Department of Computer Science

University of Reading

MARS: A Multi-models Agent to interpret Remote Sensing images in details

AUTHOR: WEIJIE CUI SUPERVISOR: DR. MUHAMMAD SHAHZAD MSC DATA SCIENCE AND ADVANCDE COMPUTING

Introduction & Problem

Remote Sensing (RS) images are vital for environmental monitoring, urban planning, and disaster response. However, their complexity makes automated interpretation challenging.

- **Current Limitations:** Single-pass AI models struggle with occlusions, small objects, and contextual reasoning.
- **Research Gap:** A lack of integrated systems that combine detection, iterative refinement, and knowledge-guided reasoning.

Research Questions

- Can the **M**ulti-models **A**gent for **R**emote **S**ensing (**MARS**) framework improve the accuracy and interpretability of RS image analysis over conventional models?

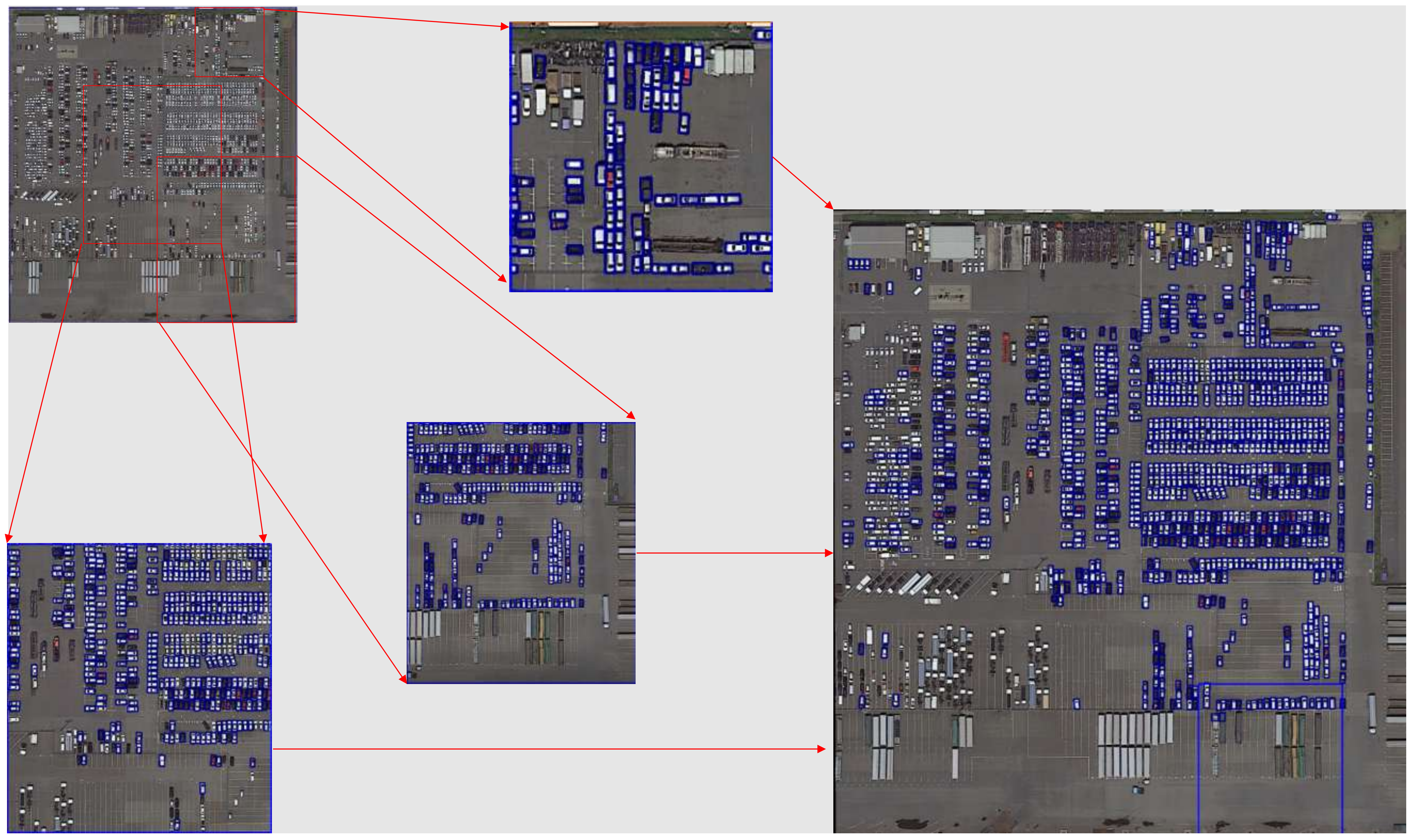
Solution: The MARS Framework

A novel multimodal AI agent that mimics human-like inspection by integrating three components:

- **Vision Model**: A frozen YOLO-OBB [1] visual model for object detection.
- **Reinforcement Learning (RL)** [2] Agent: Dynamically generates observation strategies (zooming, multi-scale patching) to focus on areas of interest.

Methodology

- **Datasets:** DOTA v1 (for object detection).
- **Development:** Design system architecture, then implement and fine-tune RL model.
- **Integration:** Combine modules into a cohesive end-to-end system.
- **Evaluation:**
  - **Quantitative:** Compare against state-of-art single-pass models using mAP and IoU.
  - **Qualitative:** Case studies on complex tasks like counting grounding small objects.



Results

- 5 search attempts with different magnifications improve the exploration of extremely small and partially obscured targets by 20%.
- For objects of normal size, it maintains the same performance as yoloV11.

Discussion

- **Contribution**: **MARS** built a visual search environment and verified the feasibility of integrating agents and visual models.
- **Limitations**: Currently, only the YOLOv11 visual model is integrated and focuses on classification and positioning tasks. Multiple searches increase runtime.
- **Practical Values**: verified the feasibility of integrating agents and visual models.

Conclusions

- An AI agent can improve the overall system's ability to process remote sensing images without increasing the complexity of the visual model.
- Can be used to integrate multiple AI models to created more functional AI systems.
- Next Step: Integrate multiple AI vision models to improve the system's recognition range.

References

1. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024). YOLOv10: Real-Time End-to-End Object Detection. https://arxiv.org/pdf/2405.14458
2. Maxin Lapan, Deep Reinforcement Learning Hands-On, (Packt Publishing Ltd.: 2024) pp. 111-194

Acknowledgements

I would like to express my deepest appreciation to my supervisor, Dr. Muhammad Shahzad, for his invaluable guidance and constructive criticism.

On a personal note, I owe my deepest gratitude to my wife Chuyu Wang and my kids, Yiming and Yidi, for their endless love and encouragement.

Finally, I would like to thank all the participants who volunteered their time for this study.