APPLeNet: Visual Attention Parameterized Prompt Learning for Few-Shot Remote Sensing Image Generalization using CLIP

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, ankitjha16, bssiitb, shirshabosecs, 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 (APPLeNet). APPLeNet 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, APPLeNet 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 APPLeNet, 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/APPLeNet`

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.

Figure 1. Performance overview of the proposed APPLeNet compared to state-of-the-art CLIP-based methods. It is shown that APPLeNet has better generalization capability irrespective of the complexity and size of the datasets in base-to-new class and across-domain generalization tasks.

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

*equal contribution

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-

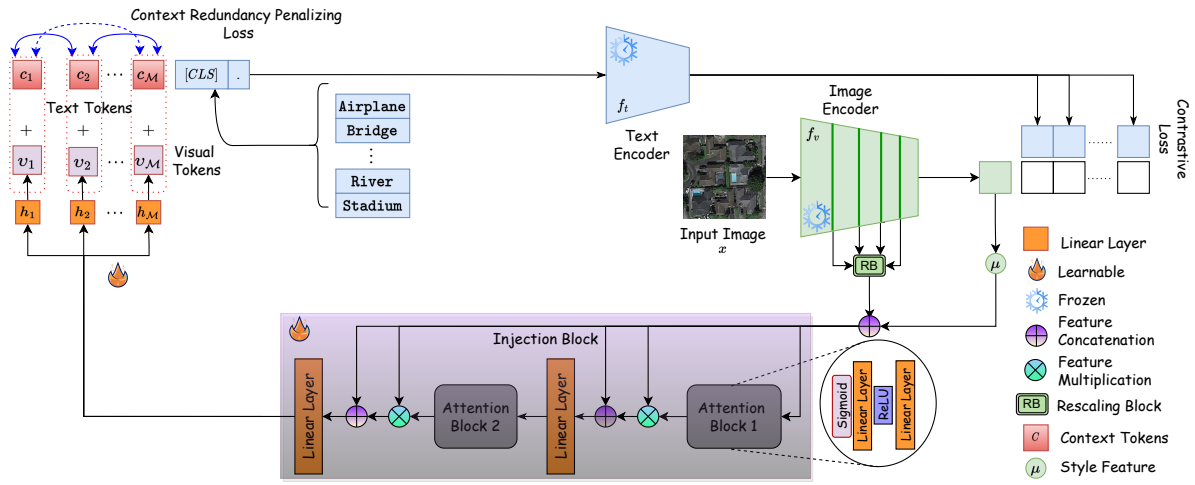

Figure 2. APPLeNet is composed of a text encoder f_t, an image encoder f_v, and an injection block designed for multi-scale visual feature refinement. The f_v produces multi-level visual content features, and the batch statistics μ for a domain as the style features, that are passed through a residual attention-based injection block. These features are then sent to individual projector networks $\{h_m\}_{m=1}^{\mathcal{M}}$ to derive the visual tokens $\{v_m\}_{m=1}^{\mathcal{M}}$. The visual tokens are added to the learnable text token embeddings $\{c_m\}_{m=1}^{\mathcal{M}}$ 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 ($\mathbf{L_{CRP}}$) among the context vectors in $\{c_m\}_{m=1}^{\mathcal{M}}$. The model is trained using a multi-task loss that comprises the contrastive loss between the image and prompt embeddings and the CRP regularizer.

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 $\mathcal{D}_s = \{\mathcal{D}_s^i\}_{i=1}^n$ 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, $P(\mathcal{D}_s^i)$, may differ for all

$i \in 1, \cdots, n$. During training, we use the labels $\mathcal{Y}_{Seen}$ from $\mathcal{D}_s$, while during testing, we use $\mathcal{Y}_{Unseen}$ from a target test domain $\mathcal{D}_t$ with $P(\mathcal{D}_t) \neq P(\mathcal{D}_s^i), \forall i \in 1, \cdots, n$. For base-to-new class generalization, we set $\mathcal{Y}_{Seen} \cap \mathcal{Y}_{Unseen} = \emptyset$. In contrast, for domain generalization (DG), we consider single-source DG and assume that the label sets for both domains are identical ($\mathcal{Y}_{Seen} \cap \mathcal{Y}_{Unseen} = \mathcal{Y}_{Seen} \cup \mathcal{Y}_{Unseen}$).

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 $\mathcal{M}$ learnable context vectors $c_1, c_2, \cdots, c_{\mathcal{M}}$ that have the same dimensionality as the word embeddings and optimizing them through back-propagation. It's important to note that $\mathcal{M}$ is a hyperparameter that determines the context length and may differ between tasks. For any given class y, the prompt can be represented as $t_y = \{[c_1], [c_2], \cdots, [c_{\mathcal{M}}], [CLS_y]\}$.

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 $\mathcal{M}$ meta-tokens, denoted as π. These meta-tokens are concatenated with context vectors to create prompts $t_y = \{[c_1(x)], [c_2(x)], \cdots, [c_{\mathcal{M}}(x)], [CLS_y]\}$, 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 μ 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 $\mathcal{B}_\phi$, with learnable parameters ϕ. This block emphasizes concatenated embeddings of style and multi-level visual features. Moreover, APPLeNet comprises $\mathcal{M}$ learnable linear layers that generate $\mathcal{M}$ visual tokens $\{v_m\}_{m=1}^{\mathcal{M}}$ given the outputs from $\mathcal{B}$. These visual tokens are then combined with the corresponding text tokens $\{c_m\}_{m=1}^{\mathcal{M}}$ which is further appended with the class token $[CLS_y]$ to generate prompt t_y. In the following sections, we elaborate on each component of APPLeNet.

Encoding style and multi-scale content features into prompts: To incorporate the multi-scale visual features from f_v into $\mathcal{B}$, we propose using global average pooling (GAP) to collapse the spatial dimensions of each channel. This produces $\hat{f}_v^l(x) \in \mathbb{R}^{C \times 1}$, where $f_v^l \in \mathbb{R}^{W \times H \times C}$ represents the output responses from the l^{th} layer. Here, (W, H) represents the spatial dimensions of the feature maps. Using this approach, we define $\hat{F}(x) = [\hat{f}_v^1(x); \cdots; \hat{f}_v^L(x)]$ 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 $F(x) = [\hat{F}(x_i); \mu_i]$, which captures both multi-scale content and the style information.

Injection block: The attention modules within $\mathcal{B}$ are denoted by $\mathcal{A}_q(\cdot)$, where $q \in 1, \cdots, \mathcal{Q}$. For the first attention block ($q = 1$), we denote the attended output features as $\mathcal{O}_1 = F(x) \odot \mathcal{A}_1 \oplus F(x)$. These features are then fed as input to $\mathcal{A}_2$ and so on, as follows:

$$\mathcal{O}_q = \begin{cases} [F(x) \odot \mathcal{A}_q(F(x)) + F(x)], \text{if } q = 1 \\ [\mathcal{O}_{q-1} \odot \mathcal{A}_q(\mathcal{O}_{q-1}) + \mathcal{O}_{q-1}], \text{otherwise} \end{cases} \quad (1)$$

We subsequently pass $\mathcal{O}_{\mathcal{Q}}$ through $\mathcal{M}$ light-weight projector networks $\{h_m\}_{m=1}^{\mathcal{M}}$ which generate $\mathcal{M}$ visual tokens $\{v_1, \cdots, v_{\mathcal{M}}\}$: $v_m = h_m(\mathcal{O}_{\mathcal{Q}})$. 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 $c'_m = c_m + v_m$.

The generated prompt is represented as:

$$t_y = \{[v_1 + c_1], \cdots, [v_{\mathcal{M}} + c_{\mathcal{M}}], [CLS_y]\} \quad (2)$$

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 $\mathbf{L_{ce}}$, which ensures proper mapping between the visual feature representation $f_v(x)$ and the textual feature 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 $\mathbf{L_{CRP}}$. This loss ensures that the token embeddings in the set $c_1 + v_1, \cdots, c_{\mathcal{M}} + v_{\mathcal{M}}$ 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,

$$p(y|x) = \frac{\exp(\text{sim}(f_v(x), f_t(t_y(\mathcal{B}_\phi(x)))/\tau))}{\sum_{k=1}^{|\mathcal{Y}|} \exp(\text{sim}(f_v(x), f_t(t_k(\mathcal{B}_\phi(x)))/\tau))} \quad (3)$$

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

$$\mathbf{L_{ce}} = \underset{\mathcal{B}_\phi, \{h_m\}}{\arg\min} \underset{(x,y)\in\mathcal{P}(\mathcal{D}_s)}{\mathbb{E}} - \sum_{k=1}^{\mathcal{Y}_{Seen}} y_k log(p(y_k|x)) \quad (4)$$

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

$$\mathbf{L_{CRP}} = \underset{\mathcal{B}_\phi, \{h_m\}}{\arg\min} \underset{(x,y)\in\mathcal{P}(\mathcal{D}_s)}{\mathbb{E}} \left| c'_j(x) \cdot c'_l(x) - \mathcal{I} \right|, \quad \forall j, l \in \{1, 2, \cdots, \mathcal{M}\}, j \neq l, c'_j = c_j + v_j \quad (5)$$

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

$$\mathbf{L_{total}} = \underset{\mathcal{B}_\phi, \{h_m\}}{\arg\min} [\mathbf{L_{ce}} + \lambda * \mathbf{L_{CRP}}] \quad (6)$$

Where λ is the weighting hyper-parameter. In the inference stage, we compute the cosine similarity between the images $x_t \in \mathcal{D}_t$ and prompt embeddings for all the classes in $\mathcal{Y}_{Unseen}$. The class with a high probability value is selected.

$$\hat{y_t} = \underset{y\in\mathcal{Y}_{Unseen}}{\arg\max}\, p(y|x_t) \quad (7)$$

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, $\mathcal{B}_\phi$ 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 $\mathcal{B}$, if required. Further, each h_m is designed as a single dense layer, which converts $\mathcal{B}_\phi(x)$ 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 λ (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 APPLeNet 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**.

Method	PatternNet			RSICD			RESISC45			MLRSNet			Avg. of all		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CLIP [40]	63.67	64.37	64.02	54.61	55.33	54.97	56.32	55.38	55.85	51.43	51.92	51.67	56.51	56.75	56.63
CoOp [72]	91.62	62.23	74.12	92.52	56.08	69.83	89.04	55.75	68.57	75.21	53.64	62.62	87.10	56.93	68.85
CLIP-Adapter [12]	82.15	63.26	71.48	78.93	55.44	65.13	81.67	56.23	66.60	71.64	53.19	61.05	78.60	57.03	66.10
CoCoOp [71]	92.39	63.34	75.16	93.18	58.67	72.00	89.78	57.18	69.86	76.32	52.75	62.38	87.92	57.99	69.88
ProGrad [75]	92.65	62.48	74.63	93.44	58.15	71.69	90.13	57.89	70.50	75.96	52.23	61.90	88.05	57.69	69.70
APPLeNet	**94.89**	**65.57**	**77.55**	**95.26**	**60.71**	**74.16**	**91.24**	**60.46**	**72.73**	**78.53**	**56.41**	**65.66**	**89.98**	**60.79**	**72.56**

ined prompt learning techniques, including CoOp [72], CoCoOp [71], CLIP-Adapter [12], and ProGrad [75].

Base-to-New (B2N) class generalization: Table 1 presents the experimental results for B2N class generalization on the four RS datasets, where the harmonic mean (H) between the classification accuracies of the Base and New classes is computed. For all the datasets, we randomly and equally divide the datasets into two groups to define the source (with the base classes) and the target (with the novel classes) domains. Compared to the CLIP's zero-shot approach, APPLeNet achieves better generalization scores, with a considerable margin of 33.47% on seen classes and 4.04% on unseen classes over all datasets (on average). We also compare APPLeNet with referred context optimization-based methods, where it outperforms CoOp and CoCoOp on the PatternNet [25], RSICD [32], RESISC45 [5], and MLRSNet [38] datasets by $3.4\%, 4.3\%, 4.2\%$, and 3.0%, and 2.4%, 2.2%, 2.9%, and 3.3%, respectively. In PatternNet, APPLeNet consistently beats CoCoOp by huge margins of $5.8\%, 5.2\%$ and 6.4% in *river*, *storage tank* and *tennis court* classes. Among all the referred methods, only CoCoOp and ProGrad show the second and third-best performance scores on generalizing the unseen classes over all RS datasets.

Cross-Dataset (CD) generalization: Table 2 presents the results of our evaluation of APPLeNet on the CD setup. In this regard, we train the model on the PatternNet [25] dataset (source domain) and report zero-shot inference results on the remaining RS datasets (target domains). Our APPLeNet outperforms the source and target classification performance by significant margins of 26.5% and 13.9%, respectively, compared to zero-shot (CLIP) and non-learnable prompt (CLIP-Adapter) methods. Besides, APPLeNet outperforms CoCoOp by 1.3%, 1.4%, and 2.1% for the unseen RS domains, namely RSICD [32], RESISC45 [5], and MLRSNet [38] datasets, respectively. Also APPLeNet beats CoCoOp by $3.6\%, 5.2\%, 5.4\%$ and 4.7% in *desert*, *mountain*, *port* and *school* of RSICD dataset. Finally, AppLeNet is better than ProGrad than at least 2.5% on all the target tasks. Based on these results, our results establish that APPLeNet successfully narrows the generalization gap between a single source and multiple targets with domain and label shifts in the CD transfer technique.

Single source multi-target domain generalization (DG):

Table 2. Comparison of APPLeNet 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**.

Method	Source	Target		
	PatternNet	RSICD	RESISC45	MLRSNet
CLIP [40]	61.72	43.25	48.56	45.13
CoOp [72]	85.23	42.53	49.34	44.50
CLIP-Adapter [12]	74.27	42.57	49.07	44.17
CoCoOp [71]	85.95	43.61	49.53	44.72
ProGrad [75]	86.14	41.25	48.26	44.12
APPLeNet	**88.17**	**44.87**	**50.97**	**46.83**

We tested the generalization performance of our proposed APPLeNet 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 APPLeNet are presented in Table 3.

The results show that ProGrad outperformed other referenced prompting techniques by at least 0.6%, while APPLeNet 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. APPLeNet 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, APPLeNet 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, APPLeNet 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 [54] visualization of the image embeddings generated by APPLeNet and compare them with CoCoOp [71] on the MLRSNetv2 dataset for the SSMT-DG task. The visualization clearly demonstrates that APPLeNet can accurately cluster each class, while the cluster points of many classes get overlapped in CoCoOp. This confirms the discriminability of APPLeNet.

Table 3. Comparing APPLeNet with state-of-the-art methods for single-source multi-target domain generalization on our released (2^{nd} version) benchmark RS datasets. We use the accuracy metric as the performance measure. Best results are shown in **bold**.

Method	Source	Target		
	PatternNetv2	RSICDv2	RESISC45v2	MLRSNetv2
ERM [56]	73.69	61.40	61.59	61.13
CLIP [40]	78.04	72.15	75.42	67.78
DANN [11]	93.56	75.49	76.18	70.53
CoOp [72]	94.25	76.50	77.87	70.97
CLIP-Adapter [12]	92.36	79.17	79.76	71.04
CoCoOp [71]	94.41	79.33	80.43	71.67
ProGrad [75]	95.18	77.46	80.65	72.29
APPLeNet	**96.63**	**81.03**	**82.23**	**74.03**

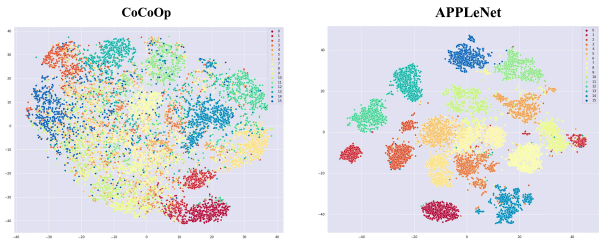

Figure 3. t-SNE plots [55] for the image feature extracted from the Meta-Net of CoCoOp and the Injection Block (IB) of APPLeNet for the SSMT domain generalization task on the MLRSNetv2 dataset. The legends represent the class labels.

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**.

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

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 $[1, \mathcal{M}]$ 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.

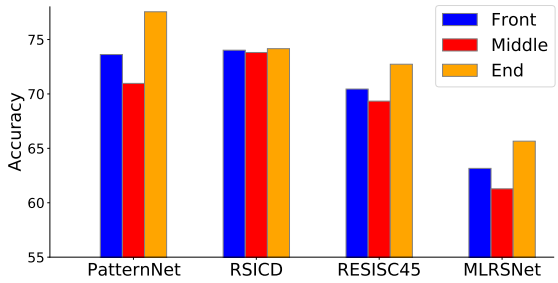

Figure 4. Classification performance on changing position of the class tokens in APPLeNet, i.e., 'Front','Middle', and 'End' for the B2N class generalization task on the four RS datasets. We consider the harmonic mean (H) of base and new classes for comparison.

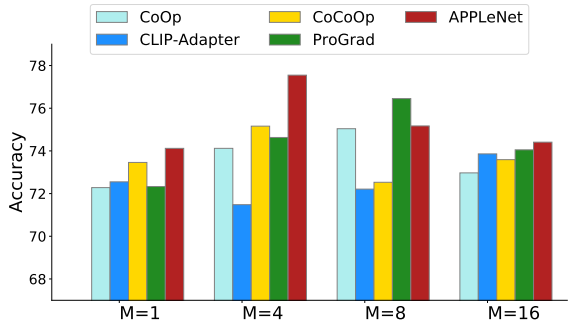

Figure 5. Classification performance of APPLeNet by varying the context length (M) for the B2N class generalization on PatternNet dataset and compared with the SOTA methods. The harmonic mean (H) of base and new classes are considered for comparison.

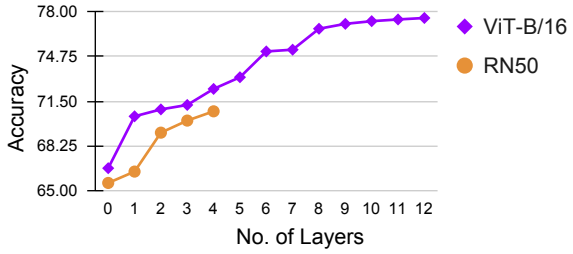

Figure 6. Performance of APPLeNet with different layers of ViT-B/16 and RN50 backbones to extract multi-scale features on PatterNet dataset. Harmonic mean (H) of base and new classes are shown as accuracy.

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**.

Context Vectors	PatternNet	RSICD	RESISC45	MLRSNet
manual initialization	77.55	**74.16**	**72.73**	**65.66**
random initialization	**81.61**	68.58	70.80	61.45
no initialization	67.90	69.57	69.42	59.16

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

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 μ.

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 (π) 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 μ, and $F(x)$ to the injection block $\mathcal{B}$. 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**.

Context Vectors	PatternNet	RSICD	RESISC45	MLRSNet
MS-CoCoOp	75.83	72.31	69.92	62.64
APPLeNet (with MS)	77.34	73.96	72.51	65.02
APPLeNet (with μ)	76.04	72.19	69.53	63.95
APPLeNet (with MS & μ)	**77.55**	**74.16**	**72.73**	**65.66**

Effect of CRP loss ($\mathbf{L_{CRP}}$): Table 7 shows the results of ablating $\mathbf{L_{CRP}}$ in Equation 6 over two loss functions, namely CRP loss ($\mathbf{L_{CRP}}$) and cross-entropy loss ($\mathbf{L_{ce}}$). Interestingly, we observed that our APPLeNet model achieved an average improvement of approximately $1-3\%$ across all datasets on the B2N generalization task in the presence of $\mathbf{L_{CRP}}$. This result justifies the significant role of $\mathbf{L_{CRP}}$ in ensuring the distinctiveness in $[c'_1, \cdots, c'_{\mathcal{M}}]$ so that they do not convey redundant information.

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

APPLeNet	PatternNet	RSICD	RESISC45	MLRSNet
without $\mathbf{L_{CRP}}$	75.34	72.89	71.63	62.15
with $\mathbf{L_{CRP}}$	**77.55**	**74.16**	**72.73**	**65.66**

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**.

	Source	Target		
No. of AMs	PatternNetv2	RSICDv2	RESISC45v2	MLRSNetv2
0	93.33	76.81	80.92	72.60
1	95.94	79.61	81.42	73.56
2	96.63	**81.03**	**82.23**	**74.03**
3	**96.77**	78.85	78.21	73.92

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

[1] Dalal Alajaji, Haikel S. Alhichri, Nassim Ammour, and Naif Alajlan. Few-shot learning for remote sensing scene classification. In *2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, pages 81–84, 2020.

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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

[4] Mingyang Chen, Wen Zhang, Wei Zhang, Qiang Chen, and Huajun Chen. Meta relational learning for few-shot link prediction in knowledge graphs. *arXiv preprint arXiv:1909.01515*, 2019.

[5] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

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

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[8] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in data science and information engineering*, pages 877–894, 2021.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

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

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[12] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

[13] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Miyuki Hino, Elinor Benami, and Nina Brooks. Machine learning for environmental monitoring. *Nature Sustainability*, 1(10):583–588, 2018.

[16] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[18] Hong Ji, Zhi Gao, Yongjun Zhang, Yu Wan, Can Li, and Tiancan Mei. Few-shot scene classification of optical remote sensing images leveraging calibrated pretext tasks. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[19] 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 Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 18–24 Jul 2021.

[20] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8484–8493, 2020.

[21] Juwon Kang, Sohyun Lee, Namyup Kim, and Suha Kwak. Style neophile: Constantly seeking novel styles for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7130–7140, 2022.

[22] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *arXiv preprint arXiv:2210.03117*, 2022.

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

[24] Haifeng Li, Zhenqi Cui, Zhiqing Zhu, Li Chen, Jiawei Zhu, Haozhe Huang, and Chao Tao. Rs-metanet: Deep meta metric learning for few-shot remote sensing scene classification. *CoRR*, abs/2009.13364, 2020.

[25] Hongzhi Li, Joseph G Ellis, Lei Zhang, and Shih-Fu Chang. Patternnet: Visual pattern mining with deep neural network. In *Proceedings of the 2018 ACM on international conference on multimedia retrieval*, pages 291–299, 2018.

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

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

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

[29] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

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

[31] Bing Liu, Xuchu Yu, Anzhu Yu, Pengqiang Zhang, Gang Wan, and Ruirui Wang. Deep few-shot learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(4):2290–2304, 2018.

[32] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017.

[33] Massimiliano Mancini, Samuel Rota Bulo, Barbara Caputo, and Elisa Ricci. Best sources forward: domain generalization through source-specific nets. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 1353–1357, 2018.

[34] Debabrata Pal, Valay Bundele, Biplab Banerjee, and Yogananda Jeppu. Spn: Stable prototypical network for few-shot learning-based hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

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

[36] Claudio Persello and Lorenzo Bruzzone. Relevant and invariant feature selection of hyperspectral images for domain generalization. In *2014 IEEE Geoscience and Remote Sensing Symposium*, pages 3562–3565. IEEE, 2014.

[37] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[38] Xiaoman Qi, Panpan Zhu, Yuebin Wang, Liqiang Zhang, Junhuan Peng, Mengfan Wu, Jialong Chen, Xudong Zhao, Ning Zang, and P Takis Mathiopoulos. Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:337–350, 2020.

[39] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020.

[40] 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.

[41] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[42] Marc Russwurm, Sherrie Wang, Marco Korner, and David Lobell. Meta-learning for few-shot land cover classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[43] Marc Rußwurm, Sherrie Wang, Marco Körner, and David B. Lobell. Meta-learning for few-shot land cover classification. *CoRR*, abs/2004.13390, 2020.

[44] Floyd F Sabins. Remote sensing for mineral exploration. *Ore geology reviews*, 14(3-4):157–183, 1999.

[45] Sudipan Saha, Shan Zhao, and Xiao Xiang Zhu. Multitarget domain adaptation for remote sensing classification using graph neural network. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[46] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

[47] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *arXiv preprint arXiv:2209.07511*, 2022.

[48] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[49] Shaoyue Song, Hongkai Yu, Zhenjiang Miao, Qiang Zhang, Yuewei Lin, and Song Wang. Domain adaptation for convolutional neural networks-based remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 16(8):1324–1328, 2019.

[50] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019.

[51] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[52] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE geoscience and remote sensing magazine*, 4(2):41–57, 2016.

[53] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Recent advances in domain adaptation for the classification of remote sensing data. *arXiv preprint arXiv:2104.07778*, 2021.

[54] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[55] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[56] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[57] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

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

[59] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 834–843, 2021.

[60] Chen Wu, Liangpei Zhang, and Bo Du. Kernel slow feature analysis for scene change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 55(4):2367–2384, 2017.

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

[62] Zhaohui Xue, Yiyang Zhou, and Peijun Du. S3net: Spectral–spatial siamese network for few-shot hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022.

[63] Zhongjie Yu, Lin Chen, Zhongwei Cheng, and Jiebo Luo. Transmatch: A transfer-learning scheme for semi-supervised few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12856–12864, 2020.

[64] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[65] Qingjie Zeng and Jie Geng. Task-specific contrastive learning for few-shot remote sensing image scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 191:143–154, 2022.

[66] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022.

[67] Ce Zhang, Isabel Sargent, Xin Pan, Huapeng Li, Andy Gardiner, Jonathon Hare, and Peter M Atkinson. Joint deep learning for land cover and land use classification. *Remote sensing of environment*, 221:173–187, 2019.

[68] Pei Zhang, Yunpeng Bai, Dong Wang, Bendu Bai, and Ying Li. Few-shot classification of aerial scene images via meta-learning. *Remote Sensing*, 13(1):108, 2021.

[69] Juepeng Zheng, Wenzhao Wu, Shuai Yuan, Haohuan Fu, Weijia Li, and Le Yu. Multisource-domain generalization-based oil palm tree detection using very-high-resolution (vhr) satellite images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[70] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[71] 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.

[72] 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.

[73] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pages 561–578. Springer, 2020.

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

[75] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*, 2022.