Supplementary for FrogDogNet: Fourier frequency Retained visual prompt Output Guidance for Domain Generalization of CLIP in Remote Sensing

Hariseetharam Gunduboina[1] Muhammad Haris Khan[2] Biplab Banerjee[1]
[1]Indian Institute of Technology Bombay, India
[2]Mohamed Bin Zayed University of Artificial Intelligence, UAE
hariseetharam552@gmail.com, muhammad.haris@mbzuai.ac.ae, getbiplab@gmail.com

A. Introduction

- We provide a comprehensive description of the datasets used to evaluate the effectiveness of the proposed FrogDogNet in section B. Additionally, we adopt the Version 2 **v2** dataset, which has been derived from existing datasets, specifically for the Single source multi target domain generalization (DG) setup.
- In Section C, we discussed various state-of-the-art (SOTA) prompting methods based on CLIP [9] that were used to compare the performance of FrogDogNet.
- In Figure 1, we show the t-SNE [11] plots for the outputs from meta-net of the CoCoOp and Fourier Frequency Block (FFB) of the FrogDogNet on the three mentioned benchmark **v2** target datasets in the Single source multi target domain generalization (DG) setup and the explanation is in section D.
- In Section E, we provided a tabular analysis and explanation of the optimal number of low-frequency components of visual feature embeddings to retain by evaluating all generalization tasks across all datasets.
- In Section F, we presented a comparative analysis of hyper-parameter Λ variation in the base-to-new (B2N) generalization setting across all datasets, using the harmonic mean (HM) of base and new accuracies as the performance metric.

B. Datasets

We evaluate the proposed FrogDogNet on four well-established remote sensing benchmark datasets: **PatternNet** [6], **RSICD** [7], **RESISC45** [2], and **MLRSNet** [8]. A detailed overview of each dataset is provided below:

- **PatternNet** consists of 38 classes, with each class containing 800 images of size 256 × 256 pixels. The dataset comprises high-resolution images sourced from Google Earth imagery, focusing on various urban areas across the United States. It is primarily used for remote sensing image retrieval tasks.

- **Remote Sensing Image Captioning Dataset (RSICD)** contains 30 classes with a total of 10,000 images, each measuring 224 × 224 pixels. The number of images per class varies. Additionally, every image in this dataset is accompanied by five descriptive sentences, making it suitable for automatic image captioning tasks. However, in our work, we utilize only the images, as the captions are learnable within FrogDogNet's approach.

- **Remote Sensing Image Scene Classification (RESISC45)** comprises 45 classes, with each class having 700 images of size 256 × 256 pixels. The dataset features images with a wide range of spatial resolutions, spanning from 20 cm to over 30 meters.

- **Multi-label High Spatial Resolution Remote Sensing Dataset (MLRSNet)** includes 46 classes and a total of 109,161 images, each sized at 256 × 256 pixels. On average, each class contains approximately 2,000 images, with spatial resolutions varying between 0.1 m and 1 m. MLRSNet is commonly used for tasks such as image retrieval, segmentation, and classification.

To further enhance our study, we extend our work to a single-source, multi-target domain generalization setup. For this purpose, we adopt a revised version of the aforementioned datasets [10], specifically tailored to this setting.

B.1. Version 2 Datasets for Single-Source Multi-Target Domain Generalization (DG)

To evaluate the effectiveness of FrogDogNet under domain generalization (DG) settings, we adopt the PatternNetv2, RSICDv2, RESISC45v2, and MLRSNetv2 datasets, which follow the same format as prior DG studies in remote sensing [10]. These v2 datasets provide a standardized benchmark for assessing the generalization capabilities of models across multiple remote sensing domains.

Key Features of the v2 Datasets

Common Class Subset for Consistency Each v2 dataset is curated to include a fixed set of 16 shared classes, ensuring a controlled evaluation setup. These classes include: *Baseball, Beach, Bridge, Dense Residential Area, Desert, Field, Forest, Harbor, Intersection, Meadow, Overpass, Parking,*

Railway, River, Sparse Residential Area, Stadium, and Storage Tank.

Standard Single-Source Multi-Target DG Setting

- A model is trained on a single dataset (e.g., PatternNetv2) and tested on the remaining datasets (e.g., RSICDv2, RESISC45v2, and MLRSNetv2) without additional adaptation.
- This follows the single-source multi-target DG framework, a widely used evaluation setup for studying domain shifts.

Domain Shift Challenges in Remote Sensing

Since the same class appears across different datasets with distinct characteristics, these v2 datasets naturally introduce domain shifts due to:

- Variations in Spatial Resolution and Scale (e.g., aerial vs. satellite imagery)
- Different Capture Conditions (e.g., viewpoint, orientation, seasonal changes)
- Scene Composition Differences (e.g., background clutter, land cover types)

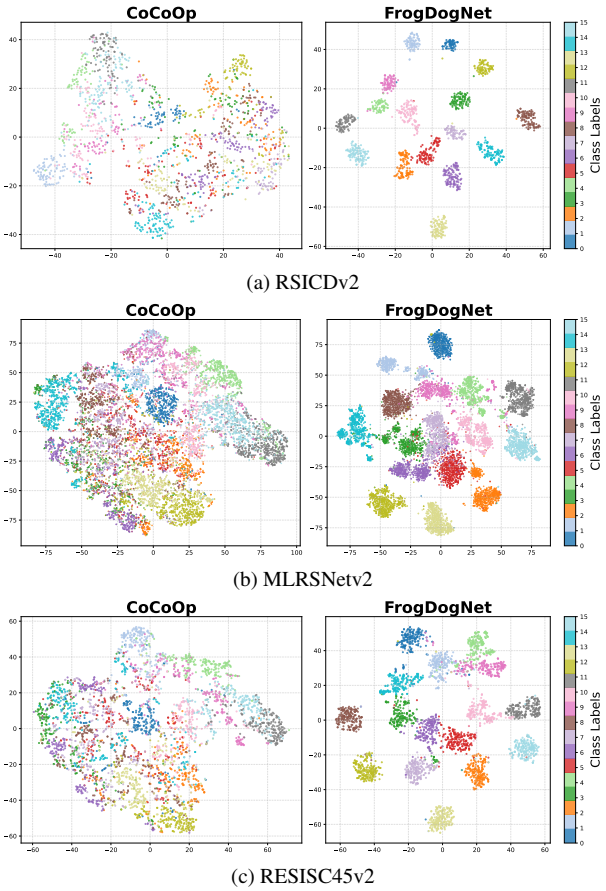
By leveraging these v2 datasets, we align our evaluation protocol with prior domain generalization benchmarks while ensuring a rigorous assessment of model robustness across diverse remote sensing environments.

C. Baseline models

To evaluate the performance of FrogDogNet, we compared this model with some of the related state-of-the-art (SOTA) prompting baseline methods based on CLIP. As a baseline, we employed Zero-shot CLIP [9]. Additionally, we explored empirical risk minimization (ERM) [13], which utilizes a trainable linear classifier on top of CLIP features, and the domain adaptation technique DANN [3], which adapts CLIP embeddings for improved generalization. Furthermore, we examined various prompt learning approaches, including CoOp [15], CoCoOp [14], CLIP-Adapter [4], ProGrad [16], MaPLe [5], APPLeNet [10], and StyLIP [1] to analyze their effectiveness in enhancing CLIP's adaptability to remote sensing datasets.

D. t-SNE visualization

Figure 1 presents t-SNE [11] visualizations comparing the image embeddings generated by FrogDogNet and CoCoOp [14] across three remote sensing target datasets—RSICDv2, MLRSNetv2, and RESISC45v2—for the domain generalization (DG) task. In all cases, FrogDogNet exhibits well-separated clusters, indicating strong feature discriminability, whereas CoCoOp shows notable overlap between class clusters. Specifically, in RSICDv2 (Figure 1a), FrogDogNet maintains distinct boundaries between categories, while CoCoOp struggles with inter-class separation. Similarly, MLRSNetv2 (Figure 1b) demonstrates



(a) RSICDv2

(b) MLRSNetv2

(c) RESISC45v2

Figure 1. t-SNE plots [12] depicting the image features extracted from the Meta-Net of CoCoOp and the FFB of FrogDogNet for the domain generalization (DG) task on three remote sensing target datasets. Legends indicate class labels.

FrogDogNet's robustness in capturing meaningful visual distinctions, unlike CoCoOp's overlapping clusters. Lastly, in RESISC45v2 (Figure 1c), FrogDogNet continues to outperform CoCoOp by forming compact and well-separated feature distributions, further validating its generalization capability across diverse datasets.

E. Fourier frequency retention analysis

The table 4 presents the average target accuracy across three different domain generalization tasks: Cross-Dataset (CD) in table 2, Domain Generalization (DG) in table 3, and B2N HM in table 1. Each row corresponds to a different number of retained low frequency components (LFCs) of visual feature embeddings, ranging from 50 to 500. The CD column represents the average target accuracy when the model is trained on PatternNet and evaluated on RSICD, RESISC45, and MLRSNet, showing how well the model generalizes to unseen datasets. The DG column reflects average target performance in a domain generalization setting, where the model is expected to handle distribution shifts more effec-

Retained LFCs	PatternNet			RSICD			RESISC45			MLRSNet			Avg. of all		
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM
50	86.8	47.4	61.12	87.7	42.3	57.11	79.4	39.2	52.52	72.7	33.9	46.22	81.65	40.7	54.24
100	91.8	56.1	69.57	92.9	44.0	59.43	85.1	53.8	65.68	77.9	37.5	50.49	86.93	47.85	61.29
150	94.0	52.5	67.13	93.3	57.0	70.69	87.3	53.8	66.46	82.6	47.2	60.12	89.3	52.63	66.10
200	94.4	57.7	71.76	92.8	62.8	75.02	89.0	64.0	74.67	79.7	47.5	59.49	88.98	58.00	70.24
250	94.3	65.2	77.31	95.1	65.4	77.61	89.9	61.6	72.99	80.8	53.9	64.67	90.03	61.53	73.14
300	94.5	65.3	77.46	94.0	53.7	68.55	90.2	66.7	76.74	83.5	55.4	66.61	90.55	60.28	72.84
350	95.5	77.6	85.63	95.7	64.1	76.68	90.6	65.0	75.69	84.9	57.5	68.56	91.67	**66.05**	**76.64**
400	95.9	74.3	84.03	94.8	60.6	74.09	91.2	63.9	75.32	85.4	52.0	64.50	91.08	62.70	74.48
450	96.1	70.7	81.31	96.1	62.2	75.59	91.4	63.6	75.01	86.7	54.9	67.22	**92.08**	62.85	74.78
500	96.1	69.6	80.32	95.7	66.6	78.74	91.6	65.6	76.58	85.6	49.7	62.76	92.25	62.88	74.60

Table 1. Comparison of base and new class performance across different retained Low frequency components (LFCs) of visual features. HM denotes the harmonic mean, providing a balanced measure of trade-off performance. The last row presents the average performance across all datasets.

Retained LFCs	Source	Target			
	PatternNet	RSICD	RESISC45	MLRSNet	Avg. Target
50	78.4	31.4	37.3	35.1	34.6
100	86.0	32.4	44.2	40.8	39.13
150	88.8	47.2	47.1	44.5	46.27
200	90.4	47.6	53.6	49.4	50.2
250	90.3	46.4	48.6	48.4	47.8
300	91.0	47.2	53.5	49.4	50.03
350	91.6	53.1	56.3	52.3	53.9
400	92.0	51.9	54.9	50.1	52.3
450	91.8	54.0	59.5	52.2	55.23
500	91.4	51.8	58.2	52.6	54.2

Table 2. Performance comparison across different retained Low frequency components (LFCs) of visual features for Cross-data (CD) generalization setting using PatternNet as the source dataset and RSICD, RESISC45, and MLRSNet as the target datasets. The last column represents the average accuracy across all target datasets.

Retained LFCs	Source	Target			
	PatternNetv2	RSICDv2	RESISC45v2	MLRSNetv2	Avg. Target
50	90.8	60.6	68.7	63.1	64.13
100	94.1	70.8	78.3	71.5	73.53
150	95.7	74.0	81.6	76.3	77.3
200	96.4	79.7	85.5	78.9	81.37
250	96.6	80.1	85.6	79.6	81.77
300	96.7	81.8	86.0	80.6	82.8
350	96.8	82.5	86.9	80.7	83.37
400	97.0	82.0	86.6	81.3	83.3
450	97.0	82.7	87.3	81.5	83.83
500	97.0	83.4	87.6	81.5	84.17

Table 3. Performance comparison across different retained Low frequency components (LFCs) of visual features for single-source multi-target domain generalization (DG) task, using PatternNetv2 as the source dataset and RSICDv2, RESISC45v2, and MLRSNetv2 as the target datasets. The last column represents the average accuracy across all target datasets.

tively. The B2N HM column shows the average HM of all datasets and likely evaluates the model's ability to transition from base categories to novel ones. The last column provides an overall average of these three domain generalization tasks, giving a holistic view of the model's performance across different generalization scenarios. As the number of retained LFCs increases, there is a steady improvement in accuracy across all tasks, indicating that retaining 350 LFCs enhances generalization ability. The overall average follows an increasing trend, peaking at 350 retained LFCs, suggesting that generalization improves with a richer feature representation.

Retained LFCs	Average Target Accuracy (%)			
	CD	DG	B2N HM	Overall Avg.
50	34.6	64.1	54.2	50.97
100	39.1	73.5	61.3	57.97
150	46.3	77.3	66.1	63.23
200	50.2	81.4	70.2	67.27
250	47.8	81.8	73.1	67.57
300	50.0	82.8	72.8	68.53
350	53.9	83.4	76.6	**71.30**
400	52.3	83.3	74.5	70.03
450	55.2	83.8	74.8	71.27
500	54.2	84.2	74.6	71.00

Table 4. Comparison of average target accuracy across different domain generalization tasks: Cross-Dataset (CD), Domain Generalization (DG), and B2N HM with different retained Low frequency components (LFCs) of visual features. The last column represents the overall average accuracy across all three tasks.

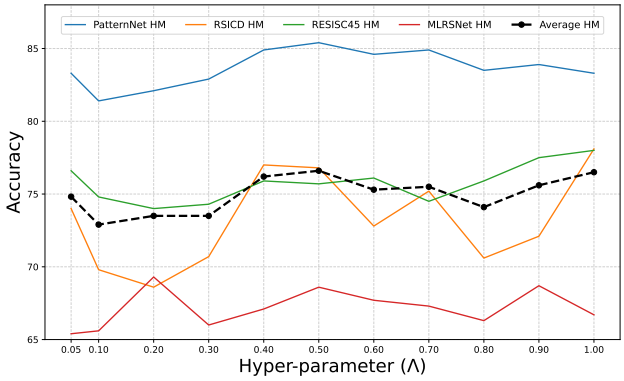


Figure 2. A comparative analysis of hyperparameter Λ variation in the base-to-new (B2N) generalization setting was conducted across all datasets, employing the harmonic mean (HM) of base and new accuracies as the performance metric.

F. Hyper-parameter variation

The effect of varying the hyper-parameter Λ on base-to-new (B2N) generalization was analyzed across four datasets:

PatternNet, RSICD, RESISC45, and MLRSNet, using the harmonic mean (HM) of base and new class accuracies as the evaluation metric. The results indicate that the choice of Λ significantly influences generalization performance. PatternNet exhibits a steady increase in HM, peaking at $\Lambda = 0.5$ with 85.4 before stabilizing. RSICD follows a more fluctuating trend, with a drop at lower Λ values, a recovery at $\Lambda = 0.4$, and a final peak at $\Lambda = 1.0$ with 78.1. RESISC45 remains relatively stable, with a gradual increase leading to a maximum HM of 78.0 at $\Lambda = 1.0$. MLRSNet shows moderate variations, peaking at $\Lambda = 0.2$ with 69.3 before stabilizing. The average HM across datasets follows a similar trend, increasing from 74.82 at $\Lambda = 0.05$ to a peak of 76.6 at $\Lambda = 0.5$, indicating that moderate values of Λ tend to enhance generalization, while excessively high or low values may result in suboptimal performance.

References

[1] Shirsha Bose, Ankit Jha, Enrico Fini, Mainak Singha, Elisa Ricci, and Biplab Banerjee. Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization. *arXiv preprint arXiv:2302.09251*, 2024. Accepted in WACV 2024. 2

[2] 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. 1

[3] 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. 2

[4] 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. 2

[5] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *arXiv preprint arXiv:2210.03117*, 2023. Accepted at CVPR 2023. 2

[6] 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. 1

[7] 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. 1

[8] 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. 1

[9] 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. 1, 2

[10] Mainak Singha, Ankit Jha, Bhupendra Solanki, Shirsha Bose, and Biplab Banerjee. Applenet: Visual attention parameterized prompt learning for few-shot remote sensing image generalization using clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2

[11] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 1, 2

[12] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 2

[13] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998. 2

[14] 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. 2

[15] 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. 2

[16] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*, 2022. 2