The 8th
ISPRS Geospatial Conference 2025
University of Tehran

Paper ID

A Quantitative Evaluation Framework for Explainable AI in Semantic Segmentation

Ensuring Trust and Transparency in Remote Sensing AI Models

"Reem Hammoud[1], Abdul Karim Gizzini[2], Ali J. Ghandour[3]

Affiliations:

1- American University of Beirut, Lebanon

2- SogetiLabs Research and Innovation (Capgemini), France

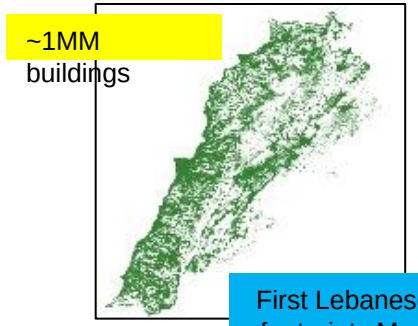
3- National Center for Remote Sensing, CNRS-L, Lebanon"

aghandour@cnrs.edu.lb

GEOspatial Artificial Intelligence (GEOAI) group

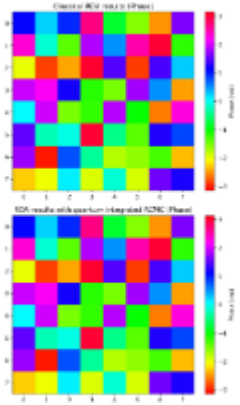
- Lebanese National Center for Remote Sensing - CNRS
- established in April 2015
- rely on openly available satellite imagery
- geogroup.ai
- know-how (deep learning, time-series imagery, in-house python library)
- GEOAI offers: AI-based solutions for Earth Observation (EO) across several verticales.

GeoUrban-AI

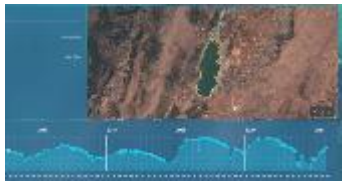


Solar rooftop potential map

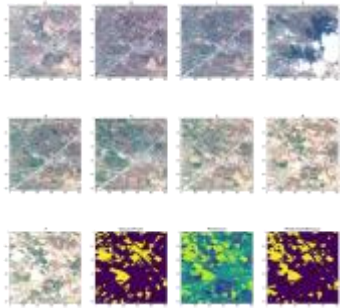
Quantum EO



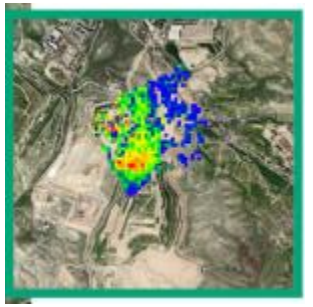
Water Body Assessment



Crop Monitoring



Methane Mapping



Introduction & Problem Statement:

- **The Rise of AI in Geospatial:** AI models are increasingly used for critical tasks like building extraction and scene understanding.
- **The "Black Box" Problem:** These models lack transparency, making it difficult to trust their decisions or debug errors.
- **The Gap:**
  - Explainable AI (XAI) exists, but evaluation methods are designed primarily for image classification, not semantic segmentation.
  - Current evaluation is often qualitative (visual inspection), which is subjective and unreliable.

- **Objective:** To propose a comprehensive quantitative evaluation framework specifically for semantic segmentation.

Proposed Methodology: The Framework

We evaluate XAI methods by manipulating input images based on "heatmaps" and measuring the impact on the model's output.

Key Strategies

- S1 — **Background Only:** Assessing the importance of identified features by removing them.
- S2 — **Highlighted Only:** Isolating features to test their direct influence.
- S3 — **Predicted Mask & Ground Truth (PMGT):** Evaluating spatial accuracy against the model's prediction and the actual ground truth.

Quantitative Evaluation Metrics:

To move beyond visual "sanity checks," we utilize pixel-level metrics:

- **Pixel Classification:**
  - **True Positives (TP):** Correctly identified important pixels.
  - **False Positives (FP):** Irrelevant pixels highlighted as important (Noise).
  - **False Negatives (FN):** Important pixels missed by the explanation.

- **Derived Metrics:**
  - **Precision:** reflects the ability to **pinpoint relevant features** (avoiding false positives)
  - **Recall:** signifies the **completeness** in capturing all crucial features (avoiding false negatives)
  - **IoU (Intersection over Union):** Spatial overlap accuracy.

$$IoU = \frac{TP}{TP + FP + FN}$$

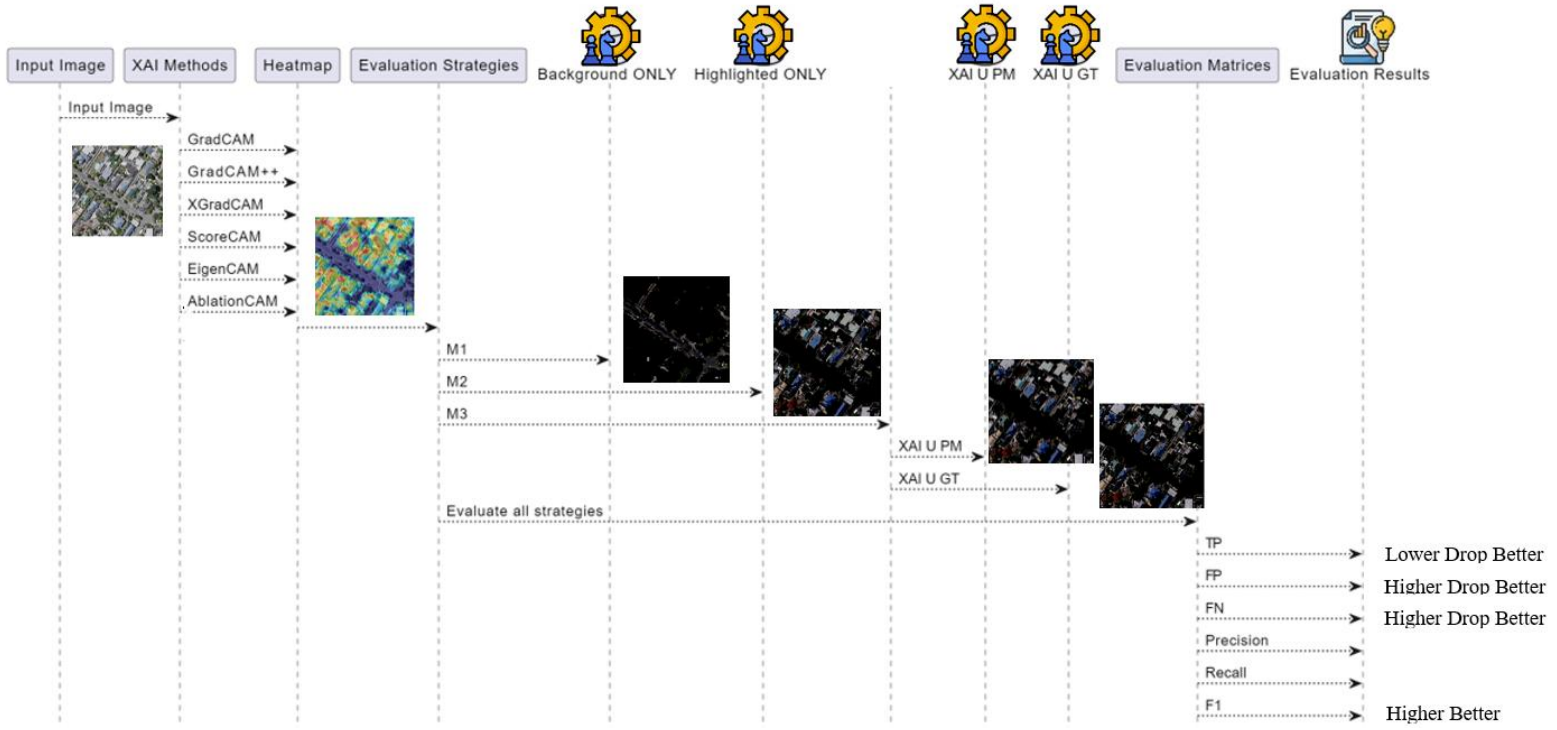
Experimental Setup:

- **Dataset:** WHU Building Dataset (High-quality building extraction for remote sensing).
- **Task:** Semantic Segmentation of buildings.

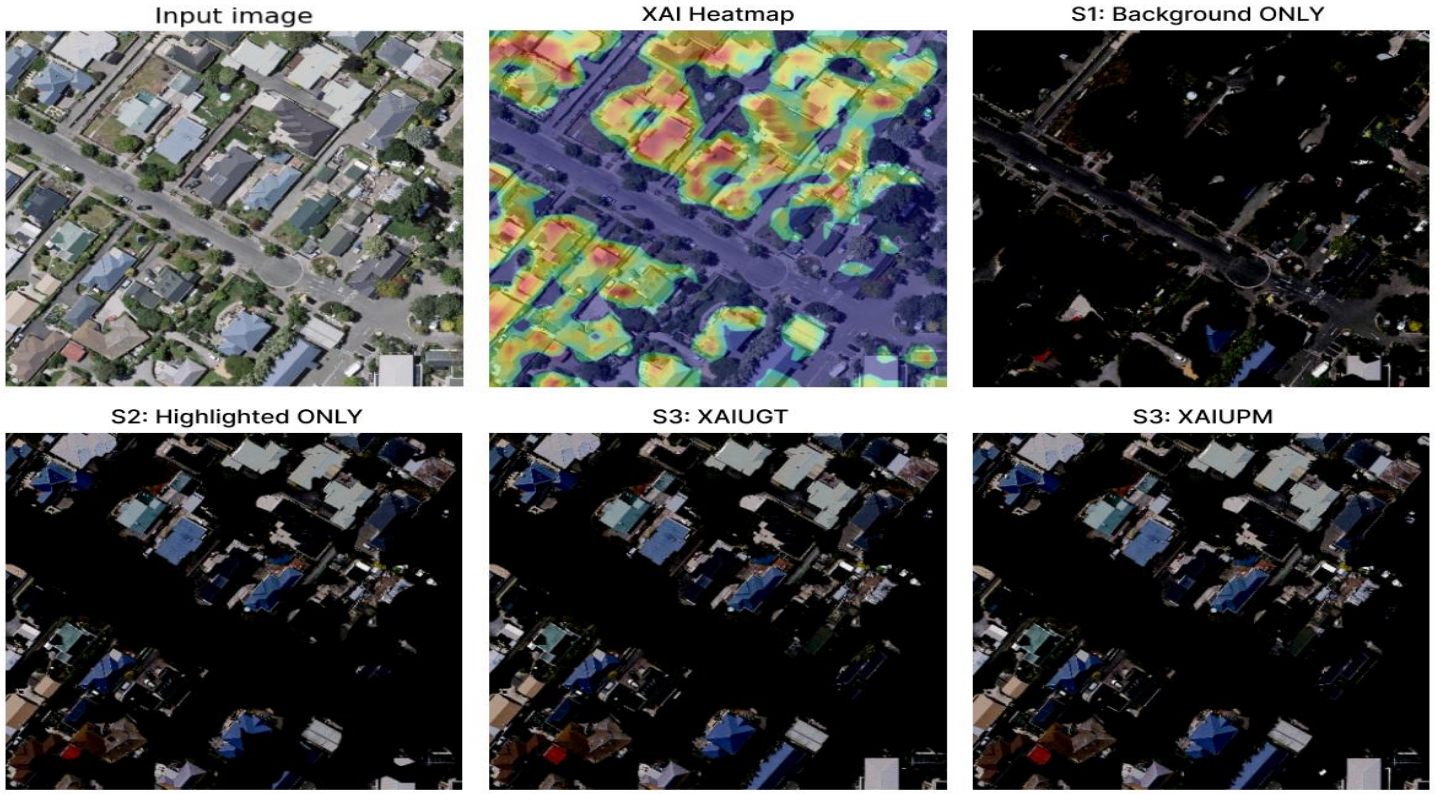
- **Evaluated XAI Methods:**
  - Grad-CAM, Grad-CAM++, XGrad-CAM (Gradient-based)
  - Score-CAM (Perturbation-based)
  - Eigen-CAM, Ablation-CAM.

- **Environment:** CUDA 12 GPU-powered environment.

Framework:


Input Image
XAI Methods
Heatmap
Evaluation Strategies
Background ONLY
Highlighted ONLY
XAI U PM
XAI U GT
Evaluation Matrices
Evaluation Results
Input Image
GradCAM
GradCAM++
XGradCAM
ScoreCAM
EigenCAM
AblationCAM
M1
M2
M3
XAI U PM
XAI U GT
Evaluate all strategies
TP
FP
FN
Precision
Recall
F1
Lower Drop Better
Higher Drop Better
Higher Drop Better
Higher Better

Visual Results:



Quantitative Results:

Threshold	XAI/Matrix	Model	Grad-CAM	Grad-CAM++	XGrad-CAM	Score-CAM	Eigen-CAM	Ablation-CAM
0.4	TP Pixels	49431	45259	39680	45238	25906	46990	45228
	TP Pixels (%)	93.58	85.68	75.12	85.64	49.05	88.96	85.62
	Drop %(**Higher better**)		7.90	18.46	7.94	**44.53**	4.62	7.96
	FP Pixels	2886	2806	4275	2847	3846	5622	3095
	FP Pixels (%)	1.38	1.34	2.04	1.36	1.84	2.69	1.48
	Increase % (**Higher better**)		-0.04	0.66	-0.02	0.46	**1.31**	0.10
	FN Pixels	3390	7562	13141	7583	26915	5832	7593
	FN Pixels (%)	6.42	14.32	24.88	14.36	50.95	11.04	14.38
	Increase % (**Lower better**)		7.90	18.46	7.94	**44.53**	4.62	7.96

Table 1: Pixel-level performance evaluation of S1.

Thresholds	Method	Model	Grad-CAM	Grad-CAM++	XGrad-CAM	Score-CAM	Eigen-CAM	Ablation-CAM
0.4	IoU (Micro)	0.89	0.81	0.69	0.81	**0.46**	0.80	0.81
	Precision	0.94	0.94	0.90	0.94	**0.87**	0.89	0.94
	Recall	0.94	0.86	0.75	0.86	**0.49**	0.89	0.86
	F1	0.94	0.90	0.82	0.90	**0.63**	0.89	0.89

Table 2: Pixel-level performance evaluation of S1. Lower metric value signifies better XAI scheme and vice versa.

Quantitative Results:

Threshold	XAI/Matrix	Model	Grad-CAM	Grad-CAM++	XGrad-CAM	Score-CAM	Eigen-CAM	Ablation-CAM
0.4	TP pixels	49431	28948	25871	27938	39505	25067	27003
	TP Pixels (%)	93.58	54.80	48.98	52.89	74.79	47.46	51.12
	Drop %(**Lower better**)		38.78	44.6	40.69	**18.79**	46.12	42.46
	FP pixels	2886	152690	123876	148112	95918	84419	141878
	FP pixels (%)	1.38	72.94	59.18	70.76	45.82	40.33	67.78
	Increase % (**Lower better**)		71.56	57.80	69.38	44.44	**41.71**	66.40
	FN pixels	3390	23873	26950	24883	13316	27754	25818
	FN pixels (%)	6.42	45.20	51.02	47.11	25.21	52.54	48.88
	Increase % (**Lower better**)		38.78	44.60	40.69	**18.79**	58.96	42.46

Table 3: Pixel-level performance evaluation of S2.

Threshold	Method	Model	Grad-CAM	Grad-CAM++	XGrad-CAM	Score-CAM	Eigen-CAM	Ablation-CAM
0.4	IoU (Micro)	0.89	0.14	0.15	0.14	**0.27**	0.18	0.14
	Precision	0.94	0.16	0.17	0.16	**0.29**	0.23	0.16
	Recall	0.94	0.55	0.49	0.53	**0.75**	0.47	0.51
	F1	0.94	0.25	0.26	0.24	**0.42**	0.31	0.24

Table 4: Pixel-level performance evaluation of S2. Higher metric value signifies a better XAI scheme and vice versa.

Quantitative Results:

Threshold	XAI/Method	Model	Grad-CAM	Grad-CAM++	XGrad-CAM	Score-CAM	Eigen-CAM	Ablation-CAM
0.4	TP pixels	49431	29475	31745	29512	45253	32331	29913
	TP Pixels (%)	93.58	55.80	60.10	55.87	85.67	61.21	56.63
	Drop %(**Lower better**)		37.78	33.48	37.71	**7.91**	32.37	36.95
	FP pixels	2886	106346	102591	105874	77490	59245	105573
	FP pixels (%)	1.38	50.80	49.01	50.58	37.02	28.30	50.44
	Increase % (**Lower better**)		49.42	48.04	49.2	35.64	**26.92**	49.06
	FN pixels	3390	23346	21076	23309	7569	20490	22908
	FN pixels (%)	6.42	44.20	39.90	44.13	14.33	38.79	43.37
	Increase % (**Lower better**)		37.78	33.48	37.71	**7.91**	32.37	36.95

Table 7: Pixel-level performance evaluation of S3: XAI-PM.

Threshold	Method	Model	Grad-CAM	Grad-CAM++	XGrad-CAM	Score-CAM	Eigen-CAM	Ablation-CAM
0.4	IoU (Micro)	0.89	0.19	0.20	0.19	**0.35**	0.29	0.19
	Precision	0.94	0.22	0.24	0.22	**0.37**	0.35	0.22
	Recall	0.94	0.56	0.60	0.56	**0.86**	0.61	0.57
	F1	0.94	0.31	0.34	0.31	**0.52**	0.45	0.32

Table 8: Pixel-level performance evaluation of S3: XAI-PM. Higher metric value signifies a better XAI scheme and vice versa.

Key Findings:

1- The framework provides a well-rounded way to assess XAI methods, moving beyond qualitative visual inspection.

2- In semantic segmentation, **contextual pixels** around the target object play a fundamental and significant role in the model's performance.

3- **Score-CAM** stood out as the most effective and reliable XAI method, consistently delivering high true positive rates, low false positives, and strong precision and IoU scores across all strategies.

Conclusion:

•**Contribution**: Our work offers a novel, rigorous quantitative evaluation framework that is essential for developing transparent, trustworthy, and accountable semantic segmentation models.

•**Impact:** The framework ensures AI transparency and trust in safety-critical domains like geospatial analysis and remote sensing.

•**Future Work:** This evaluation framework is not limited to remote sensing; its generic methodology is built to address the **fidelity** and **robustness** of explanations and can be applied to a wide range of applications, such as medical imaging.
We believe this work advances the development of reliable XAI for any critical application requiring trustworthy AI-based semantic segmentation.