SEMANTIC SEGMENTATION XAI EVALUATION

Project Overview

Motivation:

Deep learning models for remote sensing (RS) image segmentation (like building detection) offer high accuracy but act as "black boxes"—obscuring how predictions are made. This limits trust, transparency, and scientific validation in critical RS applications.

Goal:

To improve **explainability** in RS segmentation tasks by:

1. **Adapting CAM-based XAI methods** (originally developed for image classification) to semantic segmentation.
2. **Proposing evaluation metrics (M1, M2, M3)** to quantify model uncertainty.

Key Contributions

1. Methodological Innovations

- Adapted five CAM-based XAI methods to segmentation:
  - Seg-Grad-CAM
  - Seg-Grad-CAM++
  - Seg-XGrad-CAM
  - Seg-Score-CAM
  - Seg-Eigen-CAM
- Introduced a new **entropy-based evaluation framework** (M3), improving over standard metrics like drop in confidence (M1, M2).

2. Evaluation Strategies

The project defines and uses three evaluation modes:

- **M1 (Background Only):** Tests how much performance drops when important regions are removed.

- **M2 (Highlighted Only):** Tests how well the model performs using only the "important" parts of the image.
- **M3 (Highlighted + Target):** Uses entropy to measure how confident the model is when seeing the explanation region and the ground truth together.

3. Dataset & Model

- Model: U-Net based segmentation model.
- Dataset: WHU high-resolution satellite images for building rooftop segmentation.

Task 1: Analyze the Impact of Heatmap Threshold on Evaluation (M1, M2, M3

Objective

The aim of Task 1 is to understand how the **thresholding of CAM-generated heatmaps** influences the quality and reliability of explanations in a **semantic segmentation context**. CAM methods generate continuous-valued saliency maps, which must be binarized to isolate regions considered important by the model. This task systematically evaluates how different **threshold levels** (ranging from coarse to strict) affect the model's response when these masked inputs are reintroduced.

To quantify the impact of these thresholds, three evaluation strategies—**M1, M2, and M3**—are used:

METHOD	DESCRIPTION	GOAL
M1	Background-only input	High confidence drop = good explanation
M2	Highlighted-region-only input	Low confidence drop = good explanation
M3	Highlighted + Target-class input	Low entropy = low uncertainty = better explanation

This task seeks to determine the **optimal threshold range** for generating binary masks from heatmaps and assess which CAM methods remain **consistent and meaningful** across different thresholds.

Threshold Effects – Best Method For Each Threshold Value

M1 – Background Only

Threshold	Method	Confidence	Entropy
0.1	grad_cam_pp	0.0902	0.00108
0.2	score_cam	0.0324	0.00029
0.3	score_cam	0.0399	0.00066
0.4	score_cam	0.1970	0.00165
0.5	score_cam	0.5355	0.00297

Interpretation:

- As the **threshold increases**, **confidence increases** and **entropy rises**, indicating that **larger portions of meaningful input** are removed.
- At **low thresholds (0.1–0.3)**, Score-CAM suppresses confidence effectively (e.g., 0.032 at 0.2), implying strong explanation.
- However, at **0.5**, confidence is much higher (0.535), showing that not enough critical regions were masked — the explanation becomes less effective.
- **Best-performing point**: Score-CAM at threshold 0.2 — lowest confidence (0.032) and lowest entropy (0.00029).

M2 – Highlighted Only

Threshold	Method	Confidence	Entropy
0.1	score_cam	0.9044	0.00142
0.2	score_cam	0.8985	0.00135
0.3	score_cam	0.8693	0.00145
0.4	grad_cam	0.7251	0.00261

Threshold	Method	Confidence	Entropy
0.5	grad_cam	0.7682	0.00297

Interpretation:

- For **Score-CAM**, even at low thresholds (0.1–0.3), confidence remains **high (~0.9)** — this means the retained highlighted regions are indeed relevant.
- As we move to **Grad-CAM (0.4, 0.5)**, performance degrades (confidence ~0.72–0.77), indicating **less precise explanations**.
- **Best-performing configuration**: Score-CAM at 0.1 — retains high model confidence (0.904) with low entropy (0.00142).

M3 – Highlighted + Target

Threshold	Method	Confidence	Entropy
0.1	grad_cam_pp	0.9017	0.00139
0.2	score_cam	0.9050	0.00138
0.3	grad_cam	0.0351	0.00118
0.4	x_grad_cam	0.0337	0.00112
0.5	grad_cam_pp	0.0335	0.00112

Interpretation:

- **0.1–0.2 thresholds** (Score-CAM, Grad-CAM++) preserve **high confidence and low entropy** — the model is confident when given the explanation + GT.
- At higher thresholds (0.3–0.5), explanations become **too sparse**, dropping confidence **sharply** to near-zero (~0.03), even though entropy remains stable.
- **Best-performing configuration**: Score-CAM at 0.2 — highest confidence (0.905) with low entropy (0.00138), indicating the most effective focus on relevant regions.

Conclusion: **Score-CAM** shows the most consistent and effective performance across all metrics and thresholds.

Interpretation

Task 1 reveals that **heatmap thresholding is not a trivial preprocessing step**, but a critical factor that directly affects the quality and utility of visual explanations in segmentation models.

Through the M1, M2, and M3 evaluation framework, the task demonstrates that:

- **Faithful explanations** result in measurable and interpretable model behavior changes when their highlighted regions are masked or emphasized.
- **Robust CAM methods** produce saliency maps that remain meaningful across a range of threshold values, making them more reliable for real-world interpretation.
- **Entropy-based evaluation (M3)** provides deeper insight into the model's uncertainty and confidence, especially in spatially correlated tasks like segmentation—where masking a single region can influence predictions elsewhere in the image.

Task 2: Evaluating Explanation Quality Using Drop in Segmentation Performance

---

Objective

The goal of Task 2 is to assess the **faithfulness** of different CAM-based explainability methods by measuring how **critical their highlighted regions are to the model's segmentation output**. Rather than relying solely on visual interpretation or probability drop (as in classification), this task focuses on a concrete, **performance-based metric**: the **drop in segmentation accuracy**, typically measured by Intersection over Union (IoU), when input images are masked using CAM-generated saliency maps.

The hypothesis is:

> If a CAM method correctly identifies the regions most important for the model's prediction, then **removing those regions from the input should lead to a significant drop in segmentation performance**.

This performance degradation is treated as **evidence of explanation quality**—a faithful explanation should highlight truly influential regions.

Results (Faithfulness Ranking by Drop in IoU)

Rank	XAI Method	IoU Drop	Interpretation
1	**Score-CAM**	**−0.3058**	Highest drop → explanation removed most crucial regions
2	Grad-CAM++	−0.2407	Very good alignment with model focus
3	Grad-CAM	−0.2386	Comparable to Grad-CAM++
3	XGrad-CAM	−0.2386	Tied with Grad-CAM in effectiveness
5	**Eigen-CAM**	**+0.0273**	Slight **improvement** in performance when explanation was removed → indicates poor faithfulness

---

Interpretation

Task 2 provides a **practical, model-centered view of explanation quality** by moving beyond visual saliency into measurable consequences by proving if removing explanation actually break the model.

This is powerful, because it:

- **Validates explanation relevance based on model performance**, not visual appearance.
- Helps distinguish between **truly causal features** and coincidental correlations.
- Gives decision-makers a clear, quantitative way to compare XAI methods based on how **disruptive** their explanations are to the model's output.

The task also highlights that **explanation methods should not be judged by visual sharpness alone**—some methods may generate sharp-looking maps that have little effect on model performance when perturbed. Using segmentation IoU drop as a metric ensures that **only functionally meaningful explanations are rewarded**.

Task 3: Evaluating CAM Explanations Using the 2nd Decoder Block

---

Objective

The objective of Task 3 is to investigate **how the depth of the decoder layer** used for CAM generation affects the quality and informativeness of explanations in semantic segmentation. More specifically, this task re-applies the evaluation pipeline of **Task 1 (threshold vs. M1, M2, M3)** and **Task 2 (IoU drop-based evaluation)**, but this time using **CAM heatmaps derived from the 2nd decoder block** of the segmentation network (rather than the final decoder layer).

This task tests the hypothesis:

> Can intermediate decoder layers provide more robust or meaningful explanations than final-layer activations?

It explores whether explanations generated earlier in the decoding hierarchy retain spatial and semantic detail that might improve XAI outcomes, particularly in challenging segmentation cases where over-summarization at deeper layers may hinder interpretability.

Repeating Task 1 & Comparing results between Decoder 1 and 2

Grad-CAM

- **Interpretation**: Clear and consistent improvement across all metrics. Decoder 2 enhances Grad-CAM's ability to both suppress irrelevant regions (M1) and reduce prediction uncertainty (M3), making explanations sharper and more faithful.

Grad-CAM++

- **Interpretation**: Performs better with Decoder 2 in nearly all aspects. Gains in M1 and M3 suggest stronger focus and better certainty. Slightly less consistent in M2, but overall, explanations are more precise and meaningful.

XGrad-CAM

- **Interpretation**: One of the most robust improvements. All three metrics benefit significantly from deeper decoder features, especially in reducing entropy and filtering background noise. A reliable choice when using intermediate layers.

Score-CAM

- **Interpretation**: Mixed results. While M2 and M3 improve (i.e., better confidence and certainty), M1 worsens in most cases — suggesting Decoder 2 causes Score-CAM to highlight less critical or overly broad regions. Best used with caution on deeper layers.

Eigen-CAM

- **Interpretation**: Shows improvement in background suppression and uncertainty (M1, M3), but still fails to consistently retain target-relevant information (M2). Decoder 2 helps, but Eigen-CAM remains the least faithful overall.

Repeating Task 2 & Comparing Results

IoU drop has increased (worsened) in Task 3 compared to Task 2 which means that the model's segmentation performance got worse when explanations from **Decoder 2** were removed, compared to Decoder 1 in Task 2.

Interpretation

Task 3 underscores that **layer choice is a critical design decision** in the explainability pipeline of segmentation models.

Key takeaways include:

- **Deeper decoder layers** (closer to output) offer more **class-discriminative and compact explanations**, as they summarize semantically relevant content.
- **Intermediate decoder layers** provide explanations that are **spatially richer but semantically less precise**, which can be useful in applications requiring **boundary-level detail**.
- The **optimal CAM layer is method-dependent**: some CAM variants thrive on low-level spatial features (e.g., Eigen-CAM), while others benefit from high-level abstraction (e.g., Grad-CAM++).

Ultimately, Task 3 emphasizes the value of **layer-wise analysis** in explainability research. It reveals that a single-layer CAM perspective may miss important trade-offs between **semantic precision and spatial fidelity**, and that **multi-layer or hybrid CAM strategies** may offer improved explanatory power.