RIS-LAD: A Benchmark and Model for Referring Image Segmentation in Low-Altitude Drone Imagery

Kai Ye[1], YingShi Luan[1], Zhudi Chen[1], Guangyue Meng[1], Pingyang Dai[1], Liujuan Cao[1]*

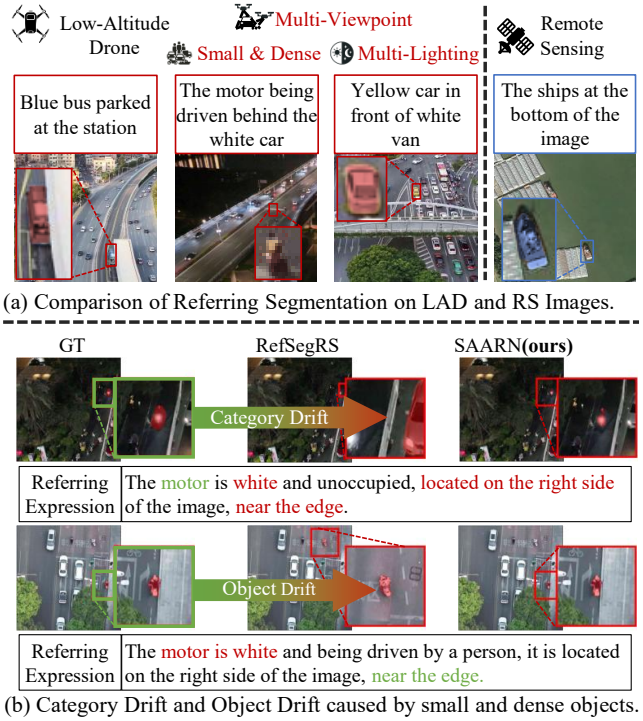
[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China.

Abstract

Referring Image Segmentation (RIS), which aims to segment specific objects based on natural language descriptions, plays an essential role in vision-language understanding. Despite its progress in remote sensing applications, RIS under Low-Altitude Drone (LAD) scenarios remains underexplored, as existing datasets and methods are typically designed for high-altitude and static-view imagery. They struggled to handle the unique characteristics of LAD views, such as diverse viewpoints and high object density. In this paper, we propose RIS-LAD, the first fine-grained RIS benchmark tailored for LAD scenarios, featuring 13,871 meticulously annotated image-text-mask triplets collected from real-world drone footage with emphasis on small, densely cluttered objects and multi-view perspectives. Additionally, we propose the Semantic-Aware Adaptive Reasoning Network, which decomposes and adaptively routes semantic information to different network stages rather than uniformly injecting all linguistic features. Specifically, the Category-Dominated Linguistic Enhancement aligns visual features with object categories during early encoding, while the Adaptive Reasoning Fusion Module dynamically selects semantic cues across scales to enhance reasoning in complex scenes. Extensive experiments reveal that RIS-LAD presents substantial challenges to state-of-the-art RIS algorithms, and also demonstrate the effectiveness of our proposed model in addressing these challenges.

Introduction

Low-Altitude Drones (LAD), which typically operate below 200 meters, have become widely utilized in real-world applications due to their flexible deployment and high versatility (Banafaa et al. 2024; Casanova et al. 2025; Li et al. 2025). This trend has sparked increasing research interest in vision tasks under LAD scenarios, leading to the development of benchmarks such as VisDrone (Zhu et al. 2021) and UAV123 (Mueller, Smith, and Ghanem 2016), which support object detection, tracking and other related tasks (Wen et al. 2021; Barekatain et al. 2017; Du et al. 2018). Recently, visual understanding in LAD scenarios (Sun et al. 2025; Li and Zhao 2025; Lin et al. 2025) has attracted growing interest, especially in multi-modal tasks. Among them, Referring Image Segmentation (RIS) (Liu and Li 2025; Wang et al. 2025; Huang et al. 2025) is a fundamental perception task that aims to segment objects from images based on descriptions. Incorporating RIS into LAD systems enables LADs to better accommodate a wider range of practical applications.



(a) Comparison of Referring Segmentation on LAD and RS Images.

(b) Category Drift and Object Drift caused by small and dense objects.

Figure 1: Illustration of the challenges in referring low-altitude drone image segmentation (RLADIS). The method shown in (b) is RefSegRS (Chen et al. 2025), one of the current SOTA approaches for RRSIS.

However, existing RIS research mainly focuses on conventional scenes or high-altitude remote sensing imagery, leaving the distinctive properties of LAD imagery largely unexplored in the current literature. Although previous Referring Remote Sensing Image Segmentation (RRSIS) studies (Shi and Zhang 2025; Chen et al. 2025; Pan et al. 2024) have provided valuable insights, they are mostly based on imagery taken from satellites or helicopters. As shown in

*Corresponding author.

Fig. 1(a), Referring Low-Altitude Drone Image Segmentation (RLADIS) presents several unique challenges compared to RRSIS: (1) diverse viewpoints due to lower and more flexible camera positions; (2) diverse illumination conditions; and (3) smaller and more densely distributed objects. These challenges introduce a substantial domain gap that hinders the effective generalization of current RRSIS methods to RLADIS tasks.

To bridge this gap, we propose RIS-LAD, the first dataset designed for fine-grained RLADIS, accompanied by a semi-automatic annotation pipeline that ensures high-quality referring expression generation and annotation efficiency. The dataset contains 13,871 image-text-mask triples, providing a benchmark for evaluating RLADIS methods. Furthermore, our analysis reveals two key challenges that hinder the direct adaptation of RRSIS methods to RLADIS: **category drift** and **object drift**. As shown in Fig. 1(b), category drift arises when the object occupies only a small region of the image, misleading the model to focus on larger, semantically similar objects. Object drift, on the other hand, results from a high density of instances belonging to the same category, making it difficult for the model to correctly identify the specific object referenced in the expression.

Motivated by these drift issues, we propose the Semantic-Aware Adaptive Reasoning Network (SAARN), which decouples linguistic features and injects them at appropriate stages of the network to ensure semantic alignment. Unlike existing decoupling-based approaches that uniformly inject all linguistic components across modules (Lei et al. 2024; Zhang et al. 2025), SAARN employs a more targeted strategy. Specifically, the Category-Dominated Linguistic Enhancement (CDLE) module injects only class-level linguistic features into the encoder. This injection aligns early visual representations with accurate category-level semantics. Meanwhile, global linguistic features are selectively integrated via class-guided gating, reinforcing semantic consistency. In the multi-scale fusion stage, the Adaptive Reasoning Fusion Module (ARFM) performs scale-aware enhancement by dynamically weighting semantic cues across multi-scale features. This mechanism simulates a reasoning process, guiding the model to infer the most semantically consistent instance among densely packed objects of the same category.

The main contributions of this paper are as follows:

- We propose RIS-LAD, the first fine-grained referring image segmentation dataset tailored for low-altitude drone scenarios. It offers diverse LAD scenes with rich textual descriptions and high-quality segmentation masks.
- To address category drift and object drift in RLADIS, we propose the Semantic-Aware Adaptive Reasoning Network (SAARN). This model focuses on aligning semantic categories and selectively integrates the most relevant linguistic features at optimal feature scales.
- We conducted extensive experiments to investigate the challenges inherent in RLADIS and construct a dedicated benchmark for this task. Our proposed SAARN achieves state-of-the-art performance on the RIS-LAD benchmark.

Related Work

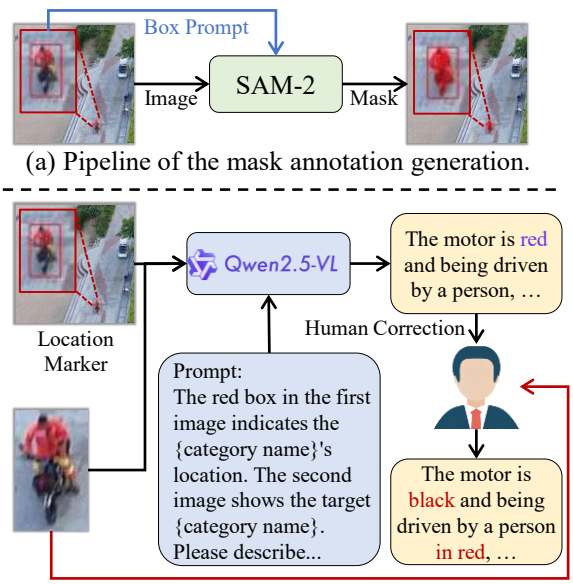
Referring Image Segmentation (RIS) aims to localize target objects at the pixel level in an image based on natural language expressions (Hu, Rohrbach, and Darrell 2016). A representative line of methods focuses primarily on enhancing vision-language alignment to improve segmentation performance (Chng et al. 2024; Wang et al. 2024; Shah, VS, and Patel 2024). VATEX (Nguyen-Truong et al. 2025) leverages a CLIP-based prior module to generate heatmaps, which enforces semantic consistency and improves contextual alignment between language and vision. ASDA (Yue et al. 2024) employs a dual-alignment mechanism with dynamic feature selection to better align visual and linguistic modalities. Other approaches explore novel architectural designs for RIS (Yang et al. 2024; Chen et al. 2024; Xia et al. 2024). IteRPrimE (Wang et al. 2025) refines visual activation maps through an iterative Grad-CAM refinement strategy and adopts a primary-word emphasis module to improve the identification of key semantic components in language expressions. Although these approaches demonstrate promising performance on conventional imagery, their effectiveness deteriorates notably when applied to LAD.

Referring Remote Sensing Image Segmentation (RRSIS) is a domain-specific extension of the RIS task for remote sensing images, first introduced by (Sun et al. 2022). Given the high resolution and small object sizes in remote sensing images, many recent approaches have adopted multi-scale feature fusion (Ma et al. 2025; Lu et al. 2025). FIANet (Lei et al. 2024) disentangles various components of the referring expression and incorporates a text-aware multi-scale enhancement module to improve multi-modal discrimination. RMSIN (Liu et al. 2024) designs rotation-aware modules to precisely segment objects based on the unique properties of remote sensing imagery. Meanwhile, pre-trained large vision models have drawn growing interest in this field (Dong et al. 2025). RSRefSeg integrates a CLIP-based (Radford et al. 2021) module to capture implicit visual activations and uses them as guidance prompts for SAM (Kirillov et al. 2023). For benchmarks, existing RRSIS datasets (Dong et al. 2024; Liu et al. 2024; Yuan et al. 2024) are mainly constructed from high-altitude sources such as Google Earth and GF-2, which feature scenes with larger objects and fixed viewpoints. Although some studies have explored low-altitude scenarios (Sun et al. 2025), challenges such as tiny objects and high instance density persist due to diverse viewpoints. Existing LAD-based datasets like UAVid-RIS and VDD-RIS (Li and Zhao 2025) remain limited in object categories and expression granularity, restricting their ability to support fine-grained RIS tasks.

RIS-LAD Dataset

This section introduces the construction process and key characteristics of the RIS-LAD.

Image and Category Selection The suitable images for the RIS task are selected from the publicly available CO-Drone dataset (Ye et al. 2025), which is designed for LAD-oriented object detection. The selected images encompass



(a) Pipeline of the mask annotation generation.

(b) Pipeline of referring expression generation.

Figure 2: Illustration of the semi-automatic annotation pipeline for RIS-LAD. Referring expression generation employs Qwen2.5-VL to produce initial descriptions conditioned on location markers and cropped object instances.

Figure 3: Word cloud for referring expressions in the RIS-LAD dataset.

Dataset	Image Source	Shooting Angle	Nighttime Scene
RefDIOR	Google Earth	fixed	✗
NWPU-Refer	Google Earth	fixed	✗
RISBench	Google Earth, GF-2, JL-1	fixed	✗
RefSegRS	Helicopter (above 1000m)	fixed	✗
RRSIS-D	Google Earth	fixed	✗
RIS-LAD	Drone (30m-100m)	30°–60°	✓

Table 1: Comparison of RIS-LAD and publicly available referring remote sensing image segmentation datasets.

diverse real-world urban scenes under various lighting conditions and from different viewpoints. Following common procedures for constructing RRSIS datasets (Yuan et al. 2024; Dong et al. 2024; Yang et al. 2025), the images are divided into uniform 1080×1080 patches, thus preserving fine-grained details. A manual screening step is performed to ensure that the retained tiles exhibit representative characteristics of the LAD imagery.

Following mainstream LAD object detection benchmarks (Zhu et al. 2021; Du et al. 2018; Mueller, Smith, and Ghanem 2016), we select eight representative object categories: people, car, motor, bicycle, tricycle, truck, bus and boat. These categories are commonly encountered and crucial for various drone-based applications. To better reflect real-world challenges, we apply a size constraint to ensure that more than 90% of the annotated instances possess a mask coverage ratio below 0.1. This constraint highlights the prevalence of small and visually inconspicuous objects in LAD scenarios.

Mask and Expression Generation Pipeline Segment Anything Model 2 (SAM-2) (Ravi et al. 2024) is employed to automatically generate high-quality instance masks, where the oriented bounding boxes from CODrone provide precise prompts to guide the segmentation process.

As shown in Fig. 2, referring expressions are generated using the multi-modal large language model (MM-LLM) Qwen2.5-VL (Bai et al. 2025). To capture both instance-level appearance and spatial attributes, each object is cropped and upsampled to improve visual clarity. The cropped instance and the corresponding full image, marked with a red bounding box that indicates the location, are then fed into Qwen2.5-VL. This dual-input strategy addresses challenges associated with small object sizes by the MM-LLM's ability to capture fine-grained visual features and generate accurate spatial descriptions when relying solely on raw coordinates. As shown in Fig. 3, we further visualize the linguistic diversity through word clouds generated from the refined referring expressions.

Manual Refinement and Dataset Splitting All automatically generated annotations are carefully verified and refined to ensure strong semantic alignment among images, referring expressions, and masks. This process yields a total of 13,871 high-quality image-text-mask triplets, randomly divided into training, validation, and test sets with a 7: 1: 2 ratio.

Comparison of RIS-LAD and RRSIS datasets. As summarized in Table 1, RIS-LAD exhibits distinct characteristics that distinguish it from existing RRSIS datasets (Dong et al. 2024; Yang et al. 2025; Yuan et al. 2024; Liu et al. 2024; Lu et al. 2025). First, in contrast to RRSIS datasets that rely on fixed top-down viewpoints, RIS-LAD contains imagery captured from low-altitude drones at oblique angles between 30° and 60°, thus offering greater variation in perspective and object appearance. Secondly, RIS-LAD uniquely includes nighttime scenes, enabling evaluation under challenging illumination conditions. Furthermore, it employs a semi-automatic annotation pipeline that integrates MM-LLMs with manual verification, ensuring both annotation quality and efficiency.

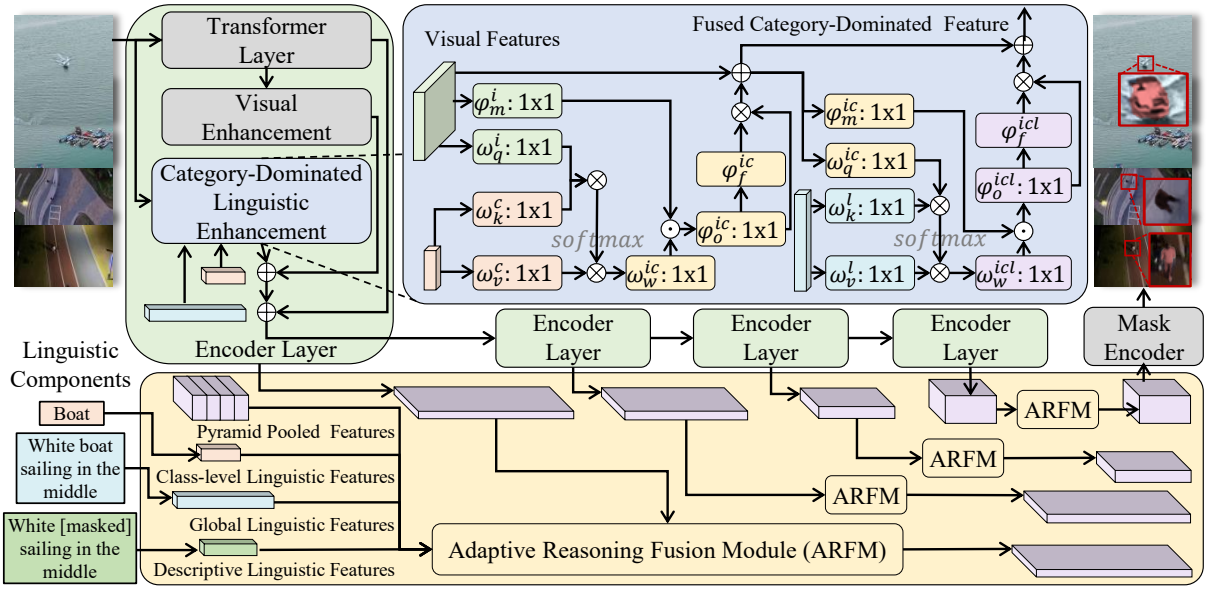


Figure 4: Illustration of the proposed Semantic-Aware Adaptive Reasoning Network (SAARN). The framework comprises two core components: Category-Dominated Linguistic Enhancement (CDLE, top) and Adaptive Reasoning Fusion Module (ARFM, bottom). To mitigate early-stage semantic misalignments, CDLE injects class-level linguistic features at the encoder stage, while ARFM adaptively integrates multi-scale and linguistic features to reason about the correct object in scenes with dense object distributions. ω and φ denote linear and nonlinear mappings, respectively.

Method

Overview

As shown in Fig. 4, we propose the Semantic-Aware Adaptive Reasoning Network (SAARN), a framework designed to address category drift and object drift in RLADIS. SAARN incorporates two key components: the Category-Dominated Linguistic Enhancement (CDLE) and the Adaptive Reasoning Fusion Module (ARFM).

Given an input image and its corresponding referring expression, a Swin Transformer (Liu et al. 2021) encoder extracts visual features, and the expression is decomposed into global, class-level, and descriptive components, yielding three types of linguistic features: l, c, and d. As shown in the bottom-left of Fig. 4, these linguistic features correspond to the complete expression, the category name, and the descriptive content excluding the category term, respectively. Such disentanglement facilitates targeted modeling of semantic intent, object class information, and detailed spatial or attribute cues, which aligns with the intuitive human reasoning process of first locating the described region or identifying all category-matching objects, and then selecting the target based on specific attributes.

At the encoder stage, CDLE injects class-level guidance through c to align early visual representations with the correct category. Subsequently, it integrates l to construct category-dominated fused representations, which are then combined via residual fusion with output from the previous layer and the Visual Enhancement module (Liu et al. 2024) for multi-scale fusion. In the multi-scale fusion stage, ARFM employs an adaptive integration of pyramid pooled features with l, c, and d across different scales. It assigns adaptive weights based on semantic alignment and spatial resolutions, allowing features at each scale to emphasize the most relevant linguistic cues. The final fused representation is then passed to a mask decoder to generate the segmentation output.

Category-Dominated Linguistic Enhancement

In order to address category drift, we propose the Category-Dominated Linguistic Enhancement (CDLE) module, which selectively injects class-level linguistic features to precisely align early visual representations with the correct categories. Descriptive linguistic components are deliberately excluded to prevent alignment with visually similar but incorrect objects, thus mitigating category drift. This module is integrated into each stage of the visual encoder to enable fine-grained cross-modal interaction.

The visual features, or the output from the previous encoder layer, are denoted as $x \in \mathbb{R}^{B\times HW\times D}$, where B is the batch size, HW is the flattened spatial resolution, and D is the dimension of the feature. The class-level linguistic features $c \in \mathbb{R}^{B\times D_b\times N}$ are derived by extracting the category token from the referring expression and encoding it using a pre-trained BERT model (Devlin et al. 2019). D_b and N denote the hidden dimension and the number of linguistic tokens, respectively. Inspired by (Yang et al. 2022), we project both x and c into a shared embedding space to compute the scaled dot-product attention to align category

cues and visual features:

$$\alpha^c = \text{softmax}\left(\frac{\omega_q^i(x)\cdot(\omega_k^c(c))^\top}{\sqrt{d}}\right)\cdot\omega_v^c(c), \quad (1)$$

where each ω denotes a 1×1 convolutional layer applied along the channel dimension, and d is the scaling factor.

Although category-guided attention α^c effectively captures semantic alignment between input and class-level linguistic cues, it tends to over-emphasize category semantics while suppressing other informative visual contexts. To offset this, a residual gate mechanism is employed, which facilitates adaptive weighting between α^c and the original feature x:

$$z^c = \varphi_o^{ic}(\omega_w^{ic}(\alpha^c)\odot\varphi_m^i(x)), \quad (2)$$

$$f^c = x + z^c\cdot\varphi_f^{ic}(z^c), \quad (3)$$

where $\varphi_m^i(\cdot)$ and $\varphi_o^{ic}(\cdot)$ denote the convolutional layer followed by GELU. And $\varphi_f^{ic}(\cdot)$ denotes a two-layer linear projection with ReLU (Agarap 2018) and Tanh activations.

Given the global linguistic features $l\in\mathbb{R}^{B\times D_l\times N}$ and the category-guided feature f^c, we again compute the scaled dot-product attention and residual gate,

$$a^l = \text{softmax}\left(\frac{\omega_q^{ic}(f^c)\cdot(\omega_k^l(l))^\top}{\sqrt{d}}\right)\cdot\omega_v^l(l), \quad (4)$$

$$z^l = \varphi_o^{icl}(\omega_w^{icl}(\alpha^l)\odot\varphi_m^{ic}(f^c)), \quad (5)$$

$$f^l = f^c + z^l\cdot\varphi_f^{icl}(z^l), \quad (6)$$

This design ensures that l is selectively activated in regions corresponding to c, thus providing a safeguard mechanism for the fusion between l and f^c. The output f^l retains the discriminative semantics of the referring expression while maintaining a strong alignment with the correct category. This effectively mitigates attention drift toward visually or semantically similar but incorrect instances.

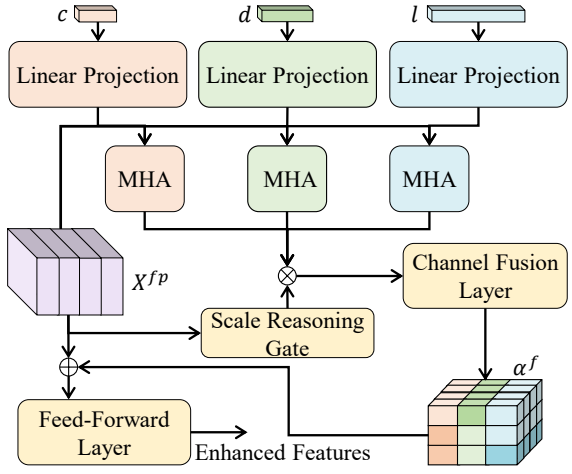


Figure 5: The architecture of the Adaptive Reasoning Fusion Block. Three types of linguistic features are enhanced via Multi-Head Attention (MHA).

Adaptive Reasoning Fusion Module

The Adaptive Reasoning Fusion Module (ARFM) has been proposed to address the problem of object drift induced by densely packed same-category instances in LAD scenarios. Inspired by the human perception process, in which reasoning typically begins with coarse spatial localization and gradually refines toward specific object attributes, ARFM adopts a progressive fusion strategy. It establishes a dynamic weighting scheme between the fused multi-scale features $X^f = \{x_1^f, x_2^f, x_3^f, x_4^f\}$ and the linguistic features l, c, and d. Thus, the model can capture scale-specific semantic cues more effectively, enabling semantic-aware attention across spatial resolutions. Fine-grained attributes like color and shape are better preserved at higher resolutions, whereas coarse spatial cues such as "top-left corner" become more distinguishable at lower resolutions.

X^f is first aligned using pyramid pooling, which is then downsampled to a shared resolution and channel dimension, resulting in $X^{fp}\in\mathbb{R}^{B\times C'\times H'\times W'}$. These features are then fed into the Adaptive Reasoning Fusion Block (ARFB), which performs cross-modal reasoning across multiple branches, guided by linguistic features. As shown in Fig. 5, ARFB contains three parallel linguistic branches that attend to c, d and l, respectively. X^{fp} is projected into a shared embedding space and interacts with each linguistic feature. Formally, for each linguistic feature $s\in\{l,d,c\}$, we compute:

$$q_s = W_s^q X^{fp} + b_s^q, \quad k_s = W_s^k s + b_s^k, \quad (7)$$

$$v^s = W_s^v s + b_s^v, \quad \alpha_s = \text{MHA}(q_s, k_s, v_s) \quad (8)$$

where W_s^q, W_s^k, and W_s^v are learnable projection matrices for the query, key, and value associated with the linguistic branch s, and $\text{MHA}(\cdot)$ denotes the standard multi-head attention (Vaswani et al. 2017). The resulting responses α_s encode the cross-modal correlation between X^{fp} and each linguistic feature.

To adaptively modulate the contribution of each semantic branch, a Scale Reasoning Gate (SRG) module based on X^{fp} is introduced:

$$[w_l, w_d, w_c] = \text{Softmax}(\text{SRG}(X^{fp})) \quad (9)$$

$\text{SRG}(\cdot)$ consists of a global average pooling layer followed by two convolutional layers with ReLU activation, which produce three normalized fusion weights. The weighted attention is computed as,

$$\alpha^f = \text{Fuse}(w_l\alpha_l,\ w_d\alpha_d,\ w_c\alpha_c) \quad (10)$$

where $\text{Fuse}(\cdot)$ denotes the channel fusion layer, which consists of the channel concatenation followed by convolutional projection for compression and alignment.

Finally, α^f is combined with X^{fp} via residual addition and further enhanced through a feed-forward network (Vaswani et al. 2017), before being passed into the subsequent scale-aware gate fusion module (Liu et al. 2024). ARFM adaptively incorporates different linguistic features into the main semantic path, placing varying emphasis on each feature to enable it to focus on the regions to which it is most sensitive. This improves context consistency and the quality of object boundary prediction.

Method	Publication	P@0.5		P@0.6		P@0.7		P@0.8		P@0.9		oIoU		mIoU	
		Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
LAVT	CVPR 2022	33.95	31.67	31.16	27.04	26.29	23.02	19.41	17.55	9.89	8.58	44.03	41.97	32.25	30.14
ASDA	MM 2024	35.74	33.19	29.44	26.85	23.14	20.16	13.54	12.78	4.01	3.94	38.70	37.53	35.46	33.33
VATEX	WACV 2025	19.27	16.07	14.76	11.87	9.74	8.47	6.02	4.67	2.08	1.63	24.83	24.27	20.32	18.53
LGCE	IEEE TGRS 2024	28.22	26.38	24.07	21.82	19.99	17.55	15.11	13.32	7.88	7.06	41.72	40.75	27.68	26.17
FIANet	IEEE TGRS 2024	42.91	40.21	37.61	35.22	32.59	29.93	26.43	23.71	14.83	13.21	45.24	43.39	39.61	37.44
RMSIN	CVPR 2024	45.85	43.36	40.97	38.11	35.24	32.36	28.01	25.08	16.33	13.75	50.17	48.82	42.08	39.60
RSRefSeg	IGARSS 2025	46.35	44.73	43.48	40.21	39.76	36.30	35.17	30.91	26.58	21.46	50.04	47.71	43.42	41.16
CADFormer	JSTARS 2025	45.34	42.20	38.90	36.99	33.60	31.92	26.86	23.60	15.40	12.63	47.37	46.47	41.36	39.32
SAARN(**ours**)		47.06	45.02	43.12	39.34	38.25	33.37	31.81	26.24	19.27	15.31	**51.54**	**49.60**	**44.30**	**41.67**

Table 2: Comparison with state-of-the-art RIS and RRSIS methods on the proposed RIS-LAD dataset.

Experiments

Implementation Details

The model is trained on the proposed RIS-LAD dataset for 50 epochs using the AdamW (Loshchilov and Hutter 2019) optimizer, with an initial learning rate of 3×10^{-5} and a weight decay of 0.01. A polynomial decay schedule is employed to progressively reduce the learning rate. All experiments were performed on four NVIDIA RTX 3080 GPUs in a batch size of 8. The param size of SAARN is 254.703M, and its memory overhead is 14,214 MB during inference.

For metrics, we report Precision@0.5–0.9 (P@X), Overall Intersection-over-Union (oIoU), and Mean Intersection-over-Union (mIoU). While P@X highlights accurate predictions, it may overestimate performance on small objects due to their ease of enclosure. Thus, oIoU and mIoU are used as primary metrics to evaluate segmentation performance.

Comparison with SOTA Methods

As shown in Table 2, the proposed SAARN achieves the best performance across both primary metrics: oIoU and mIoU. Compared to RMSIN, the state-of-the-art among previous methods (Yuan et al. 2024; Liu, Jiang, and Zhang 2025) that does not incorporate pre-trained large vision models (LVMs), our approach increases oIoU from 50.17 to 51.54 on the validation set and further achieves a significant mIoU improvement of 2.07 on the test set. Notably, SAARN achieves substantial gains under stricter localization thresholds, including +18.0% in P@0.9 and +13.6% in P@0.8.

By incorporating CLIP and SAM, RSRefSeg achieves competitive performance on Precision@0.8–0.9. However, it still underperforms compared to SAARN on the core segmentation metrics. Visualization analysis reveals that the inconsistent performance of RSRefSeg due to its tendency to produce over-extended masks, as shown in Fig. 6. Although these masks may cover the referred object, their boundaries exhibit severe region overgeneralization, leading to lower oIoU and mIoU. In contrast, SAARN generates more boundary-accurate masks, resulting in stronger segmentation performance. Compared to FIANet, which also utilizes linguistic decoupling and multi-scale fusion, SAARN performs notably better, owing to its superior instance-level perception in dense distributions of same-class objects.

CDLE	ARFM	oIoU		mIoU	
		Val	Test	Val	Test
✗	✗	49.77	48.32	42.08	39.60
✗	✓	51.31	**49.70**	43.97	40.68
✓	✗	49.82	49.28	43.31	41.02
✓	✓	**51.54**	49.60	**44.30**	**41.67**

Table 3: Ablation study of CDLE and ARFM.

l	c	d	oIoU		mIoU	
			Val	Test	Val	Test
✗	✗	✗	49.82	49.28	43.31	41.02
✓	✗	✗	50.64	49.44	43.52	41.43
✓	✗	✓	49.90	49.30	43.47	41.17
✓	✓	✗	51.06	49.52	43.98	41.61
✓	✓	✓	**51.54**	**49.60**	**44.30**	**41.67**

Table 4: Ablation study of linguistic components.

Conventional RIS methods, such as ASDA and VATEX, face significant performance degradation under the LAD setting, mainly due to the small size of objects and complex background. In particular, VATEX shows relatively limited performance, partially because it relies on a frozen CLIP image encoder for visual enhancement. This leads to negative transfer because of the difference in domain between LAD imagery and the CLIP pre-training distribution. This observation underscores the fundamental domain gap between RLADIS and conventional RIS.

Ablation Study

We conduct ablation studies on the two core modules of SAARN, CDLE and ARFM, to validate their effectiveness. As shown in Table 3, adding CDLE significantly improves mIoU. By reinforcing category semantics early in the process, CDLE helps the model focus on the correct object categories, resulting in better instance-level segmentation precision. In addition, integrating ARFM significantly improves oIoU, increasing it from 49.82 to 51.54 on the validation set. This improvement results from ARFM's ability to distinguish densely distributed objects across scales, facilitat-

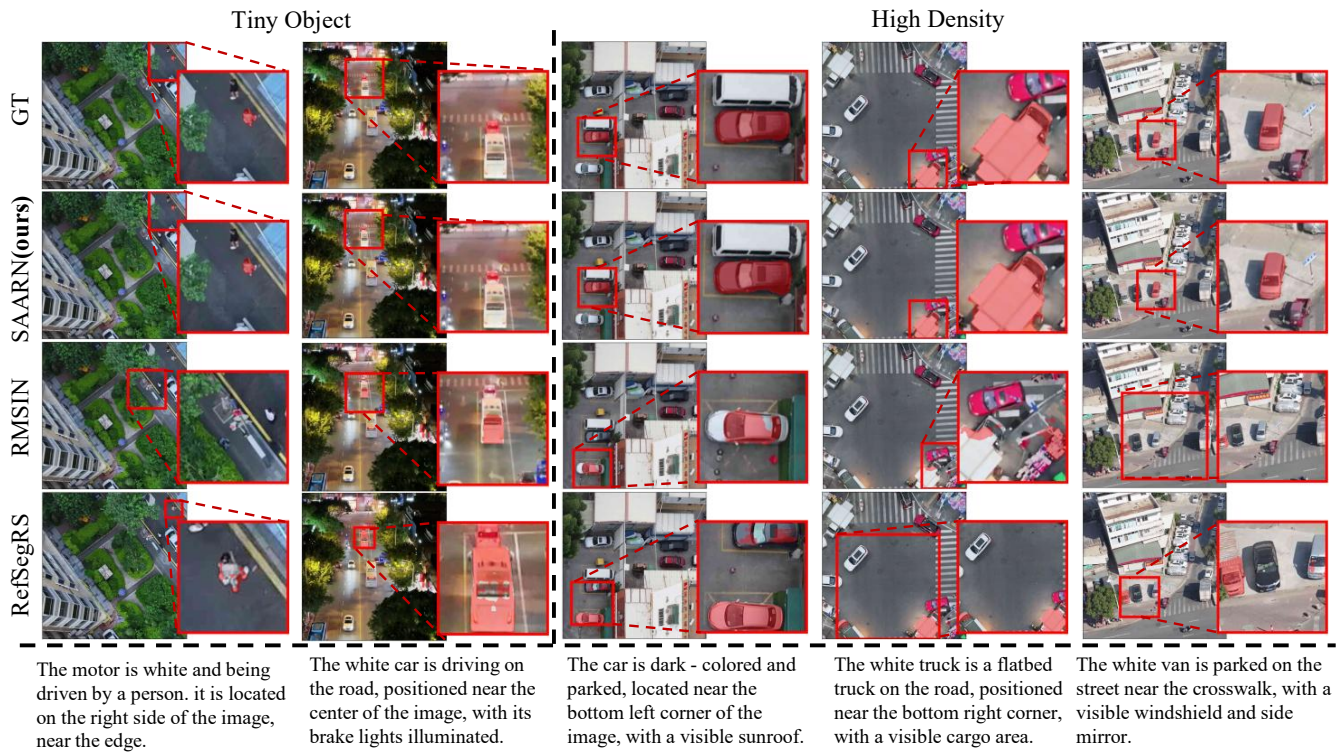


Figure 6: Qualitative comparisons between SAARN and the previous SOTA RMSIN and RefSegRS. The left part illustrates the predictions on tiny-object examples, where category drift easily occurs. The right part shows predictions in high-density scenarios, which are prone to object drift.

ing more precise localization and significantly reducing false activations.

Linguistic Components Ablation As shown in Table 4, we further conduct an analysis of the linguistic components used in the ARFM. Specifically, we evaluate the contributions of the global, class-level and descriptive linguistic feature l, c and d. Adding l improves the oIoU on validation set from 49.82 to 50.64, indicating its role in providing global semantic guidance to enhance alignment across multiple scales. Interestingly, performance drops slightly when d is introduced alongside l. The decrease suggests that, without class-level guidance, the model becomes more susceptible to semantic interference from fine-grained descriptive cues, which may disrupt category-consistent representation learning, especially at coarser resolutions. In contrast, introducing c significantly improves performance on both oIoU and mIoU. This highlights the critical role of c in semantic alignment, as it provides strong category-specific constraints.

Finally, combining all three features achieves the best overall performance. This indicates that each linguistic component injected in ARFM plays a distinct but complementary role in the reasoning process, collectively enhancing the model's ability to accurately localize the referred object.

Visualization and Quantitative Results

We present a qualitative comparison with two of the SOTA RRSIS methods to better illustrate the effectiveness of our proposed model. As shown in Fig. 6, under typical LAD scenarios, our model accurately identifies the correct object category and effectively distinguishes the object from nearby same-category instances, even in complex backgrounds and under varying drone viewpoints. In contrast, the compared RRSIS methods exhibit noticeable drift, often leading to misclassification and confusion among closely clustered objects of the same class.

Conclusion

In this work, we propose RIS-LAD, the first fine-grained dataset designed for RLADIS. Through detailed analysis, we identified two core challenges specific to this setting: category drift and object drift, both of which significantly degrade the performance of existing RES and RRSIS models. To address these issues, we propose a semantic-aware adaptive reasoning network which comprises two key modules: CDLE and ARFM. CDLE enhances early-stage category alignment, while ARFM performs adaptive multi-scale reasoning to mitigate object drift. The experiments demonstrate the strong performance of our framework and underscore the challenges posed by RIS-LAD, setting a new benchmark for LAD scenarios.

Acknowledgments

This work was supported by the National Science Fund for Distinguished Young Scholars (No.62025603 and No.62525605), National Natural Science Foundation of China (No. U21B2037, U22B2051, No. U23A20383, No. 62176222, No. 62176226, No. 62272401, No. 62576300).

References

Agarap, A. F. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Banafaa, M. K.; Pepeoğlu, Ö.; Shayea, I.; Alhammadi, A.; Shamsan, Z. A.; Razaz, M. A.; Alsagabi, M.; and Al-Sowayan, S. 2024. A comprehensive survey on 5G-and-beyond networks with UAVs: Applications, emerging technologies, regulatory aspects, research trends and challenges. *IEEE access*, 12: 7786–7826.

Barekatain, M.; Martí, M.; Shih, H.-F.; Murray, S.; Nakayama, K.; Matsuo, Y.; and Prendinger, H. 2017. Okutama-action: An aerial view video dataset for concurrent human action detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 28–35.

Casanova, J. J.; Bergmann, N. T.; Kalin, J. E.; Heineck, G. C.; and Burke, I. C. 2025. A comparison of protocols for high-throughput weeds mapping. *Smart Agricultural Technology*, 101076.

Chen, K.; Zhang, J.; Liu, C.; Zou, Z.; and Shi, Z. 2025. RSRefSeg: Referring Remote Sensing Image Segmentation with Foundation Models. *arXiv preprint arXiv:2501.06809*.

Chen, Y.-C.; Li, W.-H.; Sun, C.; Wang, Y.-C. F.; and Chen, C.-S. 2024. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *European Conference on Computer Vision*, 323–340. Springer.

Chng, Y. X.; Zheng, H.; Han, Y.; Qiu, X.; and Huang, G. 2024. Mask grounding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26573–26583.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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, volume 1 (long and short papers)*, 4171–4186.

Dong, Z.; Sun, Y.; Liu, T.; and Gu, Y. 2025. DiffRIS: Enhancing Referring Remote Sensing Image Segmentation with Pre-trained Text-to-Image Diffusion Models. *arXiv preprint arXiv:2506.18946*.

Dong, Z.; Sun, Y.; Liu, T.; Zuo, W.; and Gu, Y. 2024. Cross-modal bidirectional interaction model for referring remote sensing image segmentation. *arXiv preprint arXiv:2410.08613*.

Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; and Tian, Q. 2018. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, 370–386.

Hu, R.; Rohrbach, M.; and Darrell, T. 2016. Segmentation from natural language expressions. In *European conference on computer vision*, 108–124. Springer.

Huang, J.; Xu, Z.; Liu, T.; Liu, Y.; Han, H.; Yuan, K.; and Li, X. 2025. Densely Connected Parameter-Efficient Tuning for Referring Image Segmentation. *arXiv preprint arXiv:2501.08580*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Lei, S.; Xiao, X.; Zhang, T.; Li, H.-C.; Shi, Z.; and Zhu, Q. 2024. Exploring fine-grained image-text alignment for referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*.

Li, F.; Ren, Z.; Pan, C.; Ren, H.; Jin, J.; Wang, Q.; and Wang, J. 2025. Cooperative Sensing and Communication Beamforming Design for Low-Altitude Economy. *arXiv preprint arXiv:2506.20244*.

Li, R.; and Zhao, X. 2025. AeroReformer: Aerial Referring Transformer for UAV-based Referring Image Segmentation. *arXiv preprint arXiv:2502.16680*.

Lin, J.; Hu, Y.; Shen, J.; Shen, Y.; Cao, L.; Zhang, S.; and Ji, R. 2025. What You Perceive Is What You Conceive: A Cognition-Inspired Framework for Open Vocabulary Image Segmentation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2841–2850.

Liu, M.; Jiang, X.; and Zhang, X. 2025. CADFormer: Fine-Grained Cross-modal Alignment and Decoding Transformer for Referring Remote Sensing Image Segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Liu, S.; Ma, Y.; Zhang, X.; Wang, H.; Ji, J.; Sun, X.; and Ji, R. 2024. Rotated multi-scale interaction network for referring remote sensing image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26658–26668.

Liu, T.; and Li, S. 2025. Hybrid Global-Local Representation with Augmented Spatial Guidance for Zero-Shot Referring Image Segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29634–29643.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Lu, X.; Sun, L.; Li, L.; Jiao, L.; Yang, Y.; Huang, Z.; Chai, J.; Liu, X.; Liu, F.; Ma, W.; et al. 2025. RRSECS: Referring remote sensing expression comprehension and segmentation. *IEEE Geoscience and Remote Sensing Magazine*.

Ma, Q.; Li, L.; Lu, X.; Jiao, L.; Liu, F.; Ma, W.; Liu, X.; and Sun, L. 2025. LSCF: Long-term Semantic-guidance ConvFormer for Referring Remote Sensing Image Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*.

Mueller, M.; Smith, N.; and Ghanem, B. 2016. A benchmark and simulator for UAV tracking. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 445–461. Springer.

Nguyen-Truong, H.; Nguyen, E.-R.; Vu, T.-A.; Tran, M.-T.; Hua, B.-S.; and Yeung, S.-K. 2025. Vision-aware text features in referring image segmentation: From object understanding to context understanding. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4988–4998. IEEE.

Pan, Y.; Sun, R.; Wang, Y.; Zhang, T.; and Zhang, Y. 2024. Rethinking the Implicit Optimization Paradigm with Dual Alignments for Referring Remote Sensing Image Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2031–2040.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.

Shah, N. A.; VS, V.; and Patel, V. M. 2024. Lqmformer: Language-aware query mask transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12903–12913.

Shi, L.; and Zhang, J. 2025. Multimodal-Aware Fusion Network for Referring Remote Sensing Image Segmentation. *IEEE Geoscience and Remote Sensing Letters*.

Sun, Y.; Feng, S.; Li, X.; Ye, Y.; Kang, J.; and Huang, X. 2022. Visual grounding in remote sensing images. In *Proceedings of the 30th ACM International conference on Multimedia*, 404–412.

Sun, Z.; Liu, Y.; Zhu, H.; Gu, Y.; Zou, Y.; Liu, Z.; Xia, G.-S.; Du, B.; and Xu, Y. 2025. RefDrone: A Challenging Benchmark for Referring Expression Comprehension in Drone Scenes. *arXiv preprint arXiv:2502.00392*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, W.; Yue, T.; Zhang, Y.; Guo, L.; He, X.; Wang, X.; and Liu, J. 2024. Unveiling parts beyond objects: Towards finer-granularity referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12998–13008.

Wang, Y.; Ni, J.; Liu, Y.; Yuan, C.; and Tang, Y. 2025. Iterprime: Zero-shot referring image segmentation with iterative grad-cam refinement and primary word emphasis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8, 8159–8168.

Wen, L.; Du, D.; Zhu, P.; Hu, Q.; Wang, Q.; Bo, L.; and Lyu, S. 2021. Detection, tracking, and counting meets drones in crowds: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7812–7821.

Xia, Z.; Han, D.; Han, Y.; Pan, X.; Song, S.; and Huang, G. 2024. GSVA: Generalized Segmentation via Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3858–3869.

Yang, Y.; Ma, C.; Yao, J.; Zhong, Z.; Zhang, Y.; and Wang, Y. 2024. ReMamber: Referring Image Segmentation with Mamba Twister. *European Conference on Computer Vision (ECCV)*.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18155–18165.

Yang, Z.; Yao, H.; Tian, L.; Zhao, X.; Li, Q.; and Wang, Q. 2025. A Large-Scale Referring Remote Sensing Image Segmentation Dataset and Benchmark. arXiv:2506.03583.

Ye, K.; Tang, H.; Liu, B.; Dai, P.; Cao, L.; and Ji, R. 2025. More Clear, More Flexible, More Precise: A Comprehensive Oriented Object Detection benchmark for UAV. arXiv:2504.20032.

Yuan, Z.; Mou, L.; Hua, Y.; and Zhu, X. X. 2024. RRSIS: Referring Remote Sensing Image Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*.

Yue, P.; Lin, J.; Zhang, S.; Hu, J.; Lu, Y.; Niu, H.; Ding, H.; Zhang, Y.; Jiang, G.; Cao, L.; et al. 2024. Adaptive Selection based Referring Image Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1101–1110.

Zhang, T.; Wen, Z.; Kong, B.; Liu, K.; Zhang, Y.; Zhuang, P.; and Li, J. 2025. Referring remote sensing image segmentation via bidirectional alignment guided joint prediction. *arXiv preprint arXiv:2502.08486*.

Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; and Ling, H. 2021. Detection and tracking meet drones challenge. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7380–7399.