Normally, we could adopt the dropout or mutation operations to solve this. However, these methods are actually local sampling strategies or point-based modifications. This means the prompt learning network is still searching for solutions in the oscillation-limited domain. The inference process in diffusion models (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021) would add random noise to escape the trivial solutions and search for better solutions in complex space. Inspired by this, we also randomly sample the noise to help the model escape the suboptimal solution and search for the solutions with the global domain oscillate. Particularly, given the Gaussian noise z, we first compute the adaptive adjusting factor σ_m by computing the mean of image embeddings: $\sigma_m = Mean(x)$. Then, the adaptive adjusting factor is leveraged to scale the sampled Gaussian noise. To directly handle the overfitting problem, the noise would be directly added to the domain-adaptive images features $x_d(R_b)$:

$$\widehat{x}_d(R_b) = x_d(R_b) + \sigma_m z \tag{7}$$

Finally, the probability of the image belonging to $i-$th category name is changed from equation 3 to:

$$p(y|x) = \frac{\exp(sim(\widehat{x}_d(R_b), t_y(R_b))/\tau)}{\sum_{i=1}^{C} \exp(sim(\widehat{x}_d(R_b), t_i(R_b))/\tau)} \tag{8}$$

Experiments

To assess the effectiveness of proposed method, we conducted extensive experiments using RSIs and medical images as examples, covering three distinct problem settings:

	Base	Novel	HM
CLIP	71.19	71.33	70.63
CoOp	87.61	70.84	78.03
Co-CoOp	91.82	68.98	78.43
MaPLe	93.12	71.71	80.42
Ours (ViTAE)	93.07	73.79	81.81
Ours (ViT)	**93.77**	**75.81**	**83.36**

(a) Average over 8 datasets

	Base	Novel	HM
CLIP	64.50	**60.30**	62.33
CoOp	79.37	58.90	67.62
Co-CoOp	83.30	59.50	69.42
MaPLe	85.23	59.60	70.15
Ours (ViTAE)	86.30	58.47	69.71
Ours (ViT)	**87.05**	59.30	**70.54**

(b) MLRSNet

	Base	Novel	HM
CLIP	70.60	62.60	66.36
CoOp	87.30	64.20	73.99
Co-CoOp	93.70	59.90	73.08
MaPLe	95.30	57.90	72.03
Ours (ViTAE)	95.33	62.07	75.19
Ours (ViT)	**95.93**	**64.60**	**77.21**

(c) PatternNet

	Base	Novel	HM
CLIP	66.70	95.30	78.48
CoOp	84.80	89.13	86.91
Co-CoOp	90.97	90.00	90.48
MaPLe	91.67	93.70	92.67
Ours (ViTAE)	87.87	92.13	89.95
Ours (ViT)	**91.67**	**95.37**	**93.48**

(d) RSSCN7

	Base	Novel	HM
CLIP	73.50	70.40	71.92
CoOp	87.63	70.37	78.06
Co-CoOp	92.63	65.73	76.89
MaPLe	92.73	74.57	82.66
Ours (ViTAE)	**93.33**	75.13	83.25
Ours (ViT)	92.90	**76.03**	**83.62**

(e) AID

	Base	Novel	HM
CLIP	71.50	60.20	65.37
CoOp	88.43	60.20	71.63
Co-CoOp	92.37	58.80	71.86
MaPLe	93.93	56.27	70.38
Ours (ViTAE)	94.93	62.83	75.61
Ours (ViT)	**95.03**	**64.64**	**76.94**

(f) RSICD

	Base	Novel	HM
CLIP	80.60	68.00	73.77
CoOp	93.60	74.53	82.98
Co-CoOp	95.23	71.57	81.72
MaPLe	97.70	70.90	82.17
Ours (ViTAE)	97.00	75.43	84.87
Ours (ViT)	**98.00**	**80.00**	**88.09**

(g) UCM

	Base	Novel	HM
CLIP	73.10	90.80	80.99
CoOp	95.20	82.40	88.34
Co-CoOp	97.10	77.00	85.89
MaPLe	97.70	88.03	92.61
Ours (ViTAE)	**98.80**	91.10	94.79
Ours (ViT)	98.77	**93.70**	**96.17**

(h) WHURS19

	Base	Novel	HM
CLIP	69.00	63.00	65.87
CoOp	84.53	66.97	74.73
Co-CoOp	89.27	69.37	78.07
MaPLe	90.70	72.70	80.71
Ours (ViTAE)	**90.97**	**73.23**	**81.14**
Ours (ViT)	90.80	72.80	80.81

(i) NWPU

Table 1: Comparison with existing methods in base-to-novel generalization on 8 remote sensing recognition datasets. The best results are shown in bold.

1) base-to-novel class generalization within a dataset, 2) cross-dataset transfer, and 3) domain generalization. Due to space limitations, more detailed special domain experiments are illustrated in supplementary materials. In this section, we offer a comprehensive overview of the utilized datasets and the evaluation metrics employed. Furthermore, we provide detailed insights into implementation specifics of our experiments. Subsequently, we conduct an in-depth analysis of our method's performance in each of aforementioned problem settings. Additionally, we performed ablation experiments to elucidate the effectiveness of our proposed approach.

Experimental Details

The proposed method was evaluated on eight remote sensing datasets, namely MLRSNet (Qi et al. 2020), PatternNet (Zhou et al. 2018), RSSCN7 (Zou et al. 2015), AID (Xia et al. 2017), RSICD (Lu et al. 2017), UCM (Yang and Newsam 2010), WHURS19 (Dai and Yang 2011), and NWPU (Cheng, Han, and Lu 2017). Consistent with previous methods (Khattak et al. 2023), we employed accuracy and Harmonic Mean (HM) as evaluation metrics. The HM is computed as follows:

$$HM = \frac{2 \times Acc_{base} \times Acc_{novel}}{Acc_{base} + Acc_{novel}} \quad (9)$$

Here, Acc_{base} denotes the accuracy for base category, and Acc_{novel} denotes the accuracy for novel category. It is critical to note that the reported results are averaged over three runs. For the base-to-novel generalization setting, experiments were conducted on all eight remote sensing datasets. In the cross-dataset generalization and domain generalization settings, MLRSNet was used as the source dataset, while the remaining datasets served as the target datasets. We implemented our method based on MaPLe and adopted similar training details. All experiments were conducted using a few-shot training strategy with 16 shots, randomly sampled for each class. Pre-trained ViT-B/16 CLIP model is used as the basis for prompt tuning. The training process for all models lasted for 5 epochs, employing a batch size of 4 and a learning rate of 0.0035. We utilized the SGD optimizer and trained models on a single NVIDIA A100 GPU. The template for the word embeddings is 'a photo of $category$'. We kept the hyperparameters consistent across all datasets to ensure fair comparisons. The language and visual control networks were implemented as two independent networks with the same architecture. Each network consisted of two linear layers followed by a ReLU activation layer.

Generalization from Base-to-Novel Classes

Prompt learning aims to ease the application of large-scale models to various tasks, emphasizing effective generalization from familiar to unfamiliar classes. To study it, we conducted comprehensive experiments on remote sensing recognition datasets. Utilizing LRSM capabilities, we employed two pre-trained models, ViT and ViTAE, to integrate

	Source	Target							
	MLRSNet	PatternNet	RSSCN7	AID	RSICD	UCM	WHURS19	NWPU	Average
CoOp	72.53	66.97	69.03	67.30	63.50	77.57	85.47	70.43	71.60
Co-CoOp	71.70	65.67	68.80	66.63	62.57	76.40	85.33	70.30	70.92
MaPLe	76.83	**68.53**	71.43	65.13	59.53	**79.90**	85.23	72.80	72.42
Ours	**77.73**	67.13	**71.60**	**68.73**	**64.17**	78.50	**86.97**	**72.87**	**73.46**

Table 2: Comparisons between our method with state-of-the-art methods for cross-dataset generalization with MLRSNet dataset as the source domain and remaining remote sensing datasets as the target domains. The best results are shown in bold.

	Source	Target							
	MLRSNet	PatternNetv2	RSSCN7v2	AIDv2	RSICDv2	UCMv2	WHURS19v2	NWPUv2	Average
CoOp	72.53	66.97	69.07	67.13	64.27	77.40	85.20	71.17	71.72
Co-CoOp	71.70	65.57	69.37	67.13	62.73	75.70	84.83	70.97	71.00
MaPLe	76.83	68.03	**72.50**	64.90	59.73	**78.93**	83.07	73.17	72.15
Ours	**77.73**	**68.27**	72.10	**68.33**	**64.57**	77.30	**85.80**	**73.37**	**73.43**

Table 3: Comparisons between our method with SOTA methods for single-source multi-target domain generalization with MLRSNet dataset as the source domain and remaining datasets as the target domains. The best results are shown in bold.

domain-specific knowledge for remote sensing. Our method was rigorously evaluated against state-of-the-art techniques like zero-shot CLIP, CoOp, CoCoOp, and MaPLe, providing a robust benchmark in Table 1. Compared to the leading MaPLe approach, our method demonstrated significant performance enhancements across both base and novel categories in remote sensing datasets. Notably, our ViT-based approach exhibited superior overall performance. For base categories, we achieved a noteworthy improvement from 93.07% to 93.7%. Particularly striking were the substantial improvements for novel categories, rising from 73.79% to 75.81%. Considering both base and novel classes, our method outperformed MaPLe by 2.94%. Noteworthy was the outstanding gain of 6.56%. on the RSICD dataset. An intriguing observation arose when comparing the ViTAE model, with its deeper architecture and greater expressive capacity. Surprisingly, while ViTAE outperformed MaPLe overall, it fell slightly short compared to the ViT-based approach. This suggests an upper limit to the utilization of remote sensing knowledge, where a deeper architecture may not always be optimal for prompt learning. Detailed analysis revealed enhanced performance for the base categories in AID, WHURS19, and NWPU datasets with our ViTAE-based method. However, relatively lower performance in novel categories consistently aligned with our earlier findings, emphasizing the intricate relationship between model depth, prompt learning, and dataset characteristics.

Cross-dataset Evaluation

To demonstrate our proposed method's capacity for cross-dataset generalization, we utilized MLRSNet for training and subsequently evaluated the model on the remaining seven datasets. Comparative results, outlined in Table 2, highlight the notable performance of our method across diverse datasets. Particularly impressive was our method's superior performance on MLRSNet, achieving a substantial improvement of nearly 1%. The RSICD dataset witnessed the most significant performance boost, emphasizing the efficacy of domain knowledge for this specific dataset. Despite less favorable outcomes on PatternNet and UCM datasets, our method surpassed all algorithms in terms of overall performance with a 1.04% performance improvement.

Methods	Base	Novel	HM
CLIP	49.83	41.83	45.18
CoOp	51.59	43.77	46.81
Co-CoOp	64.45	43.16	49.45
MaPLe	62.39	44.40	49.01
Ours	**66.11**	**48.75**	**53.08**
	+1.66	**+4.35**	**+3.63**

Table 4: Comparison (average) with existing methods in base-to-novel generalization on medical image classification datasets. The best results are shown in bold.

Domain Generalization

To further validate the generalization ability of our proposed method, we conducted an evaluation in the domain generalization setting. Our approach was compared against other state-of-the-art algorithms, and the comparative results are presented in Table 3. Remarkably, our method consistently outperforms the competing algorithms, achieving the highest average performance with a noteworthy 1.28% improvement. It is important to note that while our method may encounter challenges when applied to the RSSCN7v2 and UCMv2 datasets, it excels on the RSICDv2 dataset, showcasing an impressive performance gain of 4.84%. These findings underscore the efficacy of incorporating domain-controlled prompt learning in enhancing the generalization.

Methods	Base	Novel	HM
Baseline	97.70	70.90	82.17
Baseline+VC	97.80	76.43	85.80
Baseline+LC	97.60	73.33	83.74
Ours	**98.00**	**80.00**	**88.09**

Table 5: Ablation study of domain-controlled prompt learning in different branches. VC and LC individually denote Visual and Language domain-controlled prompt learning.

Methods	Base	Novel	HM
Baseline	97.70	70.90	82.17
Dropout(0.3)	97.78	77.83	86.67
Dropout(0.5)	97.30	77.67	86.38
Mutation(0.05)	97.60	71.67	82.65
Mutation(0.1)	97.20	71.57	82.44
Ours	**98.00**	**80.00**	**88.09**

Table 6: Ablation study of overfitting-tackling strategies.

Experiments on Other Domain

To further validate the effectiveness of our proposed method, we conducted comprehensive experiments on medical domain datasets, including BTMRI (Nickparvar 2021), CHMNIST (Kather et al. 2016), and CCBTM (Hashemi 2023). The comparative results are summarized in Table 4. Specifically, our method achieves an impressive 1.66% performance improvement for base categories and an even more substantial 4.35% improvement for novel categories. When considering the overall performance metric, Harmonic Mean (HM), our method exhibits a significant 3.63% improvement compared to other algorithms. These compelling results indicate the efficacy of our method.

Ablation Study

Domain-Controlled Prompt Learning. In order to analyze the impact of different components in domain-controlled prompt learning, we conducted separate experiments for both the visual and language branches. The evaluations were performed on the UCM datasets, and summarized in Table 5. Obviously, incorporating domain-controlled prompt learning leads to performance improvements. Specifically, controlling the visual branch yields substantial performance gains, particularly in novel categories, resulting in an overall improvement of 3.63%. Simultaneously, domain-controlled prompt learning in the language branch contributes to an overall improvement of 1.57%. These findings highlight the effectiveness of domain-controlled prompt learning in benefiting both the visual and language branches.

Different Overfitting-Tackling Strategies. We adopt the proposed noisy-adding strategy to explicitly solve the overfitting problem. Adopting dropout or mutation operations seems to be a plausible solution. Thus, we implement experiments on the UCM dataset to distinguish our method from other strategies, and the results are shown in Table 6. The dropout and mutation operations could both bring performance improvements. The dropout with a rate of 0.3 has a better performance than a rate of 0.5, and the mutation with 5 percent has a better performance than the 10 percent. Though these operations could bring some performance improvements, our noisy-adding strategy could have obviously better performance improvements. This phenomenon suggests the local sampling in dropout and point jittering in mutation are insufficient in escaping suboptimal solutions, yet our method helps the network have a broader solution exploration in a global domain oscillation manner.

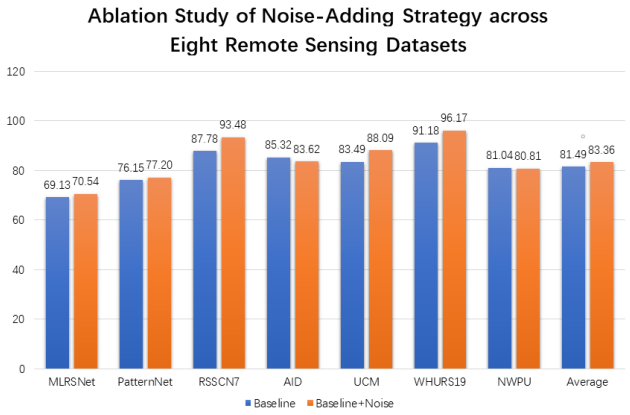

Figure 3: The ablation study of noise-adding strategy across eight remote sensing datasets.

Noise-Adding Strategy across Datasets. To comprehensively assess the impact of the noise-adding strategy, we conducted experiments across eight diverse remote sensing datasets. The performance gains achieved by incorporating the noise-adding strategy are illustrated in Figure 3. The results demonstrate that the noise-adding strategy consistently improves performance across the majority of datasets, with only minor performance decreases observed in the NWPU and AID datasets. Remarkably, the noise-adding strategy leads to an overall performance improvement of 1.87%. This observation highlights the effectiveness of the proposed strategy to mitigate overfitting and boost performance.

Conclusion

Focusing on the neglected natural-to-specific adaptation challenge, we introduce large-scale specific domain foundation models to provide specific domain knowledge and further perform domain-controlled prompt learning in both visual and language branches for specific domain images. To overcome the base-to-novel overfitting challenge, a novel noisy adding strategy is proposed to explicitly escape the suboptimal solutions in a global domain oscillation manner. To validate the effectiveness of our method, we conduct extensive experiments using specific domain datasets, producing compelling experimental results that demonstrate the superiority of our proposed approach.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (62106116, 62376156, 62322113).

References

Bangalath, H.; Maaz, M.; Khattak, M. U.; Khan, S. H.; and Shahbaz Khan, F. 2022. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35: 33781–33794.

Cao, Q.; Chen, Y.; Ma, C.; and Yang, X. 2023a. Break the Bias: Delving Semantic Transform Invariance for Few-Shot Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Cao, Q.; Chen, Y.; Ma, C.; and Yang, X. 2023b. Few-Shot Rotation-Invariant Aerial Image Semantic Segmentation. *arXiv preprint arXiv:2306.11734*.

Cheng, G.; Han, J.; and Lu, X. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10): 1865–1883.

Conde, M. V.; and Turgutlu, K. 2021. CLIP-Art: Contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3956–3960.

Dai, D.; and Yang, W. 2011. Satellite Image Classification via Two-Layer Sparse Coding With Biased Image Representation. *IEEE Transactions on Geoscience and Remote Sensing*, 8(1): 173–176.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Elhoseiny, M.; Saleh, B.; and Elgammal, A. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, 2584–2591.

Feng, C.; Zhong, Y.; Jie, Z.; Chu, X.; Ren, H.; Wei, X.; Xie, W.; and Ma, L. 2022. Promptdet: Towards open-vocabulary detection using uncurated images. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, 701–717. Springer.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems*, 26.

Fu, J.; Xu, S.; Liu, H.; Liu, Y.; Xie, N.; Wang, C.-C.; Liu, J.; Sun, Y.; and Wang, B. 2022. Cma-clip: Cross-modality attention clip for text-image classification. In *2022 IEEE International Conference on Image Processing (ICIP)*, 2846–2850. IEEE.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Ge, C.; Huang, R.; Xie, M.; Lai, Z.; Song, S.; Li, S.; and Huang, G. 2022. Domain adaptation via prompt learning. *arXiv preprint arXiv:2202.06687*.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Gong, C.; He, D.; Tan, X.; Qin, T.; Wang, L.; and Liu, T.-Y. 2018. Frage: Frequency-agnostic word representation. *Advances in Neural Information Processing Systems*, 31.

Hashemi, S. M. H. 2023. Crystal Clean: Brain Tumors MRI Dataset.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

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.

Joulin, A.; Van Der Maaten, L.; Jabri, A.; and Vasilache, N. 2016. Learning visual features from large weakly supervised data. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, 67–84. Springer.

Kather, J.; Weis, C.; Bianconi, F.; Melchers, S.; Schad, L.; Gaiser, T.; Marx, A.; and F, Z. 2016. Multi-class texture analysis in colorectal cancer histology. *Scientific Reports (in press)*.

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.

Lei Ba, J.; Swersky, K.; Fidler, S.; et al. 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, 4247–4255.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Lu, X.; Wang, B.; Zheng, X.; and Li, X. 2017. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4): 2183–2195.

Lüddecke, T.; and Ecker, A. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7086–7096.

Ma, J.; and Wang, B. 2023. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Nickparvar, M. 2021. Brain Tumor MRI Dataset.

Qi, X.; Zhu, P.; Wang, Y.; Zhang, L.; Peng, J.; Wu, M.; Chen, J.; Zhao, X.; Zang, N.; and Mathiopoulos, P. T. 2020. MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169: 337–350.

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.

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18082–18091.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Schnabel, T.; Labutov, I.; Mimno, D.; and Joachims, T. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 298–307.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. 2013. Zero-shot learning through cross-modal transfer. *Advances in Neural Information Processing Systems*, 26.

Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. 2022. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*.

Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; and Zhang, L. 2022a. Advancing plain vision transformer towards remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022b. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 139–149.

Xia, G.-S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; and Lu, X. 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7): 3965–3981.

Xu, Y.; Zhang, Q.; Zhang, J.; and Tao, D. 2021. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34: 28522–28535.

Yang, Y.; and Newsam, S. 2010. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 270–279.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. FILIP: fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Zhai, X.; Wang, X.; Mustafa, B.; Steiner, A.; Keysers, D.; Kolesnikov, A.; and Beyer, L. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18123–18133.

Zhang, R.; Fang, R.; Zhang, W.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2021. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*.

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, W.; Newsam, S.; Li, C.; and Shao, Z. 2018. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing*, 145: 197–209.

Zou, Q.; Ni, L.; Zhang, T.; and Wang, Q. 2015. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and remote sensing letters*, 12(11): 2321–2325.