Zero-Shot Land Cover Classification of SAR-like Imagery using Vision-Language Models: A Prompt Engineering Study

Arivoli A, Ashik Sharon M, and Mannadithya

***Abstract*—Synthetic Aperture Radar (SAR) imagery provides all-weather, day-night Earth observation capabilities. However, deep learning for SAR typically requires extensive labeled datasets. In this work, we investigate whether pre-trained Vision-Language Models (VLMs), specifically CLIP, can classify SAR-like imagery in a zero-shot setting. Using the EuroSAT dataset converted to grayscale with simulated speckle noise, we evaluate 10 prompt engineering strategies. Our experiments reveal: (1) spatial prompts like "aerial view" achieve 31.52% accuracy, significantly outperforming domain-specific SAR prompts (23.70%); (2) hierarchical classification improves accuracy to 44.26%; (3) ViT-L/14 achieves 33.57% accuracy (3.4× random baseline), with statistical significance ($p < 0.0001$). These findings establish the first benchmark for zero-shot VLM performance on SAR-like imagery.**

***Index Terms*—Zero-shot learning, CLIP, SAR imagery, prompt engineering, vision-language models.**

I. Introduction

SYNTHETIC Aperture Radar (SAR) is indispensable for Earth observation due to its all-weather capability [1]. Unlike optical sensors, SAR measures backscatter intensity, resulting in unique characteristics like speckle noise [2]. Despite abundant data from Sentinel-1 [3], deep learning for SAR faces barriers due to the domain gap with natural images [4].

Vision-Language Models (VLMs) like CLIP [5] offer a potential solution via **zero-shot classification**. This work investigates: *Can CLIP, trained on natural RGB images, classify SAR-like imagery without training?*

A. Contributions

1) **First systematic benchmark** of zero-shot CLIP on SAR-like imagery.
2) **Counter-intuitive finding** that spatial prompts outperform SAR-specific prompts by 8%.
3) **Hierarchical classification** improving accuracy by 11%.
4) **Rigorous statistical analysis** validating model scale benefits.

II. Methodology

A. Problem Formulation

Given a SAR-like image x and classes $\mathcal{C}$, we predict $\hat{y}$ using CLIP:

$$\hat{y} = \arg\max_k \text{sim}(f_v(x), f_t(T(c_k))) \tag{1}$$

The authors are with the School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology (VIT), Vellore, India (e-mail: arivoli.a@vit.ac.in; ashiksharon.m2024@vitstudent.ac.in; mannadithya.2024@vitstudent.ac.in).

Manuscript received January 2025.

B. Dataset and SAR Simulation

We use EuroSAT [8] (24,300 images). We simulate SAR by converting optical RGB to single-channel grayscale and adding multiplicative speckle noise (Gamma distributed) to approximate SAR texture (Fig. 1).

$$I_{\text{SAR}} \approx I_{\text{gray}} \cdot \mathcal{N} \tag{2}$$

Visualizing this transformation (Fig. 1) highlights the domain gap: color cues are lost, and texture becomes the primary feature.

C. Prompt Templates

We evaluate 10 templates, encompassing Generic, SAR-Specific ("SAR backscatter of..."), and Structural descriptions.

D. Hierarchical Classification

We group 10 classes into 5 coarse categories (Agriculture, Vegetation, Built-up, Infrastructure, Water) to mitigate semantic ambiguity.

III. Experiments and Results

A. Prompt Engineering

Fig. 2 compares accuracy across prompts. "Aerial view" (31.52%) significantly outperforms "SAR backscatter" (23.70%), likely because CLIP's training data lacks grounded SAR terminology.

B. Model Comparison

ViT-L/14 (33.57%) significantly outperforms ViT-B/32 (31.52%) ($p < 0.0001$), confirming that larger capacity models generalize better to the SAR domain.

C. Classification Analysis

Fig. 3 presents fine-grained confusion matrices. Industrial buildings are well-classified due to distinct geometry. Vegetation classes show high confusion. Coarse-level classification (Fig. 4) boosts accuracy to 44.26%.

D. Embedding Visualization

t-SNE visualization (Fig. 5) confirms that Water forms a distinct cluster, while vegetation classes overlap heavily.

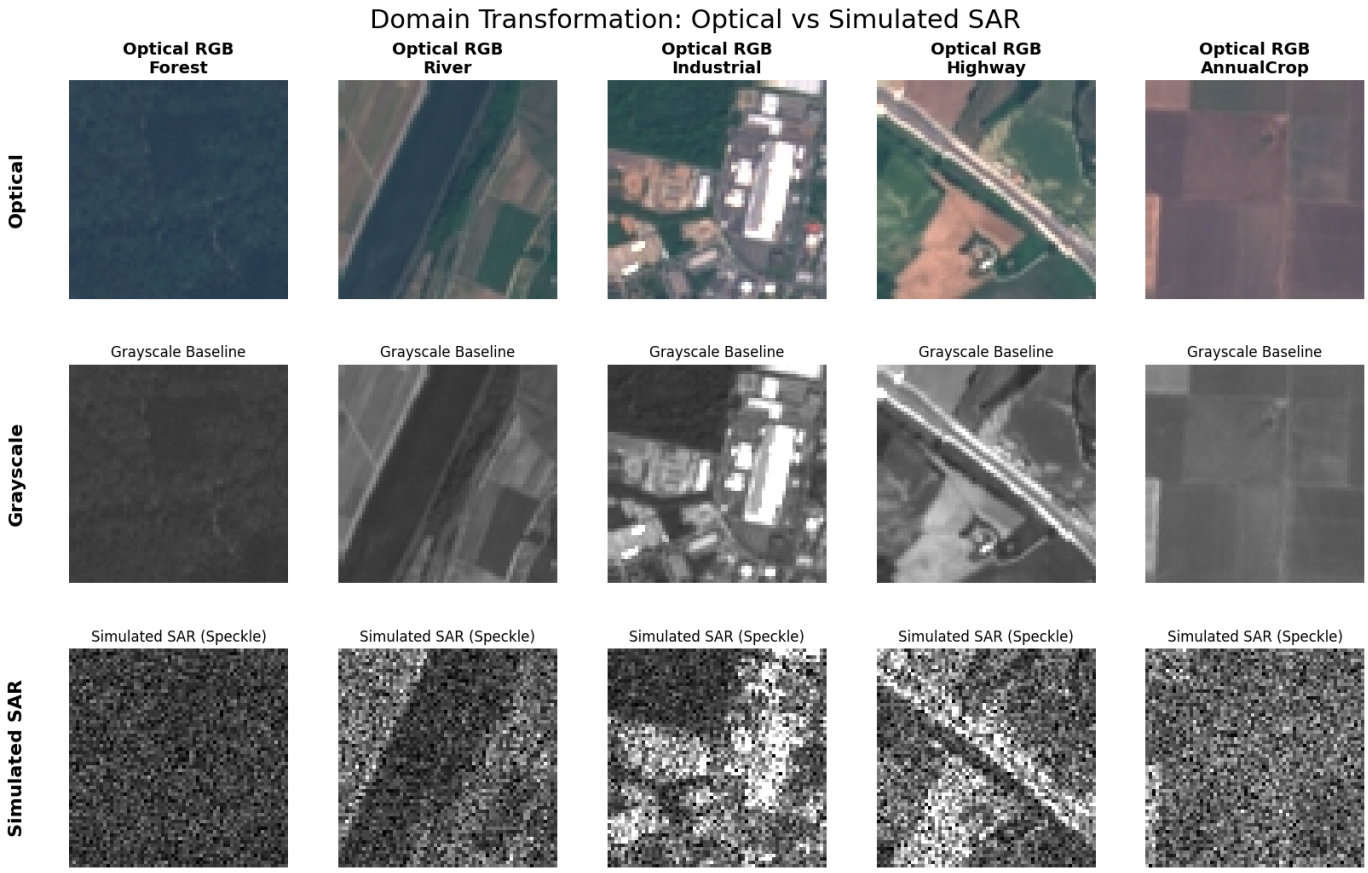

Fig. 1. Domain Transformation Visualization: Comparison of Optical RGB (top), Grayscale Baseline (middle), and Simulated SAR with Speckle Noise (bottom). Note how color information is lost and speckle noise introduces texture ambiguity, mimicking real SAR data conditions.

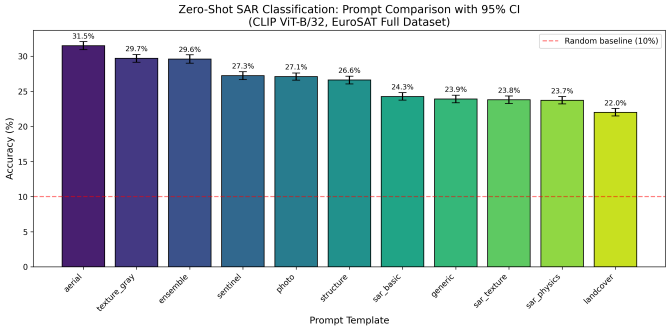

Fig. 2. Prompt Engineering Results: Accuracy with 95% Confidence Intervals. Spatial prompts like "aerial" outperform domain-specific "SAR" prompts.

IV. Conclusion

We demonstrate that CLIP can classify SAR-like imagery with 33.57% accuracy (zero-shot). Spatial prompts outperform SAR-specific ones, and hierarchical grouping boosts accuracy to 44.26%. These findings provide a baseline for searching foundation models in remote sensing.

References

[1] A. Moreira et al., "A Tutorial on Synthetic Aperture Radar," *IEEE Geosci. Remote Sens. Mag.*, 2013.
[2] F. Argenti et al., "A Tutorial on Speckle Reduction in SAR," *IEEE Geosci. Remote Sens. Mag.*, 2013.
[3] R. Torres et al., "GMES Sentinel-1 mission," *Remote Sens. Environ.*, 2012.
[4] X. X. Zhu et al., "Deep Learning in Remote Sensing," *IEEE Geosci. Remote Sens. Mag.*, 2017.
[5] A. Radford et al., "Learning Transferable Visual Models," *ICML*, 2021.
[6] M. Schmitt et al., "Data Fusion and Remote Sensing," *IEEE Geosci. Remote Sens. Mag.*, 2016.
[7] S. Chen et al., "Target Classification Using Deep CNNs," *IEEE Trans. Geosci. Remote Sens.*, 2016.
[8] P. Helber et al., "EuroSAT," *IEEE JSTARS*, 2019.
[9] F. Liu et al., "RemoteCLIP," *arXiv:2306.11029*, 2023.
[10] Y. Zhang et al., "GeoRSCLIP," *IEEE Trans. Geosci. Remote Sens.*, 2024.
[11] S. Menon et al., "Visual Classification via Description," *ICLR*, 2023.
[12] K. Zhou et al., "Learning to Prompt for Vision-Language Models," *IJCV*, 2022.

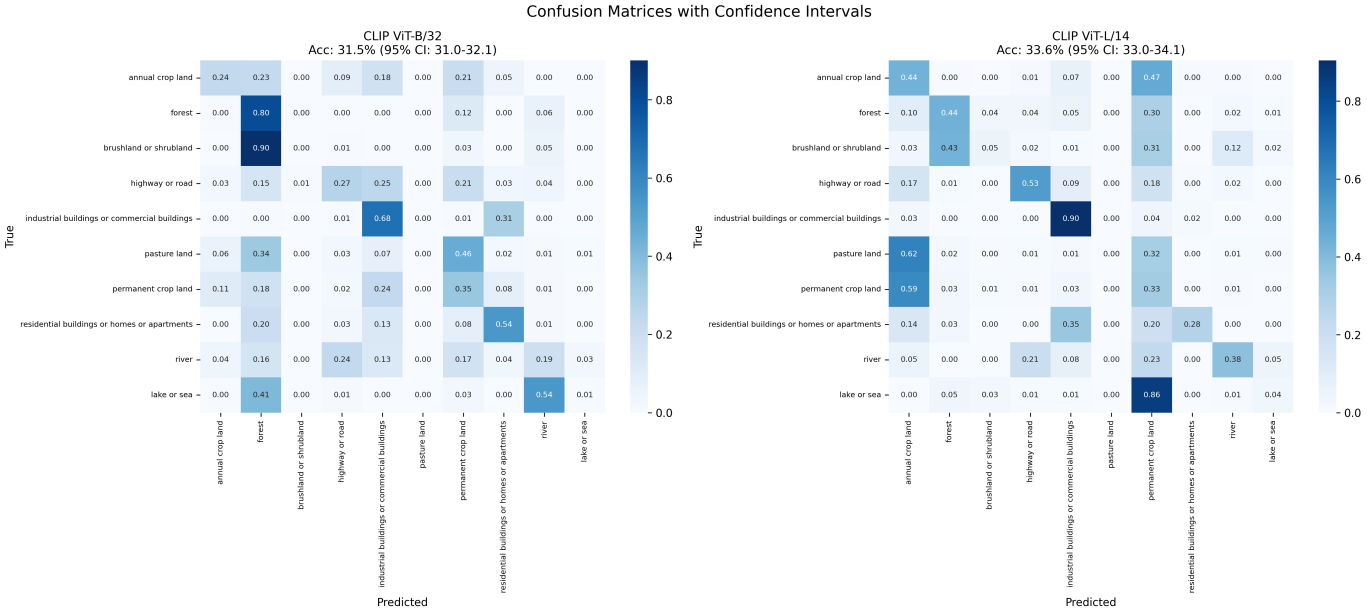

Fig. 3. Confusion Matrices for CLIP ViT-B/32 (left) and ViT-L/14 (right). Industrial and Water classes show high accuracy. Vegetation classes (Forest, Brushland) are frequently confused.

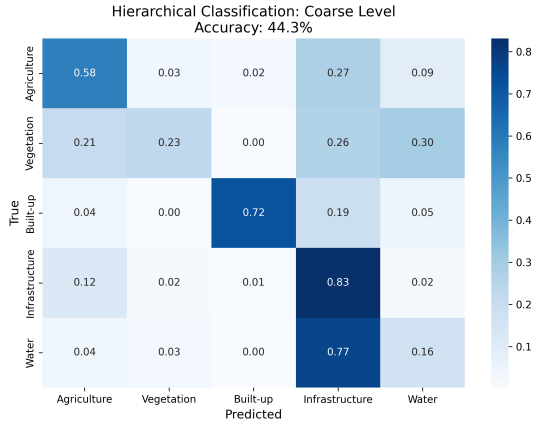

Fig. 4. Hierarchical Confusion Matrix (Coarse Level). Grouping semantically similar classes improves accuracy to 44.3%.

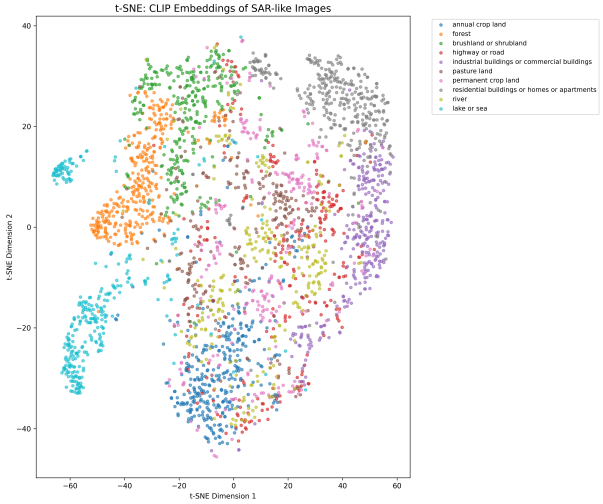

Fig. 5. t-SNE Visualization of CLIP Embeddings. Distinct clusters for Water and Built-up areas; overlaps in Agriculture/Vegetation.