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 enables models to identify and segment objects using natural language descriptions rather than predefined labels, bridging human language understanding and visual perception. While significant progress exists in natural scene segmentation, aerial imagery remains largely unexplored despite its importance in urban planning, environmental monitoring, and autonomous navigation.

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.

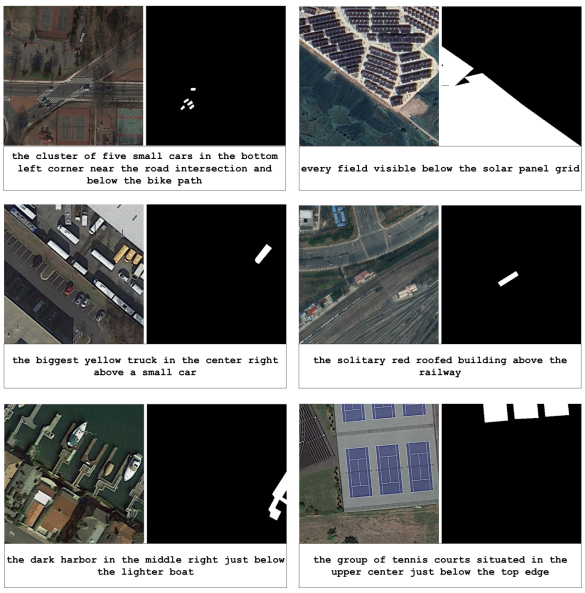


Figure 1: Representative examples from Aerial-D dataset showing diverse referring expressions with corresponding aerial images and ground truth masks.

Existing referring expression datasets, including RefCOCO, RefCOCO+, and RefCOCOg, focus ex-

clusively 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.

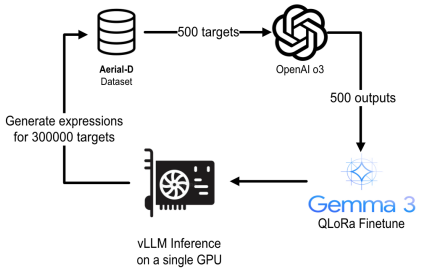


Figure 4: Knowledge distillation pipeline for scalable LLM enhancement. A small sample of 500 expressions is processed through OpenAI's O3 model to generate high-quality training targets, which are then used to fine-tune Gemma3 12B via QLora. The fine-tuned model enables cost-effective local inference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

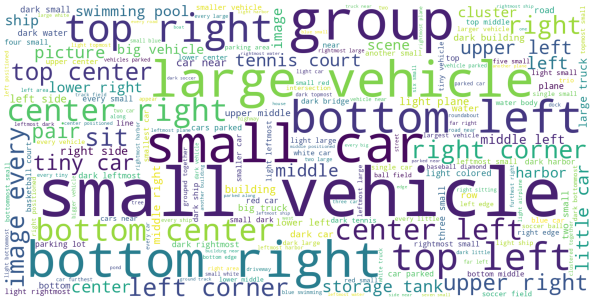


Figure 5: Word cloud visualization of the most frequent terms in Aerial-D referring expressions, highlighting the domain-specific vocabulary and spatial descriptors characteristic of aerial imagery.

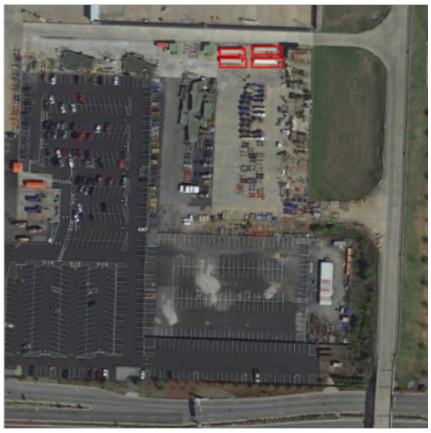
Table 1 quantifies the impact of LLM enhancement, showing that the distillation pipeline successfully tripled the dataset size from the original 506,194 rule-based expressions to 1,522,523 total expressions. The LLM enhancement process contributed nearly equal numbers of language variations (496,895) and unique visual detail expressions (519,434), demonstrating the effectiveness of the two-pronged enhancement strategy in achieving both linguistic diversity and contextual richness.

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).

Table 1: LLM Enhancement Expression Distribution

Expression Source	Train	Val	Total
Rule-Based Expressions	371,360	134,834	506,194
LLM Enhanced (Language Variations)	364,396	132,499	496,895
LLM Unique (Visual Details)	382,038	137,396	519,434
Total Expressions	**1,117,794**	**404,729**	**1,522,523**

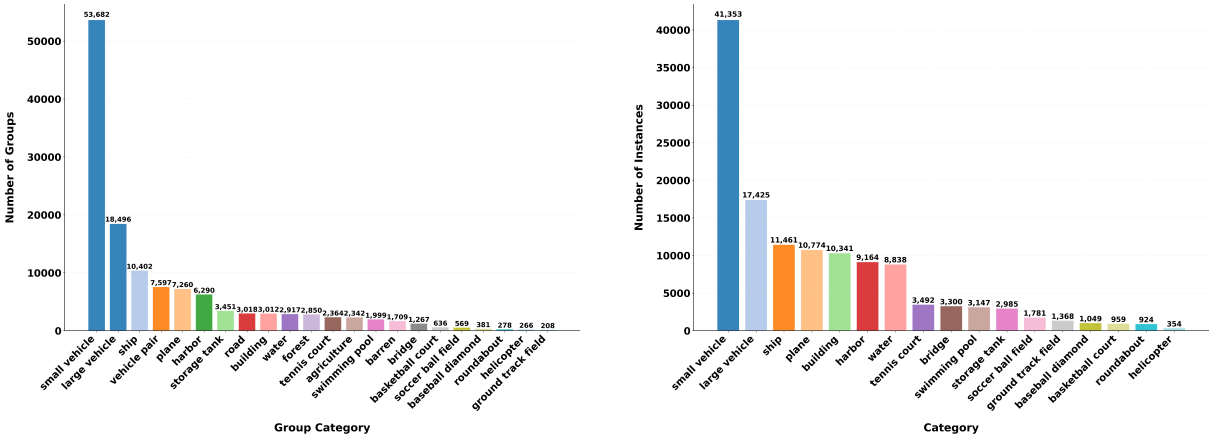


Figure 6: Category distribution analysis of Aerial-D dataset. Left: Distribution of group annotations showing the prevalence of different object categories in group-level referring expressions. Right: Distribution of individual instance annotations across semantic categories, demonstrating the dataset's coverage of aerial object types.

Table 2: Cross-Dataset Performance Evaluation - Model Trained on Aerial-D Only (Historic-filtered results in blue)

Dataset	**IoU@0.5**	**IoU@0.7**	**IoU@0.9**	**mIoU**		**oIoU**	
				Orig.	**Hist.**	**Orig.**	**Hist.**
Aerial-D	57.13%	39.54%	7.56%	49.33%	32.92%	64.30%	45.41%
RRSIS-D	32.87%	23.39%	10.34%	34.07%	32.44%	34.80%	34.33%
NWPU-Refer	25.68%	15.91%	4.02%	24.57%	20.66%	28.27%	20.12%
RefSegRS	15.55%	1.86%	0.00%	18.80%	14.75%	8.58%	4.65%

Table 3: Combined Training Performance Evaluation - Model Trained on All Dataset Train Sets (Historic-filtered results in blue)

Dataset	**IoU@0.5**	**IoU@0.7**	**IoU@0.9**	**mIoU**		**oIoU**	
				Orig.	**Hist.**	**Orig.**	**Hist.**
Aerial-D	–	–	–	–	–	–	–
RRSIS-D	–	–	–	–	–	–	–
NWPU-Refer	–	–	–	–	–	–	–
RefSegRS	–	–	–	–	–	–	–
Urban1960SatSeg	–	–	–	–	N/A	–	N/A

Table 4: Comparison with Existing RRSIS Datasets

Dataset	**Image Resolution**	**Images**	**Annotations**	**Single-object**	**Multi-object**	**Resolution**	**Annotation Generation**
RefSegRS	0.13m	4420	4420	✓	×	512	Manual
RRSIS-D	0.5m-30m	17402	17402	✓	×	800	Semi-auto
NWPU-Refer	0.12m-0.5m	15003	49745	✓	✓	1024-2048	Manual
AERIAL-D	**0.3m-4.5m**	**43,514**	**1,545,994**	✓	✓	**480**	**Automated + LLM**

Table 5: Ablation Study: Cross-Dataset Performance by Training Configuration

Training Configuration	**Samples**	**Aerial-D**			**RefSegRS**			**RRSIS-D**			**NWPU-Refer**		
		Pass@0.7	**mIoU**	**oIoU**	**Pass@0.7**	**mIoU**	**oIoU**	**Pass@0.7**	**mIoU**	**oIoU**	**Pass@0.7**	**mIoU**	**oIoU**
Rule-based Only	371K × 4	–	–	–	2.55%	3.73%	0.55%	29.89%	34.22%	36.46%	13.62%	16.78%	13.70%
Language Variations Only	364K × 4	–	–	–	3.02%	5.75%	4.99%	35.63%	41.63%	42.48%	16.90%	21.89%	16.68%
Unique Expressions Only	382K × 4	35.75%	46.54%	63.02%	2.55%	18.32%	8.37%	20.86%	31.78%	33.73%	15.91%	24.68%	29.22%
Combined All	1,118K × 2	39.54%	49.33%	64.30%	1.86%	18.80%	8.58%	23.39%	34.07%	34.80%	15.91%	24.57%	28.27%