

Open-Vocabulary Semantic Segmentation of Aerial Photos

Luis Pedro Soares Marnoto Gaspar Lopes
luis.marnoto.gaspar.lopes@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

September 2025

Abstract

Referring expression segmentation represents a fundamental challenge in computer vision that integrates natural language understanding with precise visual localization. Existing datasets for referring expression segmentation focus primarily on natural scene imagery, leaving significant limitations in aerial domain applications where objects exhibit unique spatial configurations and contextual relationships. To facilitate the development of this field, we introduce Aerial-D, the largest referring expression segmentation dataset for aerial imagery to date, comprising 37,288 image patches with over 1.5 million referring expressions covering 259,709 annotated targets across individual objects, groups, and semantic categories spanning 21 distinct classes from vehicles and infrastructure to land cover types. The dataset represents the first fully automatic construction pipeline in this field, using systematic rule-based generation followed by Large Language Model enhancement that significantly enriched both the linguistic variety and visual detail richness of the referring expressions. We demonstrate good generalization results when models trained on Aerial-D are evaluated on other aerial segmentation datasets, highlighting the dataset’s effectiveness for aerial referring expression tasks. The dataset is publicly available at <https://huggingface.co/datasets/luisml77/aerial-d>.

Keywords: Aerial imagery, referring expression segmentation, dataset, large language models, computer vision

1. Introduction

Referring expression segmentation represents a fundamental advancement in computer vision, enabling models to identify and segment objects based on natural language descriptions rather than predefined category labels. This capability bridges the gap between human language understanding and visual perception, allowing for more intuitive and flexible interaction with computer vision systems. While significant progress has been achieved in natural scene referring expression segmentation, aerial imagery remains largely unexplored despite its critical importance in applications ranging from urban planning and environmental monitoring to autonomous navigation and disaster response.

Aerial imagery presents distinct challenges that differentiate it fundamentally from natural scene photography. Objects in aerial images exhibit extreme density variations, with single images potentially containing hundreds of vehicles, buildings, or infrastructure elements. The top-down perspective creates unique spatial relationship patterns not present in ground-level photography, where traditional concepts like “above” and “below” take on different meanings within the context of geographic positioning. Additionally, aerial images capture

vast scale variations, from individual vehicles measuring mere pixels to large building complexes spanning significant portions of the image frame.

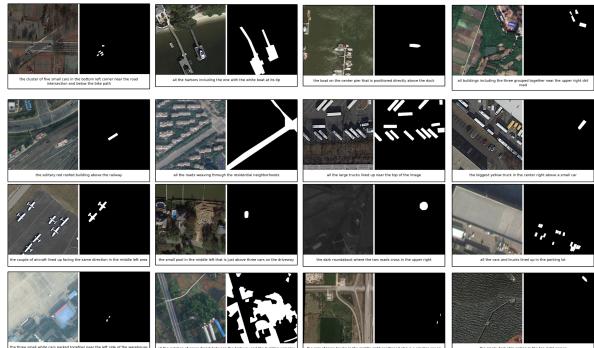


Figure 1: Representative examples from Aerial-D dataset showing diverse referring expressions with corresponding aerial images and ground truth masks.

Existing referring expression datasets, including RefCOCO, RefCOCO+, and RefCOCOg, focus exclusively on natural scenes with ground-level photography. These datasets typically contain objects with familiar human-centric spatial relationships

and conventional viewing angles. The linguistic patterns and spatial reasoning required for aerial imagery fundamentally differ from these established benchmarks, necessitating specialized dataset construction approaches that can capture the unique characteristics of overhead perspectives.

Current aerial image datasets, such as iSAID and LoveDA, provide excellent resources for traditional object detection and semantic segmentation tasks but lack the natural language component essential for referring expression applications. This limitation prevents the development and evaluation of aerial-specific referring segmentation models, creating a significant gap in the computer vision research landscape. The absence of large-scale aerial referring expression datasets has hindered progress in developing models capable of understanding complex spatial relationships and object descriptions within aerial contexts.

To address these limitations, we present Aerial-D, the first comprehensive referring expression segmentation dataset specifically designed for aerial imagery. Our dataset construction approach combines systematic rule-based expression generation with large language model enhancement to create diverse, natural, and contextually rich referring expressions. The resulting dataset contains over 1.5 million expressions across 37,288 aerial image patches, representing the largest collection of aerial referring expressions available to the research community.

Our key contributions include: (1) the introduction of Aerial-D, the first large-scale aerial referring expression segmentation dataset with over 1.5 million expressions, (2) a fully automatic dataset construction pipeline that leverages rule-based generation and LLM enhancement techniques, (3) comprehensive benchmarking results demonstrating the unique challenges of aerial referring expression segmentation, and (4) cross-dataset evaluation showing good generalization performance of models trained on our dataset.

2. Related Work

3. Aerial-D Dataset Construction

3.1. Rule-Based Expression Generation

3.2. LLM Expression Generation

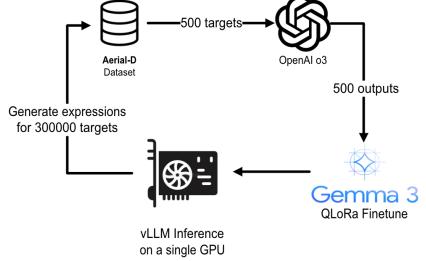


Figure 4: Knowledge distillation pipeline for scalable LLM enhancement. A small sample of 500 expressions is processed through OpenAI’s O3 model to generate high-quality training targets, which are then used to fine-tune Gemma3 12B via QLoRa. The fine-tuned model enables cost-effective local inference to enhance the full dataset of 300,000 expressions using vLLM on a single GPU.

4. Experiments

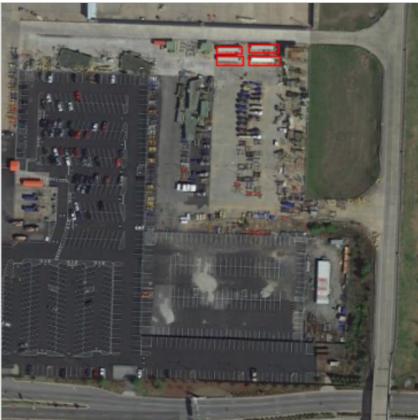
5. Conclusion and Future Work

Acknowledgements



Rule Type	Example Instance
Category	"plane"
Grid Position	"in the top right"
Extreme Position	None
Color Classification	"light"
Directional Relations	"to the bottom right of a plane" "to the top right of a plane"
Final Expressions	
"the plane in the top right" "the light plane in the top right" "the plane in the top right to the bottom right of a plane" "the light plane in the top right to the bottom right of a plane" "the plane in the top right to the top right of a plane" "the light plane in the top right to the top right of a plane"	

Figure 2: Example of rule generation for a single instance. The highlighted plane in the top right section demonstrates how the system assigns spatial, visual, and relational rules that will later be combined into referring expressions.



Expression Type	Example
Original	the group of 4 large vehicles in the top center
Enhanced	the cluster of four big vehicles near the upper middle
Unique	the four large vehicles lined up side by side just below the pale paved strip at the very top middle
Unique	the set of four big vehicles parked in a single row in the upper center beside the grassy area to the right

Figure 3: Example of LLM enhancement process showing original aerial image with group of four large vehicles (left) and corresponding expression enhancements (right).