Multi-level Fusion for Visual Question Answering on Remote Sensing Images

Soutenance de thèse – Lucrezia TOSATO
28 novembre 2025 à 09h00
Université Paris Cité, Salle du conseil, 45 rue des Saints-Pères 75006 Paris
https://u-paris.zoom.us/j/87655572084?pwd=T3BRhri0aHkijINa5VuvebomZ3fva6.1
Meeting ID: 876 5557 2084, Passcode: 870186

Devant le jury composé de :

Loïc DENIS	Université Jean Monnet	Rapporteur
Loïc LANDRIEU	Ecole des Ponts ParisTech	Rapporteur
Isabelle BLOCH	Sorbonne Université	Examinatrice
Laurent WENDLING	Université Paris Cité	Directeur de thèse
Sylvain LOBRY	Université Paris Cité	Examinateur
Flora WEISSGERBER	ONERA	Examinatrice

Résumé
In recent years, vast amounts of satellite and aerial imagery have become freely accessible thanks to public and private initiatives, such as the European Union's Copernicus program and national mapping agencies. These resources offer opportunities for environmental monitoring, disaster response, and public transparency. However, interpreting such imagery remains challenging, typically requiring expert analysis and manual effort, which limits its timely and widespread usage. Enabling broader access to actionable information requires systems that can answer natural language questions directly from remote sensing images, an emerging task known as Remote Sensing Visual Question Answering (RSVQA).

RSVQA combines computer vision with language understanding to automatically answer queries about geospatial scenes. While prior research was largely centered on optical imagery and single-resolution inputs, this work extends RSVQA by incorporating multi modal and multi-resolution data and by enhancing interpretability through segmentation-based reasoning.

The first two research questions examined the value of multi-modal and multi-resolution data. Our findings indicate that optical, Synthetic-Aperture Radar (SAR), and multi-spectral imagery contribute complementary strengths, and that combining different resolutions helps balance fine detail with contextual information. Together, these strategies improve robustness and accuracy in RSVQA. The third and fourth research questions explored whether segmentation maps could improve interpretability and explainability in RSVQA. The studies confirmed that segmentation guided attention can make reasoning more semantically grounded, even when the segmentation quality is moderate. Furthermore, explicitly linking answers to spatial regions through structured reasoning frameworks improves both transparency and user trust.

Overall, this thesis highlights that both performance and transparency in RSVQA benefit from integrating diverse modalities, multiple scales, and semantic guidance. These contributions demonstrate that combining complementary data sources with interpretable reasoning can enhance both model effectiveness and user trust, moving RSVQA toward more accessible and reliable AI systems.

Mots clés : Deep Learning, Remote Sensing, VisualQuestionAnswering,Multi-Modality, Natural Language Processing, Interpretability, Explainability, Semantic Segmentation