
Mainak Singha\(^1\)*, Ankit Jha\(^1\)*, Bhupendra Solanki\(^1\), Shirsha Bose\(^2\), Biplab Banerjee\(^1\)

\(^1\)Indian Institute of Technology Bombay, India \(^2\)Technical University of Munich, Germany

\{mainaksingha.iitb, ankitjhal6, bssiitb, shirhabosecs, getbiplab\}@gmail.com

**Abstract**

In recent years, the success of large-scale vision-language models (VLMs) such as CLIP has led to their increased usage in various computer vision tasks. These models enable zero-shot inference through carefully crafted instructional text prompts without task-specific supervision. However, the potential of VLMs for generalization tasks in remote sensing (RS) has not been fully realized. To address this research gap, we propose a novel image-conditioned prompt learning strategy called the Visual Attention Parameterized Prompts Learning Network (APPLenNet). APPLenNet emphasizes the importance of multi-scale feature learning in RS scene classification and disentangles visual style and content primitives for domain generalization tasks. To achieve this, APPLenNet combines visual content features obtained from different layers of the vision encoder and style properties obtained from feature statistics of domain-specific batches. An attention-driven injection module is further introduced to generate visual tokens from this information. We also introduce an anti-correlation regularizer to ensure discrimination among the token embeddings, as this visual information is combined with the textual tokens. To validate APPLenNet, we curated four available RS benchmarks and introduced experimental protocols and datasets for three domain generalization tasks. Our results consistently outperform the relevant literature and code is available at https://github.com/mainaksingha01/APPLenNet

**1. Introduction**

Remote Sensing (RS) images play a vital role in numerous applications, including those mentioned in [15, 44, 60, 67]. Traditional deep learning models have proven effective in recognizing complex RS images, outperforming ad-hoc machine learning techniques. However, these models tend to perform poorly in terms of generalization when faced with domain shifts. For example, Fig. 1 illustrates this issue, where a model (ERM [56]) trained on images from the PatternNet [25] dataset exhibits sub-optimal performance when applied to images from the RSICD [32] dataset, captured by two sensors with divergent spatial characteristics.

To combat such changes in data distributions between training (source) and test (target) domains, researchers have investigated domain generalization (DG) [23, 69, 70] and domain adaptation (DA) [8, 10, 45, 49, 52, 53]. DA follows a transductive setup, where the source and target domains are available simultaneously during training, while DG deals with a more realistic scenario, where a model trained on the source domain is applied to novel target domains during inference. Despite its success in computer vision literature, DG has yet to be thoroughly explored in RS.

From another perspective, few-shot learning (FSL) methods [4, 9, 18, 42, 68] have emerged as a beneficial solution for alleviating the deep learning models’ abundant data dependency for visual recognition, including in RS. FSL for different modalities, including multi-spectral and hyperspectral, has been introduced in RS [1, 31, 34, 62]. However, these models are developed solely on image data and are suboptimal regarding the semantic richness of the embedding space learned by the feature extractors. As reported by
these works, this significantly affects the FSL performance, and zero-shot transfer to novel tasks is not possible by FSL.

Large-scale pre-trained vision-language models (VLMs), such as CLIP [40], ALIGN [19], Florence [64], LiT [66], or Foundation models [2], have recently shown promising results in generalizing to various downstream target domain tasks in a zero-shot manner with minimal supervision from a different source domain. These models align image-text pairs in a shared embedding space using a contrastive learning approach, making prompt engineering a critical aspect of VLMs. However, manual prompt engineering is non-trivial, and prompt learning has received much attention to adapt CLIP for a target task. To address the generalization deficiency of the baseline prompt learning technique CoOp [72] and subsequent studies [12, 71, 75] proposed to supplement the textual prompt embeddings with visual information extracted from CLIP’s frozen vision encoder. However, while these models are validated on natural image classification, we aim to explore their potential for scene recognition from optical RS images. This task is highly challenging due to the divergent spectral and spatial artifacts that characterize these images.

Although pre-trained CLIP [40] is highly effective, it falls short when evaluated on different domains, such as RS, as evidenced by the zero-shot CLIP’s performance in Fig. 1. While prompt learning approaches like [12, 71, 72] improve on the performance of baseline CLIP, they are sub-optimal for cross-domain and cross-dataset generalization. These approaches have three critical issues: i) they only consider visual features from the deepest layer and combine them with prompt token embeddings, ignoring low and mid-level features essential for optical RS scene classification where object scales are small, and texture plays a significant role in the classification task, ii) they add the same visual information to all token embeddings, causing redundancy, and iii) existing approaches fail to disentangle domain features from content features, which is likely to aid in DG.

Drawing from these discussions, this paper seeks to address two critical research questions: i) How can we effectively utilize CLIP’s vision backbone to extract multi-scale visual content and style information for RS scenes to learn credible prompt tokens? and ii) How can we guarantee that the learned prompt tokens contain non-redundant information? We believe that tackling these issues together would result in more comprehensive and versatile prompts for optical RS scenes, as demonstrated in Fig. 1.

Our proposed APPLeNet: To address these challenges, our proposed approach, APPLeNet, makes three key contributions. Firstly, we utilize the intermediate blocks of CLIP’s vision encoder to extract multi-scale visual content information. Secondly, we calculate the average feature representation for a batch of samples from a given domain to obtain style primitives for that domain. We leverage the concept of batch-norm statistics of feature embeddings from a CNN, which carry domain-specific knowledge [29]. We combine the content and style features using an attention-based novel injection block to generate dynamic image-conditioned visual tokens combining visual content and style properties. Finally, these tokens are added element-wise to the learnable text token embeddings to generate the prompts.

Thirdly, we introduce an anti-correlation regularizer to promote discrimination among prompt tokens. This regularizer penalizes high correlation among token embeddings. As a result, APPLeNet achieves more generalizable prompts than existing methods such as [12, 71, 72]. Notably, APPLeNet demonstrates strong performance even with extremely limited training data and is effective in different domain generalization scenarios with domain and label shifts.

We highlight our major contributions as,

- We propose a solution to the few-shot optical RS scene recognition and generalization problem by using pre-trained CLIP and introducing lightweight injection blocks in a model we call APPLeNet. The key innovations of APPLeNet are leveraging multi-scale visual content and style information from CLIP’s vision encoder to learn prompt tokens and an anti-correlation regularizer that ensures the distinctiveness of the learned tokens.
- To validate our approach, we conduct extensive experiments on four optical RS image classification benchmarks and test for three essential and challenging generalization tasks: base-to-new class, cross-dataset, and single-source multi-target. We also introduce experimental protocols for these tasks, which have not been widely studied in RS.

Our experimental results demonstrate that APPLeNet outperforms the relevant literature substantially for all tasks by at least 2% in mean classification scores.

2. Related Works

FSL in general and in RS: Broadly speaking, existing few-shot learning (FSL) methods can be divided into transfer-learning based [50, 63], meta-learning based [9, 16], and metric-learning based [48, 51] approaches, respectively. Transfer learning fine-tunes the base model on each novel task but can underperform if the base and novel classes are from drastically different distributions. Alternatively, meta-learning-based supervised FSL approaches [24, 43] have gained attention because they can learn more generalizable features through episodic training. Metric-learning-based methods like matching networks [57], prototypical networks [48], and relation networks [51] focus on similarity optimization in a learnable fashion in episodes.

While all these approaches find their applicability in few-shot learning RS data, meta-learning-based algorithms [1, 24] have been predominantly explored [40, 65]. How-
ever, these models are prone to training class bias. One solution to tackle this is through uncertainty optimization, together with the FSL objective [34].

**Domain generalization:** Deep learning models often face the challenge of domain shift between training and test distributions, which makes Domain Generalization (DG) a critical task for learning generalizable features from the training set that can be applied to any novel domain during inference. DG has three variants: single-source DG, multi-source DG, and heterogeneous DG [13, 21, 30, 39, 58, 59, 73]. Initially, DG techniques focused on learning domain-invariant representations by considering data from multiple source domains through Domain Adaptation (DA) objectives [20, 26, 27, 43]. Other approaches explored self-supervised learning [3], ensemble learning [61], domain-specific networks [33], and meta-learning [35]. However, training deep learning models for DG with limited samples may affect performance. Researchers have used augmentation methods [28, 73], and generative models [21, 74] to augment the source domains with diversified style primitives. However, DG for Remote Sensing (RS) image classification has received limited attention to date [36, 69].

In contrast to existing FSL works in RS, which are trained from visual feature extractors, we are interested in leveraging the semantic superiority of CLIP-based foundation models for various target domain generalization tasks from a few source domain training samples.

**Prompt learning for CLIP:** Prompt learning is a widely used technique in natural language processing (NLP) [37], which has recently made its way into the computer vision field. The main goal of prompt learning is to leverage pre-trained language models, such as BERT [6], to provide valuable information for downstream tasks through prompts. Recent research has focused on automating the prompt generation process to eliminate manual interventions. One such approach is AutoPrompt [46], which explores tokens with the most significant gradient changes in the label likelihood. CoOp [72] optimizes prompts by fine-tuning CLIP for few-shot image classification. CoCoOp [71] proposes learning conditional prompts based on image features, partially improving CoOp’s generalization capability.

In contrast, CLIP-Adapter [12] proposes fine-tuning feature adapters in both visual and language branches. ProGrad [75] follows a similar approach to CoCoOp and explicitly ensures that the network remembers the knowledge learned from the foundation model. In [47], consistency among multiple views of the same image is used as supervision for prediction.

However, while [71, 75] utilize visual information to improve prompts, they do not account for low to mid-level visual properties and visual style information in the prompt. Additionally, the learned tokens may contain redundant information. In contrast, our proposed APPLeNet addresses these issues and proposes a more comprehensive prompt learning strategy that is well-suited for handling RS scenes.

### 3. Proposed Methodology

Let $D_s = \{D^n_s\}^n_{i=1}$ denote $n$ source domains, each with input data $x^i \in \mathcal{X}^i$ and corresponding label space $y^i \in \mathcal{Y}^{Seen}$. It is important to note that the probability distribution of each domain, $P(D^n_s)$, may differ for all

![Figure 2. APPLeNet is composed of a text encoder $f_t$, an image encoder $f_v$, and an injection block designed for multi-scale visual feature refinement. The $f_v$ produces multi-level visual content features, and the batch statistics $\mu$ for a domain as the style features, that are passed through a residual attention-based injection block. These features are then sent to individual projector networks $\{h_m\}^M_{m=1}$ to derive the visual tokens $\{v_m\}^M_{m=1}$. The visual tokens are added to the learnable text token embeddings $\{c_m\}^M_{m=1}$ and, together with the class embeddings, are forwarded to $f_t$. To reduce redundancy among the tokens, we introduce a novel Context Redundant Penalizing (CRP) loss $(L_{CRP})$ among the context vectors in $\{c_m\}^M_{m=1}$. The model is trained using a multi-task loss that comprises the contrastive loss between the image and prompt embeddings and the CRP regularizer.](image-url)
During training, we use the labels $Y_{\text{seen}}$ from $D_s$, while during testing, we use $Y_{\text{unseen}}$ from a target test domain $D_t$ with $P(D_i) \neq P(D_j)$, $\forall i, \ldots, n$. For base-to-new class generalization, we set $Y_{\text{seen}} \cap Y_{\text{unseen}} = \emptyset$. In contrast, for domain generalization (DG), we consider single-source DG and assume that the label sets for both domains are identical ($Y_{\text{seen}} \cap Y_{\text{unseen}} = Y_{\text{seen}} \cup Y_{\text{unseen}}$). Before presenting our proposed APPLeNet, we briefly introduce some important baselines, such as CLIP [40], CoOp [72], and CoCoOp [71].

3.1. Relevant baselines

CLIP: CLIP [40] is a remarkable foundational model that learns an embedding space by seamlessly integrating visual and semantic knowledge. The model comprises two encoder heads: a visual encoder $f_v$ (either ResNet [14], or ViT [7]) for processing input images $x$, and a text encoder $f_t$ (BERT [6]) that considers the corresponding text prompt $t_y$ structured as "a photo of [CLS]$_y$" where [CLS]$_y$ denotes the word embeddings for the class $y$. By means of contrastive training on a dataset of 400 million image-text pairs, CLIP strives to maximize the similarity between the image and the correct class prompt embeddings.

CoOp: CoOp [72] offers a solution to the issue of prompt engineering by replacing the manually created prompts with those generated through learning. This is achieved by utilizing a set of $M$ learnable context vectors $c_1, c_2, \ldots, c_M$ that have the same dimensionality as the word embeddings and optimizing them through back-propagation. It’s important to note that $M$ is a hyperparameter that determines the context length and may differ between tasks. For any given class $y$, the prompt can be represented as $t_y = \{c_1, c_2, \ldots, c_M\}, [\text{CLS}]_y$.

CoCoOp: Despite the effectiveness of prompt learning, CoOp is susceptible to the domain-shift problem. CoCoOp [71] conditions prompt learning on visual features to mitigate this issue. This is achieved by introducing a meta-network that generates $M$ meta-tokens, denoted as $\pi$. These meta-tokens are concatenated with context vectors to create prompts $t_y = \{c_1(x), c_2(x), \ldots, c_M(x), [\text{CLS}]_y\}$, where $c_m(x) = c_m + \pi(x), c_m$ is the $m$th text token. During CoCoOp’s training, both the meta-network and context vector parameters are updated simultaneously.

Important insight: There have been various prompt learning techniques developed after CoOp and CoCoOp, such as [22, 75]. However, these methods neglect two crucial factors in generalization tasks: the utilization of multi-scale feature composition from CLIP and the incorporation of visual style primitives into the prompts. These factors are particularly important in cases where there is a sudden change in style between source and target domains. The frozen image encoder ($f_v$) can be leveraged to encode style, while multi-scale content features can encode low, mid, and high-level visual properties, making them more transferable across categories.

3.2. Our Proposed APPLeNet

Our paper introduces a new method called APPLeNet (Attention-Parameterized Prompt Learning Network) that leverages CLIP’s visual backbone to extract multi-scale visual features and style features (mean $\mu$ of a batch of features from $f_v$) to improve text token learning. APPLeNet, depicted in Figure 2, is composed of several critical components. First, it includes CLIP’s frozen vision ($f_v$) and text ($f_t$) encoders. Additionally, it features a novel trainable Injection Block (IB), indicated as $B_\phi$, with learnable parameters $\phi$. This block emphasizes concatenated embeddings of style and multi-scale visual features. Moreover, APPLeNet comprises $M$ learnable linear layers that generate $M$ visual tokens $\{v_m\}_{m=1}^M$ given the outputs from $B$.

Encoding style and multi-scale content features into prompts: To incorporate the multi-scale visual features from $f_v$ into $B$, we propose using global average pooling (GAP) to collapse the spatial dimensions of each channel. This produces $\hat{f}_v(x) \in \mathbb{R}^{C \times 1}$, where $f_v^m \in \mathbb{R}^{W \times H \times C}$ represents the output responses from the $i$th layer. Here, $(W, H)$ represents the spatial dimensions of the feature maps. Using this approach, we define $F(x) = \{f_1^v(x); \ldots; f_l^v(x)\}$ as the concatenated multi-scale features obtained from all the $L$ encoder layers of $f_v$, where $[;]$ denotes feature concatenation.

Furthermore, the average feature statistics corresponding to the batch of features from a domain act as the indicator for the style primitives. In this regard, let $\mu_i = f_v(X^i)$ represent the style for the $i$th domain.

Together we produce $F(x) = [F(x_1) + \mu_1]$, which captures both multi-scale content and the style information.

Injection block: The attention modules within $B$ are denoted by $A_q(\cdot)$, where $q \in 1, \ldots, Q$. For the first attention block ($q = 1$), we denote the attended output features as $O_1 = F(x) \odot A_1 \odot F(x)$. These features are then fed as input to $A_2$ and so on, as follows:

$$O_q = \left\{ \begin{array}{ll}
[F(x) \odot A_q(F(x)) + F(x)], & \text{if } q = 1 \\
[O_{q-1} \odot A_q(O_{q-1}) + O_{q-1}], & \text{otherwise}
\end{array} \right. \quad (1)$$

We subsequently pass $O_Q$ through $M$ light-weight projector networks $\{h_m\}_{m=1}^M$, which generate $M$ visual tokens $\{v_1, \ldots, v_M\}$: $v_m = h_m(O_Q)$. We add the $m$th visual token embedding with the $m$th textual token embedding $c_m$ to obtain the $m$th prompt token embedding $c'_m = c_m + v_m$. 

2027
The generated prompt is represented as:

$$t_y = \{ [v_1 + c_1], \cdots, [v_M + c_M], [CLS_y] \} \quad (2)$$

### 3.3. Training and Inference

We adopt a multi-task approach to train APPLeNet using two loss functions. The first one is the supervised contrastive loss, denoted as $L_{ce}$, which ensures proper mapping between the visual feature representation $f_v(x)$ and the textual representation $f_t(t_y)$. This loss is formulated based on the cross-entropy approach.

In addition, we introduce a context redundancy penalizing loss, denoted as $L_{CRP}$. This loss ensures that the token embeddings in the set $c_1 + v_1, \cdots, c_M + v_M$ do not carry redundant information. This helps the model learn a diverse set of tokens. In this regard, the prediction probability for $x$ to belong to the label $y$ is denoted by,

$$p(y|x) = \frac{\exp(sim(f_v(x), f_t(t_y(B_o(x))))/\tau)}{\sum_{y'} \exp(sim(f_v(x), f_t(t_y(B_o(x))))/\tau)} \quad (3)$$

Here, $sim$ denotes the cosine similarity, and $\tau$ is the temperature hyper-parameter. The cross-entropy loss ($L_{ce}$) is computed between the prediction probabilities of each input image and their corresponding class labels as follows:

$$L_{ce} = \min_{B_o, \{h_m\}} \frac{1}{|D_s|} \sum_{y_{\text{seen}}} \sum_{k=1}^{M} y_k \log(p(y_k|x)) \quad (4)$$

Simultaneously, we seek to decorrelate pairwise the token embeddings using $L_{CRP}$ as,

$$L_{CRP} = \min_{B_o, \{h_m\}} \frac{1}{|D_s|} \sum_{y_{\text{seen}}} \sum_{k=1}^{M} y_k \log(p(y_k|x)) \quad (5)$$

Hence, the total loss ($L_{total}$) is computed as:

$$L_{total} = \min_{B_o, \{h_m\}} [L_{ce} + \lambda \cdot L_{CRP}] \quad (6)$$

Where $\lambda$ is the weighting hyper-parameter. In the inference stage, we compute the cosine similarity between the images $x_1 \in D_t$ and prompt embeddings for all classes in $y_{\text{unseen}}$. The class with a high probability value is selected.

$$\hat{y}_t = \max_{y \in y_{\text{unseen}}} p(y|x_t) \quad (7)$$

### 4. Experimental evaluations

#### Dataset descriptions:

Our experimental evaluation involves four datasets: PatternNet [25], RSICD [32], RESISC45 [5], and MLRSNet [38].

PatternNet comprises 38 classes, with each class containing 800 images of size 256 × 256 pixels. RSICD includes 30 classes and a total of 10,000 images, each with a size of 224 × 224 pixels. Notably, each class has a different number of images.

RESISC45 consists of 45 classes, with each class containing 700 images of size 256 × 256 pixels. MLRSNet comprises 46 classes and a total of 109,161 images, each with a size of 256 × 256 pixels.

Furthermore, we extend our work to generate learnable prompts in the single-source multi-target domain generalization setup. In this regard, we curate new versions (v2) of the above-mentioned datasets, where we consider the 16 overlapping classes from all four datasets. Details are mentioned in the supplementary text.

**Architecture Details:** In all our experiments, $B_o$ comprises two attention modules, each followed by a linear layer. Our attention module is inspired by SE-Net [17] and has two linear layers, each followed by ReLU and Sigmoid activation functions, respectively. However, we can accommodate more attention blocks in $B$, if required. Further, each $h_m$ is designed as a single dense layer, which converts $B_o(x)$ into dimensions equal to the text embeddings.

**Training and evaluation protocols:** We train APPLeNet for 50 epochs using the stochastic gradient descent (SGD) optimizer [41] with an initial learning rate of $2e^{-4}$ and a warm-up fixed learning rate of $1e^{-7}$ during the first epoch to prevent explosive gradients. We keep ViT-B/16 as the image encoder backbone, use 16 training samples (i.e., shots) from each class, and create a batch size of 4 and $\lambda$ (Eq. 6) to be 0.1 for model training. We initialize the text prompts from the embeddings of “a photo of a [CLS]" which means the context length is four. This follows the previous literature [71, 72]. We execute the model using three seeds and report the average top-1 accuracy.

### 4.1. Comparison with the state-of-the-art methods

In this section, we discuss the performance of APPLeNet with respect to the methods from the literature for the three DG tasks, as mentioned: i) **Base-to-new class generalization**, where the training and test classes are disjoint. ii) **Cross-dataset generalization**, where the model is trained on one dataset and evaluated on novel datasets with domain and label shifts. iii) **Single source multi-target DG**, where the model is trained on a source domain and evaluated on multiple novel domains under the closed-set setting.

**Baselines:** We evaluated the performance of APPLeNet compared to existing methods from the prompting literature using CLIP. As a baseline, we used Zero-shot CLIP [40]. In addition, we explored other approaches, such as ERM [56], which involves a trainable linear model on top of CLIP. We also tested a state-of-the-art DA technique, DANN [11], in combination with the CLIP features. Finally, we exam-
Table 1. Comparison of APPLenNet with state-of-the-art methods for base-to-new (B2N) class generalization task. We indicate the validation accuracy for the Base and New classes. H denotes the harmonic mean used to generalize the trade-off performance between the base and new classes. Best results are shown in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>PatternNet</th>
<th>RSICD</th>
<th>RESISC45</th>
<th>MLRSNet</th>
<th>Avg. of all</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base</td>
<td>New</td>
<td>H</td>
<td>Base</td>
<td>New</td>
</tr>
<tr>
<td>CLIP [40]</td>
<td>63.67</td>
<td>64.37</td>
<td>64.02</td>
<td>54.61</td>
<td>55.33</td>
</tr>
<tr>
<td>CoOp [72]</td>
<td>91.62</td>
<td>62.23</td>
<td>74.12</td>
<td>92.52</td>
<td>56.08</td>
</tr>
<tr>
<td>CLIP-Adapter [12]</td>
<td>82.15</td>
<td>63.26</td>
<td>71.48</td>
<td>78.93</td>
<td>55.44</td>
</tr>
<tr>
<td>CoCoOp [71]</td>
<td>92.39</td>
<td>63.34</td>
<td>75.16</td>
<td>93.18</td>
<td>58.67</td>
</tr>
<tr>
<td>ProGrad [75]</td>
<td>92.65</td>
<td>62.48</td>
<td>74.63</td>
<td>93.44</td>
<td>58.15</td>
</tr>
<tr>
<td>APPLenNet</td>
<td>94.89</td>
<td>65.57</td>
<td>72.55</td>
<td>95.26</td>
<td>60.71</td>
</tr>
</tbody>
</table>

Table 2. Comparison of APPLenNet with state-of-the-art methods for cross-dataset generalization with PatternNet dataset as the source domain and remaining RS datasets as the target domains. We use the accuracy metric as the performance measure. Best results are shown in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PatternNet</td>
<td>RSICD</td>
</tr>
<tr>
<td>CLIP [40]</td>
<td>61.72</td>
<td>43.25</td>
</tr>
<tr>
<td>CoOp [72]</td>
<td>85.23</td>
<td>42.53</td>
</tr>
<tr>
<td>CLIP-Adapter [12]</td>
<td>74.27</td>
<td>42.57</td>
</tr>
<tr>
<td>CoCoOp [71]</td>
<td>85.95</td>
<td>43.61</td>
</tr>
<tr>
<td>ProGrad [75]</td>
<td>86.14</td>
<td>41.25</td>
</tr>
<tr>
<td>APPLenNet</td>
<td>88.17</td>
<td>44.87</td>
</tr>
</tbody>
</table>

We tested the generalization performance of our proposed APPLenNet on a Single-Source Multi-Target (SSMT) DG setup. Unlike the CD setting discussed earlier, we only considered the common classes across all datasets since SSMT is a closed-set setting. We trained the model on the PatternNetv2 dataset and evaluated it on the remaining datasets. The comparison results with the state-of-the-art (SOTA) methods and APPLenNet are presented in Table 3.

The results show that ProGrad outperformed other referred prompting techniques by at least 0.6%, while APPLenNet surpassed all of them by a minimum margin of 2.4% on the MLRSNetv2, 1.6% on RSICDv2, and 1.4% on RESISC45v2 (target domains), respectively. APPLenNet beats CoCoOp in beach, forest, and river classes with a huge margin of 5.5%, 4.8% and 5.9% on average over the target datasets. Notably, APPLenNet effectively transferred the learned classification information from the PatternNetv2 to classes such as desert, sparse residential, and river to the RESISC45v2 and outperformed the SOTA methods by at least 5.7%.

Regarding the source domain classification task, APPLenNet achieved a performance of 88.17%, which was better than the second-best by 2.03%.

4.2. Ablation analysis

t-SNE visualization: In Figure 3, we present a t-SNE visualization of the image embeddings generated by APPLenNet and compare them with CoCoOp [71] on the MLRSNetv2 dataset for the SSMT-DG task. The visualization clearly demonstrates that APPLenNet can accurately cluster each class, while the cluster points of many classes get overlapped in CoCoOp. This confirms the discriminability of APPLenNet.
Table 3. Comparing APPLeNet with state-of-the-art methods for single-source multi-target domain generalization on our released (2nd version) benchmark RS datasets. We use the accuracy metric as the performance measure. Best results are shown in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Source PatternNetv2</th>
<th>Source RSICDv2</th>
<th>Source RESISC45v2</th>
<th>Source MLRSNetv2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERM [56]</td>
<td>73.09</td>
<td>66.40</td>
<td>61.59</td>
<td>61.13</td>
</tr>
<tr>
<td>CLIP [40]</td>
<td>78.04</td>
<td>72.15</td>
<td>75.42</td>
<td>67.78</td>
</tr>
<tr>
<td>DANN [11]</td>
<td>93.56</td>
<td>75.49</td>
<td>76.18</td>
<td>70.53</td>
</tr>
<tr>
<td>CoOp [72]</td>
<td>94.25</td>
<td>76.80</td>
<td>77.87</td>
<td>70.97</td>
</tr>
<tr>
<td>CLIP-Adapter [12]</td>
<td>92.36</td>
<td>79.33</td>
<td>80.43</td>
<td>71.67</td>
</tr>
<tr>
<td>CoCoOp [71]</td>
<td>94.41</td>
<td>79.33</td>
<td>80.43</td>
<td>71.67</td>
</tr>
<tr>
<td>ProGrad [75]</td>
<td>95.18</td>
<td>77.46</td>
<td>80.65</td>
<td>72.29</td>
</tr>
<tr>
<td>APPLeNet</td>
<td>96.63</td>
<td>81.03</td>
<td>82.23</td>
<td>74.03</td>
</tr>
</tbody>
</table>

Figure 3. t-SNE plots [55] for the image feature extracted from the Meta-Net of CoCoOp and the Injection Block (IB) of APPLeNet for the SSMT domain generalization task on the MLRSNetv2 dataset. The legends represent the class labels.

Sensitivity to the variation in the number of shots: We evaluate the performance of our proposed APPLeNet by varying the number of shots from 1 to 32 for the B2N class generalization task and compare it with the state-of-the-art (SOTA) prompting techniques, as shown in Table 4. In this setting, we use a context length of 4, place the class token at the end, use ViT-B/16 as the visual feature backbone, and use a unified context vector. As CLIP is a zero-shot approach, we exclude it and only consider few-shot-based prompting methods to compare and show results on the PatternNet dataset. We are able to outperform the benchmark prompt learning-based methods by at least 0.8%, 2.4%, and 1.6% for 8, 16, and 32 shots, respectively.

Table 4. Comparison of APPLeNet with state-of-the-art methods on varying the number of shots for the B2N class generalization task with PatternNet dataset. Harmonic mean (H) of base and new classes is considered for comparison, as well as to depict the generalization trade-off. Best results are shown in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>1-shot</th>
<th>4-shots</th>
<th>8-shots</th>
<th>16-shots</th>
<th>32-shots</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoOp [72]</td>
<td>70.33</td>
<td>71.61</td>
<td>72.17</td>
<td>74.12</td>
<td>74.58</td>
</tr>
<tr>
<td>CLIP-Adapter [12]</td>
<td>69.75</td>
<td>69.95</td>
<td>70.37</td>
<td>71.48</td>
<td>71.64</td>
</tr>
<tr>
<td>CoCoOp [71]</td>
<td>71.85</td>
<td>73.61</td>
<td>74.53</td>
<td>75.16</td>
<td>74.39</td>
</tr>
<tr>
<td>ProGrad [75]</td>
<td>73.67</td>
<td>72.05</td>
<td>73.16</td>
<td>74.63</td>
<td>75.56</td>
</tr>
<tr>
<td>APPLeNet</td>
<td>72.44</td>
<td>72.46</td>
<td>75.28</td>
<td>77.55</td>
<td>77.13</td>
</tr>
</tbody>
</table>

Sensitivity to the position of the class token and the prompt initialization strategy: In this experiment, we investigate the effect of the position of the class token in the learnable context vectors in [1, M] on the performance of APPLeNet in the B2N class generalization task. We experiment with three different positions for the class token: "front", "middle", and "end", while generating the learnable prompts.

We plot the harmonic mean between the base and new classes for all the RS datasets in Figure 4. Our results show that positioning the class token at the "end" of the context vector consistently improves the performance of APPLeNet.

Table 5. Ablation of different types of initialization of context vectors in APPLeNet. Harmonic mean (H) of base and new classes are considered for comparison. Best results are shown in bold.

<table>
<thead>
<tr>
<th>Context Vectors</th>
<th>PatternNet</th>
<th>RSICD</th>
<th>RESISC45</th>
<th>MLRSNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>manual initialization</td>
<td>77.55</td>
<td>74.16</td>
<td>72.73</td>
<td>65.66</td>
</tr>
<tr>
<td>random initialization</td>
<td>81.61</td>
<td>68.58</td>
<td>70.80</td>
<td>61.45</td>
</tr>
<tr>
<td>no initialization</td>
<td>67.90</td>
<td>69.57</td>
<td>69.42</td>
<td>59.16</td>
</tr>
</tbody>
</table>
on the B2N class generalization task, with at least a 3.9%, 2.3%, and 2.5% improvement on PatternNet, RESISC45, and MLRSNet datasets, respectively, compared to positioning the token at the “front” or “middle”. However, on the RSICD dataset, we observe no significant difference in performance for the different class token positions.

Finally, we consider three different prompt initialization strategy to check their efficacy in Table 5 for B2N Generalization. It highlights that manual initialization from "a photo of a" outperforms the random initialization and no initialization strategies significantly for the target datasets. Interestingly, the random initialization outperforms other on the source domain evaluation.

Sensitivity analysis of APPLeNet to context lengths ($\mathcal{M}$): During test-time prompt generation, we varied the context length ($\mathcal{M}$) and experimented with four different context lengths: 1, 4, 8, and 16. To maintain consistency, we initialized the tokens randomly. Our results, illustrated in Figure 5, show that APPLeNet outperforms the respective state-of-the-art methods by 0.7%, 2.4%, and 0.4% for context lengths 1, 4, and 16, respectively. Additionally, we found that APPLeNet achieved the best performance with $\mathcal{M} = 4$ among all the context length settings.

Sensitivity to the multi-scale features: In this study, we aimed to assess the sensitivity of APPLeNet to visual content features obtained from multiple layers of $f_v$. We utilized two CLIP vision backbones based on ResNet-50 and ViT and increased the number of feature layers in calculating $\hat{F}(x)$. As shown in Figure 6, incorporating more visual embedding layers to extract content features yielded improved performance, with a monotonically increasing trend. It is worth noting that all experiments included consideration of the style feature $\mu$.

To further highlight the importance of the injection block for intelligent multi-scale feature aggregation, we compare APPLeNet with the multi-scale version of CoCoOp [71]. Specifically, we passed $\hat{F}(x)$ to the meta-network ($\pi$) to devise the Multi-Scale (MS) - CoCoOp (see Table 6). For APPLeNet, we considered three variants where we only pass the multi-scale content features $\hat{F}(x)$, the style features $\mu$, and $F(x)$ to the injection block $B$. Our results clearly demonstrate that our multi-scale feature aggregation approach outperforms MS-CoCoOp significantly. Additionally, the results highlight the benefits of considering style primitives.

Table 6. Ablation of multi-scale features’ sensitivity in CoCoOp and APPLeNet. Harmonic mean (H) of base and new classes are considered for comparison. Best results are shown in bold.

<table>
<thead>
<tr>
<th>Context Vectors</th>
<th>PatternNet</th>
<th>RSICD</th>
<th>RESISC45</th>
<th>MLRSNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS-CoCoOp</td>
<td>75.83</td>
<td>72.31</td>
<td>69.92</td>
<td>62.64</td>
</tr>
<tr>
<td>APPLeNet (with MS)</td>
<td>77.34</td>
<td>73.96</td>
<td>72.51</td>
<td>65.02</td>
</tr>
<tr>
<td>APPLeNet (with $\mu$)</td>
<td>76.04</td>
<td>72.19</td>
<td>69.53</td>
<td>63.95</td>
</tr>
<tr>
<td>APPLeNet (with MS &amp; $\mu$)</td>
<td>77.55</td>
<td>74.16</td>
<td>72.73</td>
<td>65.66</td>
</tr>
</tbody>
</table>

Effect of CRP loss ($L_{CRP}$): Table 7 shows the results of ablating $L_{CRP}$ in Equation 6 over two loss functions, namely CRP loss ($L_{CRP}$) and cross-entropy loss ($L_{CE}$). Interestingly, we observed that APPLeNet model achieved an average improvement of approximately 1 – 3% across all datasets on the B2N generalization task in the presence of $L_{CRP}$. This result justifies the significant role of $L_{CRP}$ in ensuring the distinctiveness in $[c_1^r, \ldots, c_M^r]$ so that they do not convey redundant information.

Table 7. Ablation of APPLeNet with and without CRP ($L_{CRP}$) loss in Equation 6. Harmonic mean (H) of base and new classes are considered for comparison. Best results are shown in bold.

<table>
<thead>
<tr>
<th></th>
<th>PatternNet</th>
<th>RSICD</th>
<th>RESISC45</th>
<th>MLRSNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>without $L_{CRP}$</td>
<td>75.34</td>
<td>72.89</td>
<td>71.63</td>
<td>62.15</td>
</tr>
<tr>
<td>with $L_{CRP}$</td>
<td>77.55</td>
<td>74.16</td>
<td>72.73</td>
<td>65.66</td>
</tr>
</tbody>
</table>

Ablation with number of attention modules: We conducted an ablation study on the injection block (IB) of our APPLeNet by varying the number of attention modules (AMs) for the single-source multi-target domain generalization task. The results are presented in Table 8. We found that APPLeNet with three AMs outperformed the others on the source domain (PatternNetv2 dataset). However, IB with two AMs reported the best performance for the target domains, with at least a numerical improvement of 0.1%. It is possible that IB with three AMs suffers from a vanishing gradient problem due to its multiple sigmoid output layers and overfits the data.

Table 8. Analysis of the number of attention modules (AMs) used in the injection block for single-source multi-target domain generalization task. We use accuracy metrics as the performance measure. Best results are shown in bold.

<table>
<thead>
<tr>
<th>No. of AMs</th>
<th>Source</th>
<th>PatternNetv2</th>
<th>RSICDv2</th>
<th>RESISC45v2</th>
<th>MLRSNetv2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>93.33</td>
<td>76.81</td>
<td>80.92</td>
<td>72.60</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>95.94</td>
<td>79.61</td>
<td>81.42</td>
<td>73.56</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>96.63</td>
<td>81.03</td>
<td>82.33</td>
<td>74.03</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>96.77</td>
<td>78.85</td>
<td>78.21</td>
<td>73.92</td>
<td></td>
</tr>
</tbody>
</table>

5. Takeaways

This paper presents a novel approach, APPLeNet, for prompt learning in CLIP based foundation model for solving three challenging DG tasks in RS. We acknowledge the challenges associated with processing remote sensing scenes, and thus, we propose leveraging the frozen vision backbone of CLIP to generate multi-scale visual content features and batch statistics to generate style properties automatically. We combine visual and learnable text tokens for prompt learning, but since adding visual information can introduce redundancy, we present an anti-correlation regularizer to ensure token distinctiveness.

Our study is the first to extensively evaluate the DG paradigm in remote sensing, and we introduce new benchmarks with comprehensive experimentation. We hope our findings will inspire further research on foundational models for remote sensing applications.
References


