CS 766 Final Project Proposal

Chen Wei

1 Problem Statement

Semantic segmentation of complex urban streetscapes remains a significant challenge in computer vision due to the open-vocabulary nature of real world environments and the lack of fine-grained annotations (Perez and Fusco 2025). While zero-shot Vision-Language Models (VLMs) show promise in general image parsing, they often fail to capture domain-specific, localized visual features (e.g., obstructed traffic signs) without specific optimization (Zhong et al. 2025). This project aims to solve this core computer vision challenge by applying Parameter-Efficient Fine-Tuning (PEFT) (Ding et al. 2023) to adapt a pre-trained VLM for specialized spatial feature extraction. Specifically, we will implement Low-Rank Adaptation (LoRA) (E. J. Hu et al. 2022) to optimize the VLM to extract structured, pixel-level streetscape indicators from multi-view Street View Imagery. The efficacy of this optimized CV pipeline will then be evaluated through a robust downstream application: interpretable driving risk assessment.

2 Motivation

The primary motivation of this project is to demonstrate advanced computer vision learning principles, specifically network weight optimization, domain adaptation, and pixel-level mask generation. Relying solely on zero-shot inference in VLMs often leads to hallucinations or homogenized outputs when analyzing factors (Z. Li et al. 2025). By implementing LoRA on a frozen VLM, this project achieves specialized segmentation accuracy without the prohibitive computational costs of full-parameter training. The downstream driving risk application provides a quantitative framework to validate our vision pipeline's feature extraction capabilities, moving beyond pure visual parsing to demonstrate how CV techniques can derive causal, actionable insights from raw imagery.

3 Proposed Solution

The project will utilize a dataset of multi-view street view imageries sampled from real-world intersections. A small, representative subset of these images will be manually annotated to highlight specific traffic safety elements to serve as the ground truth for the vision fine-tuning phase.

We will fine-tune a foundation VLM using LoRA. By freezing the pre-trained model weights and injecting trainable rank decomposition matrices into the Transformer layers (E. J. Hu et al. 2022), the model will learn to specialize in segmenting open-vocabulary urban features. We will evaluate this learning phase using standard pixel-wise segmentation loss functions.

The fine-tuned VLM will process the full SVI dataset to generate pixel-level semantic masks. Following established metrics (H. Chen et al. 2025), we will apply morphological operations to compute over ten quantitative indicators across four categories:

- Traffic Safety Risk Indicators: Background Complexity and Sight Obstruction Risk.
- Spatial Structure Indicators: Building Obstruction Ratio, Visible Obstacle Density, and Visual Openness.
- Environmental Element Indicators: Drivable Area Ratio, Emergency Space, Sidewalk Ratio, and Vegetation Coverage.
- Traffic Facility Indicators: Traffic Sign Integrity and Vehicle Density.

To evaluate the quality of the visual features extracted by our CV pipeline, the indicators will be fed into an eXtreme Gradient Boosting (XGBoost) classifier (T. Chen and Guestrin 2016) to predict accident categories. We will apply Shapley Additive Explanations (TreeSHAP) (Lundberg and Lee 2017) to extract feature attributions, identifying exactly how the segmented visual elements drive downstream predictions.

4 Expected Outcomes

By the conclusion of this project, we expect to deliver:

1. A fully implemented LoRA fine-tuning pipeline for semantic segmentation of street-view imagery.
2. A quantitative evaluation demonstrating the segmentation accuracy improvements of the fine-tuned model over a zero-shot baseline.
3. Global and local SHAP dependency plots revealing the hierarchical importance of the extracted visual features.

5 Timeline

- **Week 5 - 6:** Data preparation.
- **Week 7 - 8:** Baseline model implementation and initial training.
- **Week 9 - 10:** Full-dataset inference fine-tuning and exploring.
- **Week 11 - 12:** Comparative evaluation.
- **Week 13 - 14:** Report writing and visualization of results.

References

Chen, Huan et al. (2025). "Semantic4Safety: Causal Insights from Zero-shot Street View Imagery Segmentation for Urban Road Safety". In: *Proceedings of the 8th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pp. 146–156.

Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Ding, Ning et al. (2023). "Parameter-efficient fine-tuning of large-scale pre-trained language models". In: *Nature machine intelligence* 5.3, pp. 220–235.

Hu, Edward J et al. (2022). "Lora: Low-rank adaptation of large language models." In: *Iclr* 1.2, p. 3.

Li, Zongxia et al. (2025). "Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding". In: *arXiv preprint arXiv:2505.01481*.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30.

Perez, Joan and Giovanni Fusco (2025). "Streetscape analysis with generative AI (SAGAI): Vision-language assessment and mapping of urban scenes". In: *Geomatica*, p. 100063.

Zhong, Yifan et al. (2025). "A survey on vision-language-action models: An action tokenization perspective". In: *arXiv preprint arXiv:2507.01925*.