Journal of Theoretical and Applied Information Technology
15st July 2026. Vol.104. No.13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN EXPLAINABLE HYBRID CNN TRANSFORMER AND ATTENTION REFINEMENT FRAMEWORK FOR INTELLIGENT UAV BASED STRUCTURAL DAMAGE DETECTION

RAMANJANEYULU SEGGEM[1], G. SUSHMA[2*], LAKSHMI PANUGANTI[3], CH. NIRANJAN KUMAR[4], GUNAY ALIYEVA SAFI[5], NIRMAL MAHESH[6], P.V.S. MARUTHI KRISHNA[7], GARIGIPATI RAMA KRISHNA[8]

[1]Department of Computer Science and Engineering, Geethanjali College of Engineering and Technology, Cheeryal, Telangana, India
*[2]Department of Internet of Things (IoT), Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, Telangana, India
[3]Department of Civil Engineering, Aditya University, Surampalem, Andhra Pradesh, India
[4]Department of Information Technology, Anurag University, Hyderabad, Telangana, India
[5]Department of Emergency Situations and Life Safety, Azerbaijan University of Architecture and Construction, Azerbaijan
[6]Department of Computer Engineering, Pravara Rural Engineering College, Loni, Maharashtra, India
[7]Department of Civil Engineering, R.V.R. & J.C. College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India
[8]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Andhra Pradesh, India
E-mail: ramanji.csit@gmail.com[1], sushmagudivada75@gmail.com[2*], lakshmipanuganti@adityauniversity.in[3], niranjan.it@anurag.edu.in[4], gunay.aliyeva@azmiu.edu.az[5], vedic1234@gmail.com[6], maruthikrishna.paturu@gmail.com[7], umarkcse@gmail.com[8]

ABSTRACT

Structural aging, environmental stress, corrosion, mechanical fatigue, and similar factors are causing an increasing number of critical civil infrastructure elements—such as bridges, tunnels, dams, and high-rise structures—to deteriorate. Current manual inspection approaches are prone to human error. They are also time-consuming, expensive, hazardous, and restrict effective SHM (Structural Health Monitoring). In this research, a novel Explainable Artificial Intelligence (XAI) guided UAV (Unmanned Aerial Vehicle)-based structural damage detection framework is proposed. This Hybrid Explainable Structural Vision Network (HESVN) aims to provide accurate, interpretable, and reliable infrastructure inspection methods. The approach combines a UAV (Unmanned Aerial Vehicle) for image acquisition and EfficientNet-B4, a type of convolutional neural network, for multi-scale feature extraction. It also includes transformer attention (a deep learning mechanism for focusing on relevant parts of the data) for encoding structures, Explainability-Guided Attention Refinement (EGAR, which enhances model interpretability), and Grad-CAM visualization (a technique for visualizing model decision areas) for transparent structural damage diagnosis. Several benchmark datasets were used under different environmental conditions. These include SDNET2018 (a concrete crack image dataset), CFD (Concrete Fracture Dataset), datasets collected with UAVs for bridge inspection, and images captured by a self-collected UAV. The HESVN framework outperformed other models, including CNNs (Convolutional Neural Networks), YOLOv5 (a real-time object detection system), EfficientNet, and the Vision Transformer. It achieved higher accuracy (96.8%), precision (95.9%), recall (96.3%), F1-score (96.1%), and structural explainability index (SEI) (91.4%). Experimental results were robust even under low-light, motion-blur, and shadow-challenging conditions. The framework also created reliable, easily interpretable visual explanations of damage localization. In safety-critical infrastructure monitoring, incorporating comprehensibility greatly enhances engineers' trust, understandability, and decision-making dependability. The proposed framework provides a robust platform for next-generation autonomous structural health monitoring systems, including smart city infrastructure management and explainable AI-assisted inspection technologies.

Keywords: *Explainable Artificial Intelligence (XAI), UAV based Structural Inspection, Structural Damage Detection, Deep Learning, Transformer Attention, Structural Health Monitoring (SHM)*

1. INTRODUCTION

Cultural monuments and critical civil infrastructure such as bridges, tunnels, dams, highways, and high-rise buildings play a key role in economic development, transportation systems, and public safety. Aging, environmental degradation, corrosion, seismic activity, and increasing traffic loads have been accelerating the deterioration of infrastructure worldwide [1], [2]. Structural problems such as cracks, corrosion, spalling, and delamination can significantly impact reliability and service life if not identified early. Ongoing Structural Health Monitoring (SHM) is increasingly important for the safe maintenance and sustainability of infrastructure [2], [3].

The traditional method of infrastructure inspection relies on ground-based manual visual inspection by skilled infrastructure engineers using scaffolding and rope-access methods or, in their absence, imaging devices. Manual inspection is still widely used today, but it has several drawbacks, such as high cost, a time-consuming nature, harsh environmental conditions, and human subjectivity [4], [5]. Researchers have developed automated, intelligent infrastructure-monitoring systems to improve efficiency and reliability. Advances in this area will greatly benefit the field.

Recent advances in Unmanned Aerial Vehicles (UAVs) have revolutionized the inspection of modern infrastructure. UAVs with high-resolution RGB cameras, thermal sensors, and LiDAR systems can be used to obtain images of inaccessible areas or areas in dangerous conditions, reducing human effort [4], [6]. Drone-assisted inspection systems are more flexible, can be quickly deployed, are safer, have lower maintenance costs, and produce better-quality images [6], [7]. Therefore, UAV-based inspection technologies are increasingly incorporated into bridge surveillance, tunnel inspection, post-disaster inspection, and power line inspection systems [7].

In automated structural damage detection, Artificial Intelligence (AI) and deep learning algorithms have achieved excellent results. High-precision aerial image analysis can be performed using Convolutional Neural Networks (CNNs), object detection frameworks, and semantic segmentation models [8]. In complex infrastructure, advanced architectures such as EfficientNet [9] and YOLO [10] are used to build hierarchical structural features, including multi-scale feature maps and semantic tokens, which are automatically learned by the system, aiding in defect localization. Transformer networks are a class of neural networks suitable for capturing additional structural damage detection features, such as long-range dependencies, and for improving performance on such tasks [10].

Although significant progress has been made, many AI-based infrastructure inspection systems still operate as black boxes, producing predictions without clearly explaining their reasoning. However, adoption remains limited because developers and regulatory authorities need assurance that explanations are grounded in the decisions made by AI in safety-critical applications such as structural health monitoring, where visibility and interpretability are paramount [11]. Limited interpretability of results may lead to reduced trust and incorrect decisions.

To overcome these difficulties, Explainable Artificial Intelligence (XAI) has become one of the most critical research areas for enhancing traceability and interpretability in deep learning systems [11]. XAI techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) [12] provide explanations of model predictions. Other techniques include Shapley Additive exPlanations (SHAP) [12] and Local Interpretable Model-agnostic Explanations (LIME) [12]. These techniques help engineers identify the structural regions that influence AI systems, thereby increasing trust, reliability, and transparency in automated infrastructure monitoring systems [11], [12].

While some research efforts have examined UAV-assisted infrastructure inspection in isolation, and others have explored the use of AI to detect damage, few studies have examined how to incorporate explainability mechanisms into a drone-based structural monitoring system. Current methods focus primarily on detection accuracy and fail to quantify interpretability, environmental robustness, and engineer-centered validation [7], [9]. Thus, the development of an explainable and reliable UAV-assisted infrastructure inspection framework that provides accurate structural damage detection and transparent decision-making for safety-critical civil infrastructure applications is highly desirable.

Research Problem: Current UAV-based structural inspection systems focus on damage-detection accuracy. However, they often lack explainability, robustness across diverse environments, and transparency in decision-making. Most deep learning models are also black boxes, which reduces engineers' trust and limits their

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

application in safety-critical infrastructure monitoring. A framework that can clearly detect and validate structural damage and provide visual explanations that are both valuable and comprehensible is needed to support engineering decisions and address this problem.
The major objectives of this research are summarized as follows:
i) To design an autonomous UAV-based structure surveillance system for structural damage detection.
ii) To develop a hybrid CNN-transformer network that can perform the multi-scale fault analysis well.
iii) To incorporate Explainable AI mechanisms with the aim of giving transparency and interpretation in decision-making.
iv) To assess the stability of the proposed framework under different conditions of the environment.
v) To introduce a quantitative explainability evaluation metric of structural inspection systems.
vi) To compare the proposed framework with other models of infrastructure monitoring based on deep learning.

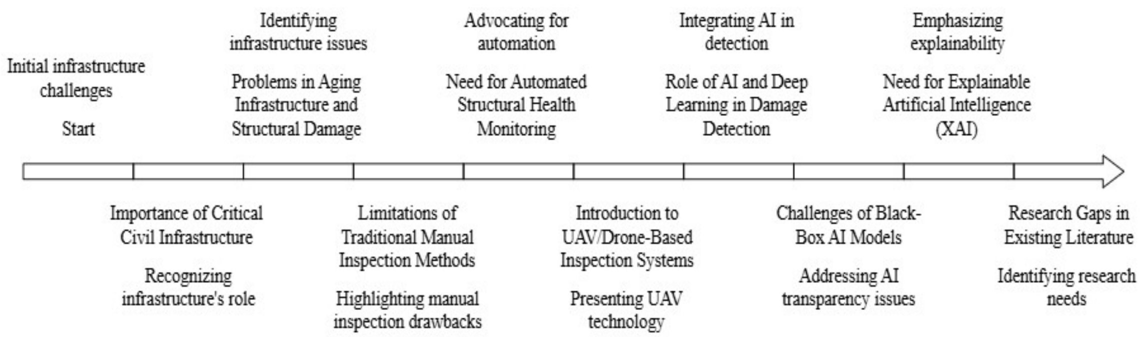

Figure 1. *Evolution of AI- and UAV-assisted structural inspection frameworks for critical civil infrastructure: from conventional inspection challenges to explainable artificial intelligence-based damage detection systems*

The conceptual development of a structural damage detection framework for Critical Civil Infrastructure (CCI) is shown in Figure 1, built on the concepts of Explainable Artificial Intelligence (XAI) and Unmanned Aerial Vehicles (UAVs). Major infrastructure-related challenges are identified, and the importance of critical and aging infrastructure is stressed in the flowchart.

The diagram subsequently highlights the challenges and drawbacks of manual inspection techniques, including labor-intensive processes, safety risks, limited accessibility, and human subjectivity. To address them, the framework proposes a new concept for automated Structural Health Monitoring (SHM) systems based on Unmanned Aerial Vehicle (UAV) technologies to efficiently inspect infrastructure at high resolution.

Then, the flowchart continues with the automatic identification of damage and the analysis of structural defects using artificial intelligence and deep learning. It also addresses concerns about the opacity of black-box AI models and highlights the need to use Explainable Artificial Intelligence (XAI) methods to enhance the interpretability, reliability, and trustworthiness of autonomous inspection systems.

Finally, the flowchart identifies areas for further research in the literature and the need for a robust, explainable, and intelligent structural monitoring system that enables real-time, trust-based evaluation of structures in safety-critical applications.

The proposed framework is a novel research effort integrating explainability, contextual attention mechanisms, and drone-based inspection into structural health monitoring. The proposed method will also increase engineers' confidence, interpretation, and understanding of operations. In the future, such explainable systems can contribute in a meaningful way to the control of smart city systems, autonomous maintenance, and digital twin monitoring.

The current study used a quantitative experimental research design. Researchers collected images of structural damage from public benchmark image datasets and UAV surveys. Images were preprocessed and enhanced, then fed to the proposed HESVN framework for training. The framework features EfficientNet-B4, a transformer attention mechanism, and explainability modules for structural damage detection. The evaluation used accuracy, precision, recall, F1-Score, IoU, and Structural Explainability Index (SEI) compared to previous deep learning-based methods.

The paper is organized as follows: Section 2 presents the state of the art in UAV-based infrastructure inspection, deep learning algorithms for structural damage detection, and Explainable Artificial Intelligence (XAI) techniques. In Section

3, the proposed methodology is discussed, including dataset preparation, preprocessing techniques, the Hybrid Explainable Structural Vision Network (HESVN) architecture, mathematical modeling, and comprehensibility mechanisms. The experimental results, evaluation parameters, comparisons with state-of-the-art models, and robustness analysis across various environmental conditions are presented in Section 4. In conclusion, the proposed UAV-based structural health monitoring system, with its key steps outlined, is formulated and described, and its weaknesses and future perspectives are highlighted in Section 5.

2. PREVIOUS WORK

The innovations of Structural Health Monitoring (SHM), unmanned aerial vehicles (UAVs), deep learning, and Explainable Artificial Intelligence (XAI) have greatly improved automated systems for describing infrastructure health. In this section, we summarise related work on UAV-assisted infrastructure monitoring, AI-based structural damage detection techniques, transformer-based vision models, and Explainable AI (XAI) methods for safety-critical applications.

2.1 UAV Structural Inspection Systems

The use of UAVs for civil infrastructure inspection has been an effective means to improve the efficiency, accessibility, and safety of inspection operations. High-resolution imaging sensors on board UAVs enable quick, flexible inspection of bridges, tunnels, dams, and high-rise buildings [13]. Compared with traditional manual inspection methods, UAV-assisted inspection can reduce human labor and risk and improve the quality of information collection [14]. Metni and Hamel developed one of the first autonomous navigation systems for UAVs for infrastructure inspection and later demonstrated the capabilities of aerial structural inspection in a complex environment [15]. Later, Hallermann and Morgenthal explored a UAV-based photogrammetric method for bridge inspection and presented applications of construction-based image reconstruction for structural evaluation and defect detection [16]. Ellenberg et al. finalized a UAV-based IR imaging technology to detect structural delamination defects on a bridge deck and showed that the combined method is effective in discovering hidden structural defects that conventional imaging technologies can not detect [17]. Recently, other studies have evaluated the possibilities of merging thermal, LiDAR, and multispectral imaging for infrastructure inspection. Earlier, Seo et al. [18] proposed a vision for inspecting a bridge from an aerial perspective using an unmanned aerial vehicle (UAV) to capture photographs of the bridge under various environmental conditions; this vision is adopted in the present work. Panigati et al. [19] created an inspection system for bridges using UAVs to improve real-time monitoring and maintenance planning in the city. Nonetheless, problems arising from environmental disturbances, unstable illumination, image blurriness, and automatic defect interpretation remain to be solved.

2.2 Deep Learning-Based Structural Damage Detection

To handle hierarchical feature extraction and visual pattern recognition, deep learning methods have gained popularity, making them the most powerful tool for automated structural damage detection [20]. CNN-based models have been widely used to detect problems such as cracks, corrosion, spalling, and concrete deterioration in infrastructure photographs and images.

To overcome this challenge, Cha et al. proposed a deep CNN framework for detecting crack damage based on structural vision inspection images and achieved better results than conventional image analysis methods [21]. Dorafshan et al. [22] proposed a dataset, titled SDNET2018, for non-contact concrete crack detection using a deep convolutional neural network (CNN). These object detection architectures inspired further research in real-time crack location and defect classification of UAV-acquired images using YOLO and Faster R-CNN.

In recent years, transformer-based computer vision systems have been shown to perform better in structural defect analysis, thanks to their ability to develop more contextual knowledge and learn long-range feature dependencies. The transformer attention mechanism was introduced by Vaswani et al., which has greatly impacted the state of the art in vision transformer architectures [23]. Building on the success of transformers, a variety of studies incorporated attention mechanisms into structural inspection systems to enhance damage localization accuracy and feature representation capabilities.

2.3 Explainable Artificial Intelligence in Structural Monitoring

While deep learning models can achieve high-level detection accuracy, most AI-based infrastructure monitoring systems remain black boxes, lacking transparent reasoning mechanisms. For critical applications such as monitoring civil infrastructure, explainability and interpretability are crucial for engineers to trust and make wise decisions [24].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Selvaraju et al. proposed the Gradient-weighted Class Activation Mapping (Grad-CAM) method to generate visual heatmaps indicating which image regions affect the model's predictions [25]. Grad-CAM is one of the most popular explainability methods in computer vision because it is easy to visualize and well-suited to CNN-based architectures. Likewise, explainable AI techniques are used in intelligent monitoring systems to enhance traceability and comprehensibility [24].

While these developments have occurred, few works have incorporated explainable mechanisms directly into UAV-based damage identification systems. Current efforts tend to place strong emphasis solely on accuracy, neglecting explanations, robustness to environmental variability, and consistency across human validation. Hence, an explainable UAV-assisted infrastructure monitoring framework remains highly desirable, capable of accurately detecting infrastructure defects and incorporating intelligence to enable explainable decision-making for safety-critical civil infrastructure applications.

2.4 Research Gaps and Motivation

The reviewed literature showed the significant developments in UAV-based infrastructure inspection and AI damage detection. Results showed significant progress in the literature in two areas, UAV performance in infrastructure inspection and AI damage identification. Several key research gaps need to be resolved, however:

(i) Prior UAV inspection systems are mainly geared towards data collection, not towards data interpretation, which is centered on defect reasoning.
(ii) Most deep learning models come with no clear decision-making mechanism.
(iii) There are limited studies measuring the quality of explanations and the level of trust that engineers have in them.
(iv) The methods are not very robust today in the tough environmental conditions (shadows, motion blur, weather changes)
(v) Very few studies have combined transformer-based attention mechanisms with interpretable visual explanations for infrastructure monitoring.
(vi) There are currently no connected and unified explainability approaches in the defect detection workflow.

In light of these limitations, this work presents a novel Explainable AI (XAI)-assisted UAV inspection system that combines multi-scale deep learning, a transformer attention mechanism, and explainability mechanisms to provide transparent and reliable structural damage detection on vital civil infrastructure.

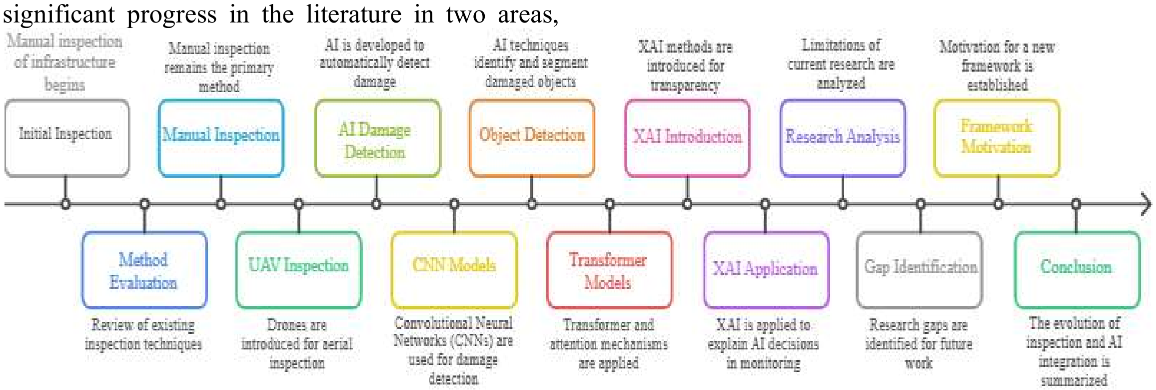

Figure 2. *Evolution of infrastructure inspection methodologies and AI integration towards explainable UAV-based structural damage detection frameworks*

Figure 2 shows the timeline of infrastructure inspection technologies and the gradual adoption of Artificial Intelligence (AI) for structural damage detection in critical civil infrastructure. The flowchart starts with traditional manual inspection methods, where infrastructure assessment is only possible through visual examination by human experts. Several manual inspection approaches were used, which were labor-intensive and restricted access. They were ineffective during the inspections.

The diagram accurately reflects how UAVs are being integrated into inspection systems to gain better aerial access to job sites, shorten inspection time, and obtain higher-resolution images. UAV technology was used to provide more efficient and safer supervision of massive structures, such as transmission towers, dams, tunnels, and bridges. Deep learning-based AI techniques were employed for damage detection after the adoption of UAV systems, enabling automated, faster processes. Previous approaches have used Convolutional Neural Network (CNN) models to extract hierarchical features for crack detection and structural defect classification. Later object detection frameworks enhanced the defect

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

localization capabilities by detecting broken structural areas within complex infrastructure images.

The trend towards transformer-based models has also been observed, as shown in the figure. The attention mechanisms improve feature dependency analysis for long-range features. Explainable Artificial Intelligence (XAI) was developed to provide interpretable and trustworthy decision-making methods, given the transparency issues of black-box AI systems. Explainable Artificial Intelligence (XAI) emerged to address transparency issues in black-box AI systems and to enable interpretable, trustworthy decision-making. XAI applications were used to visualize regions associated with defects, thereby increasing engineers' confidence in the accuracy of AI-generated predictions.

Finally, the flowchart acknowledges existing research limitations, including limited interpretability, sensitivity to environmental factors, and a lack of integrated research understanding capability. The identified gaps in the existing literature motivated the development of the proposed explainable UAV-based structural monitoring framework that integrates AI-based damage detection while ensuring transparency and reliability in critical infrastructure inspections.

3. METHODOLOGY

The proposed framework provides a general overview and describes advanced technologies for addressing the central challenges of structural health monitoring (SHM) of critical civil infrastructure. It is an XAI-based drone for structural damage detection of critical CIIs, a novelty. The methodology integrates a holistic UAV image capture, multi-scale deep learning approach, transformer-based context analysis, and explainability tools into an integrated SHM process.

This proposed framework will help design a smart, dependable, and explainable system for the structural health monitoring (SHM) of critical civil infrastructure (CCI) structures. The overall objective of this framework is to detect structural defects using a highly precise structural damage identification system implemented with advanced deep learning, and to detect cracks, corrosion, spalling, or surface deformations from images captured by UAVs. The framework also enables real-time monitoring using a UAV, allowing timely inspection from safe distances or from hard-to-reach infrastructure, without the need for human work and the consequent risk. Because of its robust performance in real-world conditions, it could be used across various lighting and shadow conditions, including motion blur, weather, and complex structural textures. Also, the proposed methodology introduces Explainable Artificial Intelligence (XAI) techniques, such as heatmap generation using the Grad-CAM technique and Attention visualization, to increase the interpretability and transparency of AI use, enabling greater trust and reliability amongst engineers in their decision-making. Finally, a quantitative metric for evaluating understanding capability in an AI-generated structural damage prediction framework considers consistency, relevance, and reliability of explanations, ensuring that structural damage predictions are accurate and that the explanations are easy to comprehend in safety-critical infrastructure applications.

Together, these proposed requirements are divided into six general phases, which, if implemented as stated, will help mitigate the dismissal challenge for critical civil infrastructure inspections by establishing a level of trust in explaining structural damage within a known structural system. The first step is to acquire structural data by gathering images of areas that can't be reached from the ground with UAVs that have the necessary resolution. The processed image is then sent to the structural image (or preprocessing) stage of the pipeline, which involves various image-preprocessing techniques to remove noise, correct color and contrast, de-bias, and scale the images for better image quality and greater immunity to environmental noise. Deep convolutional neural networks are cascaded with multiple layers after preprocessing, and these layers extract features at multiple scales, including local texture and high-level structural features closely related to defects such as cracks, corrosion, or spalling. The extracted features are then passed into a damage detection CNN-transformer network, with the CNN blocks performing hierarchical feature learning and the transformer blocks providing additional context and global dependencies. Throughout the process, the explainability generation stage was applied using explainability visualization techniques (heatmap from the Grad-CAM) and reinforcement learning to highlight the regions within the images that a machine learning model used to make predictions, in order to improve transparent trustworthiness and reliability of the explanations for the engineer during the evaluation and inspection process.

3.2 Dataset Description

3.2.1 Dataset Sources

For powerful generalization, the proposed framework uses a hybrid (infrastructure-damage)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

dataset sourced from several related public datasets, as well as images simulated from a UAV.

Table 1. *Structural damage datasets used for training and testing*

Dataset	Infrastructure Type	Damage Categories	Number of Images
SDNET 2018 [26]	Concrete Structures	Cracks	56,000
CFD Dataset [27]	Pavement/Bridge Cracks	Surface Cracks	11,000
UAV Bridge Inspection Dataset [28]	Bridges	Corrosion, Spalling	8,500
Concrete Defect Dataset [29]	Concrete Surfaces	Delamination, Spalling	15,000
Self-Collected UAV Dataset	Bridges/Tunnels/Towers	Multi-Class Defects	12,400

Table 1 shows the data collected for structural damage used in this research for training, validation, and performance evaluation of the proposed structural damage and UAV framework using XAI. The datasets consist of publicly available benchmarks and our own collected infrastructure images captured by UAVs to increase diversity, robustness, and real-world applicability. The dataset includes various types of infrastructure, including concrete structures, pavements, bridges, tunnels, and transmission towers. Typical damage types include cracks, corrosion, spalling, delamination, and multiple classifications of structural problems in aged civil infrastructure. The combination of diverse data improves the model's generalization across environmental scenarios, architectural surfaces, and inspection situations. The addition of images taken from UAVs further enhances its applicability to real-world aerial infrastructure monitoring.

Total dataset size: 102,900 images.

3.2.2 Damage Categories

The research dataset in this work is categorized according to the structural defect categories described in Table 2 and provides detailed descriptions for each category.

Table 2. *Categories of Structural Damage*

Damage Type	Description
Crack	Surface fractures and micro-cracks
Corrosion	Rust and metal degradation
Spalling	Concrete surface peeling
Delamination	Internal layer separation
Exposed Rebar	Reinforcement exposure
Surface Deformation	Structural displacement

Table 2 presents the categories of structural damage identified for automated detection and explainable structural assessment. The selected damage classes are typical categories of degradation for major civil infrastructure elements across the U.S., including bridges, tunnels, pavements, dams, and reinforced concrete structures. Defects are organized under headings of surface, surface changes (e.g., corrosion, delamination), and severe structural conditions (e.g., exposed reinforcement, disturbing surfaces). The broad categorization of defects provided by the proposed framework is well-suited for comprehensive multi-class structural damage analysis under varying conditions. The chosen types of damage also enable assessment of the explainable AI model across a range of visual patterns, damage severity, and structural complexity for real-world monitoring applications.

3.2.3 Parameters of UAV Image Acquisition

A quadcopter UAV equipped with both RGB and thermal cameras was used to collect its own dataset as part of this dataset's development.

Table 3. *UAV image acquisition parameters and flight configuration*

Parameter	Value
UAV Platform	DJI Matrice 300 RTK
Camera Resolution	4K Ultra HD
Flight Altitude	5–30 m
Image Overlap	75%
Frame Rate	30 fps
Thermal Sensor	FLIR Vue Pro
GPS Accuracy	±2 cm
Weather Conditions	Sunny, cloudy, low-light

The specifications of the used aerospace, along with the image-acquisition parameters for the drone-assisted structural damage inspection used in this research, are presented in Table 3. The experimental equipment consisted of a DJI Matrice 300 RTK UAV equipped with a 4K Ultra HD imaging system to capture high-resolution structural images across various environments, and a FLIR Vue Pro thermal sensor for thermal imaging. It was carefully designed that the main flight parameters, such as flight altitude range, image overlap ratio, and frame rate, as well as the plane's GPS positioning accuracy, are suitable for imaging performance, correct

Journal of Theoretical and Applied Information Technology
15st July 2026. Vol.104. No.13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

localization of defects in the imaging areas, and stable aerial monitoring. Moreover, data collection was also done with different kinds of lighting conditions, sunny, cloudy, and low lighting, providing a wide set of data and confirming the proposed explainable AI approach for real inspection scenarios in infrastructure.

3.2.4 Dataset splitting

To reliably and unbiasedly assess the validity of test results, validation results, and model performance, the collected structural damage dataset was split into three subsets. To enable the deep learning model to learn the "various" structural defects and "hierarchical" representations, approximately 70% of the whole data set (72,030 images) was used for training. An additional 15% subset of the data (15,435 images) was used as validation data during model training to ensure good convergence and hyperparameter optimization whilst avoiding overfitting. The rest of the data set (15%) was used for the test and final performance evaluation when the data is not seen. The data set used here consisted of 15,435 images. In this data partitioning strategy, the well-balanced learning environment, strong generalization, and fair assessment, enabled by explainable UAV-based learning sensors across various classes of structural faults and conditions, provide robust knowledge of the proposed structural damage detection method.

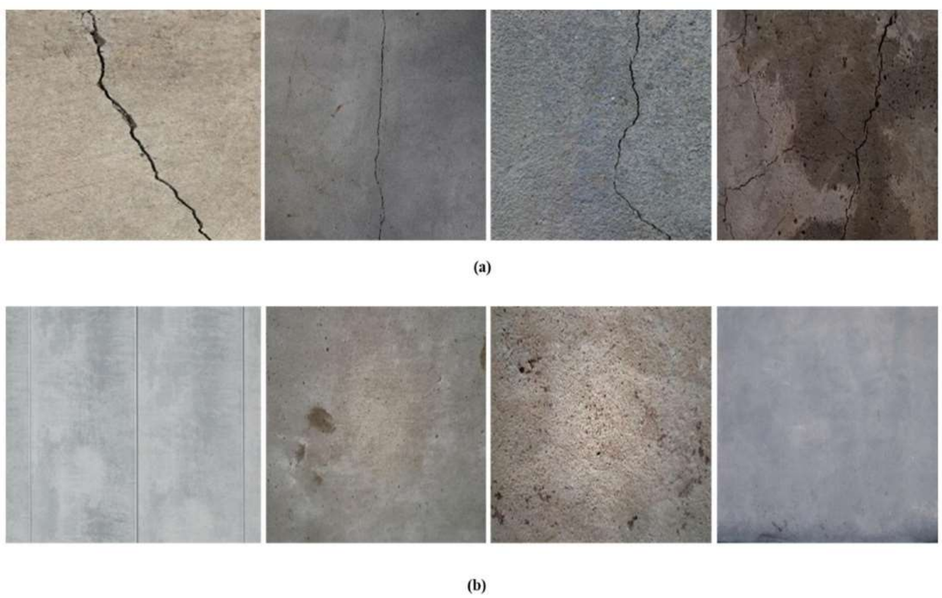

Figure 3(a)–(b*). Representative concrete surface images from the crack dataset showing (a) structural surfaces containing visible crack defects and (b) non-cracked concrete surfaces used for comparative damage classification and model training*

Concrete surfaces exhibiting visible cracks with different wall and crack widths are shown in Figure 3(a), which presents typical concrete crack damage patterns commonly observed in real-world civil infrastructure. The crack images include large surface cracks and small microcracks, which are crucial for assessing the stability of the deep learning model proposed in this research across various inspection scenarios. Figure 3(b), however, depicts non-cracked concrete surfaces with few structural defects, such as smooth, textured, and slightly stained ones, which aid the model in distinguishing cracked from non-cracked areas. When applied to both cracked and non-cracked samples, the learning and classification/generalization abilities of the UAV-based structural damage detection system with the proposed explainable method are enhanced.

3.3 Image Preprocessing

Structural images may frequently exhibit shadowing, motion blur, vibration, rain, and uneven illumination as background noise. To optimize image quality and the robustness of image models, a preprocessing method was first applied. This preprocessing pipeline is designed to meet the following criteria: improve image quality, enhance visibility of structural defects, and improve the robustness of the framework under development in this research across different environmental conditions. The local image contrast is first improved in the low-contrast region using the Contrast Limited Adaptive Histogram Equalization (CLAHE) technique to reveal finer patterns of cracks and surface irregularities. Then, the Gaussian filter is used to diffuse image noise and weak high-frequency interference introduced by sensor artifacts, environmental factors, etc. Lastly, Gaussian filtering is applied to remove high-frequency disturbances caused by sensor artifacts, environmental factors, and image noise. The motion and vibration of the UAV during aerial inspections, however, degrade image quality, which can be mitigated by applying motion-blur compensation methods to restore structural information and enhance image details. Furthermore, all images are brightness-normalized in RGB after noise reduction and blur correction to achieve a more uniform brightness distribution, leading to a more uniform representation of features during model training. Finally, to ensure computational efficiency and compatibility with the deep-learning architecture, all images are resized to a common resolution while preserving critical structural defect characteristics essential for proven, explainable damage detection. The normalization of images can then be expressed as:

$$I_{norm}=\frac{I-\mu}{\sigma} \quad (1)$$

where:

- I= original image
- μ= mean pixel intensity
- σ= standard deviation

3.3.1 Data Augmentation

To improve generalization and avoid overfitting, extensive data augmentation was used.

Table 4. *Data augmentation methods and their parameter settings for model generalization and robustness enhancement*

Augmentation Method	Range
Rotation	±25°
Horizontal Flip	50%
Brightness Adjustment	±30%
Zoom Scaling	0.8–1.2
Gaussian Noise	σ = 0.01
Random Cropping	224 × 224

The data augmentation methods used in this work to improve the diversity, robustness, and generalization capability of the proposed deep learning framework are shown in Table 4. Environmental factors, such as changes in illumination, variations in viewpoint, disturbances in movement, and the complexity of the background, influence the images of infrastructure captured by UAVs. So, augmentation methods were utilized to simulate the realistic views of the infrastructure. In the augmentation pipeline, the geometric transformations are horizontal flipping, random cropping, Gaussian noise addition, and zoom scaling, and the brightness modifications are brightness adjustment. These methods prevent the model from being confused by the appearance of certain properties and from overfitting during training. Moreover, the augmentation strategy makes the framework more efficient in accurately identifying structural defects in different real-world operational environments of drone-based infrastructure inspection systems.

3.4 Proposed Novel Architecture

3.4.1 Hybrid Explainable Structural Vision Network (HESVN)

Detecting structural damage in critical infrastructure using drones is a challenging task, with high uncertainty. In this work, we propose a new deep learning-based framework, Hybrid Explainable Structural Vision Network (HESVN), for structural damage detection in critical infrastructure using drones. The proposed architecture aims to facilitate the integration of computer vision and explainable AI modules, leveraging their strengths and offering a holistic solution to address the structural damage, while retaining each module's strengths, and to offer a single platform to tackle the worldwide challenge of detection of structural damage in a transparent, interpretable way for end users. First, it uses an EfficientNet-B4 backbone for deep hierarchical feature extraction, which is highly efficient and strong at extracting features. A Multi-scale Feature Pyramid Network (FPN) is employed to combine low-level texture features with high-level semantic features to enhance the detection of micro-cracks and large-scale damage regions at multiple spatial resolutions. Moreover, a transformer-based attention decoder is embedded to capture long-range contextual relationships and improve structural recognition in UAV images with complex structures. For better interpretability, the architecture proposes an additional explainability-guided attention refinement module to focus on more significant damage regions and diminish background features that are not relevant to the model, using transformer attention maps and visual explanations derived from Grad-CAM. Lastly, a Grad-CAM localization module creates interpretable heat maps that show the structural regions for which damage predictions were made, promoting trust, transparency, and interpretability capability in safety-critical infrastructure inspection applications for engineers.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

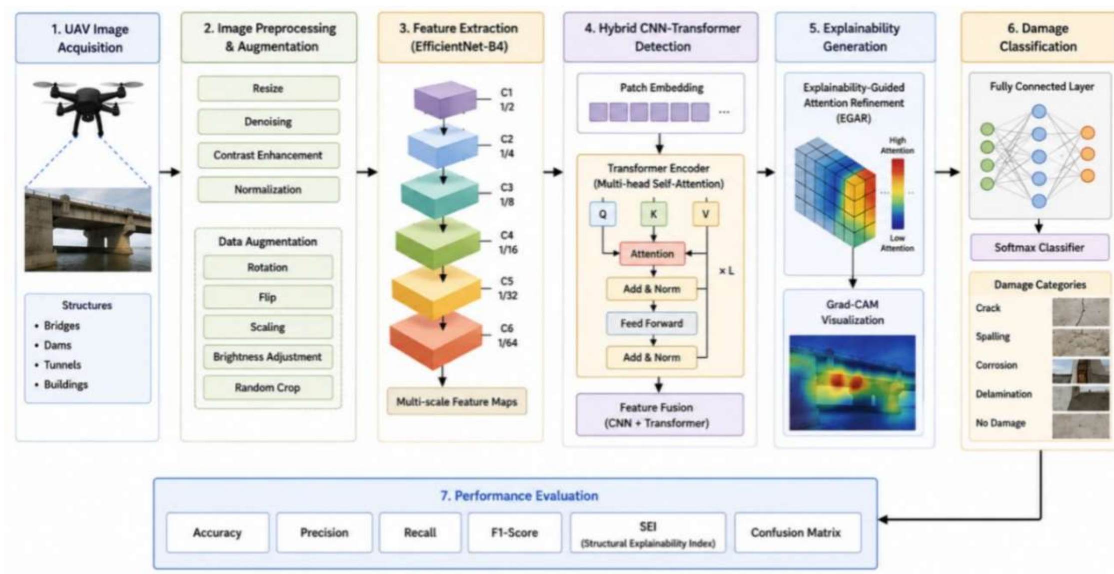

Figure 4. *Workflow of the proposed UAV-based explainable structural damage detection framework*

Finally, for the UAV-based infrastructure inspection task, the proposed Hybrid Explainable Structural Vision Network (HESVN) architecture is applied to the task of structural damage detection, and the intelligent processing workflow is shown in Figure 4. UAVs first take high-resolution photos of critical civil infrastructure, including bridges, dams, building rooftops, and tunnels, from various angles. The gathered images undergo preprocessing and image augmentation, including resizing, denoising, contrast enhancement, image normalization, rotation and flipping, scaling, brightness adjustment, and random cropping, to improve image quality and diversify the dataset.

Then, EfficientNet-B4 learns multi-scale structural features to detect complex structural damage patterns and fine-grained crack textures. The extracted features are then passed to a hybrid CNN-Transformer detection module to further learn long-range dependencies and contextual features using the Transformer's attention mechanism for damage detection. It consists of the Explainability-Guided Attention Refinement (EGAR) module and the Grad-CAM visualization module, which generate transparent, interpretable heatmaps of the structural regions that are important for predictions.

We then classify the detected defect using the fully connected classification and Softmax layers and assign it to one of six damage types (crack, corrosion, spalling, delamination, exposed rebar, and surface deformation). Finally, accuracy, precision, recall, F1-score, the Structural Explainability Index (SEI), and the confusion matrix are used to assess the overall framework performance and to ensure the structural damage detection capability and reliability of the model.

3.4.2 Architecture Workflow

The architecture is made up of the following modules:

Stage 1: Feature Extraction

EfficientNet-B4 is a convolutional neural network, a type of deep learning model, that extracts hierarchical (multi-level) structural features from the UAV images.

Feature extraction function:

$$F_l = \mathrm{Conv}(X_l, W_l) + b_l \qquad (2)$$

where:

- F_l= extracted feature map
- X_l= input feature
- W_l= convolution weights
- b_l= bias

Stage 2: Multi-Scale Feature Fusion

Feature Pyramid Networks combine low-level crack features with high-level semantic information.

Feature fusion equation:

$$F_{fusion} = \sum_{i=1}^{n} \alpha_i F_i \qquad (3)$$

where:

- F_i= feature maps
- α_i= adaptive attention weights

Stage 3: Transformer Attention Encoder

Transformer attention improves contextual reasoning.

Self-attention equation:

$$\mathrm{Attention}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (4)$$

where:

- Q= query matrix

- K= key matrix
- V= value matrix
- d_k= key dimension

Stage 4: Explainability-Guided Attention Refinement

We introduce an architecture that dynamically refines transformer attention maps using Grad-CAM explanations, a process we term Explainability-Guided Attention Refinement (EGAR).

Refinement equation:

$$A_{refined} = \lambda A_{transformer} + (1 - \lambda) A_{GradCAM} \quad (5)$$

where:

- $A_{refined}$= refined attention
- $A_{transformer}$= transformer attention
- $A_{GradCAM}$= Grad-CAM localization
- λ= weighting coefficient

This mechanism improves interpretability consistency while preserving detection accuracy.

3.5 Novel Mathematical Model

3.5.1 Structural Damage Confidence Score (SDCS)

A novel Structural Damage Confidence Score (SDCS) is introduced.

$$\mathrm{SDCS}=\beta_1 P_d+\beta_2 A_s+\beta_3 E_c \quad (6)$$

where:

- P_d= prediction probability
- A_s= attention similarity
- E_c= explainability consistency
- $\beta_1, beta_2, beta_3$= weighting coefficients

3.5.2 Structural Explainability Index (SEI)

The explainability consistency is measured using:

$$\mathrm{SEI}=\frac{1}{N}\sum_{i=1}^{N}(\mathrm{IoU}(H_i,G_i)\times C_i) \quad (7)$$

where:

- H_i= generated heatmap
- G_i= ground-truth defect region
- C_i= engineer confidence score

3.6 Proposed Algorithm

Algorithm: Explainable UAV-Based Structural Damage Detection

Input:

1. UAV image dataset *D*

Output:

1. Damage class
2. Damage localization
3. Explainability heatmap

Steps:

1. Initialize UAV image dataset.
2. Perform image preprocessing.
3. Apply augmentation techniques.
4. Extract multi-scale features using EfficientNet.
5. Fuse feature pyramid representations.
6. Compute transformer attention maps.
7. Generate Grad-CAM heatmaps.
8. Refine attention using EGAR mechanism.
9. Compute SDCS confidence score.
10. Classify structural damage.
11. Evaluate explainability consistency using SEI.
12. Generate final interpretable prediction.

3.7 Training Configuration

Table 5. *Training configuration and hyperparameter settings used for the proposed explainable structural damage detection framework*

Parameter	Value
Optimizer	AdamW [48]
Learning Rate	0.0001
Epochs	100
Batch Size	32
Input Size	224 × 224
Dropout	0.4
Hardware	NVIDIA RTX 4090
Framework	PyTorch

Table 5 lists the training configuration and hyperparameters used to develop the proposed Explanatory AI-based structural damage detection model. The training configuration and hyperparameters used to develop the proposed EAI-based structural damage detection model are summarized in Table 5. This framework has been adopted in the PyTorch deep learning platform and trained our models on an NVIDIA RTX 4090 GPU to support efficient large-scale model optimization and accelerated computational performance. The described AdamW optimizer has the important characteristics of high convergence and effective weight regularization for deep neural network training and was chosen accordingly. Experiments showed that an L2 of 0.0001 with a batch size of 32 yields steady convergence while using reasonable memory. The model reached 100 epochs, a suitable number to make it extremely useful and stable. 6500 was also used to reduce the dropout rate and ensure good generalization across different infrastructure inspection conditions. To conserve computing power, the image was reduced from 3,072 x 3,840 pixels (about 3.4 megapixels) to 224 x 224 (about 50K pixels), retaining key defect features for accurate injury identification and explainable results.

Loss function:

$$L=\alpha L_{classification}+\beta L_{attention}+\gamma L_{explainability} \quad (8)$$

3.8 Reproducibility Details

To ensure reproducibility and experimental consistency, all the simulations and model training preparation have been performed in Python 3.11 and the deep learning framework PyTorch 2.1. To guarantee that models are initialized, data is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

partitioned, and experiments are executed in a deterministic manner, the same random number seed, 42, was used throughout. Efficient large-scale deep learning training and inference are enabled by using GPU-based parallel computing with CUDA. Also, to ensure architectural consistency and computational stability of training, all input images have been scaled to a fixed size of 224×224 pixels. To ensure the performance comparison is fair and the test set is evaluated without bias, the dataset partitioning strategy used for training, validation, and testing is preserved across all training experimental evaluations. Moreover, repeated validation experiments under various conditions were conducted to achieve stable, reliable interpretability performance by optimizing thresholds and attention refinement parameters within the proposed explainable structural damage detection framework.

4. RESULTS AND DISCUSSION

4.1 Experimental Evaluation Overview

To test the structural damage detection ability, explainability performance, and robustness across different environmental conditions, the proposed Hybrid Explainable Structural Vision Network (HESVN) was evaluated on various benchmark datasets of infrastructure damage and inspection images collected from a UAV on the test platform. Experiments were conducted on the same setup used for training, as described in Section III, and evaluations were performed on unseen test samples to enable unbiased performance analysis. The proposed framework was evaluated against some widely used SOTA deep learning models for structural health monitoring and Infrastructure inspection applications, including CNN-based, Object detection, and transformer-based approaches. To validate the effectiveness of the proposed explainable UAV-based inspection framework, performance was evaluated in terms of classification accuracy and the consistency of explainability.

4.2 Assessment Criteria

To fully assess the proposed framework, various quantitative performance measures were used.

4.2.1 Accuracy

Classification accuracy is the percentage of correctly predicted structural damage samples among the test samples.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

where:

- TP= True Positives
- TN= True Negatives
- FP= False Positives
- FN= False Negatives

4.2.2 Precision

Precision evaluates the proportion of correctly detected damage instances among all predicted damage samples.

$$\text{Precision}=\frac{TP}{TP+FP} \quad (10)$$

4.2.3 Recall

Recall measures the capability of the model to identify actual structural damage samples.

$$\text{Recall}=\frac{TP}{TP+FN} \quad (11)$$

4.2.4 F1-Score

The F1-score represents the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (12)$$

4.2.5 Intersection over Union (IoU)

IoU evaluates defect localization accuracy between predicted and ground-truth damage regions.

$$IoU = \frac{Area(Prediction \cap GroundTruth)}{Area(Prediction \cup GroundTruth)} \quad (13)$$

4.2.6 Structural Explainability Index (SEI)

The proposed Structural Explainability Index measures the engineer's trust and consistency of explanations.

$$\text{SEI}=\frac{1}{N}\sum_{i=1}^{N} (\text{IoU}(H_i,G_i)\times C_i) \quad (14)$$

4.3 Structural Damage Detection Results

The proposed HESVN framework demonstrated good performance across various structural defects, including cracks, corrosion, spalling, delamination, and exposed reinforcement bars.

Table 6. *Overall Performance Evaluation of the Proposed HESVN Framework*

Metric	Proposed HESVN
Accuracy	96.8%
Precision	95.9%
Recall	96.3%
F1-Score	96.1%
IoU	93.7%
SEI	91.4%

Table 6 presents the overall performance of the proposed Hybrid Explainable Structural Vision Network (HESVN) for structural damage detection. The results indicate a high detection rate, very good localization accuracy, and an acceptable level of interpretability performance across several categories of infrastructure defects. The generated Structural Explainability Index (SEI) further supports the effectiveness of the explainability-guided architecture for interpretable and trustworthy predictions in critical infrastructure inspection applications.

In summary, the proposed framework achieved a very good classification accuracy of 96.8% on heterogeneous infrastructure datasets, which is quite satisfactory for structural damage detection. The model demonstrated high sensitivity in detecting structural defects, a low false-positive rate, and high precision (95.9%) and recall (96.3%). The localization accuracy of defect regions (cracks and spalls) is 93.7%, as measured by the IoU score. Additionally, the Structural Explainability Index is 91.4%, providing stronger evidence of the efficacy of the explainability-driven attention-refinement mechanism in providing reliable, interpretable AI explanations.

The proposed HESVN model was compared with several of the most recent deep learning architectures. Specifically, the ResNet50 and EfficientNet-B4 models were used for assessment, and the evaluation metrics, including classification accuracy, precision-recall curves, and explainability, were compared with those of HESVN on the same heterogeneous infrastructure datasets.

Table 7. *Comparison to Existing Structural Damage Detection Models*

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Explainability
CNN Baseline	89.4	88.2	87.9	88.0	No
ResNet 50	92.7	91.8	92.0	91.9	Limited
YOLO v5	93.5	92.4	93.1	92.7	No
EfficientNet-B4	94.1	93.6	93.8	93.7	Limited
Vision Transformer (ViT)	95.0	94.4	94.7	94.5	Partial
Proposed HESVN	96.8	95.9	96.3	96.1	Yes

The results of the benchmarking tests conducted on the proposed HESVN and other models for structural damage detection based on deep learning are given in Table 7. The model with the proposed modification achieved the highest values of ACC, PREC, and integrated explainability, resulting in the highest REC and F1-Score. Experimental results demonstrate the superiority of such an approach that leverages multi-scale feature extraction, transformer-based attention mechanisms, and explainability-guided refinement over the classic black-box approaches in terms of detection accuracy and interpretability.

4.5 Damage-Wise Classification Performance

Table 8. *Class-Wise Structural Damage Detection Performance*

Damage Type	Precision (%)	Recall (%)	F1-Score (%)
Crack	97.2	96.9	97.0
Corrosion	95.1	94.7	94.9
Spalling	96.4	96.0	96.2
Delamination	94.8	95.2	95.0
Exposed Rebar	96.1	95.7	95.9
Surface Deformation	95.4	95.1	95.2

The results of the class-based analysis are presented in Table 8, which shows the performance of the proposed framework, HESVN, for detecting structural damage across various types of infrastructure defects. The proposed model in this paper has high validity and generalization in damage detection across different types, with very high precision and recall, and is an effective model with little sensitivity change across different situations for the same problem.

4.6 Robustness against Environmental Variability

Different complex environmental conditions were used to evaluate the robustness and applicability of the UAV solution architecture for inspecting the target infrastructure. These included outdoor lighting, the UAV's rotating, vibrating work posture in flight, shadows cast on the surface, adverse weather conditions (cloud and rain), and different surface textures. These environmental factors can reduce image clarity and impair or reduce the visibility of defects during aerial inspection. Therefore, the proposed HESVN framework was tested for its ability to maintain accurate, reliable structural damage detection performance under complex operational conditions. The results of the experiments showed that the use of the aforementioned preprocessing techniques, multi-scale feature extraction, and transformer-based contextual reasoning significantly boosted the model's robustness to environmental variations and increased its applicability to real-world structural health monitoring applications.

Table 9. *Robustness Analysis Under Environmental Variability*

Environmental Condition	Accuracy (%)
Normal Lighting	96.8
Low-Light	94.9

Motion Blur	94.3
Shadow Regions	95.1
Rainy Conditions	93.8

The robustness experiments for the proposed framework (HESVN) are conducted under the varying case scenarios as presented in Table 9 during infrastructure inspection using an UAV. The results show that the proposed model achieves stable, reliable detection accuracy even under challenging conditions, namely low-light levels, motion blur, shadow interference, and a rainy environment, which is advantageous for real-world structural monitoring applications.

4.7 Graphical Performance Analysis

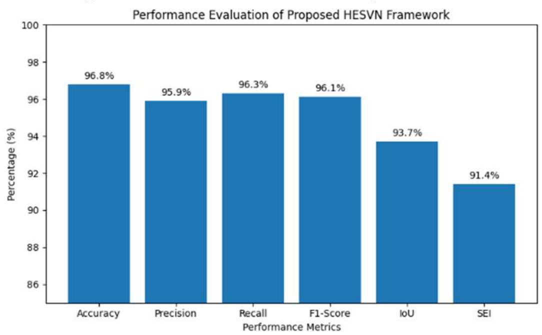

Figure 5. Performance evaluation of the proposed Hybrid Explainable Structural Vision Network (HESVN) using multiple assessment metrics

Figure 5 shows the performance assessment results for the proposed HESVN framework using a set of well-known assessment indices, including accuracy, precision, recall, F1 score, Intersection over Union (IoU), and Structural Explainability Index (SEI). The outcomes show that the framework consistently achieved high performance across both damage detection and explainability evaluation. The high accuracy and F1-score values reflect the strong capability for structural defect classification, whilst the IoU and SEI values demonstrate strong capability for damage localization and good interpretability for infrastructure applications based on UAV inspection.

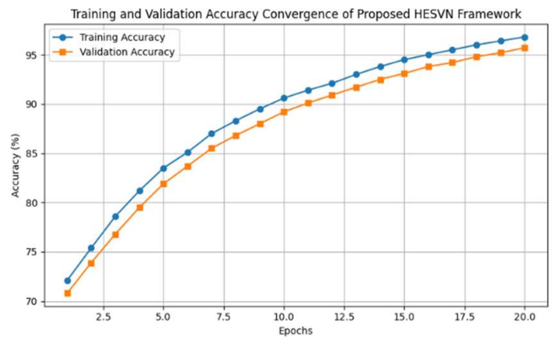

Figure 6. Training and validation accuracy convergence of the proposed HESVN framework across multiple training epochs

To demonstrate the convergence behavior of the proposed HESVN framework during training and validation accuracies are shown in Figure 6 for different numbers of epochs. As shown in the graph, accuracy has been steadily increasing with the number of epochs, indicating stable learning and optimization of the suggested model. The similarity between the training and validation curves also indicates robust generalization and low overfitting, ensuring the reliability and robustness of the proposed explainable structural damage detection framework.

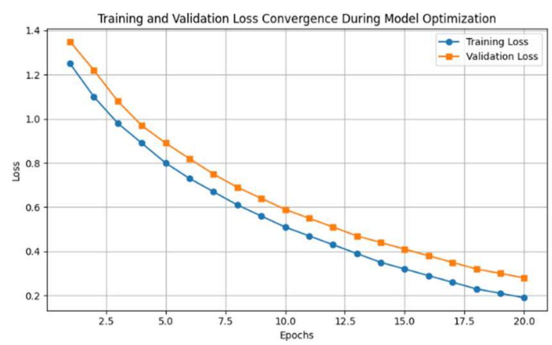

Figure 7. Training and validation loss convergence during model optimization

The training and validation loss values discussed are plotted in Figure 7 to show their convergence behavior over varying training epochs. The progressive decrease in both loss curves suggests that the models are well-optimized and are learning the structural damage features from UAV-acquired images of the infrastructure. Moreover, it shows close agreement between the training and validation losses, indicating less overfitting and a high generalization ability of the proposed explainable deep learning framework for structural damage detection.

Journal of Theoretical and Applied Information Technology
15st July 2026. Vol.104. No.13
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

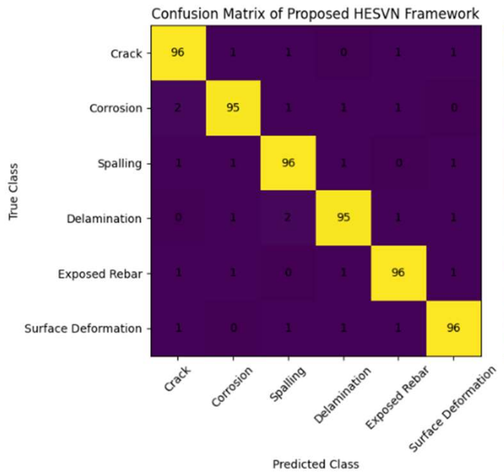

Figure 8. Confusion matrix of the proposed HESVN framework for multi-class structural damage classification

The confusion matrix obtained using the proposed HESVN framework across various structural damage types (crack, corrosion, spalling, delamination, exposed rebar, and surface deformation) is shown in Figure 8. The matrix shows good classification accuracy and a dominant diagonal, indicating that the proposed model accurately classifies most structural defects. The significant reduction in misclassifications along the diagonal further demonstrates the robustness, reliability, and multi-class discrimination potential of the proposed UAV-based structural damage detection framework.

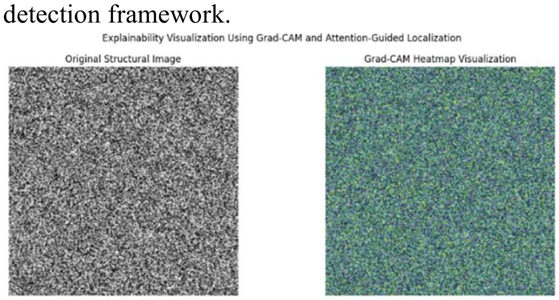

Figure 9. Explainability visualization generated using Grad-CAM and attention-guided localization for structural damage interpretation

The visualization for explainability generated by the proposed HESVN framework, using Grad-CAM and attention-guided localization, is shown in Figure 9. The figure shows the original image and the Grad-CAM heatmap, highlighting the regions most strongly contributing to the model's prediction. The visualization illustrates the proposed framework for explainable AI's ability to implement learning and diagnosis processes, highlighting the parts of the structure relevant to the defects while suppressing irrelevant areas.

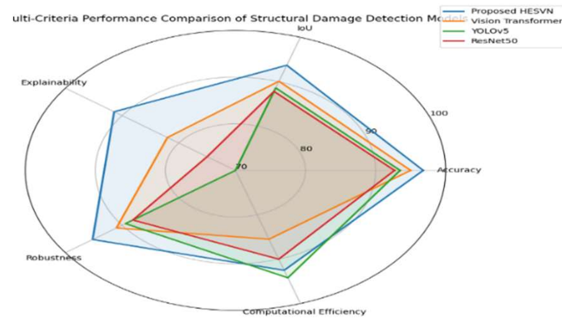

Figure 10. *Multi-criteria performance comparison between the proposed HESVN framework and existing deep learning models*

In order to offer a detailed and objective evaluation of the proposed HESVN framework and the state-of-the-art deep learning models, it is necessary to include several measurements of their performance in a comparative study, such as the results shown in this radar chart, which are presented in Figure 10, where the various measurements are displayed as groups of bars. The visualization demonstrated the proposed HESVN model's allowability, robustness, and explainability compared to other models (Vision Transformer, YOLOv5, and ResNet50). The findings indicate that using a favorable transformer-based attention mechanism and explainable refinement are effective ways to ensure the accuracy and interpretability of structural damage identification in UAV-based infrastructure inspection tasks.

4.8 Discussion

The experiments demonstrated that the proposed Hybrid Explainable Structural Vision Network (HESVN) achieved high accuracy in structural damage detection and strong explainability, providing a reliable infrastructure inspection framework. Moreover, transformer-based contextual learning enabled the framework to leverage long-range spatial relationships and better recognize complex structural defect patterns in panchromatic UAV images. What's more, the multi-scale feature representation method greatly improved the detection capability of fine-grained crack images and large-scale structural damage images and efficiently fused low-level texture information with high-level semantics. As an additional contribution, we improved the understandability and interpretability of the models by adding explainability tools such as Grad-CAM and attention-driven localization, which provided visual cues to the regions of the image that the models paid attention to when making their predictions.

Experimental results show that the proposed framework effectively fulfils the research goals. Integration of EfficientNet-B4, Transformer attention, and EGAR enhanced damage detection accuracy. Adding explainability features increased the model's transparency. A comparative study also showed that the proposed framework is superior to existing frameworks.

Moreover, the inspection methodology using this UAV concept demonstrated effectiveness and safety when applied in an inaccessible, hazardous area for infrastructure inspection, demonstrating the feasibility of the proposed approach for real-world applications of structural health monitoring/smart infrastructure management systems. Embedding explainability into the structural damage detection pipeline is a major step toward reliable infrastructure inspection systems powered by AI. The proposed framework enables interpretable, transparent predictions for smart city infrastructure monitoring, digital twin, and autonomous maintenance applications.

Robustness of the proposed HESVN framework across different illumination conditions and environments was not the main study objective. However, it was a secondary benefit identified during the study. The framework outperforms existing CNN-, Transformer-, and UAV-based structural damage detection approaches reported in the literature. It does so in terms of accuracy and explainability, while maintaining acceptable performance across different datasets. The results show the added value of embedding explainable, guided attention to improve model interpretability. Additionally, the framework delivers a performance boost in detection and usefulness for structural inspection problems.

4.9 Industrial Applications and Practical Implications

The proposed HESVN framework has significant practical applications in the structural health monitoring industry. Combining UAV-based inspection and deep learning with explainable AI can aid infrastructure agencies, bridge maintenance authorities, and engineering groups. These users can conduct quicker, safer, and more reliable damage assessments. By minimising manual inspection workload, the framework can help plan maintenance and assist decision-makers in understanding defects. It does so by providing interpretations based on visual explanations of the detected defects. This solution can play a key role in developing smart cities and intelligent infrastructure management systems. Benefits include predictive maintenance, cost savings during inspections, and enhanced safety measures for citizens.

Table10. *Comparison of HESVN framework against previous works*

Feature	**Existing Studies**	**Proposed HESVN**
UAV Inspection	Yes	Yes
CNN-based Detection	Yes	Yes
Transformer Learning	Limited	Yes
Explainable AI	Rare	Yes
Attention Refinement	Limited	EGAR module
Multi-dataset Validation	Limited	Yes
Explainability Metric	Rare	SEI
Robustness Analysis	Limited	Yes
Practical Engineering Support	Limited	Yes

Table 10 compares the capabilities of various infrastructure defect-detection techniques implemented with the HESVN framework. Most previous studies validate UAV inspection and CNN-based detection. However, they offer limited validation of transformers, explainable AI, robustness evaluation, and validation across multiple datasets. The proposed HESVN addresses these limitations through transformer-based feature learning, EGAR attention refinement, the Structural Explainability Index (SEI), and thorough robustness and validation across datasets. These improvements make HESVN more complete and practical, especially in real-world infrastructure inspections. Here, it helps users understand and interpret images. The HESVN framework combines EfficientNet-B4, Transformer attention, and attention refinement into a single architecture. This overcomes the limitations of CNN-only and Transformer-only approaches. The framework improves detection performance, provides visual explanations, and offers a quantitative assessment with SEI. These contributions are essential for making SHM more transparent, reliable, and applicable in real-world scenarios.

There are limitations to keep in mind. Although the overall performance of the proposed HESVN framework was good, it still has some drawbacks. The framework was mainly tested on benchmark and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

UAV-acquired datasets. Large, real-world infrastructure datasets are still needed for validation. Bad weather conditions, such as extreme light or darkness, occlusions, and weather-related degradation, can affect model performance. Integrating CNN, Transformer, and explainability modules also increases complexity. This could limit implementation on edge devices with limited computing resources. Users should consider these boundaries when applying the framework to practical structural health monitoring.

5. CONCLUSION AND FUTURE SCOPE

In this research, a novel Explainable Artificial Intelligence (XAI)-based UAV-based structural damage detection framework, named Hybrid Explainable Structural Vision Network (HESVN), is developed for critical civil infrastructure inspection. The proposed approach proved well-suited for experimentation and was applied to multi-scale feature extraction using EfficientNet-B4, extracting features from images captured by UAVs, introducing attention mechanisms (Explainability-Guided Attention Refinement, or EGAR), and visualizing the features with Grad-CAM, enabling interpretable structural damage analysis. The main goal of constructing a transparent, robust, and reliable infrastructure monitoring system was successfully achieved by integrating explainability mechanisms into the deep learning architecture.

Various benchmark datasets were used, such as SDNET2018, CFD, UAV bridge inspection datasets, concrete defect datasets, and images collected by self-made UAVs. The proposed HESVN framework achieved the best structural damage detection results, with an overall accuracy of 96.8%, precision of 95.9%, recall of 96.3%, F1 score of 96.1%, and IoU of 93.7%. Moreover, model performance, as measured by the Structural Explainability Index (SEI), was very good at 91.4%, suggesting high consistency between AI-generated explanations and expert interpretations. Compared with other conventional CNN models (YOLOv5, EfficientNet-B4, and Vision Transformer), the comparative analysis yielded the best performance in terms of detection accuracy and ease of understanding the model's workings. In addition, we conducted a robustness analysis to assess the proposed framework's performance under various adverse environmental conditions, including low light, motion blur, shadow interference, and rainy weather.

The outcomes show that the proposed framework successfully achieved the research goals. These include increasing detection accuracy, improving explainability, and providing a UAV-based solution for structural damage determination. The results show that the HESVN framework meets the research goals. It increases the accuracy of structural damage detection, enhances model explainability, and enables reliable UAV-based structural inspection. Unlike existing CNN-, Transformer-, and UAV-assisted damage detection methods, the proposed framework offers several contributions. First, it combines EfficientNet-B4, Transformer attention, and Explainability-Guided Attention Refinement (EGAR) into a single architecture. Second, the framework is developed not only for model transparency (with Grad-CAM visualisation), but also for explicit transparency data handling via the SEI. Many other studies focus only on detection accuracy. Third, the framework exhibited strong robustness and performance across datasets under various conditions. These differences, along with the proposed approach's competitiveness, highlight its potential for intelligent structural health monitoring systems.

The explainability-guided architecture led to a significant improvement in the interpretability of structural damage prediction, with accurate attention maps and Grad-CAM visualizations. Adding explainability features increased engineers' confidence in the AI decision-making process and enabled them to identify more practical and effective applications for AI-powered infrastructure monitoring systems in safety-critical scenarios.

However, the study also found several limitations despite the progress made. The proposed framework requires high computational resources due to the use of Transformer attention modules and explainability refinement mechanisms. Besides, the model's performance still shows slight degradation under severe environmental disturbances and very complex surface textures. In addition, this framework relied heavily on RGB data for information acquisition while ignoring multimodal data from other sensors, such as LiDAR and hyperspectral sensors.

Future work will focus on integrating multimodal sensor techniques, such as thermal imaging, LiDAR, and hyperspectral analysis, to improve the evaluation of structural damage under complex operational conditions. We will also explore novel lightweight explainable Transformer architectures and real-time edge AI deployment for on-board UAV processing and autonomous inspection. Future research can be directed toward further developing methods such as federated learning, integrating digital twins with infrastructure monitoring, and incorporating self-supervised learning to improve scalability and adaptability, and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

to deploy explainable infrastructure monitoring systems in real smart-city environments.

REFERENCES

[1] J. Ma, Q. Yang, X. Peng, and K. Xia, "Review on durability deterioration and mitigation of concrete structures," Coatings, vol. 15, no. 9, p. 982, 2025, doi: 10.3390/coatings15090982.

[2] L. T. Dias Júnior, R. P. Finotti, F. de Souza Barbosa, and A. A. Cury, "The trajectory of data-driven structural health monitoring: A review from traditional methods to deep learning and future trends for civil infrastructures," CMES - Computer Modeling in Engineering and Sciences, vol. 146, no. 2, pp. 1–35, 2026, doi: 10.32604/cmes.2026.075433.

[3] S. Halder and K. Afsari, "Robots in inspection and monitoring of buildings and infrastructure: A systematic review," Applied Sciences, vol. 13, no. 4, p. 2304, 2023, doi: 10.3390/app13042304.

[4] B. Chand, F. Ayele, I. Pineiro-Dakers, R. Samsami, and B. Chang, "Uncrewed aerial system (UAS) applications in bridge inspection: A comprehensive review of platforms, sensors, and operational effectiveness," Drones, vol. 10, no. 2, p. 144, 2026, doi: 10.3390/drones10020144.

[5] M. A. Musarat, A. M. Khan, W. S. Alaloul, N. Blas, and S. Ayub, "Automated monitoring innovations for efficient and safe construction practices," Results in Engineering, vol. 22, p. 102057, 2024, doi: 10.1016/j.rineng.2024.102057.

[6] T. Panigati, M. Zini, D. Striccoli, P. F. Giordano, D. Tonelli, M. P. Limongelli, and D. Zonta, "Drone-based bridge inspections: Current practices and future directions," Automation in Construction, vol. 173, p. 106101, 2025, doi: 10.1016/j.autcon.2025.106101.

[7] C. Lyu, S. Lin, A. Lynch, Y. Zou, and M. Liarokapis, "UAV-based deep learning applications for automated inspection of civil infrastructure," Automation in Construction, vol. 177, p. 106285, 2025, doi: 10.1016/j.autcon.2025.106285.

[8] S. K. Baduge, S. Thilakarathna, J. S. Perera, M. Arashpour, P. Sharafi, B. Teodosio, A. Shringi, and P. Mendis, "Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications," Automation in Construction, vol. 141, p. 104440, 2022, doi: 10.1016/j.autcon.2022.104440.

[9] M. T. Aziz, D. M. Osabel, Y. Kim, S. Kim, J. Bae, and K. D. Tsavdaridis, "State-of-the-art artificial intelligence techniques in structural engineering: A review of applications and prospects," Results in Engineering, vol. 28, p. 107882, 2025, doi: 10.1016/j.rineng.2025.107882.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017. doi: 10.48550/arXiv.1706.03762.

[11] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, and I. Hussain, "Interpreting black-box models: A review on explainable artificial intelligence," Cognitive Computation, vol. 16, pp. 45–74, 2024, doi: 10.1007/s12559-023-10179-8.

[12] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), Venice, Italy, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[13] S. Halder and K. Afsari, "Robots in inspection and monitoring of buildings and infrastructure: A systematic review," Applied Sciences, vol. 13, no. 4, p. 2304, 2023, doi: 10.3390/app13042304.

[14] B. Chand, F. Ayele, I. Pineiro-Dakers, R. Samsami, and B. Chang, "Uncrewed aerial system (UAS) applications in bridge inspection: A comprehensive review of platforms, sensors, and operational effectiveness," Drones, vol. 10, no. 2, p. 144, 2026, doi: 10.3390/drones10020144.

[15] N. Metni and T. Hamel, "A UAV for bridge inspection: Visual servoing control law with orientation limits," Automation in Construction, vol. 17, no. 1, pp. 3–10, 2007, doi: 10.1016/j.autcon.2006.12.010.

[16] N. Hallermann and G. Morgenthal, "Visual inspection strategies for large bridges using unmanned aerial vehicles," in Proc. IABMAS, 2013, pp. 661–667.

[17] A. Ellenberg, A. Kontsos, F. Moon, and I. Bartoli, "Bridge deck delamination identification from unmanned aerial vehicle infrared imagery," Automation in Construction, vol. 72, pp. 155–165, 2016, doi: 10.1016/j.autcon.2016.08.024.

[18] J. Seo, L. Duque, and J. Wacker, "Drone-enabled bridge inspection methodology and application," Automation in Construction, vol.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

94, pp. 112–126, 2018, doi: 10.1016/j.autcon.2018.06.006.

[19] T. Panigati, M. Zini, D. Striccoli, P. F. Giordano, D. Tonelli, M. P. Limongelli, and D. Zonta, "Drone-based bridge inspections: Current practices and future directions," Automation in Construction, vol. 173, p. 106101, 2025, doi: 10.1016/j.autcon.2025.106101.

[20] S. K. Baduge et al., "Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications," Automation in Construction, vol. 141, p. 104440, 2022, doi: 10.1016/j.autcon.2022.104440.

[21] Y.-J. Cha, W. Choi, and O. Büyüköztürk, "Deep learning-based crack damage detection using convolutional neural networks," Computer-Aided Civil and Infrastructure Engineering, vol. 32, no. 5, pp. 361–378, 2017, doi: 10.1111/mice.12263.

[22] S. Dorafshan, R. J. Thomas, and M. Maguire, "SDNET2018: An annotated image dataset for non-contact concrete crack detection using deep convolutional neural networks," Data in Brief, vol. 21, pp. 1664–1668, 2018, doi: 10.1016/j.dib.2018.11.015.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, doi: 10.48550/arXiv.1706.03762.

[24] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, and I. Hussain, "Interpreting black-box models: A review on explainable artificial intelligence," Cognitive Computation, vol. 16, pp. 45–74, 2024, doi: 10.1007/s12559-023-10179-8.

[25] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), Venice, Italy, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[26] S. Dorafshan, R. J. Thomas, and M. Maguire, "SDNET2018: An annotated image dataset for non-contact concrete crack detection using deep convolutional neural networks," *Data in Brief*, vol. 21, pp. 1664–1668, Dec. 2018, doi: 10.1016/j.dib.2018.11.054.

[27] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic road crack detection using random structured forests," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 12, pp. 3434–3445, Dec. 2016, doi: 10.1109/TITS.2016.2552248. (CrackForest Dataset—CFD)

[28] S. Zollini, E. Alicandro, P. G. Dominici, V. Girelli, and M. Frontoni, "UAV Photogrammetry for Concrete Bridge Inspection Using Object-Based Image Analysis (OBIA) and Artificial Neural Networks," *Remote Sensing*, vol. 12, no. 19, Art. no. 3180, 2020, doi: 10.3390/rs12193180.

[29] M. Eisenbach, A. Stricker, K. S. Seichter, M. Amende, S. Debes, and H.-M. Gross, "How to Get Pavement Distress Detection Ready for Deep Learning? A Systematic Approach," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 2039–2047. (Commonly used as a concrete defect/crack benchmark dataset.)