Explainable AI for Pothole Detection: Comparing YOLOv9-tiny, YOLOv10-nano, and YOLOv11-nano through LayerCAM Visualization

Leonhel V. Fortin
Department of Computer Applications
College of Computer Studies
MSU-Iligan Institute of Technology
Tibanga, Iligan City, 9200, Philippines
leonhel.fortin@g.msuiit.edu.ph

Orven E. Llantos
Department of Computer Science
College of Computer Studies
MSU-Iligan Institute of Technology
Tibanga, Iligan City, 9200, Philippines
orven.llantos@g.msuiit.edu.ph

***Abstract*—Explainable AI (XAI) is critical for deploying automated systems in safety-focused applications like infrastructure monitoring. While lightweight YOLO architectures are known for high performance in pothole detection, understanding their decision-making process is key to trusting them in real-world road assessments. This study provides a comprehensive interpretability analysis of YOLOv9-tiny, YOLOv10-nano, and YOLOv11-nano using LayerCAM visualization. Experiments on a large dataset of road surface images reveal significant differences in the attention mechanisms across these compact architectures. Quantitative analysis shows that all models achieve excellent detection performance, with mAP@0.5 scores exceeding 0.980. However, LayerCAM analysis indicates that while newer models like YOLOv11-nano produce more spatially concentrated attention hotspots, the broader attention patterns of YOLOv9-tiny exhibit a superior alignment with pothole boundaries, achieving a mean CAM IoU of 0.674 compared to 0.160 for YOLOv10-nano and 0.082 for YOLOv11-nano. These findings highlight a trade-off between focused attention and holistic object representation, establishing crucial explainability benchmarks for deploying computer vision models in infrastructure management.**

***Keywords*-lightweight computer vision, YOLO, LayerCAM, model interpretability, pothole detection, real-time object detection, explainable AI**

I. Introduction

The maintenance of road infrastructure poses a persistent and costly challenge for municipalities worldwide, with direct implications for public safety and economic activity [1]. Traditional methods of road inspection rely on manual surveys, which are slow, labor-intensive, and often subjective. The advent of deep learning has enabled the development of automated systems for pothole detection, offering a scalable and efficient alternative [2]. Among the most successful architectures for this real-time object detection task is the You Only Look Once (YOLO) family, particularly its lightweight variants, which are optimized for deployment on resource-constrained edge devices such as those mounted on inspection vehicles or drones.

In safety-critical domains such as infrastructure management, high detection accuracy alone is insufficient. Automated systems must also demonstrate transparency and interpretability to gain the trust of civil engineers and public works officials. This principle is central to Explainable AI (XAI) [3]. Stakeholders require not only confirmation that a pothole has been detected but also insight into why the model identified a specific pavement feature as defective. Such transparency is essential for model validation, debugging, and user trust, ensuring that detection results can be reliably used to allocate maintenance resources.

Although the evolution of lightweight YOLO architectures has consistently improved the balance between speed and accuracy, the impact of these architectural optimizations on model interpretability remains largely unexplored. This study addresses this critical gap by performing a comparative analysis of the attention mechanisms in three state-of-the-art lightweight models: YOLOv9-tiny, YOLOv10-nano, and YOLOv11-nano. By applying LayerCAM, a gradient-based visualization technique [4], the research investigates how these models perceive and localize potholes.

The investigation reveals a previously undocumented trade-off between architectural efficiency and interpretability. As models evolve to become more computationally efficient, their attention strategies shift from broad, context-aware focus to highly concentrated, pointer-like patterns. While all models deliver excellent detection performance, the quality and type of their explanations diverge significantly. These findings challenge the assumption that progress in model development is linear and highlight the importance of explainability-aware model selection.

The primary contributions of this paper are:

- A rigorous comparative interpretability analysis of state-of-the-art lightweight YOLO models for pothole detection.
- The identification and quantification of a fundamental trade-off between a holistic attention mechanism (faithful to object shape) and a pointer-like one (optimized for

localization precision).

- The adaptation of LayerCAM for lightweight YOLO architectures, enabling stable, layer-wise attention visualization tailored for object detection tasks.
- A comprehensive set of quantitative benchmarks for both detection performance and interpretability, establishing a baseline for future research.
- An explainability-guided deployment strategy that enables practitioners to select the most appropriate model according to specific application requirements—whether comprehensive damage assessment or real-time anomaly flagging.

In summary, this study extends beyond evaluating detection accuracy by systematically analyzing the interpretability of lightweight YOLO models. Through the adaptation of LayerCAM for object detection and the introduction of new benchmarks, it uncovers a trade-off between holistic and pointer-like attention strategies. These insights not only provide the first comparative framework for understanding interpretability in compact YOLO architectures but also establish practical guidelines for explainability-driven model selection in infrastructure monitoring.

II. Related Work

A. Explainable AI in Computer Vision

Explainable AI (XAI) aims to demystify the "black box" nature of deep neural networks, making their decision-making processes transparent. In computer vision, a prominent family of XAI techniques revolves around Class Activation Maps (CAMs), which produce heatmaps that highlight the image regions most influential to a model's prediction. The original CAM required specific global average pooling layers, limiting its applicability [5]. This limitation was overcome by Grad-CAM [6], which uses the gradients of the target class score with respect to the feature maps of a convolutional layer. This gradient-based approach made visualization possible for a wide range of CNN-based architectures without requiring architectural changes.

Further refinements led to techniques like Grad-CAM++ [7], designed to provide better localization of multiple object instances. LayerCAM [4], the technique used in this study, represents a significant advancement. It produces more detailed and higher-quality activation maps by utilizing positive gradients from multiple layers, providing a more hierarchical and faithful representation of the model's internal attention. These visualization tools are indispensable for diagnosing model failures, verifying that a model is focusing on relevant features, and building trust in automated systems.

B. Lightweight YOLO Architecture Interpretability

The YOLO family of object detectors has continuously evolved to offer an optimal balance of speed and accuracy. This evolution has included the development of lightweight variants specifically designed for deployment on edge devices with limited computational power. Each iteration introduces novel architectural concepts to enhance efficiency while preserving performance.

YOLOv9-tiny introduced Programmable Gradient Information (PGI) [8]. PGI is a novel concept that helps manage the flow of information through the network, mitigating the information loss that can occur in very deep feed-forward networks and allowing the model to learn more discriminative features with fewer parameters.

YOLOv10-nano advanced the state of the art by focusing on end-to-end detection, most notably by creating an architecture that is free of Non-Maximum Suppression (NMS) [9]. It achieves this through a dual-label assignment strategy during training, which reduces post-processing latency and simplifies the deployment pipeline.

YOLOv11-nano represents the latest iteration, incorporating further architectural enhancements to refine feature extraction and boost performance within an extremely compact model [10]. These models push the boundaries of what is possible on resource-constrained hardware. However, despite detailed architectural descriptions from their authors, few studies have systematically investigated how these efficiency-driven innovations affect the models' internal attention mechanisms and, consequently, their interpretability.

C. Quantifying Interpretability

Qualitative assessment of heatmaps is subjective. To enable rigorous comparison, the XAI community has developed quantitative metrics to evaluate the quality of attention maps. The "Pointing Game" [22] assesses localization precision. It is considered a "hit" if the point of maximum activation within a heatmap falls inside the ground-truth bounding box of the object. A high Pointing Accuracy score indicates that the model's attention is sharply focused on a correct part of the object.

To measure how well the heatmap covers the *entire* object, the Intersection over Union (IoU) between a binarized attention map and the ground-truth box is commonly used [12]. A high CAM IoU score signifies that the explanation is faithful to the object's full shape and extent. These two metrics often represent an inherent trade-off: a highly focused, "pointer-like" heatmap may excel at the Pointing Game but score poorly on CAM IoU, while a more diffuse heatmap covering the entire object might do the opposite. Our work leverages these metrics to quantify this trade-off in YOLO models.

D. Summary

The reviewed literature demonstrates significant progress in both explainable AI techniques and the evolution of lightweight YOLO architectures. Methods such as CAM, Grad-CAM, and LayerCAM have enhanced transparency in computer vision models, while YOLO variants have steadily improved detection efficiency and accuracy. However, prior research has largely overlooked how these architectural advancements influence interpretability in practice. Existing studies often emphasize performance metrics but provide limited comparative analysis of attention mechanisms, few adaptations of visualization methods tailored to object detection, and no standardized benchmarks for evaluating interpretability across lightweight YOLO models. These gaps justify the present study's focus on systematically adapting LayerCAM to YOLO architectures, quantifying interpretability trade-offs, and establishing benchmarks that inform explainability-guided deployment. In this way, the related work directly motivates and supports the contributions outlined in Section I.

III. Methodology

This section outlines the dataset preparation, experimental setup, model selection, evaluation protocol, and interpretability metrics used in this study. The methodology was designed to ensure fairness, reproducibility, and rigor.

A. Dataset and Experimental Setup

The evaluation employed a comprehensive dataset of 4,000 high-resolution road surface images. The dataset captured diverse conditions, including daylight, dusk, and overcast skies, and featured a wide range of pothole types, from incipient cracks to severe wide-area pavement defects. Each image included precise ground-truth bounding box annotations, enabling reliable quantitative evaluation.

All experiments were conducted in a reproducible Kaggle Notebook environment, accelerated by dual NVIDIA T4 GPUs. The models were implemented in the PyTorch framework with a consistent input resolution of 640x640 pixels to ensure a fair and direct comparison across all architectures [13].

B. Models

The analysis utilized lightweight YOLO models, specifically YOLOv9-tiny [8], YOLOv10-nano [9], and YOLOv11-nano [10]. These models employed pre-trained weights from prior work, trained with optimized hyperparameters to ensure stable convergence and maximal detection performance [20]. Using pre-trained, high-performing models provided a robust baseline, allowing the study to focus squarely on interpretability rather than raw detection capability.

C. Evaluation Protocol

A custom Python script automated the evaluation process using PyTorch [13] and the Ultralytics library [14]. For each model, the script processed all 4,000 images in two steps: 1) executed the model's `predict` function to obtain pothole detections, and 2) generated a corresponding attention heatmap through a tailored LayerCAM implementation.

The LayerCAM method was specifically adapted for YOLO architectures. Class activation maps were generated from the **final three convolutional layers in the model's backbone**, immediately before the detection head. To enable class-agnostic gradient backpropagation, a pseudo-score was computed as the sum of all activations in the output feature maps. The CAMs from the three layers were then **averaged to form a single, stable attention map**, representing the model's high-level semantic focus.

Interpretability metrics were computed **only for true positive detections**, defined as predictions with IoU ≥ 0.5 against ground-truth bounding boxes. This deliberate restriction ensured that the analysis concentrated on understanding the decision-making process of correct model predictions.

D. Interpretability Metrics

The generated heatmaps were evaluated using three complementary metrics:

- **Mean CAM IoU**: Measures the spatial alignment between the model's attention and the full object shape [12]. The normalized heatmap was **binarized at a threshold of 0.3** to create a segmentation mask, then compared with the ground-truth bounding box using IoU.
- **Pointing Accuracy**: Assesses localization precision by checking if the maximum activation point in the heatmap falls within the ground-truth box [22]. The score equals the ratio of hits to total true positives.
- **Mean Energy Ratio**: Calculates the proportion of the heatmap's total energy (sum of activation values) that is concentrated within the ground-truth bounding box [21]. This metric measures how well the model avoids focusing on irrelevant background regions [22]. The ratio equals the sum of activation values within the bounding box divided by the total sum of all activation values in the heatmap.

IV. Results and Interpretability Analysis

This section presents the results of the comparative evaluation of YOLOv9-tiny, YOLOv10-nano, and YOLOv11-nano, focusing on both detection performance and interpretability.

A. Detection Performance Context

As a prerequisite to interpretability analysis, the study established that all three lightweight models perform effectively in pothole detection. The results in Table I show that each model achieved excellent and broadly comparable accuracy, with mAP@0.5 scores above 0.980. This consistently high performance validates the suitability of the models for the task. More importantly, it shifts the focus of the investigation from *whether* the models can detect potholes to *how* they reach their decisions, as this is where key differences emerge and interpretability becomes the critical factor.

TABLE I

Quantitative Comparison of Detection Performance and Interpretability Metrics. Prec. denotes Precision, Rec. denotes Recall, Point. Acc. denotes Pointing Accuracy, and E. Ratio denotes Energy Ratio.

Model	**Prec.**	**Rec.**	**mAP@.5**	**CAM IoU**	**Point. Acc.**	**E. Ratio**
YOLOv9-tiny	**0.924**	**0.964**	0.983	**0.674**	0.670	0.934
YOLOv10-nano	0.892	0.956	0.980	0.160	**0.814**	**0.936**
YOLOv11-nano	0.921	0.956	**0.984**	0.082	0.787	0.932

B. Qualitative Analysis of Attention Mechanisms

The LayerCAM visualizations in Figure 1 reveal a systematic evolution in attention strategies across the YOLO generations. This qualitative evidence anchors the central thesis of the study.

YOLOv9-tiny exhibits a *holistic* attention pattern, with broad, diffuse activations that cover the full extent of potholes and their immediate context. This pattern suggests reliance on contextual cues and overall surface texture to identify defects. In the complex scene shown, this behavior yields a CAM IoU of 0.472, aligning well with the actual damaged regions.

In contrast, **YOLOv10-nano and YOLOv11-nano** demonstrate a *pointer-like* attention strategy. Their heatmaps concentrate into sharp, localized hotspots that emphasize discriminative subregions, such as sharp edges or deep shadows. While highly efficient, this approach sacrifices holistic coverage of the object. The corresponding CAM IoU scores—0.040 for YOLOv10-nano and 0.022 for YOLOv11-nano—illustrate this trade-off clearly.

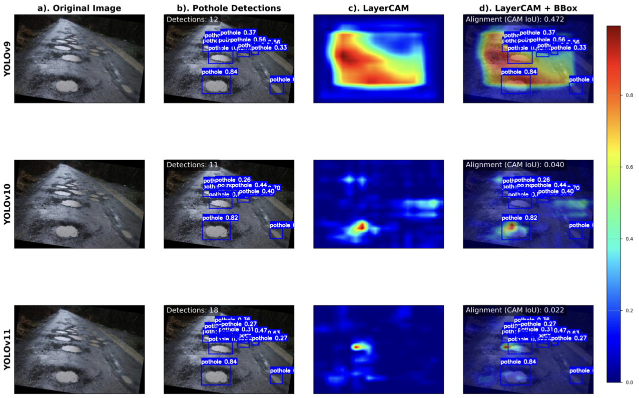

Fig. 1. LayerCAM analysis of a complex multi-pothole road scene. Each row shows: a) original image, b) detection results, c) LayerCAM attention heatmap, and d) a combined overlay. YOLOv9-tiny demonstrates a distributed attention pattern with 12 detections and high alignment (CAM IoU: 0.472). YOLOv10-nano shows more focused attention but lower alignment (11 detections, CAM IoU: 0.040). YOLOv11-nano exhibits highly concentrated hotspots that are poorly aligned with full pothole boundaries (18 detections, CAM IoU: 0.022).

C. Quantitative Interpretability Assessment

The metrics in Table I confirm the qualitative findings and highlight the nature of the interpretability trade-off.

YOLOv9-tiny excels in holistic alignment, achieving a mean CAM IoU of 0.674, over four times higher than YOLOv10-nano (0.160) and eight times higher than YOLOv11-nano (0.082). This confirms its ability to represent a pothole's complete shape and context.

YOLOv10-nano leads in pointing precision, achieving a Pointing Accuracy of 0.814, with YOLOv11-nano close behind at 0.787, both outperforming YOLOv9-tiny (0.670). These results show that while newer models excel at localizing discriminative features, they fail to capture the full object form.

Finally, all models achieved strong Mean Energy Ratio scores (≈ 0.93), showing consistent focus on the object rather than background noise.

D. Results Summary

In summary, these results confirm that all three YOLO models deliver high detection accuracy, but their interpretability profiles differ significantly. The comparative evaluation revealed a consistent trade-off between holistic and pointer-like attention strategies. The tailored LayerCAM implementation enabled stable visualization across YOLO architectures, while the use of CAM IoU, Pointing Accuracy, and Energy Ratio established reproducible interpretability benchmarks. Together, these findings form the foundation for explainability-guided deployment strategies.

V. Conclusion

This study delivered a rigorous comparative analysis of lightweight YOLO architectures, offering critical insights into the relationship between model evolution and interpretability in pothole detection. By adapting LayerCAM to object detection tasks and applying it systematically across YOLOv9-tiny, YOLOv10-nano, and YOLOv11-nano, the analysis revealed a clear distinction in their attention mechanisms. YOLOv9-tiny demonstrated a *holistic* and context-aware strategy that faithfully aligned with pothole boundaries, while YOLOv10-nano and YOLOv11-nano exhibited a *pointer-like* focus that prioritized localization precision at the expense of complete shape representation.

The evaluation established reproducible benchmarks, including CAM IoU, Pointing Accuracy, and Energy Ratio, which quantified these interpretability trade-offs and provided a foundation for future research. These results emphasize that detection accuracy alone is insufficient for model selection in safety-critical applications. Instead, model choice must also consider explainability requirements. YOLOv9-tiny is best suited for detailed human-in-the-loop damage assessment, while YOLOv10-nano and YOLOv11-nano provide advantages for fast, automated anomaly flagging. Overall, the findings highlight the importance of explainability-aware deployment strategies that align model interpretability with the practical needs of infrastructure monitoring.

VI. Future Work

Future research can extend this work in several directions. First, expanding the evaluation to multiple datasets across diverse geographic regions and road conditions would test the generalizability of the interpretability patterns observed. Second, incorporating alternative XAI techniques such as Integrated Gradients or attention rollout could provide complementary perspectives on model decision-making. Third, applying temporal analysis to video data would reveal how these models sustain attention over time in dynamic monitoring tasks. Finally, integrating interpretability objectives directly into the training process—such as attention-guided loss functions—offers a promising avenue to balance detection performance with more faithful and holistic explanations.

Use of Generative AI

Generative AI tools (ChatGPT and Claude AI) were utilized by the authors to assist with improving the grammar, clarity, and readability of the manuscript. The core research, including literature review, data analysis, and interpretation of results, was conducted independently by the authors, who reviewed, edited, and take full responsibility for the final content of this publication.

Acknowledgment

The authors express their gratitude to the Department of Science and Technology - Engineering Research and Development for Technology (DOST-ERDT) and Mindanao State University - Iligan Institute of Technology (MSU-IIT) for funding this research work.

References

[1] R. Fan, U. Özgünalp, B. Hosking, M. Liu, and I. Pitas, "Pothole Detection Based on Disparity Transformation and Road Surface Modeling," *IEEE Trans. Image Process.*, vol. 29, pp. 897–908, 2020.

[2] N. Ma, J. Fan, W. Wang, J. Wu, Y. Jiang, L. Xie, and R. Fan, "Computer Vision for Road Imaging and Pothole Detection: A State-of-the-Art Review of Systems and Algorithms," *Transportation Safety and Environment*, vol. 4, no. 4, Dec. 2022, Art. no. tdac026.

[3] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[4] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "Layer-CAM: Exploring Hierarchical Class Activation Maps for Localization," *IEEE Trans. Image Process.*, vol. 30, pp. 5875–5888, 2021.

[5] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.

[7] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, 2018, pp. 839–847.

[8] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," arXiv preprint arXiv:2402.13616, 2024.

[9] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "YOLOv10: Real-Time End-to-End Object Detection," arXiv preprint arXiv:2405.14458, 2024.

[10] R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," arXiv preprint arXiv:2410.17725, 2024.

[11] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-Down Neural Attention by Excitation Backprop," *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, 2018.

[12] J. Choe and H. Shim, "Evaluation of Localization for Weakly Supervised Object Localization," *IEEE Access*, vol. 8, pp. 207438–207450, 2020.

[13] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 8024–8035.

[14] G. Jocher et al., "Ultralytics YOLOv5," 2020, [Online]. Available: https://github.com/ultralytics/yolov5.

[15] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2017, pp. 3319–3328.

[16] S. Abnar and W. Zuidema, "Quantifying Attention Flow in Transformers," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4190–4197.

[17] J. Choe and H. Shim, "Attention-based Dropout Layer for Weakly Supervised Object Localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2019, pp. 2213–2222.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[19] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7464–7475.

[20] L. V. Fortin and O. E. Llantos, "Performance Analysis of YOLO versions for Real-time Pothole Detection," *Procedia Computer Science*, vol. 257, pp. 77–84, 2025.

[21] J. Choe and H. Shim, "Attention-based Dropout Layer for Weakly Supervised Object Localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2019, pp. 2213–2222.

[22] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-Down Neural Attention by Excitation Backprop," *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, 2018.