Open-Vocabulary Referring Segmentation of Aerial Photos

Luis Pedro Soares Marnoto Gaspar Lopes

***Abstract*—Referring expression segmentation is a fundamental task in computer vision that integrates natural language understanding with precise visual localization of target regions. Applying this to aerial imagery presents unique challenges because spatial resolution varies widely across datasets, targets often shrink to only a few pixels, and scenes contain very high object densities. This work presents Aerial-D, a large-scale referring expression segmentation dataset for aerial imagery comprising 37,288 image patches with 1,522,523 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 is constructed through a fully automatic pipeline that combines systematic rule-based expression generation with Large Language Model enhancement, enriching both the linguistic variety and visual detail within the referring expressions. As an additional capability, the pipeline produces dedicated historic counterparts for each scene, supporting real-world archival analyses such as monitoring urban change across decades. Models are trained on Aerial-D together with prior aerial datasets, yielding unified instance, semantic, and historic segmentation from text, with the historic branch demonstrating robustness to monochrome, sepia, and grainy degradations that appear in archival aerial photography. The dataset, trained models, and complete pipeline are publicly available at `luispl77.github.io/aerial-d`.**

I. Introduction

Referring expression segmentation [1], [2], [3] is a computer vision task in which a model receives a natural language description of a target region and must return the corresponding segmentation mask. Because the phrasing can reference any concept, the task is open-vocabulary and the target can be a single instance, a coherent group of instances, or an entire semantic category, such as "all roads in the patch" or "the vegetation strip along the river". The remote-sensing literature coined the term Referring Remote Sensing Instance Segmentation (RRSIS) [4] for referring expression segmentation tasks in aerial imagery, with early datasets primarily focusing on single instances. Later datasets like NWPU-Refer [5] expanded group-level coverage, while Aerial-D extends the formulation further to systematically include instances, groups, and full land-cover classes. When this formulation is applied to aerial photographs, although we refer to it simply as referring expression segmentation throughout this article, the problem becomes especially demanding because top-down perspectives compress object scales, spatial resolution varies across sensors, many targets occupy only a handful of pixels, and the scenes themselves contain extreme object densities.

INESC-ID, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal. Email: luis.marnoto.gaspar.lopes@tecnico.ulisboa.pt

In many real deployments, analysts revisit archival aerial surveys to study how cities or coastlines evolved over decades. To support such use cases, this work incorporates a historic imagery component that models the monochrome, sepia, and grainy degradations typical of mid-century aerial photography, enabling models to handle tasks such as assessing long-term urban change from degraded archival imagery.

A critical component for developing effective models for RRSIS and broader referring expression segmentation of aerial photographs is access to high-quality datasets containing aerial imagery, precise segmentation masks, and natural referring expressions. To address this need, this work presents Aerial-D, a large-scale referring expression segmentation dataset for aerial imagery comprising 1,522,523 expressions across 37,288 aerial image patches, significantly larger than prior RRSIS benchmarks [4], [6], [5]. Figure 1 highlights how this corpus spans rural and urban scenes and objects, land-cover regions, groups of multiple objects, and entire categories while retaining unrestricted, richly worded referring expressions tailored to each target.

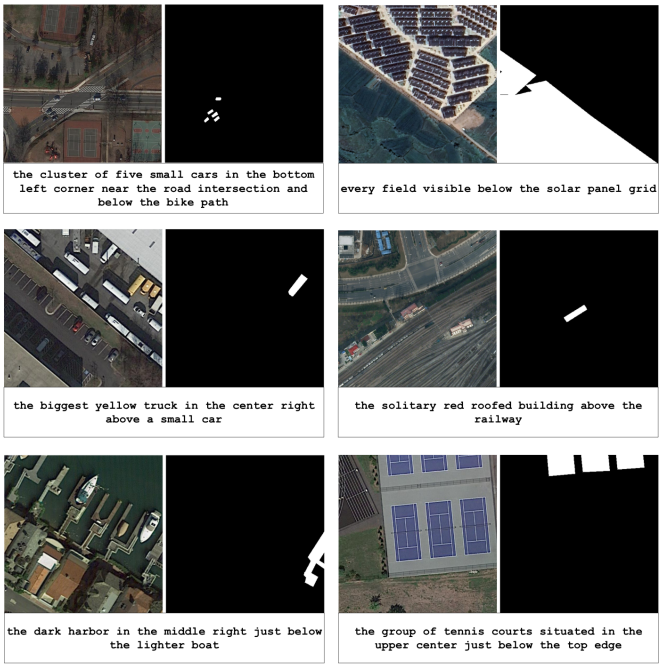


Fig. 1. Representative examples from Aerial-D dataset showing diverse referring expressions with corresponding aerial images and ground truth masks.

The key contributions of this work include: (1) a comprehensive toolchain that enables the production of complex refer-

ring expression datasets from instance/semantic segmentation datasets, including a rule-based pipeline, Large Language Model enhancement and distillation methods, and historic image data augmentation with dedicated filtering; (2) the construction of Aerial-D, a dataset comprising over 1.5 million expressions across 37,288 aerial image patches, created entirely through the proposed automatic pipeline; and (3) a unified model trained on Aerial-D alongside four additional datasets, leveraging historic transformations and other applicable components of the toolchain across the training data to deliver referring expression segmentation over instances, groups, classes, and land-cover regions while maintaining reliable performance on degraded historic imagery typical of archival aerial surveys.

II. Related Work

This section reviews the datasets and architectural developments that underpin aerial image understanding, with emphasis on instance and semantic segmentation resources, referring expression segmentation datasets, historical imagery, and architectures tailored to remote sensing.

A. Aerial Imagery Datasets

Reliable progress in aerial image understanding depends on datasets that capture both discrete objects (e.g., ships, vehicles) and continuous land-cover (e.g., roads, water, vegetation), as well as datasets that test language-based selection of specific targets. This section first summarizes instance and semantic segmentation resources that established pixel-level ground truth, and then discusses the emergence of referring expression segmentation datasets that couple images, masks, and natural language.

1) Instance and Semantic Segmentation: The iSAID dataset [7] established the foundation for instance segmentation in aerial imagery by providing 655,451 object instances across 15 categories in 2,806 high-resolution images. Building upon the DOTA dataset [8], iSAID addressed the unique challenges of aerial imagery including high object density, large scale variations, and arbitrary orientations. The dataset demonstrated that existing computer vision methods require specialized adaptation for aerial domains, as off-the-shelf approaches achieved suboptimal performance.

Complementing instance-level analysis, the LoveDA dataset [9] focused on land-cover semantic segmentation across urban and rural environments. Covering 536.15 km^2 with 0.3m resolution imagery, LoveDA enables domain adaptation research by addressing style differences between geographical environments, with urban scenes dominated by artificial objects and rural scenes containing natural elements.

2) Referring Expression Segmentation Datasets: While instance and semantic segmentation establish pixel-accurate supervision, they assume a fixed label set. Referring expression segmentation reframes the problem: given a natural language phrase, the goal is to select and segment the specific object (or group) described. In aerial imagery, this line of work began with RefSegRS [4], which introduced 4,420 image–language–mask triplets and formalized Referring Remote Sensing Instance Segmentation (RRSIS). The dataset highlighted aerial-specific challenges—small, densely packed targets and cluttered layouts—where language can disambiguate between visually similar instances.

RRSIS-D [6] expanded both scale and annotation efficiency with 17,402 image–caption–mask triplets generated through a semi-automated pipeline using the Segment Anything Model (SAM) [10]. The imagery originates from the DIOR dataset, grounding RRSIS-D in a broad remote-sensing corpus while leveraging automated mask generation. Beyond size, it targets aerial-specific phenomena—broad spatial scales and diverse object orientations—across 20 categories and seven attribute dimensions, enabling richer evaluation of language-guided selection in overhead scenes.

NWPU-Refer [5] further broadens coverage with 15,003 high-resolution images and 49,745 annotated targets spanning more than 30 countries. In contrast to semi-automated pipelines, it emphasizes purely manual annotation quality and explicitly supports single-object, multi-object, and non-object scenarios across 32 categories. Together, these datasets trace a steady shift from fixed-category segmentation toward language-conditioned, fine-grained selection in aerial imagery.

3) Historical Imagery Applications: Analysing historical aerial photographs introduces practical complications—reduced contrast, grayscale capture, film artifacts, and geometric distortions—yet these datasets are vital for studying long-term urban change. Urban1960SatSeg [11] addresses this gap with professionally annotated semantic segmentation over 1,240 km^2 of declassified mid-20th-century imagery from Xi'an, China. By focusing on degraded visual conditions, it provides a reference point for methods that must remain robust when applied to archival aerial data.

B. Architectures for RRSIS

Architectures for referring expression segmentation in aerial imagery combine vision backbones, language encoders, and fusion mechanisms that translate textual cues into pixel-level masks. The Rotated Multi-Scale Interaction Network (RMSIN) [6] is a representative design that builds bespoke processing blocks on top of a Swin Transformer [12] visual encoder and a BERT [13] language backbone. Its Intra-scale Interaction module refines fine-grained details with transformer [14] blocks, the Cross-scale Interaction module aligns multi-resolution features through cross-attention, and the Adaptive Rotated Convolution module injects rotation-aware convolutional filters to handle arbitrary object orientations.

MRSNet [5] adopts the same Swin [12]–BERT [13] backbone pairing but alters how features interact. Instead of RMSIN's triplet of modules, it employs hierarchical fusion that first consolidates information within each scale before exchanging context across scales, leading to a more progressive flow of visual detail toward the mask decoder. The shift in interaction pattern highlights how architectural variants largely differ in their feature fusion strategies rather than wholesale backbone changes.

RSRefSeg [15] illustrates the alternative of leaning fully on large vision–language models. It replaces bespoke encoders with CLIP [16] or SigLIP [17] for multimodal feature extraction and integrates SAM as the segmentation decoder. A lightweight AttnPrompter module converts language-conditioned features into sparse and dense prompts for SAM, while LoRA adapters fine-tune both CLIP/SigLIP and the SAM vision encoder to aerial imagery. Figure 2 visualizes this configuration and highlights how the prompts interface with SAM. This design achieves the strongest reported IoU on RRSIS-D [6], [15] while requiring updates only to the bridging layers and low-rank adapters.

C. Overview

The current landscape of aerial image segmentation is shaped by three pillars that enable language-guided segmentation. First, instance and semantic datasets such as iSAID and LoveDA supply pixel-accurate supervision for objects and land-cover. Second, referring expression segmentation datasets including RefSegRS, RRSIS-D, and NWPU-Refer pair images with natural expressions and masks, enabling evaluation of language-conditioned target selection at varying scales and annotation regimes. Third, architectural developments span specialized remote-sensing networks (e.g., RMSIN, MRSNet) and foundation-model designs (e.g., RSRefSeg [15]) that leverage strong vision and language backbones.

Crucially, the complementary supervision in iSAID and LoveDA presents an opportunity to construct a larger and more diverse referring expression segmentation resource by converting instance- and land-cover annotations into language-conditioned targets—motivating the creation of Aerial-D as a comprehensive dataset for aerial referring expressions.

Within this context, RSRefSeg stands out as a particularly compelling choice for referring expression segmentation: it benefits from powerful pretrained vision–language encoders (e.g., CLIP/SigLIP) and a high-capacity segmentation decoder (SAM) connected by a lightweight bridging mechanism. The resulting system establishes the leading IoU on RRSIS-D [15], demonstrating that foundation-model backbones can be adapted effectively to aerial referring expression segmentation.

III. Aerial-D Dataset Construction

This section details our comprehensive approach to constructing Aerial-D, a large-scale referring expression segmentation dataset for aerial imagery. Our methodology combines automated rule-based expression generation with a multimodal LLM expression component that is preceded by a cost-efficient distillation step, enabling both scale and linguistic diversity. We begin by establishing our source datasets, then describe our rule-based pipeline for generating referring expressions from existing annotations, followed by our distilled large language model enhancement procedure, and conclude with comprehensive dataset statistics demonstrating the scope and characteristics of the final resource.

A. Source Datasets

The Aerial-D dataset is constructed from two complementary aerial datasets with different annotation styles: iSAID (instance segmentation) and LoveDA (semantic segmentation). Together, these source datasets provide 21 distinct object and land-cover classes, with iSAID contributing 15 instance categories (e.g., vehicles, ships, buildings) and LoveDA contributing 6 semantic land-cover classes (e.g., farmland, forest, water).

We start by putting both sources into the same format so that a single model can learn from them side by side: square patches at 480×480. This size keeps small iSAID objects large enough to describe and segment, while fitting the input expectations of common vision encoders used in our model (e.g., CLIP/SigLIP image towers and SAM backbones[16], [17], [10]).

We first resize the 1024×1024 LoveDA tiles directly to 480×480 while preserving their semantic masks. iSAID has uniquely high-resolution imagery with varying aspect ratios, so we instead slide a 480×480 window with overlap across each source image and keep the patches that contain valid instances. After these resizing steps, we run connected-component analysis on LoveDA to turn buildings and water into pseudo-instance targets—these categories tend to appear as isolated structures or bounded water bodies, making them natural candidates for instance-level descriptions. The remaining land-cover classes (e.g., farmland or forest) behave as contiguous surfaces, so we keep them as semantic regions and describe them holistically ("all agricultural land in the image"). The resulting per-patch representation retains instance targets, grouped semantic regions, and the original semantic labels that feed into the subsequent rule-guided expression generation.

B. Rule-Based Expression Generation

The core challenge is figuring out how to describe these target objects using only what we know from their bounding boxes, masks, and categories. We utilize the bounding box coordinates to understand where each object sits within the image patch. As shown in Figure 3, we divide each patch into a three-by-three grid marked with dotted lines, so we can say an object is "in the top right" or "in the center". When we have multiple objects of the same type, we also check if any are in extreme positions like the topmost or leftmost instance of that category.

Since we also have the pixel masks for each object, we can analyze their colors by looking at hue–saturation–value (HSV) distributions to distinguish between light and dark objects and a controlled palette of chromatic colors. We require at least 70% dominance for achromatic labels ("light" or "dark") and a single hue to occupy at least 60% of the chromatic pixels before we commit to a specific color; When neither threshold is met, no color descriptor is assigned and the cue is discarded. This thresholds ambiguous multi-hue regions and helps us ignore noisy signals that would otherwise mislead the language generation. We also avoid using color descriptors for buildings and water since these typically show mixed colors that aren't useful for identification.

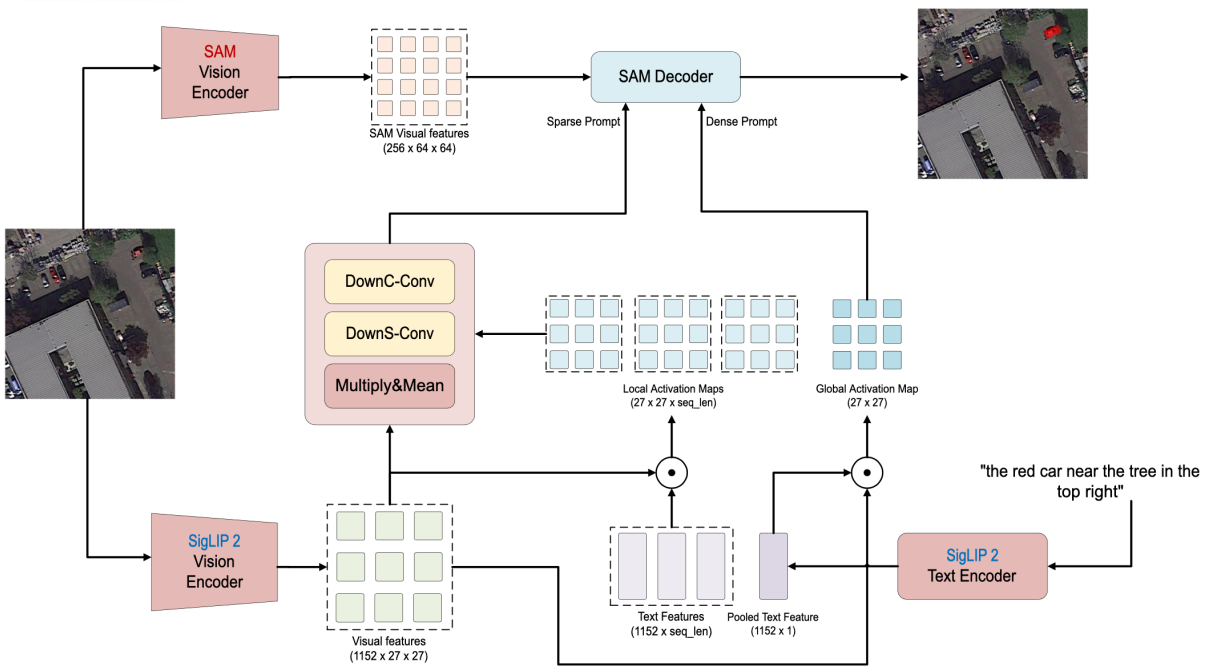


Fig. 2. Overview of the RSRefSeg architecture [15], which couples a vision–language encoder with a segmentation decoder via a learned prompting bridge.

We also create relationships between nearby objects by calculating angles between their positions, allowing us to generate expressions like "the ship to the left of the harbor" or "the vehicle above the building". The system uses eight directional relationships: above, below, to the left of, to the right of, and the four diagonal directions.

All these rules combine to generate various referring expressions for each object, as demonstrated in Figure 3 where a single plane generates multiple possible descriptions including "the plane in the top right", "the light plane in the top right", and versions with relational descriptions. However, a significant challenge emerges when multiple objects end up with identical characteristics and generate the exact same expressions, creating ambiguous references where one phrase could describe multiple different objects. We solve this fundamental problem by taking the set of all expressions for all objects and targets in each image, matching them against each other to find duplicates, and when we find expressions that are identical, we cancel both expressions out and discard them as ambiguous. This filtering stage is crucial: it removes targets that cannot be uniquely verbalized using the rules alone, ensures the rule-based component only passes forward unambiguous instances, and guarantees that the downstream dataset never asks models to resolve intentionally ambiguous descriptions.

C. LLM Expression Generation

While rule-based expression generation provides a solid foundation for referring expression data, these expressions suffer from significant limitations in language variation and visual detail coverage. The rule-based approach produces linguistically constrained expressions with limited wording variations and lacks the ability to reference contextual elements beyond predefined source dataset categories.

To address these limitations, we employ a multimodal Large Language Model (LLM) to enhance our dataset by providing both images and expressions as input, enabling the model to rewrite and improve the original referring expressions. We prompt the LLM with two complementary tasks, as shown in Figure 4. The first task focuses on linguistic variation, creating natural language alternatives for each rule-based expression without heavy reliance on visual cues. The second task uses visual information, where the model examines surrounding features in the image around the target object.

We overlap the target region with red bounding boxes to guide the model during the first task and pair each prompt with a focused close-up crop so that small or dense targets stay visible. For land-cover categories that lack crisp bounding boxes, we supply a dual-image prompt consisting of a masked overlay and the clean image, which helps the model anchor the relevant region. This combination of bounding box overlays, dual images, and mask prompts keeps the enhancement grounded on the correct area of the scene.

This dual-task prompting transforms basic expressions like "the group of 4 large vehicles in the top center" into linguistically diverse alternatives such as "the cluster of four big vehicles near the upper middle" and visually detailed descriptions like "the four large vehicles lined up side by side just below the pale paved strip at the very top middle", as shown in Figure 4. The model identifies and references contextual elements not captured in the original datasets, such as the "pale paved strip" and the "grassy area".

However, the full dataset contains approximately 300,000 captured targets including both objects and groups. To generate expressions, we process each target individually, mean-

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"

Fig. 3. 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.

ing we would need 300,000 separate LLM requests. Using production-grade LLMs at this scale—for example, OpenAI's o3 model[18] with strong visual capabilities—would cost thousands of dollars; Table VI reports the exact breakdown, making direct application prohibitively expensive for research-scale dataset construction.

To address this scalability challenge, we employ a knowledge distillation [19] approach, as illustrated in Figure 8. We utilize OpenAI's o3 model[18] and compare it against a much more lightweight open-weights model, Gemma3[20]. We obtain 500 high-quality outputs from o3 on a representative random subset of targets from the initial dataset. These outputs serve as training data for supervised fine-tuning using the parameter-efficient QLoRA method[21] on Gemma3-12B.

During fine-tuning we apply LoRA adapters across both the text decoder and the SigLIP-derived vision stack embedded in Gemma3, which improves instruction adherence, suppresses hallucinations, and stabilizes the two-task output schema. The custom-tailored Gemma3 variant can then process all 300,000 targets on a single GPU while honoring the dual-task prompt structure—behavior the base Gemma3 model fails to follow reliably without distillation. Notably, the distilled model's output quality approaches o3's once fine-tuned; qualitative comparisons in Figure 9 show closely matched enhancements with markedly reduced hallucinations relative to the base Gemma3 model.

D. Historic Image Filter Augmentation

In order to improve robustness to archival image conditions, we augment training with three parametric transformations that reproduce characteristic degradations of historical aerial photographs (Figure 5). These filters are applied on the fly during training rather than baked into the dataset, so each mini-batch can include either a clean or a historically degraded view of the same patch.

Let $I_{orig}(x) \in [0, 255]^3$ denote the RGB image at pixel x, and let $\mathrm{clip}(\cdot)$ clamp values to $[0, 255]$.

We simulate grayscale capture by converting to luminance, as in Eq.(1):

$$I_{\text{bw}}(x) = 0.299\, R(x) + 0.587\, G(x) + 0.114\, B(x). \quad (1)$$

To emulate film response and grain, we first apply a mild gamma adjustment (Eq.(2)), then a linear contrast change around the mean (Eq.(3)), followed by additive Gaussian noise (Eq.(4)):

$$I_\gamma(x) = 255\,\big(I_{\text{bw}}(x)/255\big)^\gamma. \quad (2)$$

$$I_c(x) = \big(I_\gamma(x) - \mu\big)\, c + \mu. \quad (3)$$

$$I_{\text{grain}}(x) = \mathrm{clip}\big(I_c(x) + \eta(x)\big), \quad \eta(x) \sim \mathcal{N}(0, \sigma^2). \quad (4)$$

We use $\gamma = 1.2$, $c = 0.8$, and $\sigma = 0.1 \times 255$ to produce mild contrast loss and film grain.

Finally, we apply a fixed sepia transform (Eq.(5)) followed by uniform sensor/scan noise (Eq.(6)):

$$\begin{bmatrix} S_R(x) \\ S_G(x) \\ S_B(x) \end{bmatrix} = \mathrm{clip}\left(\begin{bmatrix} 0.272 & 0.534 & 0.131 \\ 0.349 & 0.686 & 0.168 \\ 0.393 & 0.769 & 0.189 \end{bmatrix} \begin{bmatrix} R(x) \\ G(x) \\ B(x) \end{bmatrix} \right). \quad (5)$$

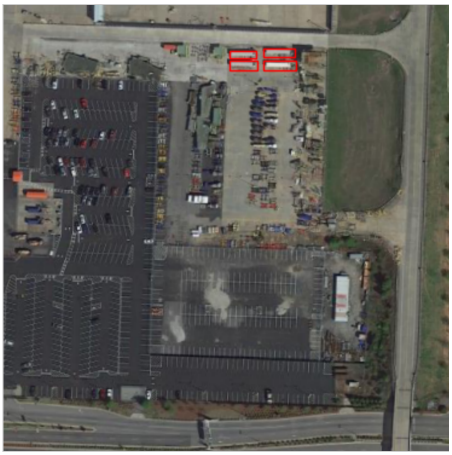
$$I_{\text{sepia}}(x) = \mathrm{clip}\big(\mathbf{S}(x) + \xi(x)\big), \quad \xi(x) \sim \mathcal{U}(0, 50). \quad (6)$$

These effects mimic tonal range reduction, lens grain, and scanning artifacts typical of mid-century aerial photography while preserving the spatial structure that segmentation relies on. Figure 5 illustrates the visual impact of each transformation.

E. Final Dataset Statistics

The rule-based generation yields 506,194 starting expressions and identifies 259,709 annotated targets across the corpus (Table II). Building on this base, the LLM enhancement is prompted to produce one language variation for each original expression and two unique visual-detail expressions for each target, adding 496,895 and 519,434 expressions respectively and resulting in 1,522,523 total expressions. Of these, 1,278,453 expressions describe discrete instances or groups, while 244,070 cover the land-cover regions that remain at the semantic level. Figure 6 illustrates how this process expands the vocabulary.

Table I compares Aerial-D with prior RRSIS datasets and shows how it scales along three axes: images, targets per image, and expressions per target. First, Aerial-D contains



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

Fig. 4. Example of LLM enhancement process showing original aerial image with group of four large vehicles (left) and corresponding expression enhancements (right).

Fig. 5. Comparison of original aerial image patch with three historic filter transformations: grayscale conversion, sepia toning, and Gaussian noise addition. These filters simulate common degradation patterns in historical aerial photography.

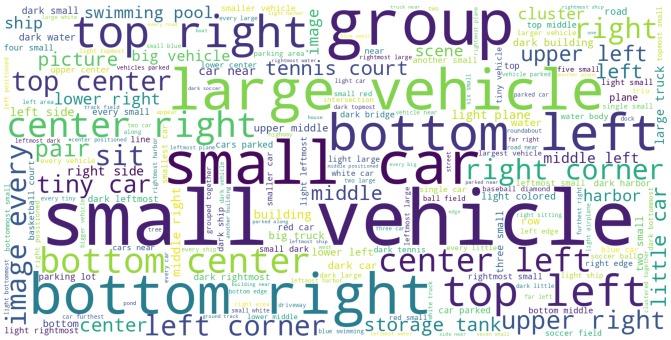


Fig. 6. Word cloud visualization of the most frequent terms in Aerial-D referring expressions, highlighting the domain-specific vocabulary and spatial descriptors characteristic of aerial imagery.

nearly three times as many images as previous datasets. Second, each image typically includes many segmented targets. Third, each target is paired with multiple referring expressions. The table also clarifies how the corpus splits between the 1.28 million instance-level expressions and the 244 thousand semantic expressions that describe the land-cover categories, a combination that is absent from earlier datasets. Together, these factors yield more than 1.5 million referring expressions, positioning Aerial-D among the largest publicly available RRSIS resources. Beyond scale, Aerial-D relies on a fully automatic pipeline that combines rule-based generation with LLM enhancement, supports both single-object and multi-object references, and preserves the original category balance as illustrated in Figure 7.

IV. Experiments

This section presents a comprehensive experimental evaluation of Aerial-D that spans model training, cross-dataset generalization, and targeted ablations. We begin by outlining the RSRefSeg backbone and training configuration, then report cross-dataset results on established aerial referring expression segmentation datasets. Beyond aggregate performance, we also include: (i) ablation of expression enhancement strategies; (ii) ablation of historic-filter training; and (iii) qualitative comparison across language models (o3, base Gemma3, and our distilled Gemma3-Aerial model), coupled with a cost analysis of these alternatives.

A. Model Architecture

Evaluating Aerial-D demands a model that preserves the link between natural-language instructions and precise masks while handling densely packed aerial targets. RSRefSeg meets these requirements and already demonstrated state-of-the-art results on RRSIS-D[6]. We reimplemented the architecture in PyTorch [23] and verified those RRSIS-D gains before extending the system to Aerial-D, which confirmed that the design transfers reliably to new datasets.

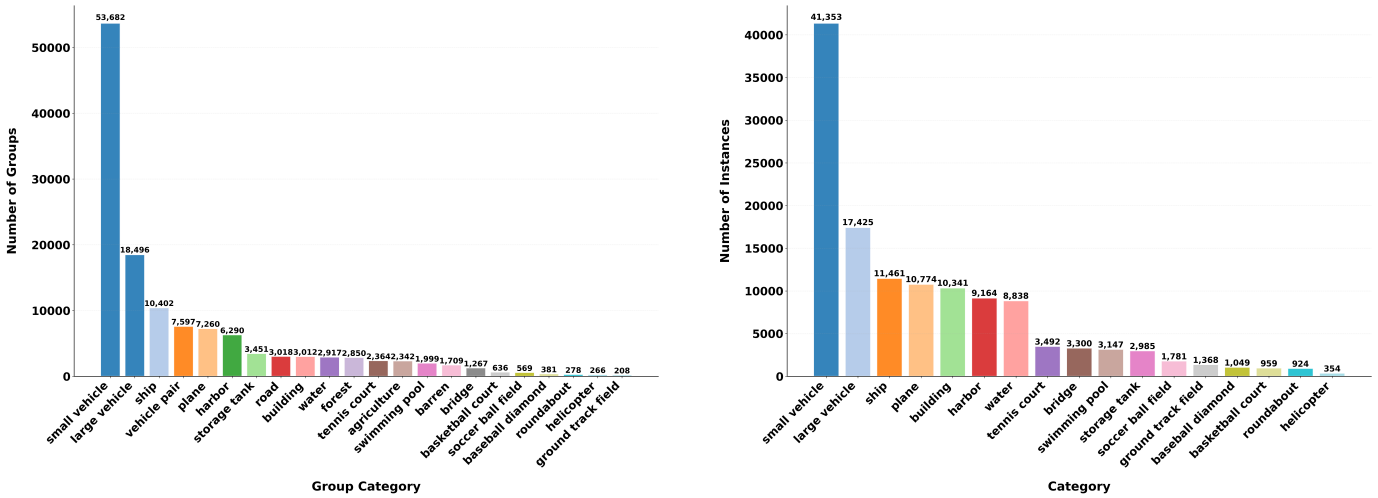


Fig. 7. Category distribution analysis of the Aerial-D dataset. Left: Distribution of grouped and semantic targets showing the prevalence of region-level categories in the corpus. Right: Distribution of individual instance annotations across semantic categories, demonstrating the dataset's coverage of aerial object types.

TABLE I
COMPARISON WITH EXISTING RRSIS DATASETS

Dataset	Image Resolution	Images	Instance Expr.	Semantic Expr.	Single-object	Multi-object	Patch Size	Annotation Generation
RefSegRS	0.13m	4,420	4,420	–	✓	×	512	Manual
RRSIS-D	0.5m–30m	17,402	17,402	–	✓	×	800	Semi-auto
NWPU-Refer	0.12m–0.5m	15,003	49,745	–	✓	✓	1,024–2,048	Manual
Aerial-D	**0.3m–4.5m**	**37,288**	**1,278,453**	–	✓	✓	**480**	**Automated + LLM**
			–	**244,070**	×	✓		

TABLE II
EXPRESSION DISTRIBUTION BY SOURCE

Source	Train	Validation	Total
Rule-Based	371K	135K	506K
LLM Language Variations	364K	133K	497K
LLM Visual Variations	382K	137K	519K
Total	**1,118K**	**405K**	**1,523K**

Our reimplementation of RSRefSeg[15] mirrors the original component pairing: SigLIP2[24] supplies the image–text encoder and SAM[10] provides the mask decoder. We fine-tune both modules with Low-Rank Adaptation (LoRA)[25] layers placed on the query and value projections of each Vision Transformer[26] encoder block and on the query, key, value, and output projections of the text encoder. Two checkpoints appear throughout the experiments. RSRefSeg-b keeps SAM-ViT-Base and LoRA rank $r = 16$ for a lighter configuration, whereas RSRefSeg-l upgrades to SAM-ViT-Large with rank $r = 32$ to maximize capacity while preserving the same training recipe.

B. Experimental Setup

Training adheres to the RSRefSeg recipe so that differences arise from the data rather than from custom optimization. Batches contain four samples with gradient accumulation of two steps, yielding an effective batch size of eight. All experiments run on a single NVIDIA RTX A6000 GPU. Both RSRefSeg-b and RSRefSeg-l follow the same training recipe, with RSRefSeg-b using SAM-ViT-Base and LoRA rank $r = 16$ for a lighter configuration, while RSRefSeg-l uses SAM-ViT-Large with rank $r = 32$ to maximize capacity. The combined model uses `SigLIP2-SO400M` for language–vision encoding at 384×384 resolution and `SAM-ViT-Large` at 1024×1024 for mask decoding, with LoRA rank $r = 32$ steering the adaptation. We employ AdamW[27] with learning rate

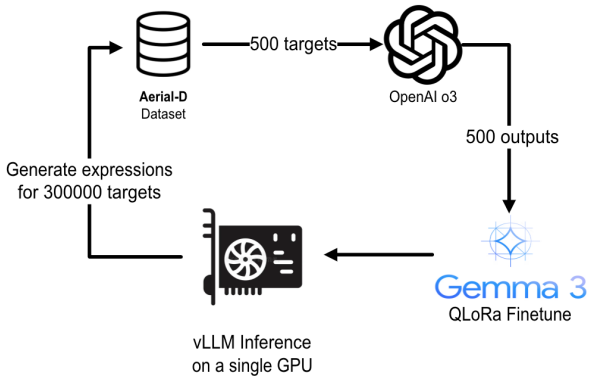


Fig. 8. Knowledge distillation pipeline for scalable LLM enhancement. A small sample of 500 expressions is processed through OpenAI's o3 model[18] to generate high-quality training targets, which are then used to fine-tune Gemma3-12B[20] via QLoRA[21]. The fine-tuned model enables cost-effective local inference to enhance the full dataset using vLLM[22] on a single GPU.

1×10^{-4}, weight decay 0.01, polynomial decay (power 0.9), mixed precision[28], and gradient clipping at 1.0 to mirror the original training dynamics. During the data preparation phase of training, we apply one of the three historic filters (selected with equal probability) to 20% of training images in each non-historic dataset. Note that Urban1960SatSeg is inherently historic imagery, so it receives no additional filtering and provides direct supervision for archival conditions.

In order to keep the cross-dataset mix balanced, Aerial-D contributes only the *LLM Visual Variations* subset highlighted in Section IV-D. That ablation shows this subset carries the strongest signal while still covering every target; limiting Aerial-D to these expressions keeps the millions of available sentences from overwhelming the four public datasets, each of which supplies only tens of thousands of annotations. The combined run therefore spans Aerial-D (LLM Visual Variations), RRSIS-D[6], NWPU-Refer[5], RefSegRS[4], and Urban1960SatSeg[11]. Validation follows the official splits, and Aerial-D relies on the full validation split (405K expressions), which is further broken down into instance targets and semantic regions in Table III. To probe robustness we also evaluate fully filtered validation sets for Aerial-D, RRSIS-D, NWPU-Refer, and RefSegRS by converting every image with one of the three historic filters. Scores for those variants appear in the "Hist." columns.

C. Evaluation Results

Tables III and IV report validation results for the combined model trained jointly on all datasets. Both tables focus on mean IoU [29] and overall IoU for the original validation splits alongside their historic-filtered counterparts, highlighting aggregate overlap rather than thresholded pass rates.

Table III isolates Aerial-D and reports validation performance by target type. Instance targets correspond to explicit objects or groups, while the "All Targets" column aggregates both instance- and semantic-level references across the full 405K validation expressions. These values serve as baseline checkpoints for future work that evaluates new architectures on Aerial-D.

Across the external datasets in Table IV, the RSRefSeg-b and RSRefSeg-l checkpoints remain competitive with previously published results despite being trained within a single unified pipeline. That schedule blends five datasets, applies historic filters to more than 20% of the training imagery, and supervises both instance- and semantic-level targets drawn from Aerial-D. Historic columns illustrate that the LoRA-tuned RSRefSeg-l retains accuracy under simulated degradations, closing the gap between clean and historic imagery without requiring dataset-specific adjustments. The same table also foregrounds how the larger checkpoint strengthens overall IoU on RefSegRS and NWPU-Refer while maintaining the strong RRSIS-D performance reported in prior work.

The performance gap on RefSegRS deserves particular attention: while RMSIN achieves 59.96% mIoU, RSRefSeg-l reaches 44.52%. This gap reflects distribution shift, as RefSegRS contains referring patterns (e.g., vehicles along specific roads) that differ significantly from Aerial-D, RRSIS-D, and NWPU-Refer. Multi-dataset training biases the model toward the majority distribution, causing lower performance on out-of-distribution RefSegRS samples.

The combined evaluation therefore remains competitive with published references across every dataset while providing reproducible Aerial-D baselines. The modest gap between the original and historic evaluations confirms that injecting filtered imagery during training delivers robustness without eroding accuracy on contemporary photographs.

D. LLM Expression Generation Ablation

To measure how synthetic language affects segmentation, we retrain RSRefSeg on Aerial-D while isolating the different expression sources. The goal is to contrast the original rule-generated sentences against those produced by the LLM pipeline and determine which variants provide the strongest supervision. We adopt the lighter RSRefSeg-b configuration—SAM-ViT-Base with LoRA rank $r = 16$—so each run completes quickly while preserving the optimization settings from Section IV-B.

Using this setup, we train four separate models, each exposed to a distinct slice of Aerial-D: (i) *Rule-based Only* retains the deterministic descriptions produced by the rule system; (ii) *LLM Language Variations* relies on rewrites that diversify the wording while preserving the target; (iii) *LLM Visual Variations* selects LLM augmentations that inject alternative visual cues; and (iv) *Combined All* unites the three sources. Figure 4 illustrates how the enhanced variants expand the phrasing beyond the rule-based baseline. We evaluate the resulting checkpoints on four validation sets without additional tuning: Aerial-D (using the 405K-expression validation split) and three external datasets—RefSegRS, RRSIS-D, and NWPU-Refer—to observe how each expression type supports generalization beyond the training distribution.

Table V summarizes the results and explicitly reports both the number of *Samples* and *Epochs* used per configuration. Looking across datasets, several consistent patterns emerge. On Aerial-D, the *Combined All* configuration achieves the best accuracy, which is expected given that the validation distribution matches that training mixture. For the three external datasets, different subsets yield the strongest generalization: on RRSIS-D, the *LLM Language Variations* run delivers the highest scores; on NWPU-Refer, emphasizing varied visual cues through *LLM Visual Variations* is most beneficial; and on RefSegRS, uniting the three sources provides the best results. These outcomes highlight the breadth of ways to phrase referring expressions and show that leveraging LLMs to introduce targeted variety improves cross-dataset generalization.

In order to keep early stopping responsive to each subset, we monitor validation loss across the four runs and halt training as soon as it begins to rebound. Because the *Combined All* subset is roughly three times larger than the others, its validation loss ticks upward immediately after the second epoch, so we stop that run at two epochs. The three smaller subsets continue improving through the fourth epoch before showing the same rise, allowing four full passes over their data. This schedule means that the language-variation and

TABLE III
AERIAL-D SUPERVISION VARIANTS EVALUATED ON THE VALIDATION SPLIT (HISTORIC SCORES IN BLUE).

Model	Instance Targets		Semantic Targets		All Targets	
	mIoU	oIoU	mIoU	oIoU	mIoU	oIoU
RSRefSeg-b (ours)	49.49% / 40.98%	62.25% / 56.61%	54.26% / 47.00%	64.01% / 57.34%	47.10% / 39.77%	62.40% / 56.60%
RSRefSeg-l (ours)	51.82% / 42.70%	62.32% / 56.26%	55.48% / 49.01%	64.72% / 58.89%	49.81% / 42.18%	63.95% / 58.41%

TABLE IV
CROSS-DATASET VALIDATION RESULTS FOR RSREFSEG VARIANTS (OURS) AND PUBLISHED BASELINES (HISTORIC SCORES IN BLUE; "–" INDICATES METRICS NOT REPORTED IN THE CITED WORK).

Model	RefSegRS		RRSIS-D		NWPU-Refer		Urban1960SatSeg	
	mIoU	oIoU	mIoU	oIoU	mIoU	oIoU	mIoU	oIoU
RSRefSeg-b (ours)	24.81% / 17.54%	40.89% / 29.41%	64.37% / 61.16%	76.83% / 75.44%	39.42% / 33.15%	59.52% / **56.56%**	**70.65%**	**88.86%**
RSRefSeg-l (ours)	44.52% / **36.03%**	55.74% / **45.74%**	**65.37%** / **62.61%**	76.33% / **76.03%**	**45.75%** / **39.11%**	62.75% / 55.29%	69.74%	88.73%
RMSIN[6], [5], [15]	**59.96%**	**76.81%**	62.27%	76.50%	41.75%	62.66%	–	–
RSRefSeg-b[15]	–	–	63.68%	76.05%	–	–	–	–
RSRefSeg-l[15]	–	–	64.67%	**77.24%**	–	–	–	–
MRSNet[5]	–	–	–	–	44.86%	**63.59%**	–	–
Urban1960SatUSM[11]	–	–	–	–	–	–	68.80%	–

visual-variation models actually process fewer total samples than the combined run, yet they surpass it on RRSIS-D and NWPU-Refer. The outcome reveals that curated expression subsets can be more sample efficient than the full mixture, which is why the combined model in Section IV-C draws on the LLM Visual Variations split of Aerial-D. Beyond sample efficiency, constraining Aerial-D to that subset keeps the cross-dataset mix from being dominated by a single corpus whose expression pool would otherwise reach into the millions.

E. Distillation Ablation: Gemma3 vs. o3 Model Comparison

This ablation measures how the generator choice inside the LLM enhancement stage affects expression quality and overall cost. We evaluate three options for producing the enhanced expressions: OpenAI's o3, the off-the-shelf Gemma3-12B, and the distilled Gemma3-Aerial model described in Section III-C (see Figure 8).

All three models are prompted and decoded in the same way so that differences stem from the generator rather than the interface. The base Gemma3 baseline frequently ignores the dual-task schema, hallucinates objects that are not present, and references the guiding bounding boxes inside the expression, which undermines segmentation training. Distillation sharply reduces those errors, producing grounded descriptions that resemble the o3 outputs while remaining accessible on local hardware.

Table VI quantifies the practical impact. Running o3 across ∼300,000 targets would cost roughly $6.2K, whereas the distilled Gemma3 produces comparable guidance for about $26—roughly 238× cheaper because inference runs locally instead of via a commercial API. Figure 9 illustrates why the student is worth training: the base Gemma3 hallucinates a second baseball diamond that does not exist, while the distilled variant stays aligned with the image and mirrors the grounded detail that o3 provides.

F. Historic Filter Ablation Study

In order to understand how much the historic-image filters described in Section III-D contribute to robustness, we repeat the combined training without injecting those filters. Models that only encounter clean, contemporary imagery typically falter when historic photographs suddenly introduce monochrome toning, contrast loss, or sepia casts. We therefore run an ablation that removes the filters from the training mix and compares the resulting model against the full recipe, using the RSRefSeg-b variant because its SAM-ViT-Base backbone trains faster and lets us report results alongside the base model in Table IV.

Table VII summarises both setups. Each row lists the clean mIoU (Orig.), the historic counterpart (Hist., in blue), and the drop relative to the full-training baseline. Removing filters keeps the Orig. values aligned with Table IV yet costs five to nine points on the historic splits. Excluding Urban1960SatSeg is far more damaging: the model recovers less on the filtered datasets, collapses to roughly 18% mIoU on Urban1960SatSeg, and loses the only source of direct historic supervision. Synthetic filters therefore help, but real historic imagery remains necessary because Urban1960SatSeg introduces different land-cover labels—such as treating buildings as area classes—and includes considerably more noise than the other datasets.

Values in parentheses denote percentage-point change relative to the baseline combined model in Table IV; blue marks historic-filtered validation scores.

V. CONCLUSION AND FUTURE WORK

This work introduces Aerial-D together with an end-to-end methodology that converts existing aerial segmentation datasets into a large-scale repository of referring expressions. The pipeline begins with a rule-driven generator that translates masks into prose grounded on location, appearance, and relational cues, then refines that language through LLM

TABLE V
EXPRESSION ENHANCEMENT ABLATION ACROSS FOUR DATASETS

Training Configuration	Samples	Epochs	Aerial-D		RefSegRS		RRSIS-D		NWPU-Refer	
			mIoU	oIoU	mIoU	oIoU	mIoU	oIoU	mIoU	oIoU
Rule-based Only	371K	4	34.57%	39.31%	3.73%	0.55%	34.22%	36.46%	16.78%	13.70%
LLM Language Variations	364K	4	46.45%	56.99%	5.75%	4.99%	**41.63%**	**42.48%**	21.89%	16.68%
LLM Visual Variations	382K	4	46.54%	63.02%	18.32%	8.37%	31.78%	33.73%	**24.68%**	**29.22%**
Combined All	1,118K	2	**49.33%**	**64.30%**	**18.80%**	**8.58%**	34.07%	34.80%	24.57%	28.27%

Expression Type	Example
Original	the orange baseball diamond in the top left
o3	the orange baseball diamond with the light pole near home plate in the upper left
Gemma3 Base	the bright orange baseball diamond to the left of another similar baseball diamond in the top left
Gemma3-Aerial-12B	the orange baseball field with a chainlink fence surrounded by grass to the north and trees to the west

Fig. 9. Qualitative comparison between o3, the base Gemma3 model, and our fine-tuned Gemma3-Aerial-12B on the same aerial scene. Each model receives the identical rule-based prompt and decoding setup, revealing how their rewritten expressions differ under matched conditions.

TABLE VI
COST ANALYSIS: GEMMA3 VS. O3 MODEL FOR LARGE-SCALE ANNOTATION (COST CALCULATIONS BASED ON API PRICING WITH AVERAGE REQUEST TOKEN COUNTS)

Model	Cost per request	Cost for 300K requests
o3 Model	$0.020728	$6,218.32
Distilled Gemma3	$0.000087	$26.01
Savings	238× **cheaper**	**$6,192.31 (99.6%)**

rewriting while keeping costs manageable via a distilled Gemma3 annotator. The resulting corpus enables RSRefSeg to be trained jointly across five datasets, establishes reproducible baselines on Aerial-D, and remains competitive with published results on RRSIS-D, NWPU-Refer, RefSegRS, and Urban1960SatSeg. By pairing these evaluations with ablations on expression sources and historic-image filters, we demonstrate that Aerial-D delivers a harder dataset for referring segmentation in aerial photos and highlights the specific ingredients that improve robustness.

Future work can extend this foundation in three directions. First, the expression-enhancement pipeline can be applied directly to the native captions supplied with public datasets such as RRSIS-D and NWPU-Refer, enriching their language with the visual grounding cues that proved effective for Aerial-D and creating a unified, higher-variety training pool. Second, multilingual variants of these expressions can be produced while preserving full automation by pairing high-quality translation models with our distillation recipe—either prompting o3 and training a Gemma3 student to mimic those translations or seeding the process with dedicated systems such as Tower [30]. Third, emerging multimodal systems like Gemini 2.5[31] already output full segmentation masks and bounding boxes; integrating them could expand Aerial-D with additional targets derived from the same imagery, which can then be described with our expression-generation stages to unlock richer open-vocabulary supervision.

REFERENCES

[1] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, pp. 787–798.

[2] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *European Conference on Computer Vision (ECCV)*. Springer International Publishing, 2016, pp. 108–124.

[3] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *European Conference on Computer Vision (ECCV)*. Springer International Publishing, 2016, pp. 69–85.

[4] Z. Yuan, L. Mou, Y. Hua, and X. X. Zhu, "Rrsis: Referring remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[5] Z. Yang, H. Yao, L. Tian, X. Zhao, Q. Li, and Q. Wang, "A large-scale referring remote sensing image segmentation dataset and benchmark," 2025. [Online]. Available: https://arxiv.org/abs/2506.03583

[6] S. Liu, Y. Ma, X. Zhang, H. Wang, J. Ji, X. Sun, and R. Ji, "Rotated multi-scale interaction network for referring remote sensing image segmentation," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 26 648–26 658.

[7] S. W. Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. S. Khan, F. Zhu, L. Shao, G.-S. Xia, and X. Bai, "isaid: A large-scale dataset for instance segmentation in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, cVPR'19 Workshops; arXiv:1905.12886. [Online]. Available: https://arxiv.org/abs/1905.12886

TABLE VII
HISTORIC-FILTER ABLATIONS FOR THE RSREFSEG-B BASE ARCHITECTURE. EACH ROW LISTS THE CLEAN SCORE FOLLOWED BY THE HISTORIC-FILTER VARIANT (BLUE) WITH PERCENTAGE-POINT DELTAS RELATIVE TO TABLE IV; THE SECOND ROW ALSO REMOVES URBAN1960SATSEG SUPERVISION.

Training Setup	RRSIS-D (mIoU)		NWPU-Refer (mIoU)		RefSegRS (mIoU)		Urban1960SatSeg (mIoU)
	Orig.	Hist.	Orig.	Hist.	Orig.	Hist.	Orig.
No Filters	64.29%	56.88% (-4.28)	41.41%	32.40% (-0.75)	44.25%	25.47% (+7.93)	68.88% (-1.77)
No Urban1960SatSeg	64.41%	61.16% (±0.00)	40.70%	34.89% (+1.74)	44.05%	34.51% (+16.97)	4.87% (-65.78)
No Filters + No Urban1960SatSeg	-	-	-	-	-	-	-

[8] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

[9] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *NeurIPS 2021 Datasets and Benchmarks Track*, 2021, arXiv:2110.08733. [Online]. Available: https://arxiv.org/abs/2110.08733

[10] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023. [Online]. Available: https://arxiv.org/abs/2304.02643

[11] T. Hao, L. Zhang, Y. Zhang, M. Chen, J. Zhang, and H. Fu, "Urban1960satseg: Unsupervised semantic segmentation of mid-20th century urban landscapes with satellite imageries," 2025. [Online]. Available: https://arxiv.org/abs/2506.09476

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021, pp. 10 012–10 022.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[15] K. Chen, J. Zhang, C. Liu, Z. Zou, and Z. Shi, "Rsrefseg: Referring remote sensing image segmentation with foundation models," 2025. [Online]. Available: https://arxiv.org/abs/2501.06809

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[17] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023.

[18] OpenAI, "Introducing openai o3," OpenAI Blog, 2025, accessed 2025-01-15. [Online]. Available: https://openai.com/index/introducing-openai-o3/

[19] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, nIPS 2014 Deep Learning Workshop. [Online]. Available: https://arxiv.org/abs/1503.02531

[20] Google DeepMind, "Gemma models overview," Google AI for Developers, 2025, accessed 2025-01-15. [Online]. Available: https://ai.google.dev/gemma/docs

[21] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," 2023. [Online]. Available: https://arxiv.org/abs/2305.14314

[22] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*. ACM, 2023, pp. 611–626.

[23] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[24] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," 2025. [Online]. Available: https://arxiv.org/abs/2502.14786

[25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019, published at ICLR 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

[28] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed precision training," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://openreview.net/forum?id=r1gs9JgRZ

[29] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[30] D. Alves, N. M. Guerreiro, J. Alves, J. Pombal, R. Rei, J. G. C. Fernandes, A. Farinhas, L. Coheur, and A. F. T. Martins, "Tower: An open multilingual large language model for translation-related tasks," 2024. [Online]. Available: https://arxiv.org/abs/2402.17733

[31] Google DeepMind, "Gemini 2.5: Multimodal model overview," Product and research overview, 2025, accessed 2025-01-15. [Online]. Available: https://deepmind.google/technologies/gemini/