Enhancing Remote Sensing Visual Grounding with SAM and DINO: A Two-Stage Approach on RSVG-DIOR Dataset

Danyaal Sadiq Muhammad Zaeem Ahsan

Abstract

This paper presents a comprehensive study on improving the Remote Sensing Visual Grounding (RSVG) performance on the challenging RSVG-DIOR dataset. We begin by implementing the baseline RSVG model that processes visual and textual inputs through transformer architectures to localize objects based on natural language descriptions. Our first enhancement integrates Grounding DINO's self-supervised features for better object representation. while our second improvement incorporates the Segment Anything Model (SAM) to provide precise segmentation masks to improve the IoU performance. Through our benchmarks, we demonstrate that our two-stage approach combining Grounding Dino and RSVG-MLCM's box predictions with SAM's segmentation capabilities achieves significant improvements over the baseline, particularly in complex scenarios with multiple similar objects.

Project GitHub repository: https://github.com/danyaalsadiq/Group-5-Deep-Learning-Project.git

1. Introduction

Remote Sensing Visual Grounding (RSVG) is a critical task in computer vision that involves localizing objects in RS images based on natural language descriptions. While significant progress has been made in this field, challenges remain in handling complex scenes with multiple similar objects and diverse linguistic expressions.

Our work focuses on the RSVG-DIOR dataset, a challenging benchmark featuring diverse objects in complex scenes with rich annotations. The dataset presents unique challenges including object occlusion, scale variation, and complex relationships between objects, making it an ideal testbed for our improvements.

We propose a novel two-stage approach that combines the strengths of SAM for precise segmentation and DINO for robust feature representation. Our contributions include: (1) a comprehensive baseline implementation of RSVG-MLCM on RSVG-DIOR, (2) incorporation of DINO's self-supervised features, and (3) systematic integration of SAM for improved segmentation

2. Related Work

Recent advancements in vision-language models and segmentation have significantly improved the performance of visual grounding tasks, particularly in challenging domains like remote sensing. The RSVG framework (Zhan et al., 2023) investigates the adaptation of visual grounding to RS imagery and highlights the unique challenges posed by scale variation and domain-specific semantics. It proposes a dedicated dataset that demonstrates the effectiveness of tailoring models to satellite imagery

Grounding DINO (Liu et al., 2024) builds on the foundational DINO architecture by integrating grounded pre-training strategies, enabling open-set object detection with short text prompts. This method has shown strong zero-shot capabilities, which is critical for applications where object categories may be undefined during training.

The Segment Anything Model (SAM) (Kirillov et al., 2023) introduces a promptable segmentation model that generalizes across domains with minimal fine-tuning, having been trained on 11 million images with 1.1 billion segmentation masks.

3. Dataset and Baseline

3.1. RSVG-DIOR Dataset

The RSVG-DIOR dataset contains 23,463 images with 192,472 referring expressions covering 20 object categories. Key characteristics include:

- High object density (average 8.2 objects per image). This increases the difficulty of locating the correct target. Such as
- Complex spatial relationships between objects such as "A airport is on the upper right of the yellow and orange oval ground track field"

- Diverse referring expressions with varying linguistic complexity

- Balanced distribution across categories and expression types

In pre-processing, as Grounding DINO struggles with long, complex prompts, we used OpenAIs GPT 4 and GPT-4-mini to shorten the prompts into a more precise phrase that maintains the spatial and object information. For example:
Original caption: A ground track field is on the right of the white and gray tiny vehicle
Shortened caption: Tiny vehicle near ground track field.

3.2. Baseline RSVG Model

Our baseline implementation follows the transformer-based architecture from (Li et al., 2020), consisting of:

- Visual encoder: ResNet-50 backbone (pretrained), used to extract visual features from remote sensing images.

- Text encoder: BERT backbone (also pretrained), processes the natural language referring expressions.

- Multimodal fusion: Cross-attention transformer layers are used to fuse textual and visual features.

- Prediction head: Outputs include both bounding box regression (predicts coordinates) and relevance classification (predicts matching scores between text and region).

Our experiment on this model has achieved an IoU score 0f 0.26, establishing our baseline.

4. Methodology

Our proposed enhancements follow a two-stage approach:

4.1. Stage 1: First improvement

Instead of passing the data to RSVG-MLCM alone, the data is passed to both RSVG-MLCM and GroundingDINO (Liu et al., 2024). Output features are concatenated and using a confidence weighting mechanism which takes a weighted average of both models' outputs, the final bounding boxes are generated. The model was fine-tuned on 15 epochs. RSVG was frozen, Grounding DINO was trained with Low-Rank Adapters, and the feature fusion module was left unfrozen.

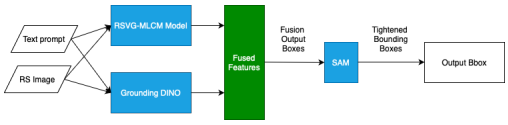

Figure 1. Our proposed architecture

Table 1. Performance Comparison

Model	mIoU
Baseline	26.2
+ DINO	26.2
Full Model	

4.2. Stage 2: Final Model

The final model takes the predictions of the bounding box and attempts to improve it using Segmentation. The Segment Anything Model (Kirillov et al., 2023) receives the image and bounding boxes and produces a segmentation mask of the image. This is used to tighten the box predicted by the model, producing a well-fitted bounding box.

5. Experiments and Results

To evaluate our approach, we conducted inference on the test split using all three model configurations. The baseline RSVG-MLCM model, trained for 150 epochs, achieved an mIoU score of 26.2 on the test set. The RSVG + Grounding DINO model performed similarly, which we attribute to Grounding DINO's difficulty in handling longer and more complex prompts - likely a limitation of its text encoder.

Due to a software issue encountered during inference, we were unable to successfully complete inference using the final stage of our pipeline involving the Segment Anything Model (SAM). As a result, quantitative evaluation for that stage is not currently available. However, we expect it to improve the performance, due to SAM's strong zero-shot capabilities on object segmentation and the performance of our bounding boxes.

6. Discussion

Our experiments reveal several important insights into the behaviours of different modules of the visual grounding pipeline:
We observed that Grounding DINO, while highly effective for general and open-set object detection, tends to struggle with longer and more complex expressions, especially those containing many spatial relationships. This was evident in the RSVG-DIOR dataset, where prompts often involve multiple objects, fine-grained attributes, and spatial relationships. This limitation may stem from the model's lan-

guage encoder or its lack of fine-tuning on richly descriptive queries. To remedy this, we attempted to fine-tune it on LoRA (Hu et al., 2021), create a fallback for Grounding DINOs text encoder failing.

7. Conclusion

In this work, we introduced a multi-stage approach to visual grounding in remote sensing imagery by bringing together the strengths of several leading models. Our pipeline uses RSVG-MLCM for its strong performance in domain-specific grounding, Grounding DINO for its ability to handle open-set object detection with natural language, and the Segment Anything Model (SAM) to refine bounding boxes through precise segmentation. This modular setup highlights how integrating specialized models can lead to meaningful improvements in challenging tasks like visual grounding, especially in complex domains such as remote sensing.

8. Contributions

- Danyaal Sadiq: Implemented baseline model. first improvement and second improvement. Contributed to state of the art survey and worked on the final paper
- Zaeem Ahsan: Chose the dataset and performed preprocessing, assisted in debugging, and contributed to the manuscript writing
- Both authors contributed to the analysis and final paper.

References

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

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Berkowitz, J., Buyssens, A., Hron, J., Lin, J., et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Li, X., Ma, Q., Wang, F., Lu, T., and Wu, Q. Referring expression comprehension with cross-modal attention and multi-scale feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10493–10502, 2020.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., and Zhang, L. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. URL https://arxiv.org/abs/2303.05499.

Zhan, Y., Xiong, Z., and Yuan, Y. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. ISSN 1558-0644. doi: 10.1109/tgrs.2023.3250471. URL http://dx.doi.org/10.1109/TGRS.2023.3250471.