LoRA-CLIP: Efficient Low-Rank Adaptation of Large CLIP Foundation Model for Scene Classification

Mohamad Mahmoud Al Rahhal, Senior *Member, IEEE*, Yakoub Bazi *, Senior *Member, IEEE*, and Mansour Zuair

***Abstract*—Scene classification in remote sensing (RS) imagery has been extensively investigated using both learning-from-scratch approaches and fine-tuning of ImageNet pre-trained models. Meanwhile, CLIP (Contrastive Language-Image Pretraining) has emerged as a powerful foundation model for vision-language tasks, demonstrating remarkable zero-shot capabilities across various domains. Its image encoder is a key component in many vision instruction-tuning models, enabling effective alignment of text and visual modalities for diverse tasks. However, its potential for supervised remote sensing (RS) scene classification remains unexplored. This work investigates the efficient adaptation of large CLIP models (containing over 300M parameters) through Low-Rank Adaptation (LoRA), specifically targeting the attention layers. By applying LoRA to CLIP's attention mechanisms, we can effectively adapt the vision- model for specialized scene classification tasks with minimal computational overhead, requiring fewer training epochs than traditional fine-tuning methods. Our extensive experiments demonstrate the promising capabilities of LoRA-CLIP. By training only on a small set of additional parameters LoRA-CLIP outperforms models pre-trained on ImageNet, demonstrating the clear advantages of using image-text pretrained backbones for scene classification.**

***Index Terms*—Scene classification, CLIP foundation model, vision transformer encoder, Low rank adaptation (LoRA).**

I. Introduction

Scene classification in remote sensing (RS) imagery is a fundamental task with wide-ranging applications, including urban planning, and agriculture management. Accurate categorization of land cover types from aerial or satellite images enable informed decision-making and efficient resource allocation. The advent of deep learning era has significantly advanced scene classification, making it one of the earliest and most extensively explored applications of these methods in the RS domain. Early approaches predominantly involved training convolutional neural networks (CNNs) from scratch or fine-tuning models pre-trained on large-scale datasets such as ImageNet to leverage transferable features [1].

Over the years, numerous solutions have been proposed to enhance the performance of RS scene classification. These include hierarchical architectures, attention mechanisms, and graph-based models that capture both global and spatial features [2], [3]. In addition, the vision Mamba model has emerged as another promising solution through its innovative architecture [4].

While these advancements have significantly improved RS scene classification, there remains potential to further elevate performance by exploiting more versatile encoding models. The emergence of foundation models like CLIP (Contrastive Language–Image Pre-training) [5] has established them as powerful models for vision-language tasks, demonstrating impressive zero-shot capabilities for many applications including RS [6],[7]. Unlike traditional ImageNet models that were trained solely on image classification tasks, CLIP is trained on a vast corpus of image-text pairs, enabling it to understand and align visual and textual modalities effectively. This distinctive training paradigm enhances its versatility. This aspect is confirmed with numerous approaches leveraging its capabilities including text-image retrieval [8], and instruction-tuning models that adapt mainly the CLIP vision encoder for diverse tasks [9], [10]. The vision encoder of CLIP, typically based on transformer architectures, has been central in these advancements, allowing models to generalize across multiple tasks with minimal fine-tuning. However, despite its widespread adoption in various domains, the potential of the CLIP vision transformer encoder for supervised RS scene classification remains unexplored compared to its counterparts pretrained on ImageNet.

Building on the versatility of CLIP foundation models, this work addresses the gap in leveraging its vision transformer encoder for supervised RS scene classification. Specifically, we adapt the 300M parameter using Low-Rank Adaptation (LoRA) [11]. We introduce a small set of trainable parameters to the attention layers while keeping the original model frozen. This approach efficiently tailors the model's capabilities to the RS domain, significantly reducing computational overhead and storage requirements compared to full fine-tuning. LoRA enables faster convergence, requiring fewer training epochs, and its lightweight parameter updates make it particularly appealing for resource-constrained environments. By training only a small set of additional parameters, LoRA minimizes the need for large-scale storage and computational resources, offering scalable solutions for a variety of deployment scenarios, including potential applications in decentralized or federated learning settings [12]. Extensive experiments validate the effectiveness of this methodology, with LoRA-CLIP outperforming ImageNet-pretrained models.

Mohamad Mahmoud Al Rahhal is with the Applied Computer Science Department, College of Applied Computer Science, King Saud University, Riyadh 11543, Saudi Arabia. mmaalrahhal@ksu.edu.sa. Yakoub Bazi, and Masnsour Zuair are with the Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia (e-mail: ybazi@ksu.edu.sa, zuair@ksu.edu.sa), corresponding Aouther: ybazi@ksuedu.sa

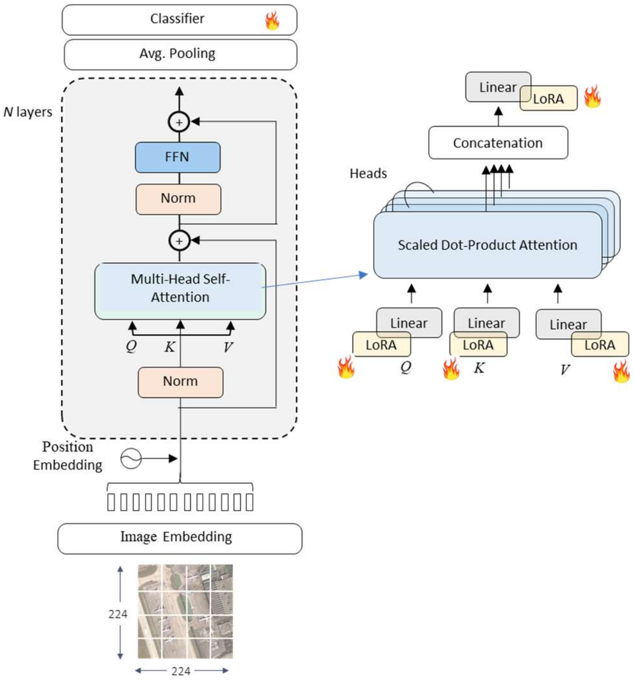


Figure 1. LoRA-CLIP: Incorporating LoRA weights into the linear layers of the multi-head self-attention in the transformer blocks of the CLIP vision encoder.

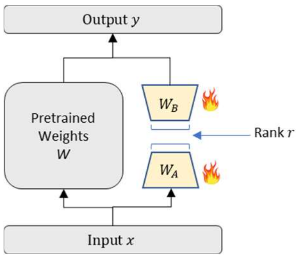


Figure 2. LoRA Mechanism: Injecting trainable weights into pretrained weight matrices in the transformer blocks while keeping the original model frozen.

By transferring knowledge from a contrastively pretrained model to a specialized RS task, this work demonstrates the scalability and adaptability of LoRA-CLIP models, providing a robust solution for RS scene classification.

II. LoRA-CLIP

A. CLIP Model

CLIP uses a dual-encoder architecture pairing (Vision Transformer (ViT) or Residual Networks (ResNet) with transformer text encoders, trained on 400M image-text pairs. Its contrastive learning paradigm enables zero-shot transfer by learning aligned visual-semantic representations. Unlike ImageNet models focused on classification, CLIP serves as a foundation model with broader visual understanding. Its scalability is demonstrated through models ranging from ResNet-50 (25M parameters) to ViT-L14 (300M parameters). This work will focus on leveraging ViT-L14 for scene classification.

B. LoRA-CLIP

Let S = $\{(X_i, y_i)\}_{i=1}^{N}$ represent a set of N images, where X_i denotes the image and $y_i \in \{1,2,\dots,m = C\}$ is its corresponding class label, with C representing the number of defined classes. The objective of is to efficiently map sequences of image patches to their semantic labels by incorporating LoRA weights into the attention mechanism of the vision encoder.

Figure 1 shows the overall architecture of the classification pipeline. It begins by dividing each input image into smaller patches of size (14×14 pixels) for an image of (224×224×3 pixels). These patches are linearly projected into embeddings of dimension 1024 and enriched with positional embeddings to encode spatial information. The vision encoder consists of $N = 24$ transformer blocks, each containing a Multi-Head Self-Attention (MHSA), with Query (Q), Key (K), and Value (V) projections of dimension 1024 and an Output (O) projection of the same dimension. The Feed-Forward Network (FFN) within each block has a two-layer structure: the first linear layer expands the dimensionality to 4096, followed by a GELU activation, and the second linear layer reduces it back to 1024. Each block includes layer normalization and residual connections to enhance training stability and gradient flow. For the classification task, we augment the encoder output (of dimension 1024) with average pooling and a classifier head to predict the semantic label.

Instead of fine-tuning the entire vision encoder with 300M parameters, we freeze its pre-trained weights and introduce LoRA modules into the (Q, K , V, and O) projection linear layers of MHSA for each transformer block. These LoRA modules add learnable low-rank updates, enabling task-specific adaptation with a minimal increase in parameters.

Specifically, for each weight matrix $W \in \mathcal{R}^{1024\times1024}$ in the attention mechanism as shown in Figure 2, we introduce low-rank matrices with learnable weights W_A and W_B such that the modified weight matrix is:

$$W_{new} = W + \alpha W_A W_B \tag{1}$$

Here W represents the original pre-trained weight matrix (e.g., for the queries W_Q, keys W_K, values W_V, or output projection W_O). $W_A \in \mathcal{R}^{1024\times r}$ and $W_A \in \mathcal{R}^{r\times1024}$ are the learnable low-rank matrices, and α is a scaling factor that controls the magnitude of the adaptation. The rank r of A and B is a hyperparameter, typically kept small (e.g., 4, 8, 16, or 32) to ensure computational efficiency, and reduce the number of additional parameters introduced during training. For example, ranks r equal to 4,8,16, and 32 add only 0.27%, 0.53%, 1.04% and 2.04% parameters, respectively. The scaling factor α is typically set to $2r$ to balance the adaptation strength and stability during fine-tuning, ensuring efficient training without overfitting.

During training, we minimize well-known cross-entropy loss between predicted and ground-truth labels:

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i log\big(p(y_i/X_i)\big) \tag{2}$$

where $p(y_i/X_i)$ is the softmax probability for the correct class. We optimize only the LoRA parameters (W_A, W_A matrices for Q, K , V, and O) in addition to the classifier mounted on the top of the vision encoder. This adaptation approach efficiently transfers CLIP's strong visual representations to RS scene classification tasks with minimal computational cost, using lightweight fine-tuning to avoid extensive retraining.

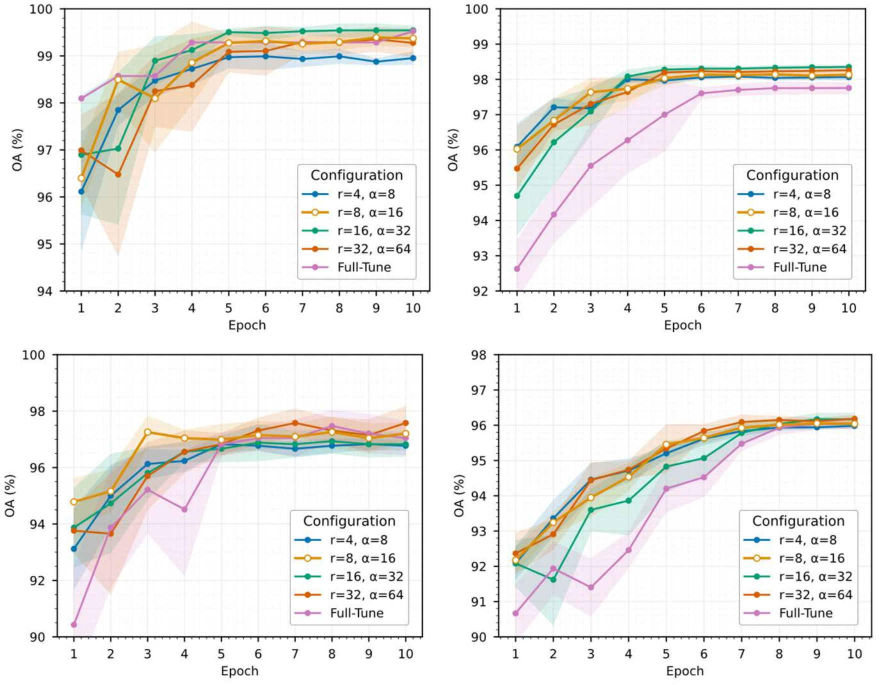
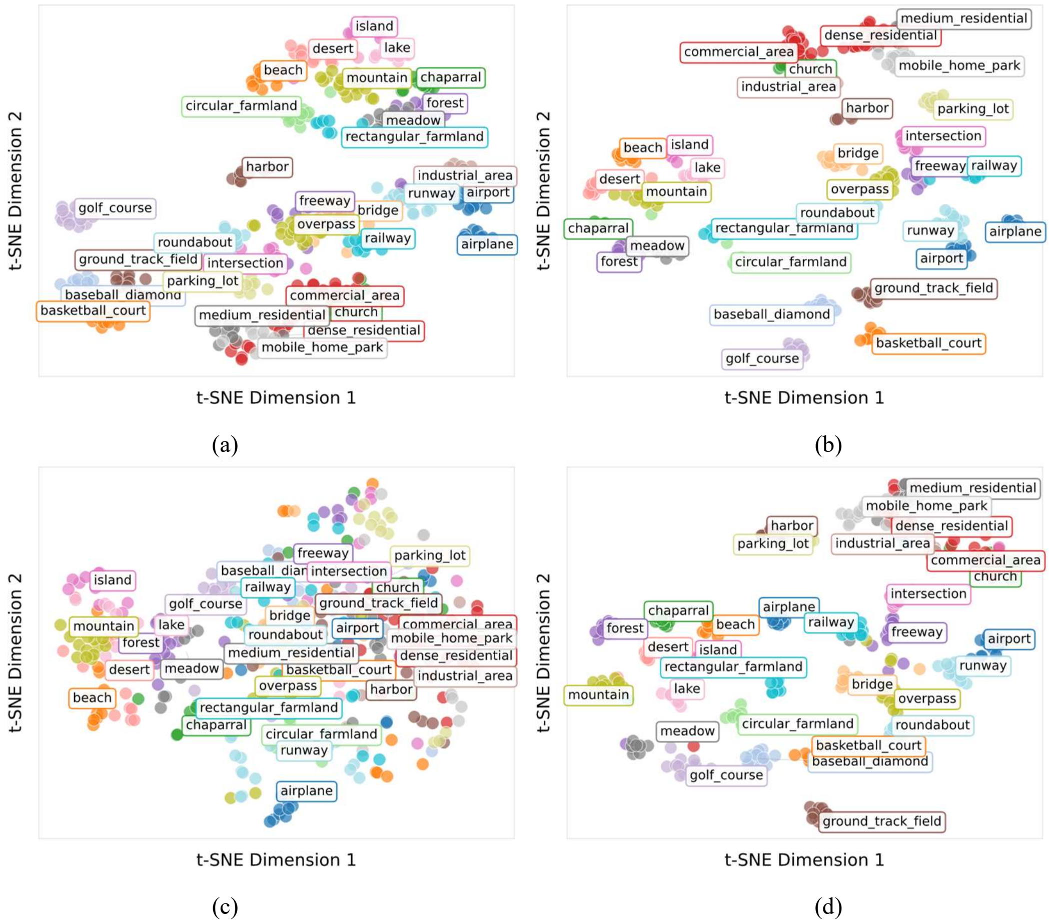


Figure 3. LoRA-CLIP Overall accuracy versus the number of epochs for: a) UCM (first row left), b) AID (first row right) c) Optimal-31 (second row left)), and d) NWPU (second row right) datasets.



(a)

(b)

(c)

(d)

Figure 4. t-SNE representation for Optimal31 dataset: LoRA-CLIP (a) before and (b) after training; and LoRA-Mamba (c) before and (a) after training.

III. EXPERIMENTAL RESULTS

A. Dataset Description and Experiment Settings

We utilize four well-known benchmark datasets for scene classification: UCM-Merced [13], AID [14], Optimal-31 [15], and NWPU-RESISC45 [16]. The UCM dataset contains 21 scene classes with 100 images per class, totalling 2,100 images. The AID dataset consists of 30 scene classes with approximately 200 to 400 images per class, amounting to 10,000 images. The Optimal-31 dataset includes 31 scene classes, each represented by 60 images sourced from Google Earth. Lastly, the NWPU-RESISC45 dataset is one of the largest benchmarks, featuring 45 scene classes with 700 images per class. For consistency, all images from these datasets are resized to 224 × 224 pixels to match the input requirements of the vision encoder.

We adopt AdamW optimization with 2e-4 initial learning rate decaying via cosine schedule to 2e-6 over 10 epochs. Data augmentation includes random horizontal/vertical flips (p=0.3). Following standard benchmarking protocols, we use established train/test splits (Merced: 80/20, AID: 20/80 and 50/50, Optimal-31: 80/20, and NWPU: 10/90 and 20/80). Experiments were conducted on an NVIDIA RTX A6000 (48GB GPU) with five repetitions. Results are reported as overall accuracy (OA) ± standard deviation.

B. Main Results

In examining the overall accuracy (OA) trajectories across epochs, the LoRA-CLIP model demonstrates rapid adaptation across all tested configurations (r, α) . The convergence patterns illustrated in Figure 3 reveal that regardless of the dataset (UCM, AID, Optimal-31, or NWPU), the model achieves substantial performance gains within the initial three epochs. Notably, higher rank configurations (r=16 and $r = 32$) consistently exhibit stronger early-epoch performance, while maintaining competitive convergence rates. This efficient convergence behaviour is particularly notable when compared to the full fine-tuning baseline, which typically demonstrates slower adaptation, especially evident in the AID and NWPU datasets. The comparative analysis at epoch 10 (Table 1) reveals the superior performance of LoRA-CLIP across multiple datasets: with r=16, α=32 achieving 98.35±0.05% on AID (20% training) and 98.88±0.12% on AID (50% training), while r=32, α=64 configuration reaches the highest performance on NWPU (95.30±0.28% and 96.18±0.11 at 10% and 20% training, respectively) and OPTIMAL-31 (97.58±0.59% at 80% training). These results consistently outperform established state-of-the-art methods, including, while maintaining significantly lower computational overhead through parameter-efficient fine-tuning.

C. Impact of Pretraining Strategies on Model Performance:

To comprehensively evaluate the effectiveness of CLIP-based pretrained models, we conduct extensive experiments comparing different architectures and pretraining paradigms, as shown in Table 2. Our evaluation includes ViT-B-16 (86M), VIT-L-14 (300M), and vision Mamba-B-16 (60M) and Mamba-L-14 (300M) models [17], a recent architecture leveraging state space modeling for efficient sequence processing. Qualitative analysis using t-SNE visualizations (Figure 4) demonstrates CLIP's strong zero-shot performance, with clear class clustering observed for example for OPTIMAL-31 dataset before fine-tuning. Although both CLIP and Mamba exhibit improved class separation following training, CLIP retains its initial clustering advantage. Quantitatively, the results shown in Table II demonstrate the superior performance of CLIP-based models. For example, with LoRA adaptation (r=8, α=16), across diverse RS datasets. Notably, CLIP-L achieves significant improvements over traditional ImageNet-pretrained models. The substantial performance advantage is even more pronounced in low-data regimes, where CLIP-L with LoRA adaptation achieves 95.18% accuracy on NWPU with only 10% training data, outperforming VIT-L by 4.55%.

This marked improvement can be attributed to two key factors: (1) the effectiveness of contrastive language-image pretraining in learning more robust and transferable visual representations, and (2) the scalability benefits of larger model architectures when combined with efficient adaptation strategies. The consistent superior performance across different datasets and training settings suggests that the text-image matching pretraining paradigm offers a more effective foundation for scene classification compared to traditional image-only pretraining approaches.

TABLE I OVERALL CLASSIFICATION ACCURACY IN (%)±STANDARD DEVITATION OBTAINED BY LORA-CLIP COMPARED TO STATE-OF-THE-ART METHODS.

Method	MERCED	AID		NWPU		OPTIMAL-31
	80% train	20% train	50% train	10% train	20% train	80% train
ResNet50+EAM [3]	98.98±0.37	94.26±0.11	97.06±0.19	91.91±0.22	92.95±0.09	96.45±0.28
MBLANet [18]	99.52±0.23	94.62±0.19	96.56±0.24	91.93±0.20	94.33±0.01	96.50±0.25
EMSCNET(ResNet50) [19]	99.44±0.16	95.13±0.10	96.96±0.10	92.16±0.07	94.08±0.20	--
CGINet [20]	**99.84±0.16**	95.35±0.14	97.10±0.24	92.28±0.17	94.38±0.13	97.35±0.14
HGTNet [2]	--	96.91±0.14	98.47±0.18	94.52±0.06	95.75±0.10	96.33±0.15
MopNet-GCN-ResNet50 [21]	--	95.53±0.11	97.11±0.07	--	--	95.34±0.31
LoRA-CLIP						
r=4, α=8 (0.5%)	98.95±0.13	98.06±0.09	98.65±0.05	94.97±0.21	95.98±0.06	96.77±0.17
r =8, α =16 (0.5%)	99.37±0.18	98.14±0.09	98.51±0.12	95.18±0.15	96.04±0.15	97.20±0.40

r =16, α =32 (0.5%)	99.54±0.09	**98.35±0.05**	**98.88±0.12**	95.15±0.22	96.17±0.18	96.83±0.43
r =32, α =64, (0.5%)	99.28±0.30	98.26±0.04	98.48±0.08	**95.30±0.28**	**96.18±0.11**	**97.58±0.59**
CLIP Full tuning	99.52±0.08	97.71±0.07	98.20±12	94.88±0.12	96.06±0.16	97.04±0.54

TABLE II PERFORMANCE COMPARAISION BETWEEN IMAGENET PRETRAINED MODELS AND CLIP MODELS FINTUNED USING LORA WITH r=8, AND α=16.

	Merced	AID		NWPU		OPTIMAL31
Model	*80% train*	*20% train*	*50% train*	*10% train*	*20% train*	*80% train*
LoRA-VIT-B (ImageNet)	97.94±12	92.30±0.44	95.97±0.26	88.53±0.52	92.35±0.32	89.52±1.50
LoRA-VIT-L (ImageNet)	98.33±0.52	94.05±0.23	96.70±0.11	90.63±0.36	93.67±0.10	91.72±1.39
LoRA-Mamba-B (ImageNet)	99.19±0.24	95.47±0.21	97.29±0.27	91.34±0.27	93.95±0.23	92.90±0.97
LoRA-Mamba-L (ImageNet)	99.24±0.21	95.59±0.25	97.20±0.31	90.38±1.15	93.95±0.20	93.44±1.23
LoRA-CLIP-B	99.38±0.29	95.08±0.19	97.92±12	93.48±0.11	94.99±0.16	94.95±0.46
LoRA-CLIP-L	**99.37±0.18**	**98.14±0.09**	**98.51±0.12**	**95.18±0.15**	**96.04±0.15**	**97.20±0.40**

IV. CONCLUSIONS

In this work, we introduced LoRA-CLIP, a method for efficient RS scene classification that adapts CLIP's vision encoder using low-rank updates to its attention layers. This approach demonstrates several key advantages, including consistent performance gains over ImageNet-pretrained models across multiple datasets, rapid convergence within three epochs, and minimal parameter overhead (ranging from 0.27% to 2.04%). Experimental results show superior accuracy on standard benchmarks, surpassing state-of-the-art methods. t-SNE visualizations further confirm CLIP's strong zero-shot transfer capabilities, suggesting that text-image pretraining offers a more effective foundation than image-only pretraining for scene classification. Future research could explore advanced LoRA variants, multimodal fusion with CLIP's text embeddings, and instruction-tuned models for universal RS scene classification. Additionally, investigating federated learning paradigms would be valuable, leveraging LoRA's minimal parameter overhead for scalable distributed training with reduced communication costs.

ACKNOWLEDGMENT

This research was supported by the researchers supporting project number (RSPD2025R995), King Saud University, Riyadh, Saudi Arabia.

REFERENCES

[1] Y. Bazi, M. M. Al Rahhal, H. Alhichri, and N. Alajlan, “Simple Yet Effective Fine-Tuning of Deep CNNs Using an Auxiliary Classification Loss for Remote Sensing Scene Classification,” *Remote Sensing*, vol. 11, no. 24, p. 2908, Jan. 2019, doi: 10.3390/rs11242908.

[2] Z. Li *et al.*, “A Hierarchical Graph-Enhanced Transformer Network for Remote Sensing Scene Classification,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 20315–20330, 2024, doi: 10.1109/JSTARS.2024.3491335.

[3] Z. Zhao, J. Li, Z. Luo, J. Li, and C. Chen, “Remote Sensing Image Scene Classification Based on an Enhanced Attention Module,” *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 11, pp. 1926–1930, Nov. 2021, doi: 10.1109/LGRS.2020.3011405.

[4] K. Chen, B. Chen, C. Liu, W. Li, Z. Zou, and Z. Shi, “RSMamba: Remote Sensing Image Classification With State Space Model,” *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024, doi: 10.1109/LGRS.2024.3407111.

[5] A. Radford *et al.*, “Learning Transferable Visual Models From Natural Language Supervision,” *arXiv:2103.00020 [cs]*, Feb. 2021, Accessed: Jan. 09, 2022. [Online]. Available: http://arxiv.org/abs/2103.00020

[6] Z. Zhang, T. Zhao, Y. Guo, and J. Yin, “RS5M and GeoRSCLIP: A Large-Scale Vision- Language Dataset and a Large Vision-Language Model for Remote Sensing,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–23, 2024, doi: 10.1109/TGRS.2024.3449154.

[7] M. M. Al Rahhal, Y. Bazi, H. Elgibreen, and M. Zuair, “Vision-Language Models for Zero-Shot Classification of Remote Sensing Images,” *Applied Sciences*, vol. 13, no. 22, p. 12462, 2023.

[8] F. Liu *et al.*, “RemoteCLIP: A Vision Language Foundation Model for Remote Sensing,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024, doi: 10.1109/TGRS.2024.3390838.

[9] “RS-LLaVA: A Large Vision-Language Model for Joint Captioning and Question Answering in Remote Sensing Imagery.” Accessed: Jan. 01, 2025. [Online]. Available: https://www.mdpi.com/2072-4292/16/9/1477

[10] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, “GeoChat: Grounded Large Vision-Language Model for Remote Sensing,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27831–27840. Accessed: Jan. 01, 2025. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Kuckreja_GeoChat_Grounded_Large_Vision-Language_Model_for_Remote_Sensing_CVPR_2024_paper.html

[11] E. J. Hu *et al.*, “LoRA: Low-Rank Adaptation of Large Language Models,” Oct. 16, 2021, *arXiv*: arXiv:2106.09685. doi: 10.48550/arXiv.2106.09685.

[12] B. Büyüktaş, G. Sumbul, and B. Demir, “Federated Learning Across Decentralized and Unshared Archives for Remote Sensing Image Classification: A review,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 12, no. 3, pp. 64–80, Sep. 2024, doi: 10.1109/MGRS.2024.3415391.

[13] Y. Yang and S. 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*, in GIS ’10. San Jose, California: Association for Computing Machinery, Nov. 2010, pp. 270–279. doi: 10.1145/1869790.1869829.

[14] G. Xia *et al.*, “AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017, doi: 10.1109/TGRS.2017.2685945.

[15] Q. Wang, S. Liu, J. Chanussot, and X. Li, “Scene Classification With Recurrent Attention of VHR Remote Sensing Images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019, doi: 10.1109/TGRS.2018.2864987.

[16] G. Cheng, J. Han, and X. Lu, “Remote Sensing Image Scene Classification: Benchmark and State of the Art,” *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017, doi: 10.1109/JPROC.2017.2675998.

[17] A. Hatamizadeh and J. Kautz, “MambaVision: A Hybrid Mamba-Transformer Vision Backbone,” Jul. 10, 2024, *arXiv*: arXiv:2407.08083. doi: 10.48550/arXiv.2407.08083.

[18] S.-B. Chen, Q.-S. Wei, W.-Z. Wang, J. Tang, B. Luo, and Z.-Y. Wang, “Remote Sensing Scene Classification via Multi-Branch Local Attention Network,” *IEEE Transactions on Image Processing*, vol. 31, pp. 99–109, 2022, doi: 10.1109/TIP.2021.3127851.

[19] “EMSCNet: Efficient Multisample Contrastive Network for Remote Sensing Image Scene Classification | IEEE Journals & Magazine | IEEE Xplore.” Accessed: Dec. 03, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10086539

[20] “Co-Enhanced Global-Part Integration for Remote-Sensing Scene Classification | IEEE Journals & Magazine | IEEE Xplore.” Accessed: Dec. 03, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10440366

[21] “Multi-Output Network Combining GNN and CNN for Remote Sensing Scene Classification.” Accessed: Dec. 03, 2024. [Online]. Available: https://www.mdpi.com/2072-4292/14/6/1478