Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Domain Prompt Learning with Quaternion Networks (Extended Abstract)*

Qinglong Cao[1,2] , **Zhengqin Xu**[1] , **Yuntian Chen**[2] , **Chao Ma**[1] , **Xiaokang Yang**[1]
[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China
[2] Ningbo Institute of Digital Twin, Eastern Institute of Technology, China
{caoql2022, fate311}@sjtu.edu.cn, ychen@eitech.edu.cn, {chaoma, xkyang}@sjtu.edu.cn

Abstract

Foundational vision-language models (VLMs) like CLIP have revolutionized image recognition, but adapting them to specialized domains with limited data remains challenging. We propose Domain Prompt Learning with Quaternion Networks (DPLQ), which leverages domain-specific foundation models and quaternion-based prompt tuning to effectively transfer recognition capabilities. Our method achieves state-of-the-art results in remote sensing and medical imaging tasks. This extended abstract highlights the key contributions and performance of DPLQ.

1 Introduction

Foundational Vision-Language Models (VLMs) such as CLIP have demonstrated strong generalization capabilities across natural image datasets by leveraging massive paired image-text data during training. However, these models often underperform when applied directly to specialized domains like remote sensing or medical imaging, due to substantial domain gaps. Prompt learning has emerged as a lightweight and efficient method for adapting VLMs without retraining the entire network. Nevertheless, existing prompt learning methods primarily focus on manipulating language prompts, neglecting the visual modality and domain-specific characteristics. To bridge this gap, we propose Domain Prompt Learning with Quaternion Networks (DPLQ), a novel framework that introduces external domain knowledge into both vision and language branches through quaternion representations. By explicitly modeling orthogonal cross-modal relationships, DPLQ facilitates a more robust adaptation of VLMs to specialized fields.

2 Method Overview

Aims to prompt VLMs efficiently from a generalized domain to specific domains like remote sensing and medical images, domain prompt learning with quaternion networks is proposed to facilitate the integration of domain-specific knowledge from large-scale foundation models into VLMs. Illustrated in Figure 1, the quaternion network enables the identification of cross-modal relationships between domain-specific vision features from the foundation model and generalized contextual embeddings from the language branch. Subsequently, this information is utilized to map the generalized contextual embeddings into the specialized domain. Furthermore, well-aligned vision-language relationships in pre-trained VLMs are leveraged to propagate domain-specific information from the specialized language branch into the vision branch.

Our DPLQ framework consists of three key stages:

- **Domain-Specific Feature Extraction:** We utilize large-scale pretrained domain-specific foundation models to extract rich visual features pertinent to the target domain. These features serve as external guidance to steer prompt learning towards domain-relevant representations.
- **Quaternion-Based Prompt Fusion:** The extracted domain-specific visual features and the learnable language context embeddings are fused within a quaternion hidden space. In this space, visual features are placed on the imaginary axis while language prompts occupy the real axis, enabling orthogonal modeling of cross-modal relationships. Additionally, Gaussian noise scaled by the domain features is injected into the quaternion space to enhance robustness against overfitting.
- **Dual-Branch Prompt Injection:** Prompt tokens derived from the quaternion space are hierarchically injected into both the language and vision encoders of the VLM. This dual-branch prompting ensures that domain-specific knowledge is effectively propagated throughout the network, enhancing the alignment between visual and textual representations.

This design enables efficient domain adaptation by exploiting the strong alignment properties of pre-trained VLMs while introducing targeted modifications driven by external domain knowledge.

3 Experimental Validation

We conduct extensive experiments to validate the effectiveness of DPLQ on a range of domain-specific datasets, and the

*This paper is an extended abstract of the work published at CVPR 2024, titled *Domain Prompt Learning with Quaternion Networks* [Cao *et al.*, 2024].

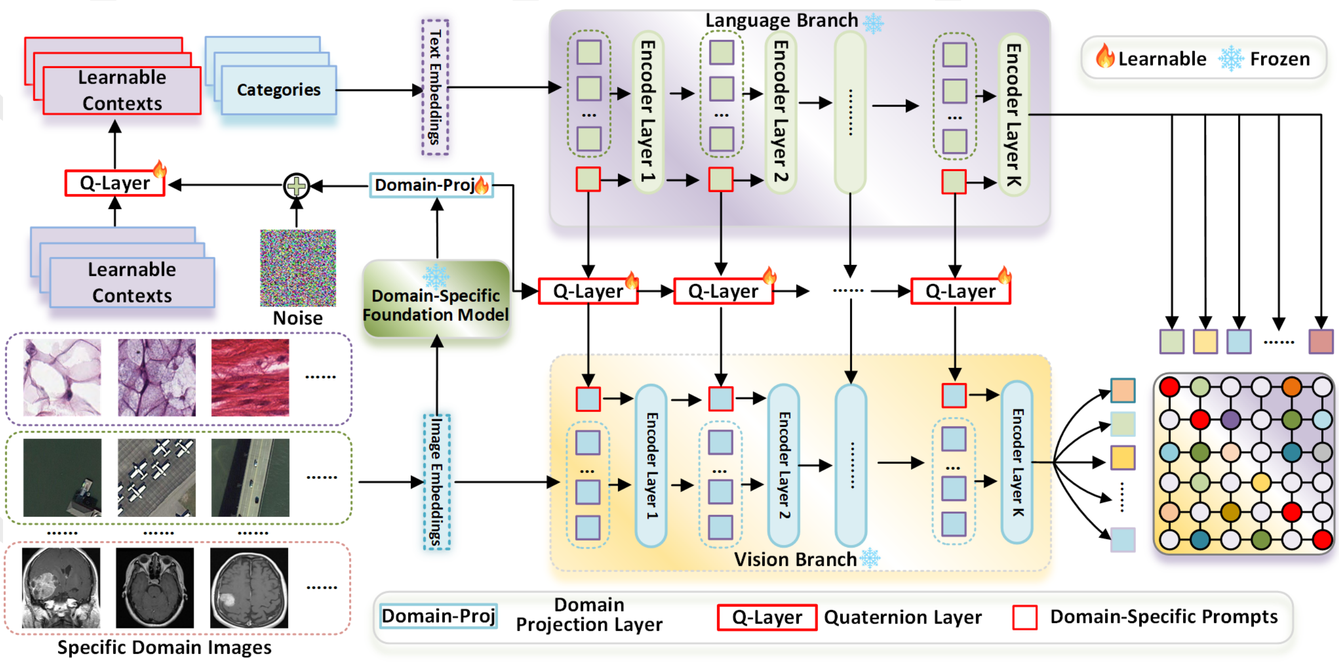

Figure 1: Overview of our proposed Domain Prompt Learning. We use the large-scale domain-specific foundation model as guidance, and exploit quaternion networks to mine the intermodal relationships between domain-specific vision features from the domain-specific foundation model and contextual embeddings from the language branch. Based on the stable vision-language matching relationships in pre-trained VLMs, the domain-specific information is hierarchically forwarded from the language branch to the vision branch.

Method	MLRSNet			PatternNet			RSSCN7		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	64.50	60.30	62.33	70.60	62.60	66.36	66.70	95.30	78.48
CoOp	79.37	58.90	67.62	87.30	64.20	73.99	84.80	89.13	86.91
CoCoOp	83.30	59.50	69.42	93.70	59.90	73.08	90.97	90.00	90.48
MaPLe	85.23	59.60	70.15	95.30	57.90	72.03	91.67	93.70	92.67
Ours (ViTAE)	88.96	57.10	69.56	97.07	62.37	75.94	91.53	94.53	93.01
Ours (ViT)	87.07	59.00	70.34	95.80	66.20	78.30	91.20	95.57	93.33

Method	AID			RSICD			UCM		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	73.50	70.40	71.92	71.50	60.20	65.37	80.60	68.00	73.77
CoOp	87.63	70.37	78.06	88.43	60.20	71.63	93.60	74.53	82.98
CoCoOp	92.63	65.73	76.89	92.37	58.80	71.86	95.23	71.57	81.72
MaPLe	92.73	74.57	82.66	93.93	56.27	70.38	97.70	70.90	82.17
Ours (ViTAE)	94.03	74.97	83.43	94.57	65.20	77.19	97.10	72.10	82.75
Ours (ViT)	94.50	75.77	84.10	95.67	64.83	77.29	97.90	73.30	83.83

Method	WHURS19			NWPU			Avg (8)		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	73.10	90.80	80.99	69.00	63.00	65.87	71.19	71.33	70.63
CoOp	95.20	82.40	88.34	84.53	66.97	74.73	87.61	70.84	78.03
CoCoOp	97.10	77.00	85.89	89.27	69.37	78.07	91.82	68.98	78.43
MaPLe	97.70	88.03	92.61	90.70	72.70	80.71	93.12	71.71	80.42
Ours (ViTAE)	98.80	89.90	94.41	91.60	71.23	80.14	94.28	73.43	82.05
Ours (ViT)	98.80	90.80	94.63	91.70	75.03	82.53	94.08	75.06	83.50

Table 1: Comparison with SOTA methods on 8 remote sensing datasets. **Red** and **blue** indicate the best and second-best.

results are shown in Table 1 and Table 2. Our evaluations cover two major domains: remote sensing and medical imaging.

Remote Sensing. As shown in Table 1, we benchmark our method on eight remote sensing datasets, including MLRSNet [Qi *et al.*, 2020], PatternNet [Zhou *et al.*,

Average over datasets	Base	Novel	HM	BTMRI	Base	Novel	HM
CLIP	49.83	41.83	45.18	CLIP	50.60	51.20	50.89
CoOp	51.59	43.77	46.81	CoOp	48.93	53.30	51.02
CoCoOp	64.45	43.16	49.45	CoCoOp	52.37	52.80	52.58
MaPLe	62.39	44.40	49.01	MaPLe	53.67	61.60	57.36
Ours	74.36	44.74	53.36	Ours	60.97	56.30	58.54

CHMNIST	Base	Novel	HM	CCBTM	Base	Novel	HM
CLIP	31.60	27.40	29.35	CLIP	67.30	46.90	55.28
CoOp	41.70	25.67	31.78	CoOp	64.13	52.33	57.63
CoCoOp	74.30	25.30	37.74	CoCoOp	66.67	51.37	58.03
MaPLe	74.03	25.10	37.49	MaPLe	59.47	46.50	52.19
Ours	87.80	26.60	40.83	Ours	74.30	51.33	60.72

Table 2: Comparison between our method and SOTA methods for base-to-novel generalization on medical image classification datasets. Our method performs well over the compared methods. We use red and blue to indicate the first and second best scores.

2018], RSSCN7 [Zou *et al.*, 2015], AID [Xia *et al.*, 2017], RSICD [Lu *et al.*, 2017], UCM [Yang and Newsam, 2010], WHURS19 [Dai and Yang, 2011], and NWPU [Cheng *et al.*, 2017]. DPLQ consistently outperforms state-of-the-art prompt learning baselines such as CoOp [Zhou *et al.*, 2022b], CoCoOp [Zhou *et al.*, 2022a], and MaPLe [Khattak *et al.*, 2023]. Notably, with the ViT backbone, our method achieves an average harmonic mean (HM) accuracy of 83.5%, surpassing MaPLe's 80.4%.

Medical Imaging. For medical domain evaluation, as shown in Table 1, we use three datasets:BTMRI [Nickparvar, 2021], CCBTM [Hashemi, 2023], and CHMNIST [Kather *et al.*, 2016]. Guided by the MedSAM foundation model, DPLQ significantly improves performance over baseline methods, achieving an average HM of 53.36%, compared to CoCoOp [Zhou *et al.*, 2022a]'s 49.45%.

Ablation Studies. We perform detailed ablation studies to assess the contribution of each component. Results show that quaternion modeling, dual-branch prompting, and controlled noise injection all contribute to performance gains. Without quaternion networks, performance drops notably, underscoring the importance of orthogonal inter-modal fusion.

Generalization. DPLQ also demonstrates strong generalization capabilities in cross-dataset evaluations and domain generalization settings, further highlighting its robustness and adaptability. While the proposed method does not achieve top performance on each dataset, it excels with an average accuracy of 73.60%, corresponding to a 1.35% improvement over MaPLe.

4 Conclusion

In this work, we propose Domain Prompt Learning with Quaternion Networks (DPLQ) to address the challenge of adapting large-scale VLMs to specialized domains. By leveraging external domain-specific foundation models and introducing a novel quaternion-based prompt learning strategy, DPLQ effectively transfers general recognition abilities into specialized fields. Our method not only injects domain-specific information into both vision and language modalities but also models cross-modal relationships through the orthogonal structure of quaternion networks. Extensive experiments demonstrate that DPLQ achieves new state-of-the-art results on a variety of remote sensing and medical imaging datasets, confirming its effectiveness and generalization capabilities. We believe DPLQ provides a promising foundation for future research on efficient and robust domain adaptation for VLMs.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was supported by NSFC Grants (62106116, 62322113, 62376156), China Meteorological Administration Grant (QBZ202316), Natural Science Foundation of Ningbo (2023J027), and the High Performance Computing Centers at Eastern Institute of Technology and Ningbo Institute of Digital Twin.

References

[Cao *et al.*, 2024] Qinglong Cao, Zhengqin Xu, Yuntian Chen, Chao Ma, and Xiaokang Yang. Domain prompt learning with quaternion networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26637–26646, 2024.

[Cheng *et al.*, 2017] 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.

[Dai and Yang, 2011] Dengxin Dai and Wen Yang. Satellite image classification via two-layer sparse coding with biased image representation. *IEEE Transactions on Geoscience and Remote Sensing*, 8(1):173–176, 2011.

[Hashemi, 2023] Seyed Mohammad Hossein Hashemi. Crystal clean: Brain tumors mri dataset, 2023.

[Kather *et al.*, 2016] JN Kather, CA Weis, F Bianconi, SM Melchers, LR Schad, T Gaiser, A Marx, and Zollner F. Multi-class texture analysis in colorectal cancer histology. *Scientific Reports*, 2016.

[Khattak *et al.*, 2023] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[Lu *et al.*, 2017] 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.

[Nickparvar, 2021] Msoud Nickparvar. Brain tumor mri dataset, 2021.

[Qi *et al.*, 2020] 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.

[Xia *et al.*, 2017] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55:3965–3981, 2017.

[Yang and Newsam, 2010] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010.

[Zhou *et al.*, 2018] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:197–209, 2018.

[Zhou *et al.*, 2022a] 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*, 2022.

[Zhou *et al.*, 2022b] 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.

[Zou *et al.*, 2015] Qin Zou, Lihao Ni, Tong Zhang, and Qian Wang. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 12(11):2321–2325, 2015.