

# Supplementary Materials: Rethinking the Implicit Optimization Paradigm with Dual Alignments for Referring Remote Sensing Image Segmentation

Anonymous Authors

In the supplementary material, we first demonstrate the inherent challenges in the referring remote sensing image segmentation task, which shows the necessity of our proposed DANet. Then, we show more visualization results of our method in remote sensing scenes. Finally, we analyze the failure case and our future work.

## A TASK CHALLENGES

Referring remote sensing image segmentation (RRSIS) faces several challenges due to the complexity and variability of remote sensing data: **(1) Scale and Complexity:** Normal referring image segmentation tasks typically deal with smaller-scale scenes and simpler backgrounds. However, remote sensing images cover large geographical areas with diverse landscapes and can vary significantly in scale and resolution, from large-scale aerial imagery to high-resolution satellite images, as shown in Figure A. It is challenging to maintain accuracy for objects of interest (e.g., buildings, roads, and vegetation) across different scales. **(2) Semantic Variability:** Remote sensing images often contain a wide range of semantic classes, including natural landscapes, man-made structures, and various objects. The diversity of classes and their varying appearances pose challenges for accurately segmenting different types of targets, highlighting the importance of the effective use of textual guidance. **(3) Noise and Artifacts:** Remote sensing images can be affected by noise, artifacts, and occlusions due to long-distance imaging, which can interfere with the segmentation process and lead to inaccuracies in the results.

Addressing these challenges requires specialized techniques that can handle the scale, semantics, variability, and limited context of remote sensing images while effectively leveraging referring expressions for accurate segmentation.

## B VISUALIZATION RESULTS

### B.1 Qualitative Demonstration

In Figure C, we show more qualitative results of referring remote sensing image segmentation to demonstrate the superiority of our method. It can be seen that when the same class of target such as "storage tank" appears in the image, the model needs to accurately identify the correct target based on the given location or shape description such as "right" and "small". The comparison method LAVT [2] may be interfered by the same semantics and mark the wrong target. Our DANet avoids this interference by dual alignments to achieve accurate discrimination in most scenarios.

### B.2 Activation Map

As shown in Figure B, the introduction of our dual alignments effectively aids in textual and visual modal alignment, and some descriptive language that explicitly has an orientation or shape has more accurate activation in images. Without dual alignments, the

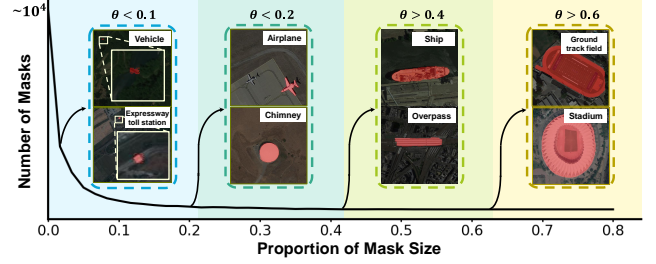


Figure A: Distribution of mask sizes in the RRSIS-D [1] dataset.

activation of text in visual images may receive interference from other similar objects at the wrong location, while the affinity explicit alignment and agent reliable alignment can effectively activate accurate foreground regions, alleviating the problems of the inter-domain gap and class-agnostic predictions.

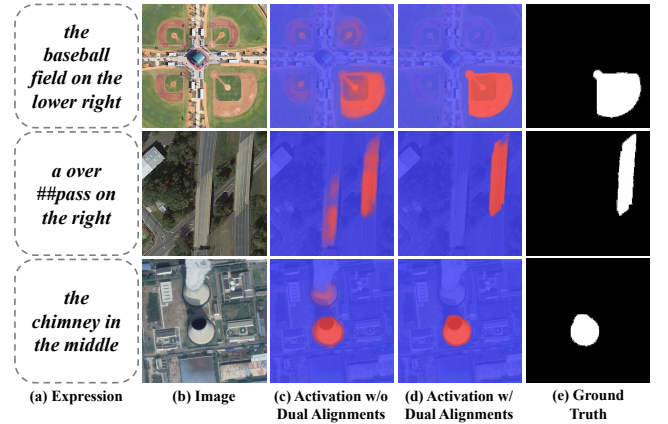


Figure B: Visualizations of activation maps w/ and w/o dual alignments.

## C FAILURE CASE AND FURTHER RESEARCH

For the failure case, as shown in Figure D, the prediction fails to correctly segment the "the airplane on the upper right" due to the ambiguity in the description and limitations in handling specific positional information. The model fuzzy prioritizes the concept of discriminating the right side and ignores the smaller targets present in the upper right. We believe that the pattern of the failure case also sheds light on the possible direction of our future work. Future research directions could include improving text understanding,

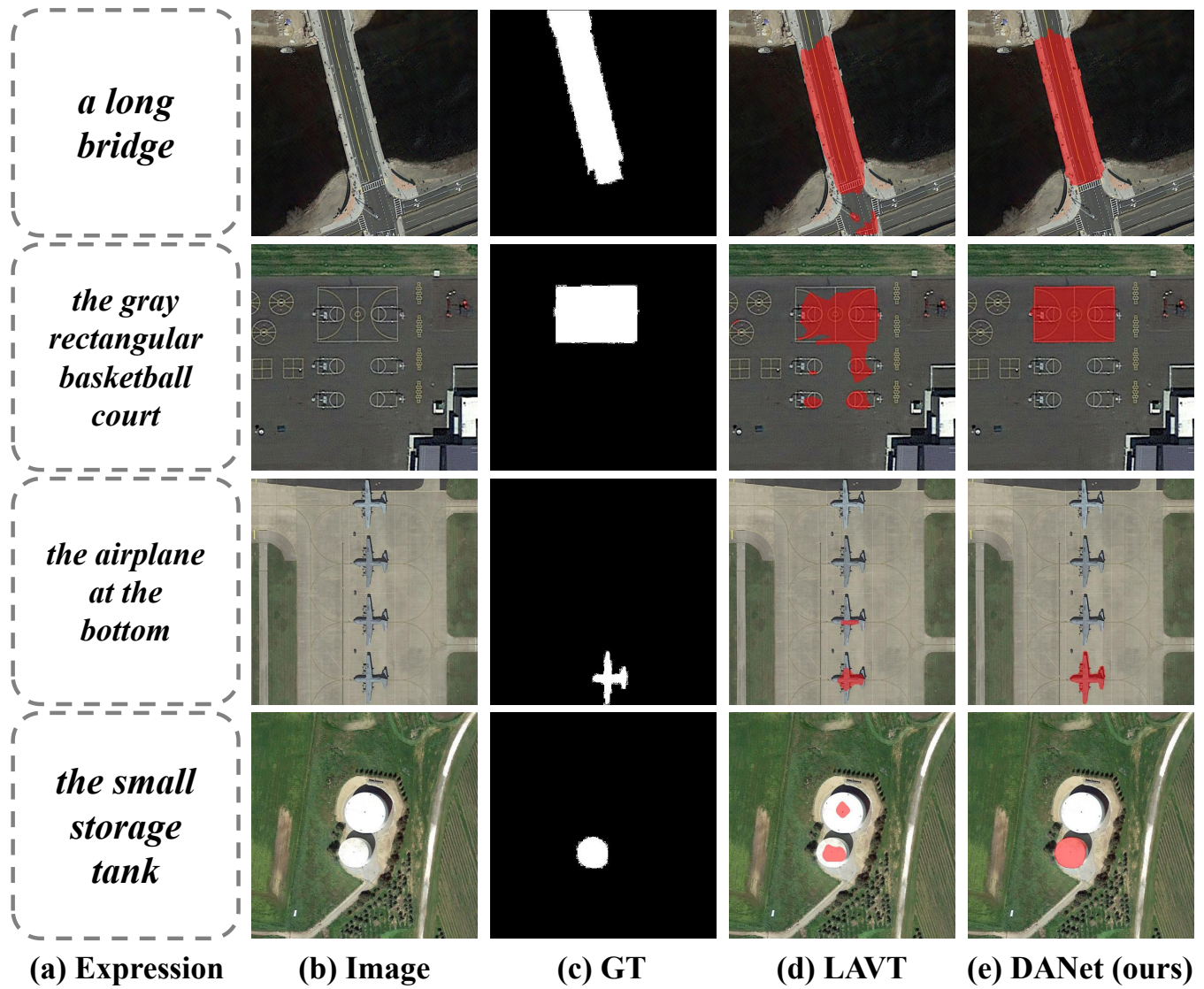


Figure C: Comparison of different RRSIS methods.

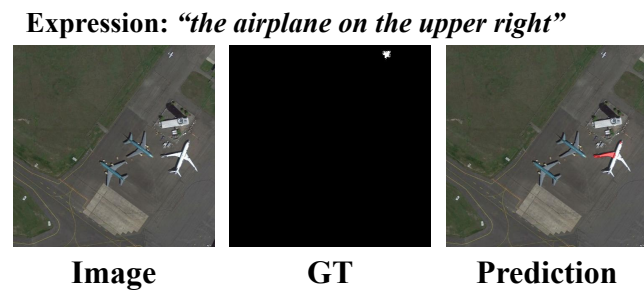


Figure D: Demonstration of the failure case.

enhancing spatial awareness in models, diversifying training data, and integrating multimodal information for better context understanding and segmentation accuracy in referring remote sensing image segmentation tasks.

## REFERENCES

- [1] Sihan Liu, Yiwei Ma, Xiaoqing Zhang, Haowei Wang, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. 2023. Rotated Multi-Scale Interaction Network for Referring Remote Sensing Image Segmentation. *arXiv preprint arXiv:2312.12470* (2023).
- [2] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. 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.