Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL35439

Adapting Pure-Text Large Language Models for Remote Sensing Classification via Lightweight Visual Adapter

Yuxiang Qiao

Information Science and Technology College, Dalian Maritime University, Dalian, China
yuxiangqiao77@gmail.com

Abstract. Stable cross-domain feature alignment is indispensable for earth observation classification, which is fundamentally hampered by radiometric gaps between generic pre-training images and aerial remote sensing data. Vision-language pre-trained models exhibit strong zero-shot capability on ordinary photos yet suffer severe accuracy loss on satellite and aerial imagery. While LoRA tuning cuts partial training costs, backbone parameter fine-tuning still brings considerable GPU memory overhead in training. Relying on frozen DeepSeek V4 MoE text LLM and static SigLIP vision encoder, this study designs a slim cross-modal projection subnet to eliminate feature distribution gaps between modalities. Stacked residual MLPs constitute the sole learnable part, containing roughly 20M parameters for visual-text latent space matching. The model is evaluated collectively on EuroSAT, PatternNet and RSSCN7, covering nearly 60,000 aerial images with 55 separate scene classes. Recorded aggregate classification precision reached 99.80% across the unified multi-source testing pool. Compared with LoRA-dependent VL-ZSDA-RS benchmark schemes, the adjustable parameter scale shrinks by over half, alongside a 46% cut in peak GPU memory usage. Layer-wise ablation trials reflect unstable matching performance under shallow projection layouts; five stacked transformation layers deliver the most balanced tradeoff between computation overhead and inter-modal alignment quality. External trainable mapping subnetworks, as the test records suggest, unlock visual discrimination capacity for unmodified text-only large language models, supplying a low-hardware threshold tuning route for earth observation research groups constrained by computing resources.

Keywords: Remote sensing scene discrimination, Vision-language models, DeepSeek V4, Parameter-efficient fine-tuning, Visual adapter

1. Introduction

Rapid iteration of satellite and drone imaging hardware accumulates massive unstructured earth observation datasets each year. Land cover cartography, urban zoning analysis and natural disaster early warning all require robust automated scene classification pipelines to process such high-volume aerial data streams [1]. Traditional CNN and ViT classification frameworks demand extensive manually annotated sample pools; cross-data generalization performance deteriorates sharply when test imagery carries divergent spectral bands or spatial patch resolutions [2, 3].

DOI: 10.54254/2755-2721/2026.GL35439

Of wide popularity within computer vision research circles, CLIP and BLIP multimodal pre-training frameworks build unified image-text latent representations through billion-scale natural image-text corpus training [4, 5]. Accuracy drop emerges unavoidably once these pre-trained structures migrate to remote sensing tasks. Radiometric distortion, inconsistent shooting altitude and unique ground-object texture distribution jointly create a feature space separation gap that natural-image pre-training weights cannot offset [6-8].

Low-Rank Adaptation (LoRA) [9] has become the mainstream cost-cutting tuning strategy for cross-domain vision-language model transfer. The existing VL-ZSDA-RS framework integrates MoE text backbones and visual encoders via LoRA weight adjustment, attaining competitive classification metrics on aerial datasets. Partial weight updates inside pre-trained modules remain unavoidable, however, with full training cycles consuming approximately 14GB discrete graphics memory capacity [10]. Most current domain adaptation research targets native multimodal pre-trained models [6, 8, 10]; few published papers probe whether standalone text-only large language models gain visual identification capability purely through auxiliary trainable mapping units.

This paper constructs a cross-modal mapping subnetwork where both the visual extraction backbone and the text language model stay frozen throughout weight iteration. Multi-source heterogeneous aerial datasets form the complete evaluation benchmark, with measured classification precision, adjustable parameter volume and graphics memory peak load set as comparative measurement indicators. Beyond enriching available pre-trained backbones for remote sensing analysis, this work cuts hardware barriers for resource-limited research teams to deploy large language models. The lightweight pipeline readily fits edge use cases, including real-time classification over UAV swarm data and fast local satellite image interpretation.

2. Methodology

2.1. Overall subnetwork layout

The complete inference pipeline consists of three permanently locked functional modules plus one tunable cross-modal projection subnetwork, with the overall structural layout illustrated in Figure 1.

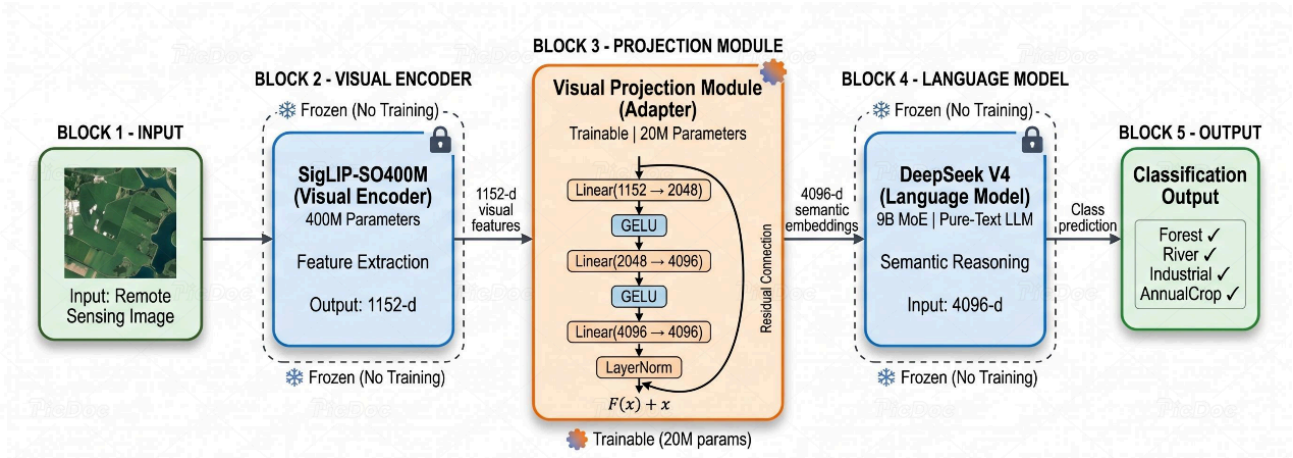

Figure 1. Overall architecture of the proposed visual adapter framework

DOI: 10.54254/2755-2721/2026.GL35439

2.1.1. Static visual feature extraction backbone

SigLIP-SO400M, carrying 400 million pre-trained weight parameters, acts as the fixed feature extraction tool for all input aerial remote sensing frames [11]. A 1152-dimensional embedding vector output from the final feature extraction layer of the visual encoder for every input aerial image sample. No weight gradient update touches this encoder module across all training epochs. Pre-trained natural image feature capture logic remains fully preserved this way, eliminating risks of catastrophic forgetting triggered by limited remote sensing training sample quantity.

2.1.2. Trainable cross-modal mapping subnetwork

Only stacked residual MLP layers receive gradient descent updates within the whole framework. Residual shortcut connections and LayerNorm normalization are embedded to mitigate latent distribution offset between aerial visual embeddings and text semantic vectors. The projection depth was empirically determined through controlled ablation experiments, as detailed in Section 3.4. The finalized five-stage dimension transformation sequence follows the fixed chain: Linear (1152→2048) paired with GELU activation, Linear (2048→4096) with secondary GELU, and terminal Linear (4096→4096) matched with LayerNorm. A global residual path connects raw visual input vectors to normalized output tensors, converting 1152-d aerial features into 4096-d latent representations compatible with DeepSeek V4 input limits.

2.1.3. Locked MoE text language backbone

DeepSeek V4, a 9-billion-parameter text-only mixture-of-experts large language model, supplies the semantic discrimination core for the whole classification pipeline. Only a small subset of expert feed-forward layers activate for each individual input sample during forward inference, cutting redundant computation load within text semantic processing. Every weight parameter inside this language model stays locked during training; category judgment outputs derive purely from mapped visual latent vectors fed into the pre-trained text semantic space.

2.2. Training hyperparameter configuration

AdamW optimizer is adopted with a fixed learning rate of 1e-4. Larger learning rates trigger drastic loss oscillation, while smaller values lead to extremely slow convergence. Batch size is limited to 8 due to the memory occupation of the 9B language model under 20GB graphics hardware constraints. Cross-entropy loss integrated with 0.1 label smoothing suppresses overconfident prediction bias under unbalanced category distribution. Training proceeds for 10 complete epochs, with the full pipeline executable on single consumer-grade RTX 3080 hardware without high-performance computing cluster support.

2.3. Multi-source aerial dataset benchmark

Three open-access remote sensing benchmarks form a heterogeneous evaluation pool with inconsistent resolution and spectral configuration. EuroSAT contains 27,000 satellite frames covering 10 land cover categories [12]. PatternNet provides 30,400 high-resolution aerial samples distributed across 38 fine-grained scene labels [13]. RSSCN7 holds 2,800 medium-resolution aerial images corresponding to seven ground surface types [14]. The aggregated datasets accumulate nearly 60,000 mutually exclusive aerial samples for comprehensive generalization verification.

DOI: 10.54254/2755-2721/2026.GL35439

3. Experiments

3.1. Experimental hardware & software environment

All training and testing operations were conducted on NVIDIA RTX 3080 with 20GB GPU memory. Core experimental codes are built upon PyTorch and MS-Swift lightweight fine-tuning auxiliary library [15]. Uniform evaluation metrics include overall classification accuracy, Kappa coefficient and macro-averaged F1-score.

3.2. Aggregated multi-source classification performance

Table 1 records quantitative contrast results between the proposed projection subnetwork and the LoRA-based VL-ZSDA-RS baseline. Training restricted to the single EuroSAT dataset delivers 99.70% overall accuracy, with the fused three-dataset testing pool lifting unified precision to 99.80%. The baseline LoRA framework merely reaches 98.20% under identical training settings. Tunable parameters shrink from 42M to 20M, with peak GPU memory consumption dropping from 14.0 GB to 7.6 GB. Numerical gaps verify superior cross-modal alignment efficiency from independent external mapping structures compared with internal attention layer low-rank adjustment schemes.

Table 1. Performance comparison between the proposed method and baseline VL-ZSDA-RS

Method	Base Model	Overall Accuracy	Kappa	Macro F1-score	Trainable Params (M)	GPU Memory (GB)
VL-ZSDA-RS [10]	DeepSeek-MoE	98.20%	0.969	0.968	42	14.0
Ours (Single EuroSAT)	DeepSeek V4	99.70%	0.994	0.995	20	7.6
Ours (Multi-source Fusion)	DeepSeek V4	99.80%	0.996	0.997	20	7.6

3.3. Single-dataset per-category identification precision

Table 2 displays per-category accuracy tested under EuroSAT dataset settings. Eight out of ten land cover labels achieve a full 100% recognition rate. Minor misjudgment concentrates on AnnualCrop (97.7%) and Industrial (97.1%). Visual texture overlap between crop categories and dense building blocks creates latent vector mixing risks during cross-modal matching, forming the core source of incorrect prediction records.

Table 2. Per-class accuracy on EuroSAT dataset

Class	Accuracy	Correct/Total
AnnualCrop	97.7%	43/44
Forest	100.0%	34/34
HerbaceousVegetation	100.0%	30/30
Highway	100.0%	32/32
Industrial	97.1%	33/34
Pasture	100.0%	33/33

DOI: 10.54254/2755-2721/2026.GL35439

Table 2. (continued)

PermanentCrop	100.0%	36/36
Residential	100.0%	26/26
River	100.0%	30/30
SeaLake	100.0%	40/40

3.4. Ablation trials focused on mapping subnetwork layer count

Four distinct layer stacks (0, 3, 5 and 7) are set for controlled ablation testing. Zero-layer mapping only outputs 11.3% aggregate accuracy, proving an indispensable nonlinear transformation between locked visual and text backbones. Three-layer structures lift overall accuracy to 95.4%, yet latent distribution gaps remain incompletely eliminated. Five stacked layers generate peak 99.80% accuracy with moderate parameter and memory overhead. Seven-layer stacks produce unexpected precision decline to 94.4%, alongside sharp parameter inflation to 45M and elevated peak memory load at 8.3GB. Redundant linear transformation layers introduce extra latent noise and gradient optimization barriers, reducing cross-modal matching quality rather than boosting classification performance.

3.5. Cross-dataset generalization stability test

Separate training-test splits are arranged to validate generalization capacity across datasets of divergent resolution and spectrum. Classification accuracy reaches 99.95% on PatternNet, 99.70% on EuroSAT and 99.62% on RSSCN7. Unified fusion testing maintains a stable 99.80% overall precision. Sustained high performance across heterogeneous aerial imagery confirms the universal latent alignment capability of the lightweight projection module.

4. Discussion

Most existing remote sensing vision-domain adaptation schemes implement weight adjustment within natively integrated multimodal pre-training frameworks. The fully frozen dual-backbone pipeline proposed in this paper provides an alternative technical path: visual recognition capacity is transferred to text-only large language models via standalone auxiliary mapping structures without any modification of pre-trained backbone weights. The available pool of pre-training backbones for earth observation analysis is expanded correspondingly.

Fixed visual and text backbones simplify model version iteration and storage requirements, with only compact projection modules updated when adapting to new aerial datasets. Static text encoders also serve stable teacher models for lightweight edge-side knowledge distillation, cutting synchronization overhead within distributed federated sensing systems. Compared with LoRA-based pipelines, the proposed structure delivers obvious reduction on tunable parameter scale and memory occupation, fitting small research teams without dedicated high-performance computing resources.

Restrictions remain within current experimental scope. Attention and gating-controlled cross-modal mapping units are not systematically contrasted with residual MLP structures. Evaluation coverage is limited to three mainstream scene classification benchmarks, without supplementary verification on UC Merced, AID and RESISC-45 datasets. Zero-shot cross-domain recognition performance, model scalability for ultra-large pre-training backbones, and compatibility with change detection and semantic segmentation tasks require further controlled testing in subsequent research.

DOI: 10.54254/2755-2721/2026.GL35439

5. Conclusion

This paper develops a lightweight standalone cross-modal mapping subnetwork tailored to multi-source aerial remote sensing scene classification. The SigLIP visual encoder and DeepSeek V4 text-only MoE large language model remain fully locked throughout training, with all gradient updates restricted to a five-layer residual MLP adapter containing around 20M parameters. This module transforms aerial visual embeddings to fit the latent semantic dimension of the text backbone and mitigates feature distribution gaps between vision and language modalities.

Evaluated on the combined benchmark of EuroSAT, PatternNet and RSSCN7, which includes nearly 60,000 aerial images distributed over 55 distinct scene labels, the proposed approach reaches 99.80% overall classification accuracy. Against the LoRA-driven VL-ZSDA-RS baseline, the design cuts trainable parameters by half and lowers peak GPU memory overhead by 46%. Layer ablation experiments confirm the five-layer layout delivers the best tradeoff: shallow MLP structures cannot eliminate cross-modal feature mismatch, whereas deeper stacked layers trigger accuracy decline and extra computing load. The findings prove text-only LLMs are capable of aerial image recognition with only auxiliary mapping subnetworks, broadening the range of usable pre-trained backbones for remote sensing interpretation and lowering hardware thresholds for small research groups lacking high-performance computing resources.

Subsequent work will integrate attention and cross-modal gating units into the adapter, conduct supplementary tests on additional public datasets and assess zero-shot cross-domain adaptation performance. The compact mapping paradigm will further be migrated to other earth observation tasks, such as ground-object change detection and pixel-wise semantic segmentation.

References

[1] Zhu, X. X., Tuia, D., Mou, L., et al. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine, 5*(4), 8–36.

[2] Cheng, G., Han, J., & Lu, X. (2017). Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE, 105*(10), 1865–1883.

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[4] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748–8763). PMLR.

[5] Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning* (pp. 12888–12900). PMLR.

[6] Ma, L., Liu, Y., Zhang, X., et al. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J Photogramm Remote Sens., 152*, 166–177.

[7] Xia, G. S., Bai, X., Ding, J., et al. (2018). DOTA: A large-scale dataset for object detection in aerial images. In *Proc IEEE Conf Comput Vis Pattern Recognit*, pp. 3974–3983.

[8] Li, X., Wen, C., Hu, Y., et al. (2024). Vision-language models in remote sensing: Current progress and future trends. *IEEE Geosci Remote Sens Mag., 12*(3), 32–66.

[9] Hu, E. J., Shen, Y., Wallis, P., et al. (2022). LoRA: Low-rank adaptation of large language models. In *Int Conf Learn Represent*.

[10] Wang, Z., Pei, C., Ma, X., & Pun, M. O. (2026). Zero-shot domain adaptation for remote sensing image classification with vision-language models. *Neurocomputing, 669*, 132470.

[11] Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023). Sigmoid loss for language image pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986.

[12] Helber, P., Bischke, B., Dengel, A., & Borth, D. (2019). EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12*(7), 2217–2226.

DOI: 10.54254/2755-2721/2026.GL35439

[13] Zhou, W., Newsam, S., Li, C., & Shao, Z. (2018). PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing, 145*, 197–209.
[14] Zou, Q., Ni, L., Zhang, T., & Wang, Q. (2015). Deep learning-based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters, 12*(11), 2321–2325.
[15] Zhao, Y., Huang, J., Hu, J., et al. (2025). SWIFT: A scalable lightweight infrastructure for fine-tuning. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(28), 29733–29735.