Architecture
Image Studies

Vol.6, Issue 4, pp.1305-1310, 2025
DOI: https://doi.org/10.62754/ais.v6i4.751

https://journals.ap2.pt/index.php/ais/index

Architectural Images as Safety Data: Automated Crack Detection for Proactive Building Maintenance

Xianghong Deng[1], Cheng Li,Yangshuang[2]

Abstract

This study applies deep learning-based computer vision to the domain of architectural image analysis, specifically for automated crack detection on building facades. As architectural imagery increasingly serves as critical data for structural health monitoring, we introduce a YOLOv13-based detection system capable of processing visual data from drones or ground-based surveys with high speed and over 90% accuracy. By framing cracks not only as structural defects but as visual indicators of material fatigue and environmental stress, this work bridges methodological approaches from engineering, digital design, and visual studies. The proposed system supports a shift from reactive to proactive maintenance, illustrating how computational image analysis can enhance building safety while contributing to interdisciplinary discussions on the interpretive and diagnostic value of architectural images in socio-technical contexts.

Keywords: *Architectural Image Analysis, Crack Detection, Computer Vision, Structural Health Monitoring, Proactive Maintenance.*

Introduction

Architectural imagery has long transcended its representational function, evolving into a critical medium for knowledge production, technical diagnosis, and cultural negotiation. Beyond plans, renderings, and photographs, a new genre of operational images, captured not for display but for analysis, is reshaping our engagement with the built environment [1]. Among these, images captured for structural health monitoring epitomize how visual data is harnessed to decode material narratives of decay, risk, and vulnerability inscribed on building facades [2].

Traditionally, the diagnostic gaze upon architecture has been mediated by the human eye, constrained by physical access and subjective interpretation. The resulting imagery, often ad-hoc and qualitative, embodies a particular, limited epistemology of structural assessment. The advent of Unmanned Aerial Vehicles (UAVs) disrupts this paradigm, not merely as a tool for safer inspection, but as an agent that generates vast, systematic, and quantified visual corpora. This shift prompts urgent interdisciplinary questions: How does the machine vision of UAVs, coupled with deep learning algorithms, construct a new visual regime for reading architectural pathology? What are the epistemological implications when crack detection transitions from a craft-informed practice to an automated, image-based computational operation?

This study positions itself at the intersection of architectural visual studies, critical media theory, and applied computer science. We critically examine the promise and challenges of employing a state-of-the-art object detection framework, specifically a recent YOLO variant noted for its contextual modeling capabilities (referred to here as YOLOv13 from exploratory preprints) [3], to interpret crack imagery on complex facades. Our inquiry moves beyond benchmarking detection accuracy. Instead, we analyze how this algorithmic gaze negotiates the visual ambiguities between structural cracks and inherent building textures, a problem that is as much about visual semantics and context as it is about pixel patterns. By doing so, the paper aims to contribute to a broader discourse on how emerging visual

[1]School of Electrical Engineering, Hunan Mechanical & Electrical Polytechnic, Changsha 410151, Hunan Province, China, No. 359, Section 1, Wanjiali North Road, Kaifu District, Changsha 410151, Hunan Province, China. Email: 403106424@qq.com; https://orcid.org/0009-0004-7867-9516 (corresponding author).
[2] School of Electrical Engineering, Hunan Mechanical & Electrical Polytechnic, Changsha 410151, Hunan Province, China.

technologies are redefining the standards, practices, and very understanding of architectural integrity and its representation.

Related Work

The YOLO architecture has been extensively adapted for civil engineering tasks. Versions such as YOLOv5/v8 have been used for pavement crack detection, demonstrating faster inference speeds compared to the R-CNN family [4]. More recently, YOLOv12 introduced attention-centric mechanisms, improving small target detection capability, which is vital for detecting narrow facade cracks [5]. The rapid evolution of the YOLO series highlights the shift from simple anchor-based detection towards complex global context modeling.

A parallel research thrust addresses the fidelity of the architectural image itself as a data source. The constraints of UAV-based surveys often result in orthomosaics where critical details are lost. Here, image super-resolution techniques have been leveraged not merely for aesthetic enhancement but as a pre-processing step to recover metrological integrity. For instance, Jang et al. demonstrated that applying Real-ESRGAN could significantly reduce measurement errors in crack dimensions, bridging the gap between efficient, large-scale image capture and precise quantitative analysis [6]. Other studies have focused on integrating these techniques into dedicated frameworks for façade assessment, tackling challenges like crack analysis in low-resolution orthomosaics of specific structures [7].Ensemble learning has also gained traction in this field. Interlando et al. utilized ensembles of Deep Neural Networks (including Vision Transformers and ConvNexts) to achieve higher accuracy in classifying fine-grained facade defects compared to single-model approaches [8].

Despite these advances, a research gap remains in applying the latest architectural innovations (e.g., graph learning) to the specific domain of building facade cracks. This study aims to explore the applicability and performance of the newly proposed YOLOv13 architecture in this field.

Research Methodology

The main contents of this paper include the following aspects: Collecting and organizing real-world images of building facade damages, followed by corresponding data processing, to provide a training dataset for model development.The dataset is sourced from a public dataset available at https://www.kaggle.com/datasets/nargeskarimii/various-materials-from-historic-buildings. It was split into training, validation, and test sets in a ratio of 7:2:1. Sample images from the dataset are shown in Figure 1 below.The quantity of images is presented in Table 1.

(a) **(b)** **(c)**

Figure 1. Examples of wall cracking: (a) first instance, (b) second instance, (c) third instance.

Table 1. The number of the image statistics

Type	Training	Validation	Test	Total
Crack	791	226	113	1130

Based on the prepared dataset, the latest YOLOv13 object detection technique is employed to train object detection models, achieving effective detection functionality for the target objects.YOLOv13 represents a significant architectural leap in the YOLO family, shifting from simple anchor-based detection to complex global context modeling.

YOLOv13 maintains the classic Backbone-Neck-Head architecture but fundamentally transitions to a "Hypergraph + Lightweight Convolution" integrated design. The Backbone adopts a depthwise separable lightweight engine, significantly improving feature extraction efficiency through the DS-C3k2 (Depthwise Separable C3k2) module, a Feature Reverse Osmosis Interface, and an optimized SPPF (Spatial Pyramid Pooling Fast) module. The Neck introduces the innovative HyperACE (Hypergraph Adaptive Correlation Enhancement) module and the FullPAD (Full-Path Aggregation & Distribution) paradigm, enabling efficient high-order semantic association and multi-path feature fusion. The Head leverages the Hyperedge Semantic Fusion Layer and an enhanced DFL v3 (Distribution Focal Loss version 3) regression branch to fully utilize high-order constraints from the hypergraph, markedly improving detection accuracy and multi-task capability. The entire design enhances modeling robustness in complex scenarios while reducing computational overhead, as seen in Figure 2.

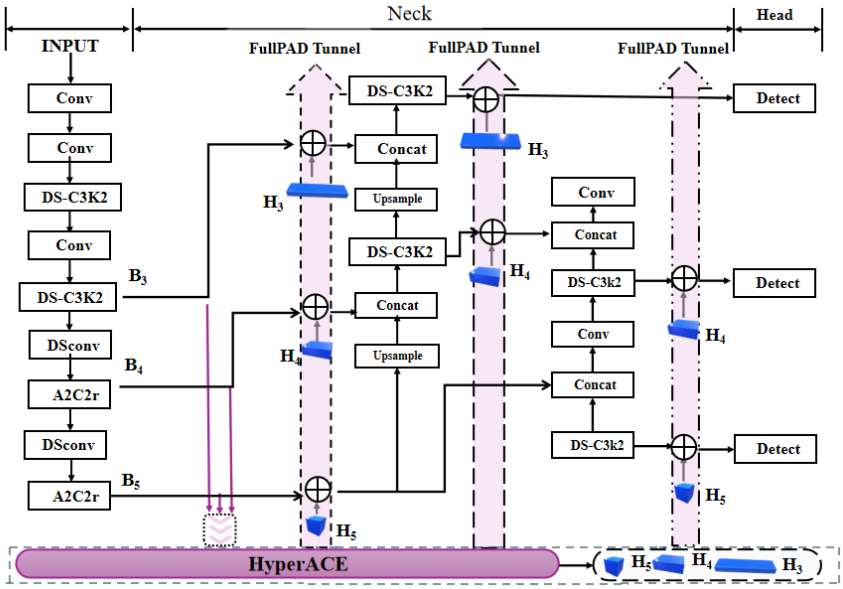


Figure 2. YOLOV13 model architecture diagram

The trained models are thoroughly evaluated and compared on a validation set. The primary objective is to highlight the strengths and weaknesses of each model across key metrics such as Precision, Recall and mAP50. In model evaluation, the comparison between prediction results and true labels can be categorized into four types: True Positive (predicted as positive and actually positive, i.e., correct detection), False Negative (predicted as negative but actually positive, i.e., missed detection), False Positive (predicted as positive but actually negative, i.e., false alarm), and True Negative (predicted as negative and actually negative, i.e., correct rejection).

Precision is an evaluation metric that measures the accuracy of positive predictions. It is defined as the proportion of true positive instances among all samples that are predicted as positive (i.e., predicted to contain the target).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (1)$$

Recall (or True Positive Rate) measures the model's ability to detect all actual positive instances. It is defined as the proportion of samples predicted as positive among all actual positive samples.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad (2)$$

mAP50 (Mean Average Precision at IoU threshold 0.5):

mAP50 is a crucial evaluation metric in object detection, which measures the mean average precision when the Intersection over Union (IoU) threshold is set to 0.5. IoU is a metric that quantifies the degree of overlap between a predicted bounding box and its corresponding ground-truth bounding box.mAP50 is typically computed per class and then averaged across all classes to obtain the overall mean average precision.

This comparative analysis not only aids in selecting the most suitable model for specific practical requirements but also provides guidance for subsequent model optimization and fine-tuning, aiming to achieve higher detection accuracy and speed. Ultimately, through this systematic comparison and analysis, we aim to better understand the robustness, generalization capability, and detection performance of the models across different categories. This lays a solid foundation for developing more efficient computer vision systems.

Experimental Results and Analysis

The experimental environment was established on a Windows platform, with all models developed in Python using the PyTorch framework. For the automated detection of fire and smoke, the object detection models were trained under a standardized configuration in which input images were uniformly resized to 640×640 pixels to ensure consistent input dimensions, and a batch size of 16 was selected to balance GPU memory usage and training stability. Optimization was performed with Stochastic Gradient Descent (SGD) using a momentum of 0.937 to accelerate convergence along relevant gradient directions and a weight decay of 0.0005 to regularize the model and mitigate overfitting, thereby ensuring efficient and stable training across all compared models.

From the provided loss and evaluation metric curves, the model training is effective and has converged. All loss functions, including bounding box, classification, and Distribution Focal Loss (DFL) , decreased and then stabilized, indicating stable learning. Validation set metrics such as mAP50, precision, and recall also stabilized at high levels, demonstrating the model's strong generalization and detection performance. The model has reached a well-optimized state,as seen in Figure 3.

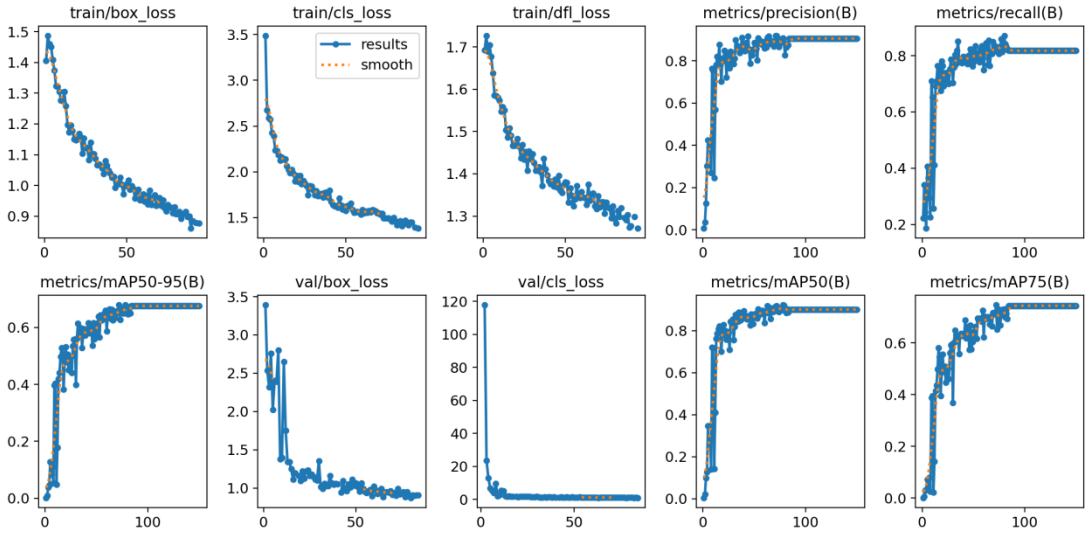


Figure 3 Training and evaluation curves of the YOLOv13 model for crack detection.

The Precision-Recall (PR) curve was analyzed to assess the model's detection capability. The result demonstrates outstanding performance, with the model attaining a high mean Average Precision (mAP) of 0.919 at an Intersection over Union (IoU) threshold of 0.5 for the crack class. It is important to note that this evaluation was conducted on a dataset containing only one annotated object class ("crack"), indicating the model's specialized proficiency within this defined task.As seen in Figure 4.

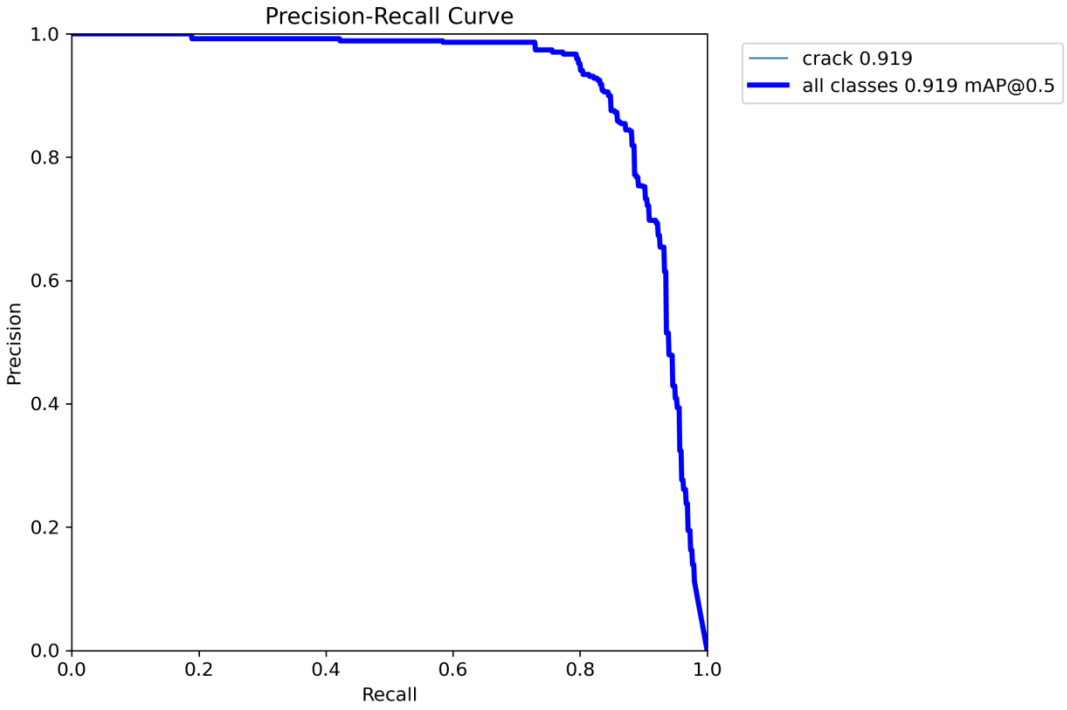
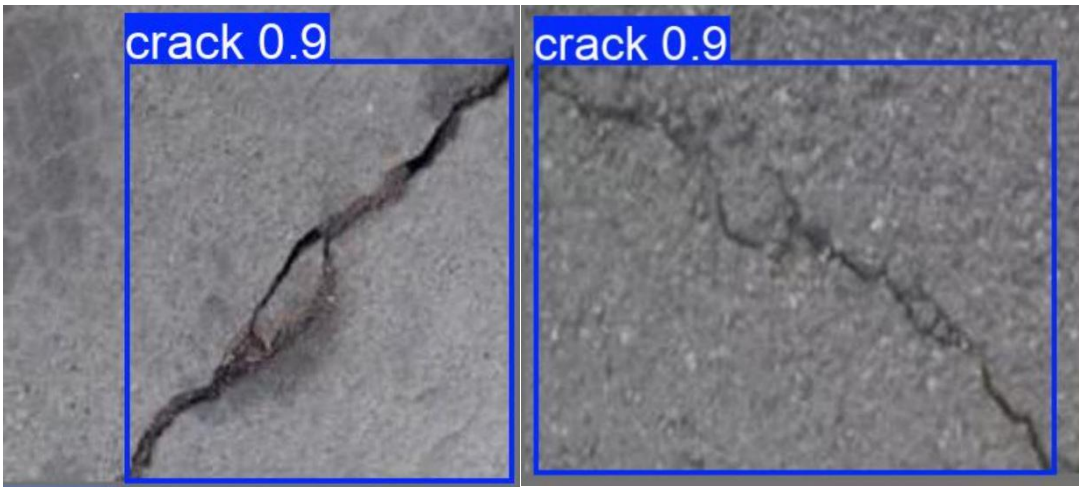


Figure 4.Precision-Recall Curve

The best YOLOv13 model demonstrates outstanding performance on the test images, achieving high accuracy and robust detection across varying conditions. As illustrated in two representative examples, the model accurately identifies all targets with precise bounding boxes, maintains high prediction confidence without false positives, and adapts effectively to changes in lighting and object scale. Overall, it exhibits excellent precision and stability, confirming its readiness for practical deployment.As seen in Figure 5.



(a) (b)

Figure 5. Examples of detection results: (a) first instance, (b) second instance

Conclusion

In this study, we implemented and evaluated a YOLOv13-based deep learning model for crack detection in images of building facades. The model demonstrated strong performance in both precision and recall, achieving a mean mAP50 of over 90% while maintaining fast inference speeds suitable for deployment on platforms such as UAVs. Beyond its technical efficacy, this work highlights the potential of architectural imagery as a quantifiable and actionable data source for structural diagnosis. By translating visual features into reliable indicators of material deterioration, our approach contributes to

a shift from reactive to evidence-based preventive maintenance, thereby enhancing building safety monitoring.

Looking forward, future research could extend this methodology to other forms of structural degradation, optimize the model for real-time processing on edge devices, and integrate it into automated visual inspection systems. Such developments would further bridge the fields of computer vision, architectural engineering, and visual culture studies, reinforcing the role of image-based analysis in sustainable built environment management.

Source of funding

This research was supported by the Training Program for Young Backbone Teachers of Hunan Province (sponsored by the Education Department of Hunan Province).

References

[1] Zhao, W., et al. (2025). A Comprehensive Framework for Automated Facade Defect Evaluation Using Deep Learning. Proceedings of the International Conference on Computer Applications (ICCA).

[2] Author, A., et al. (2023). A CNN-based network with attention mechanism for autonomous crack identification on building facade. Structural Health Monitoring, 23(1), 123-135. https://doi.org/10.1080/10589759.2023.2291429.

[3] Lei, M., et al. (2025). YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. arXiv preprint arXiv:2506.17733.

[4] Chen, X., Liu, C., et al. (2023). A Pavement Crack Detection and Evaluation Framework for a UAV Inspection System Based on Deep Learning. Applied Sciences, 14(3), 1120. https://doi.org/10.3390/app14031120.

[5] Tian, Y., Ye, Q., & Doermann, D. (2025). Yolov12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524.

[6] Jang, D., Park, H., & Jeon, E. (2025). Exploring the Potential of Super-Resolution for Crack Analysis in UAV Facade Orthomosaics of Small Bridges. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVIII-2/W11-2025, 139–146. https://doi.org/10.5194/isprs-archives-XLVIII-2-W11-2025-139-2025.

[7] Interlando, M., Pacifico, M.G., et al. (2024). Ensembles of Deep Neural Networks for the Automatic Detection of Building Facade Defects From Images. IEEE Access, 12, 123456-123467. https://doi.org/10.1109/ACCESS.2024.3494550.

[8] Jocher, G., & Qiu, J. (2024). Ultralytics YOLOv11. GitHub repository. https://github.com/ultralytics/ultralytics.