# Supplementary for 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

## 1. Introduction

- We present the detailed description of the datasets used in validating the proposed novel APPLeNet, in section 2. Furthermore, we present the curated dataset for domain generalization (DG) setup.
- In section 3, we first discuss the performance of our proposed APPLeNet and the referred SOTA prompting techniques by changing the visual backbone of the pre-trained CLIP. Secondly, we study the effect of the types of context vectors, i.e., unified context (UC) and class-specific context (CSC) in the APPLeNet.
- In Figure 1, we show the tSNE [7] plots for the outputs of the meta-net of the CoCoOp and injection block (IB) of the APPLeNet on the three mentioned benchmark RS datasets.

## 2. Datasets

We experiment with the proposed APPLeNet on four different remote sensing benchmark datasets; PatternNet [3], RSICD [4], RESISC45 [1], and MLRSNet [5]. The detailed descriptions are as follows:

**PatternNet** [3] includes 38 classes and each class has 800 images of size $256 \times 256$ pixels. The images are large-scale high-resolution images collected from Google Earth imagery based on US cities for remote sensing image retrieval.

**Remote Sensing Image Captioning Dataset (RSICD)** [4] includes 30 classes and total number of $10,000$ images of size $224 \times 224$ pixels. Each class has a different number of images. This dataset also has five sentence descriptions per image, usually used for auto-image captioning applications. Nevertheless, here we have used only the images, as the captions are learnable in our approach.

**Remote Sensing Image Scene Classification (RESISC45)** [1] dataset includes 45 classes and each class has 700 images of size $256 \times 256$ pixels. The spatial resolution of its images varies largely, ranging from 20 cm to more than 30 m.

**Multi-label High Spatial Resolution Remote Sensing Dataset (MLRSNet)** [5] includes 46 classes and a total number of $109,161$ images of size $256 \times 256$ pixels. Each class has around 2000 images and varying spatial resolution from 0.1-1m. This dataset is mainly used for image retrieval, segmentation, and classification.

We also extend our work on generating learnable prompts on the single-source multi-target domain generalization setup. To do so, we release the new version of the above-stated datasets.

**Table 1.** Details of datasets, used for B2N, CD, SSMT Domain Generalization techniques.

| Dataset for B2N and CD | Details | Dataset for SSMT-DG | Details |
|---|---|---|---|
| PatternNet [3] | 38 classes and 30.4K total images | PatternNetv2 | 16 common classes: baseball, beach,bridge dense residential area, desert, field, forest, harbor, intersection, meadow, overpass, parking, railway, river, sparse residential area, stadium and storage tank |
| RSICD [4] | 30 classes and 10K total images | RSICDv2 | |
| RESISC45 [1] | 45 classes and 31.5K total images | RESISC45v2 | |
| MLRSNet [5] | 46 classes and 109K total images | MLRSNetv2 | |

**Table 2.** Ablation of APPLeNet with different context vectors on the PatternNet and RSICD datasets. H denotes the harmonic mean used to generalize the trade-off performance between the base and new classes. Best results are shown in **bold**.

| | PatternNet | | | RSICD | | |
|---|---|---|---|---|---|---|
| Context | Base | New | H | Base | New | H |
| UC | **94.89** | **65.57** | **77.55** | **95.26** | **60.71** | **74.16** |
| CSC | 88.83 | 63.91 | 74.34 | 92.61 | 60.13 | 72.92 |

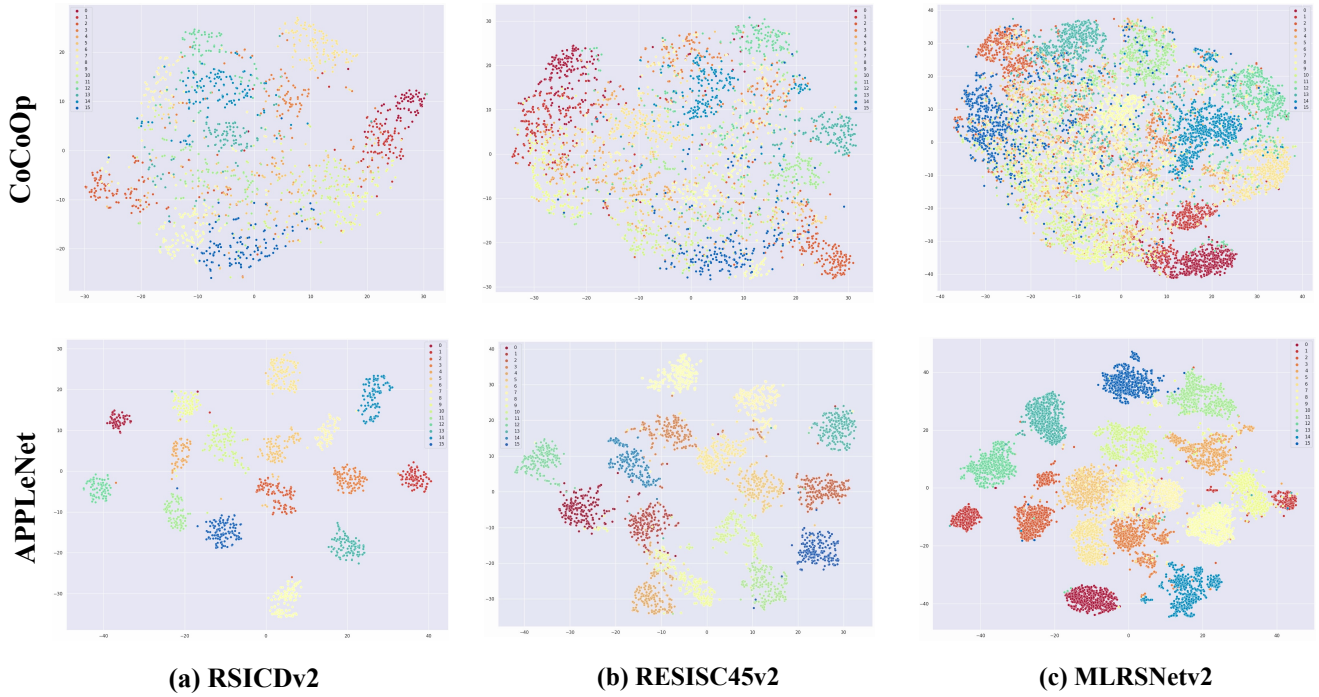
## 3. Ablation Studies

**Sensitivity to the types of context vectors:** We compare two types of context vectors used for generating learnable prompts: unified context (UC) and class-specific context

*equal contribution

**Table 3.** Comparison of APPLeNet with state-of-the-art methods on SSMT image classification task on the RS datasets. We use the accuracy metric as the performance measure. The best results are shown in **bold**.

| Method | ResNet-50 | | | | ResNet-101 | | | | ViT-B/16 | | | | ViT-B/32 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Source | Target | | | Source | Target | | | Source | Target | | | Source | Target | | |
| | PNv2 | RSv2 | REv2 | MNv2 | PNv2 | RSv2 | REv2 | MNv2 | PNv2 | RSv2 | REv2 | MNv2 | PNv2 | RSv2 | REv2 | MNv2 |
| CLIP [6] | 59.41 | 57.50 | 52.58 | 49.68 | 65.58 | 64.86 | 71.27 | 63.60 | 78.04 | 72.15 | 75.42 | 67.78 | 77.34 | 72.66 | 74.88 | 66.86 |
| CoOp [10] | 92.15 | 65.27 | 65.71 | 62.66 | 92.37 | 75.43 | 74.67 | 70.45 | 94.25 | 76.50 | 77.87 | 70.97 | 93.38 | 77.31 | 80.31 | 71.32 |
| CLIP-Adapter [2] | 76.70 | 66.71 | 65.37 | 63.17 | 82.37 | 77.13 | **78.05** | 71.63 | 92.36 | 79.17 | 79.76 | 71.04 | 87.50 | 79.93 | 82.26 | 71.47 |
| CoCoOp [9] | 92.83 | 67.84 | **70.60** | 63.77 | 91.10 | 77.70 | 77.17 | 71.56 | 94.41 | 79.33 | 80.43 | 71.67 | 93.89 | 79.83 | 81.93 | 72.34 |
| ProGrad [11] | 89.51 | 65.61 | 66.28 | 62.26 | 89.64 | 75.93 | 76.55 | 69.71 | 95.18 | 77.46 | 80.65 | 72.29 | 94.37 | 78.54 | 79.34 | 72.21 |
| APPLeNet | **94.53** | **69.90** | 68.83 | **65.80** | **95.03** | **79.23** | 77.63 | **72.15** | **96.63** | **81.03** | **82.23** | **74.03** | **96.23** | **80.97** | **83.15** | **73.56** |




**(a) RSICDv2** **(b) RESISC45v2** **(c) MLRSNetv2**

**Figure 1.** t-SNE plots [8] between the image feature extracted from the meta-net of CoCoOp ($1^{st}$ row) and the injection block (IB) of APPLeNet ($2^{nd}$ row) for the SSMT domain generalization task on the three benchmark RS datasets. The legends represent the class labels.

(CSC). The performance of SOTA methods and our proposed APPLeNet on the B2N class generalization task is evaluated using both types of context vectors. The results of the ablation study are presented in Table 2. We observe that the UC setup outperforms the CSC setup in terms of the harmonic mean (H) between the base and new classes. Specifically, on the PatternNet and RSICD datasets, the UC setup outperforms the CSC setup by 3.2% and 1.2%, respectively.

**Change in visual encoder backbone:** We conducted experiments to investigate the effect of different vision backbones of the pre-trained CLIP in generating text-prompts and extracting visual features for the cross-data generalization task. We used ResNet-50, ResNet-101, ViT-B/16, and ViT-B/32 as visual backbones, and compared their performance to APPLeNet, zero-shot CLIP, and other referred prompt learning techniques. The experimental results, shown in Table 3, indicate that APPLeNet outperforms all the other methods in the few-shot setting by a significant margin. Specifically, APPLeNet achieves better classification scores (average) of 1.0% and 1.6% on ResNet-50 and ViT-B/16 CLIP's visual backbones, respectively. We also conducted an ablation study with other visual backbones such as ResNet-101 and ViT-B/32, which can be found in the supplementary materials.

**t-SNE visualization:** In Figure 1, we present a t-SNE [7] visualization of the image embeddings generated by APPLeNet and compare them with CoCoOp [9] on the RSICDv2, RESISC45v2, and MLRSNetv2 datasets 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.

## References


[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

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

[9] 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.

[10] 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.

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