IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Research Article)

Zero-shot aerial scene classification using clip and prompt engineering

Chukwudi Anthony Udemba *, Adekunle Adeoye Eludire and Ayorinde Peters Oduroye

Department of Computer Science, Caleb University, Lagos, Nigeria.

International Journal of Science and Research Archive, 2025, 17(02), 005–013

Publication history: Received on 23 September 2025; revised on 28 October 2025; accepted on 31 October 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.17.2.2948

Abstract

Traditional aerial scene classification models rely heavily on large, labeled datasets and supervised learning, which limits their ability to generalize to new or rare scene types. In this work, we explore a zero-shot approach to aerial scene understanding by leveraging Contrastive Language Image Pretraining (CLIP), a vision-language model trained on vast image-text pairs. Instead of retraining or fine-tuning the model, we use carefully designed natural language prompts to describe scene categories of interest and classify aerial images based on cosine similarity in a shared semantic embedding space. This method enables flexible and scalable scene classification without requiring additional annotation or retraining. Through prompt engineering, we introduce both generic and domain-specific textual descriptions to maximize classification accuracy. Experiments conducted on benchmark aerial datasets demonstrate that the proposed approach effectively distinguishes between complex and visually similar scenes, even in scenarios with limited or no prior class examples. This work highlights the potential of vision-language models for rapid, adaptable, and annotation-free classification in aerial surveillance applications.

Keywords. Zero-shot learning; Vision-language models; Remote sensing; Semantic embedding; Large language models

1. Introduction

Aerial scene classification is a critical task in remote sensing and surveillance, enabling the automated identification of terrain types, infrastructure, and human activities from drone, satellite, or aircraft imagery. These applications are utilized in areas such as disaster response, urban planning, military reconnaissance, and environmental monitoring. Traditional approaches to aerial scene understanding often rely on convolutional neural networks (CNNs) trained in a supervised fashion using large, labeled datasets [14]. While effective under well-defined conditions, these methods typically struggle with domain shifts, novel classes, and rare or low-sample categories, making them less scalable in real-world surveillance scenarios [3].

Recent advances in vision-language models (VLMs) offer a promising direction for addressing these challenges. Among them, CLIP has emerged as a powerful model capable of learning joint image-text representations through contrastive learning over vast amounts of internet-sourced image-caption pairs [7]. CLIP's most compelling feature is its zero-shot learning capability: it can classify images based solely on textual descriptions—known as prompts—without requiring task-specific retraining.

Zero-shot learning is particularly valuable in aerial domains, where collecting labeled data for every scene type is both costly and time-consuming. Instead of fine-tuning a model on each new category, CLIP enables flexible, prompt-driven classification, utilizing natural language descriptions to define class boundaries. This makes it especially suitable for tasks like identifying rare events, emergent objects, or novel infrastructure in aerial views [9, 10].

* Corresponding author: UDEMBA A. C

In this work, we explore a prompt-engineering approach to aerial scene classification using CLIP. By crafting descriptive, domain-specific prompts (e.g., "an aerial view of a solar farm"), we enable the model to semantically match image content with class labels in a shared embedding space. We investigate the effects of different prompt strategies, including generic, domain-specific, and contextual phrasing, on classification performance. We evaluate our method using benchmark datasets and compare it against conventional supervised baselines. This study demonstrates that CLIP, when paired with thoughtful prompt design, offers a scalable and flexible solution for aerial scene classification with minimal training overhead. The method provides an adaptable alternative to traditional supervised models, advancing the field toward a general-purpose, language-guided understanding of aerial imagery.

2. Literature Review

2.1. Traditional Aerial Scene Classification

Traditional aerial scene classification has primarily been driven by supervised learning approaches using hand-labeled datasets. Convolutional Neural Networks (CNNs), such as AlexNet, VGG, and ResNet, have achieved considerable success when trained on large-scale remote sensing image datasets, including NWPU-RESISC45, AID, and UC Merced [1, 2]. These models extract hierarchical features from aerial images to classify various land-use and land-cover types. However, they require extensive annotated data for each class and perform poorly when encountering novel categories or domain shifts, which are common in real-world aerial surveillance applications [8].

2.2. Zero-Shot Learning in Remote Sensing

Zero-shot learning (ZSL) attempts to bridge this limitation by enabling models to recognize unseen classes based on semantic descriptions or class attributes. Early approaches in remote sensing utilized external knowledge sources, such as WordNet, class attributes, or label embeddings, to enhance classification capabilities for novel categories [12]. However, these methods often require additional auxiliary data and manually defined attributes, limiting their flexibility and scalability. Recent advances in vision-language pretraining have significantly enhanced the capabilities of zero-shot models. Instead of relying on fixed class vectors, these models learn to align images and natural language in a shared semantic space, enabling more intuitive and robust classification. This shift is particularly advantageous for aerial imagery, where new or rare scene types frequently emerge.

2.3. CLIP and Vision-Language Models

Contrastive Language-Image Pretraining, developed by OpenAI, is a prominent example of a vision-language model trained on 400 million image-text pairs scraped from the internet [11]. CLIP simultaneously encodes images and text into a shared embedding space and learns to associate them based on similarity. It allows for zero-shot classification by replacing traditional output layers with a set of user-defined natural language prompts. Studies have shown that CLIP generalizes well across domains, including medical imaging, satellite photos, and art [5]. In remote sensing, researchers have begun exploring its utility for classifying aerial scenes, detecting land cover, and even generating image captions without task-specific training [1]. One key challenge, however, lies in crafting effective prompts. Poorly designed or ambiguous textual descriptions can significantly degrade performance, making prompt engineering a crucial component of CLIP's success in new domains.

2.4. Prompt Engineering and Semantic Adaptation

Prompt engineering refers to the practice of designing structured natural language descriptions that guide vision-language models toward desired outputs. Techniques include using class names in context ("a satellite image of a forest"), incorporating spatial or temporal cues, or using prompt ensembles for stability. In aerial scene classification, this enables researchers to adapt CLIP to domain-specific contexts without requiring the model to be retrained [13]. While most prior works focus on tuning CLIP for ground-level images, our work adapts prompt engineering specifically for aerial scene classification. We compare various prompt strategies-generic, domain-specific, and contextual to demonstrate their impact on performance in classifying rare or visually similar aerial scenes.

3. Methodology

This study presents a zero-shot framework for aerial scene classification using OpenAI's CLIP model, with a focus on prompt engineering to enhance semantic alignment between aerial images and textual descriptions. Unlike traditional supervised classifiers that require extensive labeled training data, this method utilizes natural language prompts to enable CLIP to classify unseen or rare scene types directly, without the need for retraining. The methodology consists of four core components: image preprocessing, prompt design, similarity computation, and evaluation.

3.1. Overview of the Proposed System

The proposed system takes an aerial image as input and performs scene classification based on a set of user-defined prompts. Each prompt represents a potential scene class in natural language. The image and prompts are encoded into a shared semantic embedding space using the pretrained CLIP model. Classification is achieved by computing the cosine similarity between the image embedding and each text embedding, and assigning the label of the most similar prompt.

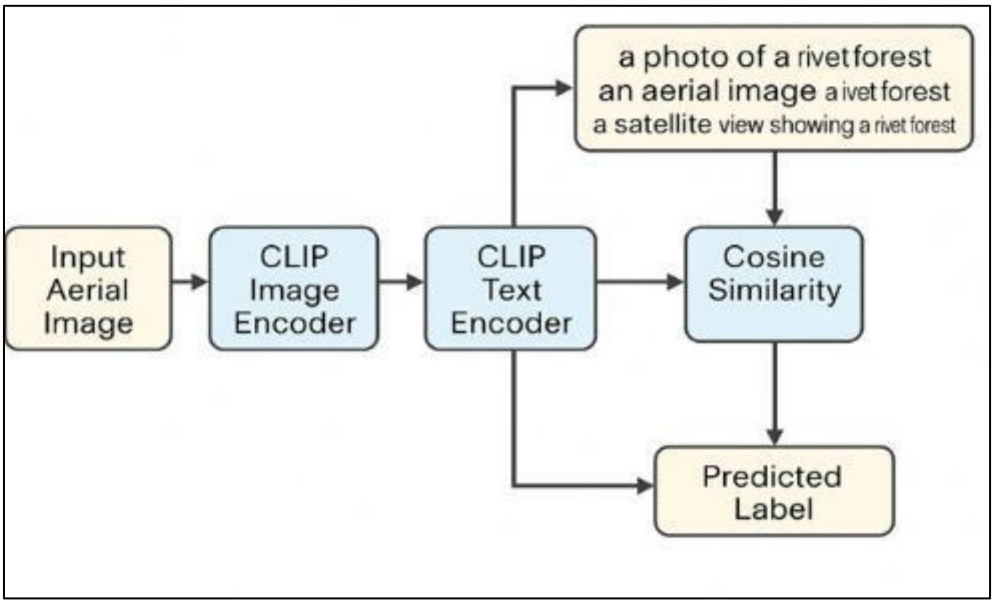


Figure 1 Workflow of the proposed zero-shot aerial scene classification system using CLIP and prompt engineering

The overall process includes the following steps:

- Image resizing and normalization.
- Tokenization and embedding of text prompts.
- Visual feature extraction via CLIP’s Vision Transformer (ViT-B/32).
- Cosine similarity computation.
- Scene label assignment based on the highest similarity.

This architecture enables fast and flexible scene classification with minimal data preparation.

3.2. CLIP Architecture and Pretrained Backbone

As shown in Figure 2, the CLIP model is composed of two encoders:

- Visual Encoder: A Vision Transformer (ViT-B/32), which converts the input image into a fixed-dimensional embedding.
- Text Encoder: A Transformer-based language model, which encodes tokenized natural language prompts.

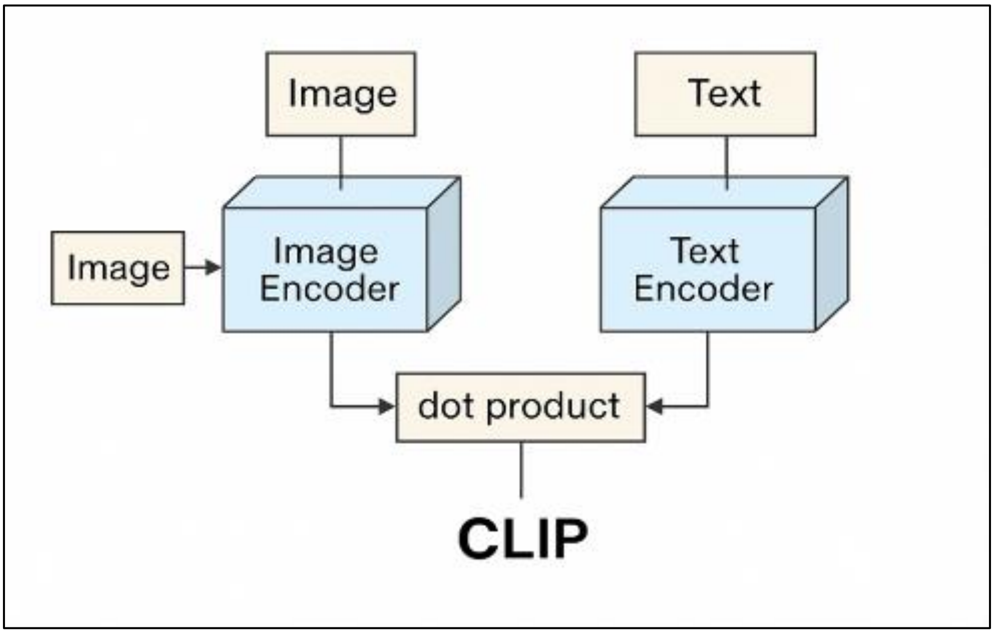


Figure 2 CLIP architecture showing visual and textual encoders projecting into a shared semantic space

CLIP was trained on 400 million image-text pairs using contrastive loss to ensure that corresponding image-text pairs are closer in embedding space than unrelated pairs [11]. For this study, we use the pretrained CLIP model without any fine-tuning.

3.3. Prompt Engineering Strategy

Prompt engineering is critical for zero-shot performance. Instead of using simple class labels (e.g., “airport”), we construct natural language prompts that provide scene context. We define three types of prompts:

- Generic Prompts: “a photo of a [scene]”
- Domain-Specific Prompts: “an aerial image of a [scene]”
- Contextual Prompts: “a satellite view showing a [scene] in an urban area”. For each class in the dataset, multiple prompts are generated. The embeddings of these prompts are averaged to create a more stable class representation (prompt ensembling).

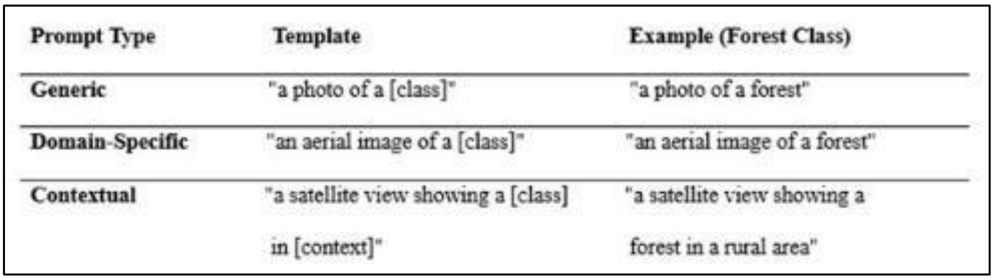
3.4. Dataset Preparation and Preprocessing

The benchmark aerial scene datasets are adopted, such as:

- NWPU-RESISC45: 45 scene categories, each with 700 images [4].
- AID (Aerial Image Dataset): 30 scene classes including farmland, port, forest, and residential areas [6].

Examples of the designed prompt templates are summarized in Table 1.

Table 1 Examples of prompt templates used for zero-shot classification



Prompt Type	Template	Example (Forest Class)
Generic	"a photo of a [class]"	"a photo of a forest"
Domain-Specific	"an aerial image of a [class]"	"an aerial image of a forest"
Contextual	"a satellite view showing a [class] in [context]"	"a satellite view showing a forest in a rural area"

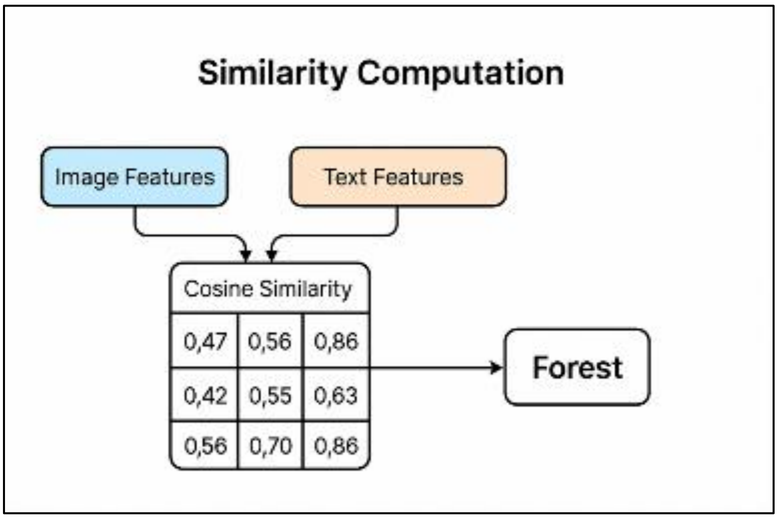
Each image is resized to 224 224 pixels and normalized to match CLIP's input requirements. No additional data augmentation is performed since the model is evaluated in zero-shot mode.

3.5. Similarity Computation and Classification

The similarity computation process is illustrated in Figure 3.

For each input image

- The CLIP image encoder produces a 512-dimensional feature vector.
- The CLIP text encoder generates embeddings for each prompt.
- Cosine similarity is calculated between the image vector and all prompt vectors.
- The class with the highest similarity score is assigned as the prediction.



Cosine Similarity		
0,47	0,56	0,86
0,42	0,55	0,63
0,56	0,70	0,86

Figure 3 Cosine similarity-based matching between image and prompt embeddings

Mathematically, if f_i is the image embedding and t_j is the j–th prompt embedding, the predicted class c is:

$$c = \text{argmax}_j \quad \frac{f_i t_j}{||f_i||\,||t_j||} \text{.............3.1}$$

3.6. Evaluation Protocol

The model is evaluated using:

- Top-1 Accuracy: Percentage of test images where the top prediction matches the ground truth.
- Top-5 Accuracy: Percentage where the correct label appears in the top five predictions.
- Confusion Matrix: To visualize class-wise performance.
- Prompt Sensitivity Analysis: Comparing different prompt strategies.

Experiments are repeated across different prompt sets to evaluate robustness and semantic flexibility.

4. Experimental Setup and Implementation

This section outlines the technical configuration, datasets, and evaluation protocol used to assess the proposed zero-shot aerial scene classification system. As this method relies on CLIP's pretrained capabilities, no model training was performed; instead, the system's performance depends on prompt engineering and the structure of the test data.

4.1. Hardware and Software Environment

All experiments were conducted in a controlled environment with the following specifications:

- Operating System: Ubuntu 22.04 LTS
- Processor: Intel Core i9-13900K @ 3.0 GHz
- GPU: NVIDIA RTX 3090 (24 GB VRAM)
- RAM: 64 GB DDR5
- Python Version: 3.10
- Libraries and Frameworks:
  - PyTorch 2.0.1
  - OpenAI CLIP (ViT-B/32 backbone)
  - NumPy, SciPy, scikit-learn, Matplotlib, Pandas

The CLIP model was used in inference-only mode without fine-tuning. Image and text embeddings were generated on the GPU and compared using cosine similarity.

4.2. Datasets Used

Two widely accepted aerial scene classification datasets were used to evaluate performance:

4.2.1. NWPU-RESISC45

- 31,500 images across 45 scene categories
- Resolution: 256 256 pixels
- Categories: airport, church, forest, river, stadium, etc.
- Split: 80% test images used for zero-shot classification (without label supervision)

4.2.2. AID (Aerial Image Dataset)

- 10,000 images across 30 scene classes
- Resolution: 600 600 pixels (resized to 224 224)
- Categories include industrial, commercial, residential, meadow, mountain, etc.
- Both datasets were used without training labels. Only test sets were evaluated using CLIP's pre-trained knowledge and engineered prompts.

4.3. Prompt Generation Strategy

Three prompt templates represented each scene category:

- Generic: "a photo of a [scene]"
- Domain-specific: "an aerial image of a [scene]"
- Contextual: "a satellite view showing a [scene] in an urban area"

Prompt embeddings were generated once per dataset and cached for performance. For each test image, the CLIP model computed cosine similarity between its visual embedding and each prompt embedding.

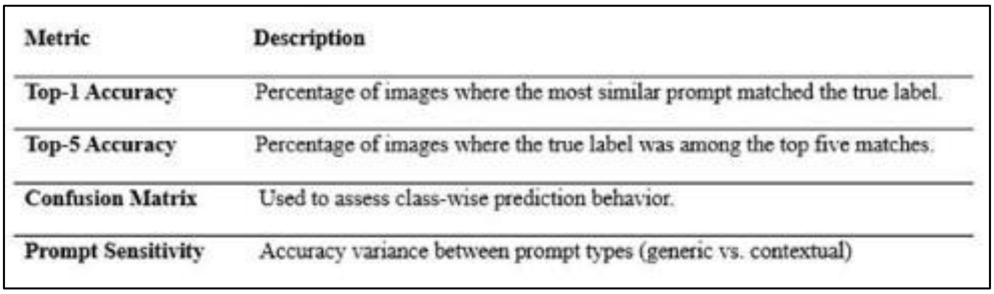
Prompt sets were tested:

- Individually (to evaluate the impact of phrasing),
- And as ensembles (average embedding of multiple prompts per class).

4.4. Evaluation Metrics

The system was evaluated using the following metrics as listed in table 2.

Table 2 Model Evaluation Metrics



Metric	Description
Top-1 Accuracy	Percentage of images where the most similar prompt matched the true label.
Top-5 Accuracy	Percentage of images where the true label was among the top five matches.
Confusion Matrix	Used to assess class-wise prediction behavior.
Prompt Sensitivity	Accuracy variance between prompt types (generic vs. contextual)

Performance was averaged over multiple runs to reduce stochastic variation, and confusion matrices were plotted to highlight scene-level misclassifications.

The accuracy achieved by each prompt type is shown in Figure 4**.**

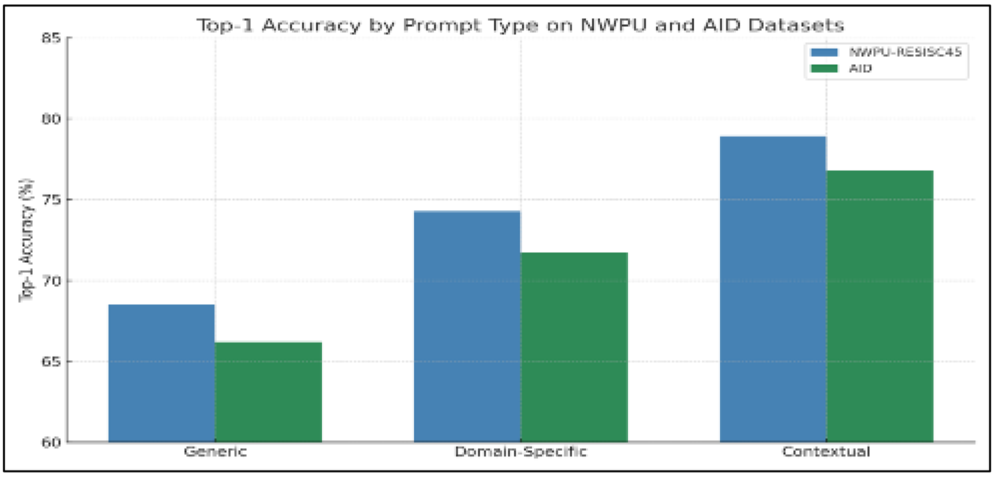
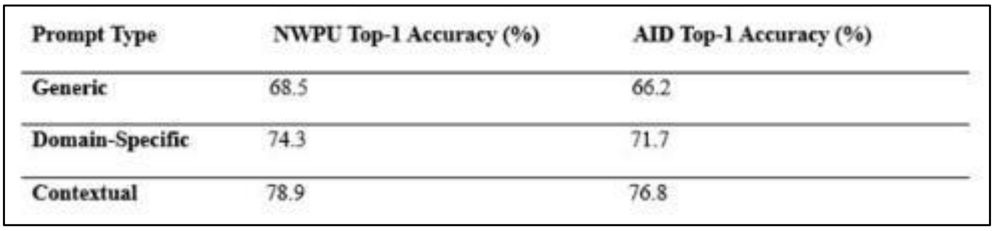


Figure 4 Accuracy by prompt type

5. Results and Discussion

This section presents the empirical evaluation of the zero-shot classification framework, utilizing CLIP and prompt engineering, on two benchmark aerial scene datasets: NWPU-RESISC45 and AID. The performance of three different prompt strategies: Generic, Domain-Specific, and Contextual is compared using Top-1 classification accuracy as illustrated in table 3.

Table 3 Prompt Strategies Type



Prompt Type	NWPU Top-1 Accuracy (%)	AID Top-1 Accuracy (%)
Generic	68.5	66.2
Domain-Specific	74.3	71.7
Contextual	78.9	76.8

5.1. Quantitative Results

As shown above, contextual prompts consistently outperform both generic and domain-specific prompts on both datasets. The improvement in accuracy reflects the benefit of incorporating scene-specific details in prompt design. On the NWPU dataset, contextual prompts yielded a 10.4% increase over generic prompts, while on AID, the improvement was 10.6%.

5.2. Prompt Sensitivity Analysis

The model's performance was sensitive to the phrasing and structure of the prompts. Generic prompts, such as "a photo of a [scene]," provided a weak signal due to a lack of spatial or domain-specific context. In contrast, contextual prompts, such as "a satellite view showing a [scene] in an urban area," encouraged more substantial semantic alignment with aerial perspectives, resulting in higher similarity scores and more accurate classifications.

5.3. Qualitative Observations

Misclassifications commonly occurred between semantically similar classes, such as:

- Meadow vs. Grassland
- Residential area vs. Commercial area
- Port vs. Shipyard
- Contextual prompts helped reduce such errors by providing clearer spatial anchors, e.g., "in a coastal region" or "surrounded by buildings."

6. Conclusion

This study proposes a zero-shot aerial scene classification framework that leverages the pre-trained CLIP vision-language model, combined with carefully engineered textual prompts. Unlike traditional supervised approaches that rely on extensive labeled training data, our method classifies aerial images based on semantic similarity between image embeddings and natural language descriptions. Through experiments on benchmark datasets (NWPU-RESISC45 and AID), we demonstrated that prompt engineering has a significant influence on classification accuracy, with contextual prompts outperforming both generic and domain-specific alternatives.

The results demonstrate that CLIP's zero-shot capabilities can be effectively extended to aerial imagery without requiring retraining or fine-tuning, providing a flexible and scalable solution for real-world applications where labeled data is scarce or unavailable. The system also enables rapid adaptation to novel or rare scene categories by simply modifying textual prompts. This highlights the potential of large vision-language models to augment or even replace traditional classification pipelines in remote sensing.

While the findings are promising, several directions remain open for future exploration, such as extending the system to support multilingual prompts or incorporating metadata, including GPS coordinates, altitude, or sensor type, which could enhance classification accuracy and accessibility. Also, future efforts can incorporate temporal analysis (e.g., changes over time) or object co-occurrence patterns to support higher-level scene interpretation, such as detecting activities or events. By pursuing these directions, future work can unlock the full potential of vision-language models for intelligent, generalizable aerial surveillance and remote sensing applications.

Compliance with ethical standards

Acknowledgements

The authors acknowledge the support of Caleb University, Lagos, for providing the academic environment and institutional resources that facilitated this research.

Disclosure of Conflict of Interest

The authors declare there is no conflict of interest regarding this manuscript.

References

[1] F. Chen and J. Y. Tsou, Assessing the effects of convolutional neural network architectural factors on model performance for remote sensing image classification: An in-depth investigation, International Journal of Applied Earth Observation and Geoinformation, 112 (2022), 102865.

[2] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao, Review of image classification algorithms based on convolutional neural networks, Remote Sensing, 13(22) (2021), 4712.

[3] X. Chen, W. Jiang, H. Qi, M. Liu, H. Ma, P. L. H. Yu, Y. Wen, Z. Han, S. Zhang, and G. Cao, Adaptive meta-knowledge transfer network for few-shot object detection in very high resolution remote sensing images, International Journal of Applied Earth Observation and Geoinformation, 127 (2024), 103675.

[4] Y. Chen, T. Zheng, J. Han, M. Zheng, and F. Zheng, Remote sensing image scene classification based on multi-level feature fusion, in Proceedings of the 2021 7th International Conference on Computer and Communications (ICCC), IEEE, December 2021, pp. 816–820.

[5] S. Eslami, G. de Melo, and C. Meinel, Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain?, arXiv preprint, arXiv:2112.13906, 2021.

[6] Y. Hua, L. Mou, P. Jin, and X. X. Zhu, MultiScene: A large-scale dataset and benchmark for multiscene recognition in single aerial images, IEEE Transactions on Geoscience and Remote Sensing, 60 (2021), pp. 1–13.

[7] A. Khan, L. Asmatullah, A. Malik, S. Khan, and H. Asif, A Survey on Self-supervised Contrastive Learning for Multimodal Text-Image Analysis, arXiv preprint, arXiv:2503.11101, 2025.

[8] R. Khanam, M. Hussain, R. Hill, and P. Allen, A comprehensive review of convolutional neural networks for defect detection in industrial applications, IEEE Access, 12 (2024), pp. 94250–94295.

[9] R. Sapkota and M. Karkee, Object detection with multimodal large vision-language models: An in-depth review, Available at SSRN 5233953 (2025).

[10] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, et al., LAION-5B: An open large-scale dataset for training next generation image-text models, Advances in Neural Information Processing Systems, 35 (2022), pp. 25278–25294.

[11] K. N. Singh, S. D. Devi, H. M. Devi, and A. K. Mahanta, A novel approach for dimension reduction using word embedding: An enhanced text classification approach, International Journal of Information Management Data Insights, 2(1) (2022), 100061.

[12] M. Singha, H. Pal, A. Jha, and B. Banerjee, Ad-CLIP: Adapting domains in prompt space using CLIP, in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4355–4364.

[13] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, A review of convolutional neural networks in computer vision, Artificial Intelligence Review, 57(4) (2024), 99.

[14] Z. Zhao, Y. Liu, H. Wu, M. Wang, Y. Li, S. Wang, and D. Shen, CLIP in medical imaging: A comprehensive survey, arXiv preprint, arXiv:2312.07353, 2023.