

PAPER SUMMARY

This paper presents CGNet, a novel instance segmentation method for remote sensing imagery. The model's core mechanism involves iteratively refining and fusing information from a contour branch and a mask branch using Gated Recurrent Units (GRU) and an attention head. To enhance feature representation, it also introduces auxiliary supervision from a Contrastive Language-Image Pre-training (CLIP) model. The method's effectiveness has been validated on the NWPU VHR-10 and SSDD datasets.

OVERALL REVIEW

The paper introduces a novel framework for remote sensing instance segmentation, and the experimental results are promising. However, the manuscript requires major revision before it can be considered for publication. Key issues include insufficient justification for the novelty of some core modules, a lack of quantitative data to support claims of being "lightweight", etc.

MAIN CONCERNS

- (1) The paper's related work section mentions that PolySnake also uses GRU for contour iteration. The authors need to more clearly articulate the unique innovation of using ConvGRU to simultaneously iterate on both contour and mask branches in CGNet. What specific designs within the proposed ConvGRU structure make it particularly suitable for this dual-branch fusion task in remote sensing, compared to existing work?
- (2) The paper claims CGNet has a relatively small number of parameters, making it suitable for smaller datasets, and emphasizes its inference speed advantages. However, the comparative experiments in Table 1 and Table 2 do not provide quantitative comparisons of model parameters, FLOPs, or inference speed (FPS) against the benchmark models. This data is essential to validate the "lightweight" and "high-efficiency" claims.
- (3) The paper mentions using a fixed text template ("a pixel of ship") to generate text features. The performance of CLIP can be sensitive to prompt engineering. Did the authors experiment with other text templates? An analysis or discussion of the sensitivity to this choice would make the study more complete.
- (4) The iterative module uses 12 stacked layers with shared weights. How was this number 12 determined? An ablation study on the number of iterations would be beneficial to understand the trade-off between performance and computational cost.
- (5) According to the results, CGNet's performance on medium and large objects is not the best on the NWPU VHR-10 dataset, yet it excels on large objects in the SSDD dataset. Can the authors discuss the potential reasons for this discrepancy? Is it related to the inherent characteristics of the datasets or does it reflect a limitation of

the model in handling different types of large objects?

- (6) The paper mentions using Smooth L1 Loss for the main task and BCE Loss for CLIP supervision. How are the multiple losses from the backbone, contour branch, mask branch, and the final fused output combined and weighted during joint training? A more detailed description of the total loss function is needed
- (7) Incorporating citations of relevant methodologies would better enhance the quality of the article. The following articles are recommended to be referred: 10.1016/j.patcog.2025.111503, doi.org/10.3390/rs17010125. The former article utilizes an auxiliary segmentation branch with geometric priors to guide and enhance the main aircraft detection task, which is methodologically comparable to this study's use of a contour branch to aid instance segmentation. The latter article offering a valuable reference for this paper's discussion on handling complex backgrounds and noise interference.
- (8) The model diagrams (e.g., Fig. 1) are quite dense and somewhat cluttered. And, The text size in Figures 2-4 is too large, making it inconsistent with the rest of the document.