VISION AGENTIC SYSTEM FOR SPATIAL METRIC QUERY ANSWERING IN REMOTE SENSING IMAGES

Yinghao Wang[1] and Cheng Wang[2]

[1]University of Cambridge, United Kingdom [2]University of East Anglia, United Kingdom

PROBLEM

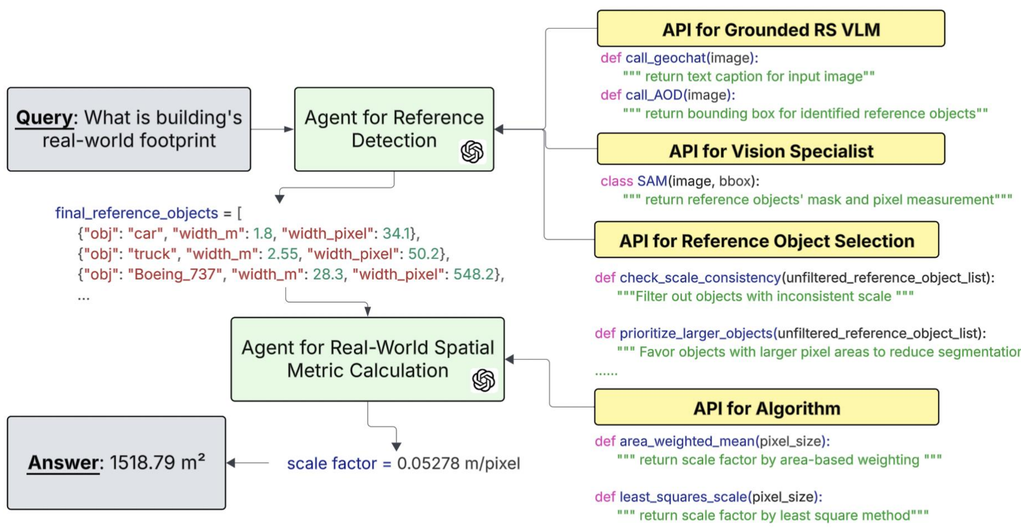
Accurately answering spatial metric query from Remote Sensing (RS) images without scale factor remains challenging for current AI systems. Traditional Vision-Language Models (VLMs) lack robust spatial reasoning capabilities, while RS-specialized VLMs are limited to predefined tasks such as image captioning. The core challenge is autonomously determining image scale factor to convert pixels to real-world measurements without explicit manual input.

OVERVIEW

We propose a visual agentic system that autonomously answers spatial metric queries in remote sensing images without requiring explicit scale information. Our approach uses a two-agent architecture that dynamically integrates specialized models through code generation. This system bridges the gap between general image understanding and precise spatial measurement by autonomously determining scale factors from reference objects in the image. Unlike end-to-end approaches, our method enhances both interpretability and adaptability, making it suitable for various real-world applications including building footprint estimation, urban planning, and geospatial analysis.

METHODOLOGY

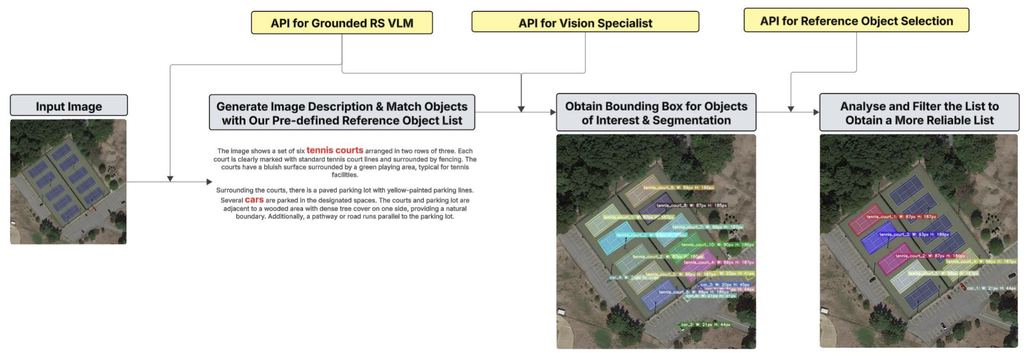
Our system employs a two-agent approach for spatial metric estimation in remote-sensing image.



Agent for Reference Detection

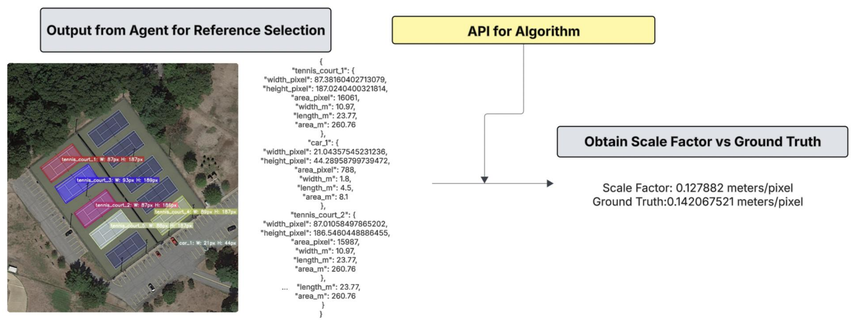
Agent for Reference Detection establishes a reliable reference object list through a three-stage process:

- It leverages Geochat's RS grounding capability (alternatively using GPT-4o plus Landing AI's agentic object detection) to identify objects with known dimensions, generating a reference object list with precise bounding boxes.
- It interfaces with SAM2 for pixel-perfect segmentation of selected references.
- It employs criteria-based filtering to dynamically select optimal reference objects for accurate scale estimation while minimizing errors from imaging artifacts and segmentation inconsistencies.



Agent for Real-World Spatial Metric Calculation

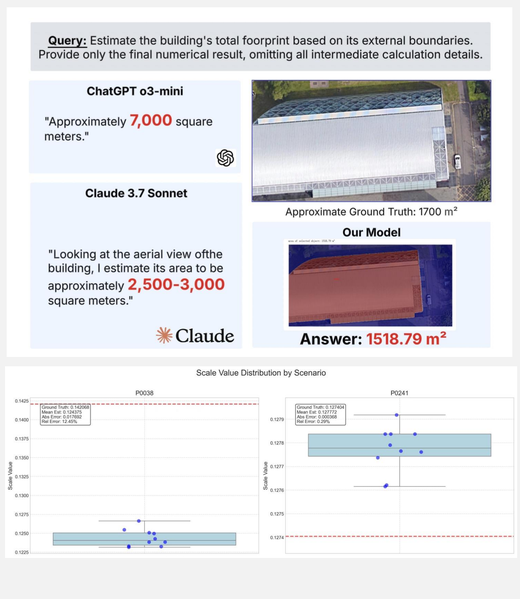
Building on reference detection results, the Spatial Metric Calculation Agent employs context-adaptive algorithms to determine optimal scale factors. The agent intelligently selects the most appropriate method based on reference object characteristics and image properties. Once the accurate image scale is established, the system processes spatial queries by combining RS image analysis with the calculated scale factor, delivering precise real-world measurements that directly address the user's original question.



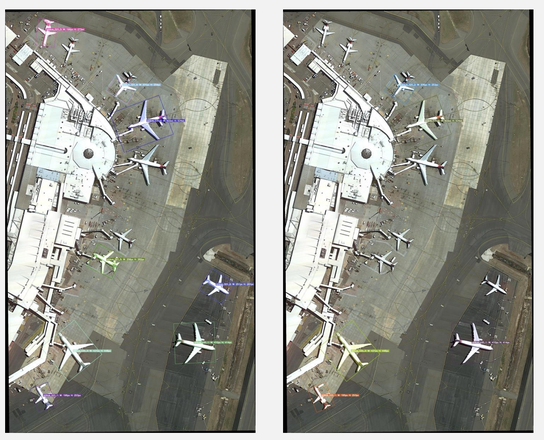
RESULTS

Our visual agentic system integrates code-generation agents to dynamically orchestrate a grounded specialized VLM and a Vision Specialist into structured subroutines. As demonstrated in the example, our system achieves higher accuracy in measuring the footprint area of buildings in remote sensing images, highlighting its potential for autonomous spatial measurement and metric query answering. Our agentic approach enhances interpretability and adaptability, unlocking new possibilities for autonomous geospatial analysis and urban planning.

To further evaluate the robustness and accuracy of scale factor estimation, we query RS images from the DOTA dataset. The examples of tennis courts and airports, along with their corresponding box plots, demonstrate that our method achieves stable scale estimation with results closely aligned to the ground truth.



More examples on agentic detection and selection of reference objects, and scale factor estimation:



Future Work

Future work includes integrating additional Vision Specialists into the APIs to expand capabilities, such as shadow detection for object height estimation, and extending spatial metric analysis to 3D measurements for enhanced real-world applicability.

AFFILIATIONS

UNIVERSITY OF CAMBRIDGE

University of East Anglia

REFERENCES

[1] ZHANG C., WANG S.: Good at captioning, bad at counting: Benchmarking GPT-4V on Earth observation data . In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (Los Alamitos, CA, USA, June 2024), IEEE Computer Society, pp. 7839–7849.

[2] KUCKREJA K., DANISH M. S., NASEER M., DAS A., KHANS., KHAN F. S.: Geochat: Grounded large vision-language model for remote sensing. The IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024).

[3] SURÍS D., MENON S., VONDRICK C.: Vipergpt: Visual inference via python execution for reasoning. Proceedings of IEEE International Conference on Computer Vision (ICCV) (2023).

[4] KIRILLOV et al.: Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2023), pp. 4015–4026.

[5] XIA et al.: DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. The IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018).