ENTROPY-CENTRIC EXPLAINABLE AI FOR REMOTE SENSING IMAGE SEGMENTATION

Ali Saleh*, Abdul Karim Gizzini†, Nadine Alameh‡, Ali J. Ghandour§
* Faculty of Engineering, Lebanese University, Beirut Lebanon.
* University of Paris-Est Créteil (UPEC), LISSI/TincNET, F-94400, Vitry-sur-Seine, France.
‡ LunateAI, Wasghinton DC, USA.
§ National Center for Remote Sensing - CNRS, Mansourieh, 22411, Lebanon.
Corresponding Author: aghandour@cnrs.edu.lb

Abstract—**Artificial intelligence (AI) has become a powerful approach to solving complex problems in critical domains. Many concerns arise regarding the decision-making process of its models, which is mainly due to deep neural networks outperforming their peers at the cost of ambiguity in feature extraction and prediction. Consequently, in critical domains like remote sensing, where we need to analyze high-resolution imagery using black-box models, the lack of transparency limits their trust and thus their adoption. In front of this reality, explaining and understanding the complex AI model's decision-making process becomes a must. Explainable AI (XAI) aims to bridge this gap by providing insights into how and why certain decisions are made. While significant progress has been achieved in explaining image classification tasks, image segmentation still offers considerable room for improvement. In this context, this paper proposes an entropy-centric XAI method for semantic segmentation. Moreover, a new XAI evaluation methodology is proposed to efficiently measure the relevance of the highlighted regions by the proposed XAI method. Experimental results demonstrate the superiority of the proposed XAI method in comparison to the recently adapted XAI methods for semantic segmentation.**

Index Terms—**Explainable AI, Remote Sensing, Image Segmenation, Entropy-Centric.**

I. Introduction

Artificial Intelligence (AI) has achieved promising success in various domains, including remote sensing, where it is employed in critical applications, such as land use monitoring [1], environmental assessment [2, 3], and disaster management [4]. AI-based solutions introduced automated feature extraction compared to traditional machine learning schemes. Feature extraction is one of the most complex and time-consuming phases in the training process for any model. The lack of transparency in extracting the desired features results in losing the interpretability of the decision-making methodology employed by the corresponding model. This raises concerns that impact human trust in AI, including its fairness and the ethical aspects of its decisions. In this context, the emergence of Explainable Artificial Intelligence (XAI), a field dedicated to uncovering the reasoning behind model predictions, become a must. Hence, providing explanations that allow humans to understand, validate, and trust AI-driven decisions.

XAI schemes can be applied to different applications, including image classification and semantic segmentation. Both of these applications can benefit from XAI in revealing the key image features that mostly influenced the model prediction, enabling more interpretability and thus responsible use of AI in remote sensing. Despite the XAI success achieved in image classification, extending this success to image segmentation is challenging due to the spatial correlation between the pixels and regions.

Among different XAI methods, Attribution methods focus on assigning importance scores to the input features, indicating their contribution to a model prediction. Thus, understanding what features are affecting its decisions[5]. XAI attribution methods are further divided into different classes: (*i*) Gradient-based, where the importance score is computed based on the gradients of backpropagation, (*ii*) Sampling-based methods that treat the model as a black box and systematically sample parts of the input and observe output changes to estimate feature importance. It is worth mentioning that evaluating the robustness and efficiency of different XAI methods is essential for validating their faithfulness and reliability [6].

In this context, this work focuses on proposing a sampling-based XAI method based on the entropy uncertainty principle. It consists of 2 sampling phases: Creation of Base sampling matrices then the perturbation one for input image perturbation, followed by a model inference to compute the target object entropy spatial scores to determine the importance scores for each reagion in the image. Moreover, to efficiently evaluate the performance of the proposed Entropy-Centric XAI method, we propose a new XAI evaluation methodology that primarily assesses whether highlighted relevant regions outside the target object are truly influential. The performance evaluation performed using the WHU dataset for building footprint segmentation [7] shows the superiority of our proposed entropy-centric method compared to the recent XAI methods adapted for semantic segmentation [8].

The remainder of this paper is organized as follows: Section II presents the proposed entropy-centric XAI methodology. The performance of the benchmarked XAI methods in terms of the proposed XAI evaluation methodology is analyzed and discussed in Sections III and IV, respectively.

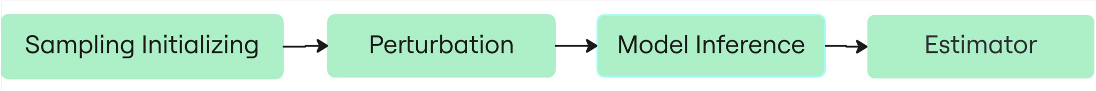

Fig. 1. Method Pipeline

II. Methodology

This section highlights the main principles of the proposed entropy-centric XAI method, as well as the proposed XAI evaluation methodology. First of all, we breifly presents the XAI Sobol method, which motivates our proposed method.

A. Conventional Sobol Method

Sobol indices [9] is a variance-based global sensitivity method that measures the contribution of input variables to the output variance of a model. For a model output $\mathbf{f}(\mathbf{x})$ with input features $\mathbf{x} = \{\mathbf{x_1}, \dots, \mathbf{x_n}\}$, the total variance D is decomposed into main and interaction effects:

$$D = \sum_{i=1}^{n} D_i + \sum_{i<j} D_{ij} + \cdots + D_{1,2,\dots,n}, \tag{1}$$

where D_i denotes the variance contribution of input i and higher-order terms represent interactions. The total Sobol index quantifying the importance of input i, including all its interactions, is defined as:

$$S_{T_i} = 1 - \frac{\mathrm{Var}_{x_{\sim i}}[\mathrm{E}_{x_i}[f(x \mid x_{\sim i})]]}{D}, \tag{2}$$

Sobol indices are estimated by systematically perturbing inputs using quasi-Monte Carlo (QMC) sampling [9], which improves coverage of the input space and convergence. To reduce the high computational cost of estimating total-order indices in high-dimensional settings, the Jansen estimator [9] is employed. It computes $\hat{S}_{T_i}$ using paired sampling matrices A, B, and C_i as:

$$\hat{S}_{T_i} = \frac{\frac{1}{2N} \sum_{j=1}^{N} (f(\mathbf{A}_j) - f(\mathbf{C}_{i,j}))^2}{\hat{V}}, \tag{3}$$

Thus, enabling efficient estimation with fewer model evaluations.

B. Proposed Entropy-Centric XAI Method

The proposed Entropy-Centric XAI method utilize Sobol XAI concept and is based on the entropy methodology that measures how the perturbation of different input patches affects the output entropy and hence identifies the regions that maintain confidence in the segmentation output. The greater the entropy change upon masking a region, the higher its attribution score. The binary entropy $H \in [0,1]^{H \times W}$ of the predicted target class probability p_k at each pixel can be expressed as:

$$H = -p_k \log(p_k) - (1 - p_k) \log(1 - p_k) \tag{4}$$

Through Entropy-Centric we adapt Sobol to image segmentation and continue with developing the entropy methodology. Adapting to image segmentation necessitates a fundamental shift from explaining scalar output values, as in classification, to handling spatially structured outputs. In segmentation tasks, the primary objective is to interpret the model's predictions at the pixel or region level for specific target objects, rather than for the image as a whole.

To achieve this, each perturbed image forward pass of the segmentation model yields a spatial score map $S_c \in \mathbb{R}^{H \times W}$, essentially a probability distribution indicating the likelihood of the object class at every pixel. The output map corresponding to the target object is isolated by multiplying the model's output by the object's mask $M \in [0,1]^{H \times W}$.

$$f_{masked}(S_c) = S_c \odot M \tag{5}$$

This allows for analyzing the contribution of input regions to the prediction of the relevant object, discarding spurious influences elsewhere in the image.

Figure 1 illustrates the Entropy-Centric pipeline. It begins by initializing the base sampling matrices A and B using quasi-Monte Carlo (QMC) methods [9], followed by generating the perturbation masks C_i based on these matrices. The perturbed images $A_j^{'}$ and $C_{i,j}^{'}$ are subsequently created by applying the upsampled A_j and $C_{i,j}$ masks according to eq. 6.

$$A_j^{'} = I \odot \mathrm{up}(A_j) \qquad C_{i,j}^{'} = I \odot \mathrm{up}(C_{i,j}), \tag{6}$$

where $I \in \mathbb{R}^{H \times W}$ denotes the input image with dimesniosn H and W. Next, model inference is performed to obtain the spatial score maps of the perturbed inputs $S_{A_j^{'}}$ and $S_{C_{i,j}^{'}}$. Then target masking them using eq. 5 to get $f_{masked}(S_{A_j^{'}})$ and $f_{masked}(S_{C_{i,j}^{'}})$. The Softmax probabilities are then computed as:

$$\begin{cases} P_{A_j} = Softmax(f_{masked}(S_{A_j^{'}})) \\ P_{C_{i,j}} = Softmax(f_{masked}(S_{C_{i,j}^{'}})) \end{cases} \tag{7}$$

Finally, the importance scores are calculated, after calculating the entropy of the target class from its softmax probabilities through eq.4, using :

$$S_i = \frac{1}{N} \sum_{j=1}^{N} \left| H\left(P_{C_{i,j}}\right) - H\left(P_{A_j}\right) \right| \tag{8}$$

where $H(C_j)$ and $H(A_j)$ denote the entropy values computed for perturbed and base samples, respectively. We note that the proposed Entropy-Centric XAI method focuses on the uncertainty rather than only the output score, which offers a different explainability perspective.

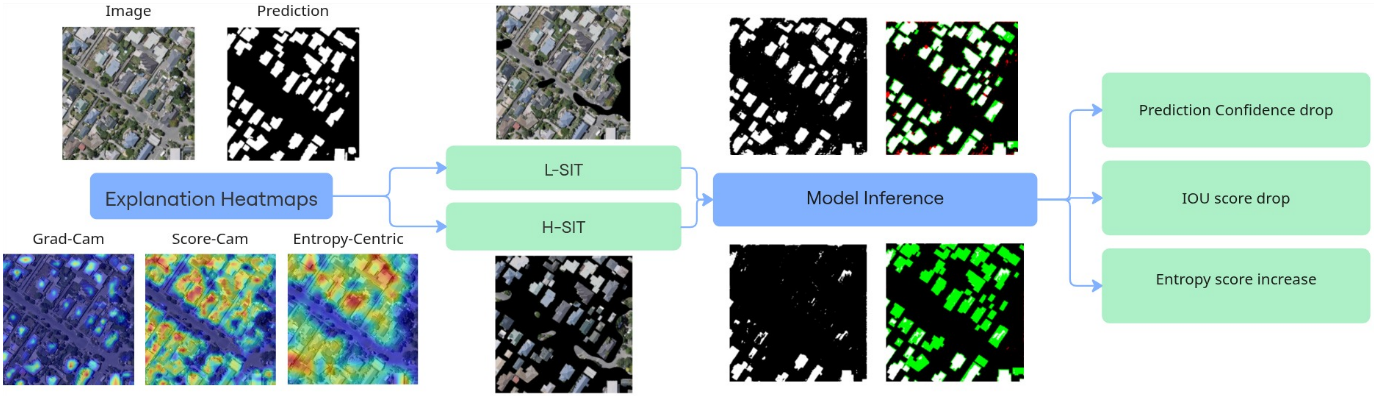

Fig. 2. Sample of XAI heatmaps generated by the 3 methods Grad-Cam,Score-Cam and Entropy-Centric with the evaluation framework proposed through Irrelative Validation and False Sailency Detection methods. The masked images, prediction and IOU maps are based on the Entropy-Centric heatmap.

C. Proposed Evaluation Methodology

XAI performance evaluation remains challenging due to the absence of a unified evaluation framework, making fair comparison between XAI methods difficult. The authors in [8] proposed an XAI evaluation methodology we refer to as Low-Salience Irrelevance Test (L-SIT) that evaluates whether regions assigned low importance in the explanation are indeed irrelevant to the model. L-SIT removes pixels with explanation values below a chosen threshold and measures the resulting change in prediction confidence and segmentation accuracy. Given the explanation heatmap E with ϕ the set of highlighted pixels above a specific threshold, and T the set of target object pixels of the input image I. The perturbed image by L-SIT can be expressed as:

$$I'_{\text{L-SIT}} = T \cup \phi. \tag{9}$$

A small performance drop confirms the irrelevance of low-salience regions outside the target object, while a large drop reveals their hidden importance. However, L-SIT cannot detect cases where unimportant regions are mistakenly given high salience, which motivates the need for a complementary evaluation strategy.

To address this limitation, we propose a new XAI evaluation methodology, denoted as High-Salience Influence Test (H-SIT), which primarily assesses whether regions outside the target object but highlighted as important are truly influential. By removing these out-of-object, high-saliency regions, the resulting performance deterioration directly reflects their actual relevance to the model's decision process. The perturbed image by H-SIT is defined as:

$$I'_{\text{H-SIT}} = T \cup \overline{\phi}. \tag{10}$$

where $\overline{\phi}$ is the set of highlighted pixels below a specific threshold. Through the proposed H-SIT XAI evaluation methodology, a large drop in the considered metric is expected since the regions most affecting the model decision are removed, enabling the identification of falsely highlighted regions that L-SIT methodology alone cannot capture.

III. Results

This section presents the performance evaluation of the proposed Entropy-Centric XAI method, Grad-Cam, and Score-Cam XAI methods [8]. The XAI evaluation is performed using the L-SIT and the proposed H-SIT XAI evaluation methodology. Moreover, the rooftop U-Net architecture segmentation model [7] trained using the WHU dataset is used as a case study.

Figure 2 shows a sample of XAI heatmaps generated by the three methods Grad-Cam, Score-Cam, and the proposed entropy-centric, out of 142 images. Unlike the Grad-CAM method, which highlights some in-target class pixels. We can notice that Entropy-Centric and Score-Cam gave the better heatmaps by showing that the majority of the highly highlighted areas consist of buildings and their surroundings. Clearly giving us insight that the most pixels that affect the model segmentation results are the building pixels.

To further quantitatively evaluate the reliability of the benchmarked XAI schemes, L-SIT and the proposed H-SIT are used in terms of the following metrics: (*i*) The drop in the model's prediction confidence inside the target object area compared to the original one when I is passed to the model, (*ii*) The drop in the Intersection over Union (IoU) score between the perturbed image prediction and the ground truth compared to the IoU score of the original prediction with the ground truth. IoU tells us how much the predicted object overlaps with the correct object. (*iii*) The increase in the Entropy of the target object between the masked image prediction and the original prediction. Table I shows the quantitative results for the benchmarked XAI methods where the threshold is 0.1.

For the L-SIT methodology, the low-importance regions defined by the heatmaps below threshold 0.1 are masked out. As expected, this results in a reduction in prediction confidence and IoU scores, and a rise in entropy, regardless of the method used. Notably, the proposed Entropy-Centric outperforms the Grad-CAM and the Score-Cam methods. This indicates that the Entropy-Centric heatmap more precisely identifies irrelevant areas, as their removal barely impacts

TABLE I

L-SIT and H-SIT quantitative results for the three metrics: (i) Mean of the Prediction Confidince Score Drop, (ii) Mean of the IOU Score Drop, (iii) and Mean of the Entropy Score Increase at 0.1 threshold over the WHU dataset.

	Prediction Confidence		IOU Score		Entropy Score	
XAI Methods	L-SIT (↓ better)	H-SIT (↑ better)	L-SIT (↓ better)	H-SIT (↑ better)	L-SIT (↓ better)	H-SIT (↑ better)
Entropy-Centric	2%	37.4%	0.7%	44%	3.4%	48.5%
Grad-Cam	27.3%	2.2%	29.7%	-0.1%	38.8%	5%
Score-Cam	0.6%	32.3%	4.1%	36.2%	8.9%	45.3%

TABLE II

Comparative Analysis: Gradient-Based vs. Sampling-Based XAI for Segmentation

XAI Schemes	Performance	Stability	Background Reliance	Interaction Awareness	Methodology	Granularity	Primary Objective
Score-CAM	+++	Low	Weak	Limited	Model-Specific	Target-based	Fast Visualization
Entropy-Centric	++++	High	Strong	High	Model-Agnostic	Patch-based	Model Auditing

model predictions, segmentation accuracy, or confidence. In contrast, Grad-Cam suffers from a considerable performance degradation when low-quality regions are masked, likely due to less discriminative attribution scores that misclassify important pixels as low-impact.

Concerning the proposed H-SIT XAI evaluation methodology, which focuses on masking high-importance regions as determined by each method, reveals complementary insights. Here, the metrics result in bad performance, reflecting the fact of masking the high-importance regions. In this context, both the proposed Entropy-Centric and the Score-Cam methods consistently yield the largest drops in prediction confidence and IoU, and the greatest increases in entropy, with the superiority of the proposed Entropy-Centric method. This confirms the efficient ability of the proposed Entropy-Centric method to accurately isolate decision-critical regions, making the model highly sensitive to their removal. We note that the Grad-CAM method barely has any drop in its metrics, which demonstrates its heatmap's deficiency in giving any importance to the regions outside the building areas. This is not always true as segmentation highly relies on the interaction between image regions.

IV. Discussion and Conclusion

Besides the quantitative performance superiority of the proposed Entropy-Centric over the Score-CAM method, Table II summarizes how both methods differ fundamentally in their computational demands, implementation requirements, and the nature of the explanations they provide for segmentation models. Score-Cam requires a small multiple of a standard forward pass, and produces fine-grained attribution maps whose spatial resolution is tied to the underlying convolutional feature maps. However, this method requires access to internal activation maps of the trained model, making it sensitive to architectural details. In addition, its implementation often relies on model-specific hooks and careful layer selection, which limits portability across architectures. In contrast, the proposed Entropy-Centric operates in a fully black-box manner and relies on systematic input perturbations, resulting in a substantially improved stability. Because it does not depend on gradients or internal representations, its explanations remain consistent across architectures and naturally extend to pipelines containing non-differentiable components. Although its explanations are patch-based and therefore limited by the chosen sampling grid, this coarser granularity enables the explicit analysis of contextual dependence and feature interactions by measuring confidence or entropy changes when regions are masked. Consequently, Score-Cam is well-suited for rapid, fine-resolution visualization, whereas the proposed Entropy-Centric method provides more robust and implementation-agnostic explanations that are better aligned with thorough model analysis and verification.

This work addresses the critical gap in explainable AI methods for image segmentation by proposing an Entropy-Centric XAI method. The proposed method calculates the final relevance scores based on the entropy uncertainty concept. In addition, we propose a robust XAI evaluation methodology denoted as H-SIT. Performance evaluation demonstrates the superior fidelity of the proposed Entropy-Centric method over the benchmarked methods. The proposed Entropy-Centric method is able to effectively isolate irrelevant regions as well as accurately identify decision-critical areas. While applied to the building segmentation use case, enabling trustworthy AI deployment in remote sensing applications, this work establishes a robust foundation for XAI methods for semantic segmentation. Future directions include multi-dataset validation, hybrid XAI methods, and diverse architectures support.

References

[1] V. Sangeetha and L. Agilandeeswari, "Artificial intelligence enabled spectral-spatial feature extraction techniques for land use and land cover classification

using hyperspectral images – an inclusive review," *The Egyptian Journal of Remote Sensing and Space Sciences*, vol. 28, no. 3, pp. 455–467, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110982325000390

[2] M. R. C. Santos and L. Cagica Carvalho, "Ai-driven participatory environmental management: Innovations, applications, and future prospects," *Journal of Environmental Management*, vol. 373, p. 123864, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301479724038519

[3] T. Miller, I. Durlik, E. Kostecka, P. Kozlovska, A. Łobodzińska, S. Sokołowska, and A. Nowy, "Integrating artificial intelligence agents with the internet of things for enhanced environmental monitoring: Applications in water quality and climate data," *Electronics*, vol. 14, no. 4, 2025. [Online]. Available: https://www.mdpi.com/2079-9292/14/4/696

[4] V. Eren and H. Duman, "Artificial intelligence support in disaster management," *Kamu Yönetimi ve Teknoloji Dergisi*, vol. 7, no. 1, p. 13–36, 2025.

[5] K. Abhishek and D. Kamath, "Attribution-based xai methods in computer vision: A review," 2022. [Online]. Available: https://arxiv.org/abs/2211.14736

[6] H. Shreim, A. Gizzini, and A. Ghandour, "Trainable noise model as an xai evaluation method: application on sobol for remote sensing image segmentation," *Environ. Sci. Proc*, vol. 1, no. 0, 2023.

[7] H. Nasrallah, A. E. Samhat, Y. Shi, X. X. Zhu, G. Faour, and A. J. Ghandour, "Lebanon solar rooftop potential assessment using buildings segmentation from aerial images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 4909–4918, 2022.

[8] A. Gizzini, M. Shukor, and A. Ghandour, "Extending cam-based xai methods for remote sensing imagery segmentation," *Environ. Sci. Proc*, vol. 1, no. 0, 2023.

[9] T. Fel, R. Cadène, M. Chalvidal, M. Cord, D. Vigouroux, and T. Serre, "Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.