International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 15 Issue 3, March-2026

Deep Learning Architectures for Crack Width Measurement: A Review of Recent Advances and Emerging Trends

Kanchan Dhapekar[1], Dr. Divya Prakash[2]

[1,2]Department of Civil Engineering, Poornima University, Jaipur – 303905, Rajasthan, India

ABSTRACT

Structural Health Monitoring (SHM) has an important role to play in ensuring the safety and maintainability of civil infrastructure. The latest advancements in the field of deep learning have allowed to detect structural damage automatically with a high degree of precision; however, traditional models are often faced with the difficulty of extracting features of damage that matter and effectively processing multiple heterogeneous multimodal data types. Attention-based deep learning methods have thus proven to be formidable solutions, as they are able to apply an adaptive weighting of features, better localization as well as an effective multimodal fusion. The article provides an extensive review of attention processes that are used in SHM which include channel, spatial, temporal and self-attention. There is a systematic taxonomy of existing architectures and multimodal methods of data fusion are strictly investigated. It has addressed salient applications and current trends in performance then addressed key research issues including data paucity, domain shift, and computational overhead. Finally, emerging opportunities, such as the physics-informed attention, an edge deployment that is lightweight, inter-domain adaptive framework, etc., are highlighted to guide the development of the next-generation intelligent SHM systems.

Keywords: Structural health monitoring, attention mechanism, deep learning, multimodal fusion, transformer, crack detection, crack measurement.

INTRODUCTION

SHM is an essential field of study towards the protection of structural integrity and the time-span of civil structures, such as bridges and buildings, pavements, and offshore platforms. Traditional inspection methods are manual and their application is not always objective and are also not useful in identifying incipient damage. The recent history of sensing technology and artificial intelligence has led to an outburst in the desire to use deep learning in SHM systems.

The convolutional neural networks (CNNs), recurrent neural networks (RNNs) and hybrid versions have demonstrated good performance in automated identification of structural damage. However, these traditional deep learning models treat the attributes of all inputs on equal basis, a factor that can reduce sensitivity to minor anomalies, predispose it to environmental noise and limits interpretation. Attention mechanisms counteract these failures by preserving neural networks which emphasize features which are relevant to damage and inhibit irrelevant data. Attention-based models are found to have significant effects on robustness and reliability when joined with the multimodal sensor data in the operational SHM scenarios. The current article is a literature review on the recent advancements of attention-based deep learning in SHM and outlines the subsequent issue of research.

Deep learning: An overview

Machine learning (ML) is a subclass to artificial intelligence (AI). The ML algorithms are designed to develop trainable models that can absorb empirical or simulated data and produce predictive answers to future events [11]. In this paradigm, deep learning (DL) is viewed as a sophisticated feature-learning machine, as well as a material subdivision of ML. SHM models based on DL are usually developed in such that they learn autonomy on complex streams of data. These are the three main classes of ML -based SHM models: reinforcement, unsupervised, and supervised learning [12]. In the supervised learning, input is in labelled training data used to tune an ML model. The models that produce discrete or categorical output are called classifiers whereas models that produce continuous output are called regressors. The unsupervised approach to learning uses unlabelled data, forming clusters given no training program; some popular clustering algorithms include spectral clustering, k-means, partitioned clustering, hierarchical clustering, and so on.

During the recent few years, DL has become a revolution concept in SHM, responding to the growing problems that face modern civil infrastructure [13]. The rapidly growing sensor technologies that present vast and rich data has enhanced the ability of DL to discover saliency on its own. SHM has been enhanced greatly with the addition of capabilities through the fusion of DL with computer vision techniques, which allows both one-dimensional vibration and strain measurements along with four-dimensional RGB video streams to be examined, thus enhancing the performance of damage detectors. It is along with these developments, in conjunction with hardware additions and the ubiquitous nature of user-friendly frameworks that have made the implementation of DL in SHM more democratic and indispensable in the context of structural safety reinforcement [14]. Unlike the conventional choice of ML techniques that are often characterized by manual feature selection and engineering tasks, DL eases these limitations since it offers more efficient and scalable paradigm. DLSHM procedures support (deep neural networks) layers that allow full feature extraction and hierarchical representation of the raw input data [15]. Now, every DNN layer temporally becomes more representative of the input response, and thus, builds an end-to-end system without the necessity of human-crafted features. As a result, the regression, pattern classification, and feature extraction parameters are combined by optimization. This unified solution makes DL-based SHM able to face a wide array of problems without having to heavily rely on a priori knowledge related to the domain and hence its pivotal position in the discussion. The conceptual mind map of machine learning is illustrated in figure 1.

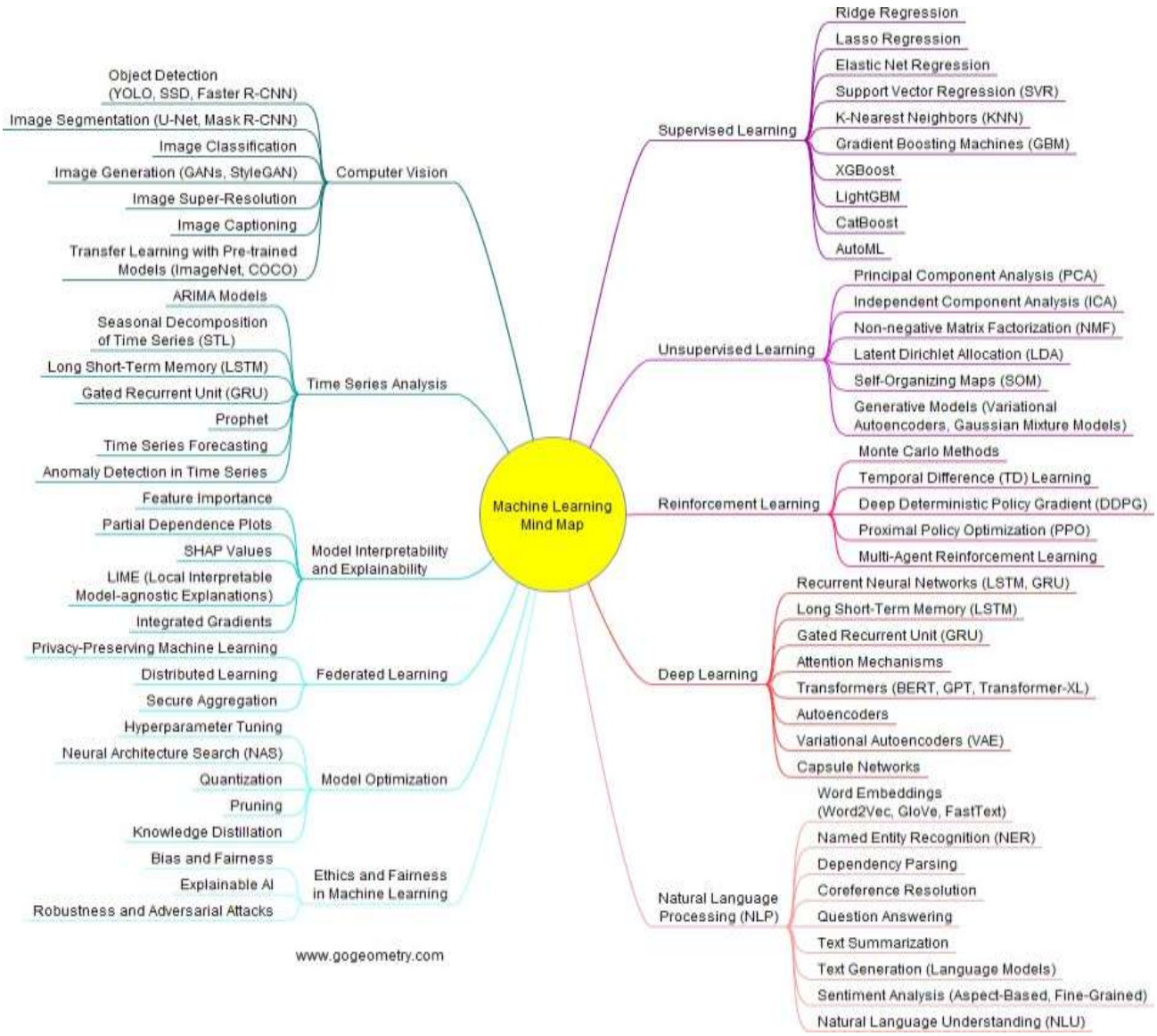


Figure 1. Machine Learning Algorithm mind map [16]

Structural Health Monitoring:

Structural Health Monitoring (SHM) is a significant field of study that is focused on the persistent evaluation of civil infrastructure safety, integrity, and performance, which capture bridges, structures, pavements, and offshore infrastructure. SHM helps to recognize the early signs of structural compromising and make nuanced maintenance decisions using an in-combination of sensor technologies and data-gathering systems as well as burdensome analysis

processes. Traditional inspection methods tend to be manual, tedious and vulnerable to human fallibility thus limiting their effectiveness in big examinations and in real-time applications. Developments in data-driven approach, specifically, deep-learning models have triggered the radical change in SHM by allowing automated feature discovery, enabling the accurate localization of damage, and assessment of the state directly on sensor and image-based measurements. Multimodal measurements are also growing and are now available in modern SHM systems and include vibration measurements, acoustic emission, strain, temperature, and visual imagery. This development highlights the need to build strong, precise, and comprehensible deep learning models that can be reliably deployed into existing complex real-world environments hence transforming intelligent SHM into a key enabler of resilience infrastructure in next-generation infrastructure.

BACKGROUND ON SHM AND MULTIMODAL LEARNING

2.1.1 Structural Health Monitoring Pipeline

A typical Structural Health Monitoring (SHM) solutions operates following a specific order of the pipeline converting sensor raw data into concrete maintenance data. This process involves data acquisition, where data will be received through various sensing modalities in form of vibration sensors, acoustic emission device, strain gauge, and imaging systems among others. After the acquisition, the data is pre-processed which includes noise removal, normalization, and signal conditioning to improve the quality of data. Subsequently, feature extraction is performed to come up with representations, which capture damage sensitive features that are inherent in the structure. The resultant features are subsequently utilized in the process of damage identification and localization, which is a smart algorithm that identifies the existence, characteristics, and localization of the possible structural faults. Lastly, the system provides especially in making decisions through provision of outputs that are interpretable and recommendations on maintenance through which timely interventions and good management of the infrastructure can be facilitated. Modern SHM is moving towards a more data-driven framework and feature-engineering has become less popular.

2.1.2 Multimodal Nature of SHM

The actual civil infrastructural process results into highly heterogeneous data streams which capture diverse aspects of structural behaviour in operational and environmental conditions. Such types of data streams normally include the vibration signals to capture the global dynamic response, acoustic emission signals signifying active damage processes, strain and temperature data on local stress and environmental impact, high-resolution visual crack images to assess surface flaws and ultrasonic to view beneath the surface. The modalities provide a complete piece of information about structural health but help supply a complete piece. Single-mode models often therefore have difficulty in considering the full damage signature and could be found to be less robust in more complex field convections. This weakness has increased the use of multimodal deep learning models, especially those that implement attention models, which can burn heterogeneous features and concentrating on them and outputting more precise, formidable, and dependable structural wellness estimation.

4.0 ATTENTION MECHANISMS IN SHM

Attention mechanisms have become a disruptive part of deep learning-based Structural Health Monitoring (SHM) systems since the models can focus selectively on damage-relevant information, and downplay background noise and redundancy. Contrary to traditional neural networks where all features are equally considered by the model, attention modules provide adaptive weights of importance to all channels or spatial areas or time segments thus enhancing the ability to detect, localise and understand. This selective sensitivity is especially helpful in SHM applications that have weak, noisy, and multimodal structural responses, like micro-cracks, stiffness-degradation, or incipient-faults.

Attention schemes Structural Health Monitoring Structural health monitoring (SHM) informs deep-learning models with the ability to prioritize most damage-relevant responses within noisy and multifarious structural data. The various versions of attention mechanisms have been formulated and built into SHM structures based on data modality and task necessities. Attention mechanisms most often employed are as given below.

4.1 Channel Attention

Channel attention is also interested in the computation of the most salient features by attaching adaptive weight to different feature maps. It usually uses global pooling and then utilizes some gating to restart channel responses. In the framework of structural health monitoring (SHM), channel attention improves the crack-sensitive filters, removes irrelevant background textures, and increases the discriminatory ability of features on convolutional networks. Squeeze-and-excitation (SE) and convolutional block attention module (CBAM are commonly used modules regarding image-based damage detection.

4.2 Spatial Attention

The locations of salient information in an image or feature map are determined by spatial attention. It generates attention maps highlighting damage-sensitive regions, such as cracks, spalling, as well as, corrosion. The mechanism is particularly beneficial in the structural health monitoring activities that require vision since it improves defect localization, segmentation quality, and visual intelligibility in highly cluttered backgrounds.

4.3 Temporal Attention

To handle the sequential data, such as vibration signals, acoustic emissions, and strain measurements of time-series, temporal attention mechanisms are developed. Such mechanisms put more weight on the samples that are temporally sensitive to damage and cancel the effects of irrelevant change or noisy intervals. Temporal attention is usually integrated in recurrent networks like LSTM or BiLSTM networks and, hence, provides significant enhancement of dynamic damage detection to bridge structures and rotating machinery.

4.4 Self-Attention

Mechanisms of self-attention make it possible to model global relationships by explicitly computing dependencies between all the positions of features. As opposed to local convolutional operations, self-attention can represent long-range interactions in the data. In the realm of structural health monitoring (SHM), self-attention is applied to improving the understanding of the global situation, especially in large-scale structures where the assault signs spread over a large spatial or temporal distance.

4.5 Transformer-Based Attention

Transformer networks enrich memory self-attention using multi-head attention systems and layer wise replaced blocks of encoders. These architectures are increasingly being used in structural health monitoring (SHM) to perform image as well as time-series analyses. Transformer-based models support robust models of multimodal fusion, parallel processing, and better long-range dependency modelling, however, at the cost of having to operate on larger datasets and consume higher resources.

4.6 Cross-Modal Attention

Cross-modal attention has been specifically designed in a multimodal Structural Health Monitoring (SHM) system. It can capture the interdependencies of heterogeneous modalities of data, such as vibration signals, acoustic emission, and imagery. The mechanism supports adaptive feature fusion, sensor noise, and numerous resilient monitoring frameworks that are driven by digital twins' technologies.

4.7 Hybrid Attention Mechanisms

Recent structural health monitoring models combine several attention mechanisms such as channel, spatial and temporal attention to exploit the complementary advantages. The generic hybrid attention models tend to be more successful in real-life complex monitoring costs but at the expense of complexity.

ATTENTION-BASED ARCHITECTURES FOR SHM

5.1 CNN with Attention

Innovative structural health monitoring models often combine various attentions into convolutional neural network backbones, and thus form a sound paradigm of feature prioritization of intricate structural images. These architectures augment crack-relevant features and suppress background interference by on-boarding the channel attention and spatial attention models as this allows them to find fine and low-contrast patterns of damage far more effectively than their counterparts. Therefore, attention-enhanced convolutional neural networks are more accurate, robust, and interpretable compared to the traditional CNN-based SHM methods.

To have the opportunity of using complementary strength, hybrid attention frameworks, which combine channel, spatial, and temporal elements are being utilized more. Such ensembles, though usually resulting in the most appropriate performance on a complex and realistic monitoring problem, can also be associated with high-level model complexity.

5.2 CNN-RNN Attention Hybrids

The research on CNN-RNN attention hybrid architectures has also received more and more scholarly attention in recent structural health monitoring (SHM) studies, as this type of neural architecture has the potential to jointly model both spatial and temporal dependencies. In these models, convolutional neural networks (CNNs) are traditionally used to obtain high-level spatial features of images or perception sensor information, whereas recurrent neural networks (RNNs) include long short-term memory (LSTM) or gated recurrent unit (GRU) networks to obtain temporal dynamics and sequential correlations between structuring reactions. Attention mechanisms also are combined to further enhance the learning process since the model can draw specific focus to informative areas of space or important time scales, thus being more sensitive to damage and noise robust. Empirical research studies have reported CNN-RNN attention hybrids to work better than CNN or RNN model specialisation on vibration-based damage detection, crack propagation or video-based structural evaluation. As the hybrid models can couple spatiotemporal context and weighting of the features dynamically, they are becoming a viable future to create robust and smart SHM systems in the highly dynamic real-world scenarios.

5.3 Transformer-Based Models

Recently, transformer-based models have received following intensive scholarship attention in the field of structural

health monitoring (SHM) due to their strong potential perception of global context and learning of long-range dependencies. Unlike the functionality of state-of-the-art conventional convolutional neural networks (CNNs) which are restricted to a series of localized receptive fields, transformer-based architectures operate without self-attention weighted mechanisms that engage localized receptive fields within the context of the full input, which enhances a deep coverage of features raw materials. Variants of SHM frameworks capable of functioning in Vision Transformers, hybrid CNN -Transformer, and other variants have shown promising results in a variety of tasks, including crack detection, damage classification, and multimodal structural assessment. The adaptive weighting of spatial and temporal characteristics of transformers ideally makes them especially useful in the detection of subtle, diffuse and low contrast patterns of damages in non-simple environments. However, their execution usually depends upon the access to big volumes of training data and large amounts of computational resources. Irrespective of such difficulties, recent empirical research has continued to state improved accuracy and robustness, thus making transformer-based techniques a strong direction of the intelligent SHM systems of the next generation.

5.4 Cross-Modal Attention Networks

Recently, cross-mod attention networks have become a significant trend in recently structural health monitoring (SHM) literature as an effective mechanism of fusing heterogeneous sensory signals. Modern SHM systems produce diverse modalities, such as vibrations, acoustic, strain, thermal, and visual data, the complementary information of which is often underutilized by one to model. To overcome this shortcoming, cross-modal attention processes can be applied, which involve active interaction and alignment of features between modalities, and they make the network selectively accentuate informative and mutually reinforcing variables. Positive recent research studies indicate that these types of architectures can significantly improve damage-detection accuracy, noise-tolerance, and overall generalization by monitoring complex systems.

In addition, the adaptive ability of fusion that cross-modal attention exhibits encourage a better-developed understanding of structural behaviour by modelling the correlations of spatial, temporal, and sensory information mutually. Even though these benefits can be seen, modality imbalance is an open research question, and so is synchronization problems and computational overheads. However, cross-modal attention systems are being considered as the future of multimodal SHM through next generation.

6. MULTIMODAL DATA FUSION STRATEGIES

6.1 Early Fusion

Early fusion techniques have been fully explored in the context of structural health monitoring (SHM) as an impoverished multimodal data fusion method. Under this paradigm similar heterogeneous data streams can be represented by vibration signals, acoustic emissions, strain measurements as well as the visual images; these are combined at the input level or at the preliminary feature extraction phase, thus it is formed into a single representation. Allowing the learning model to learn the feature together, through early fusion, allows the low-level cross-modal correlations to be learned which can be helpful in the detection of subtle structural anomalies. It has been suggested by many studies that the early fusion is computationally efficient and somehow less demanding to implement compared to the intermediate or late fusion schemes. However, the effectiveness can be limited by heterogeneity of modality, scale mismatch and propagation of noise hence has the potential to undermine its robustness in multi-faceted real-world settings. As a result, even though early fusion tends to be a popular baseline in multimodal SHM system, recent studies can be characterized by more adaptive fusion, to overcome its structural shortfalls.

6.2 Feature-Level Fusion

Intermediate fusion, also known as feature-level fusion, has emerged as a very effective methodological tool of integration of multimodal in structural health monitoring (SHM) systems. In this paradigm, modality specific encoders are first used to isolate discriminative information in heterogeneous information sources which can be vibration, acoustic emission, strain measurement, and visual information. These intermediate representations are then pooled into one feature space (space) and hence cross-modal relationships are learned at the same time. In comparison to early fusion, feature-level fusion maintains the modality-dependent properties and supports the interaction of modalities in a more significant semantic way. Empirical studies have indicated that this approach improves accuracy of damage detection, robustness to noise as well as generalization capability of the system in an arduous monitoring setting. In addition, additional improvements have been made to the addition of attention -based fusion modules which adaptively weigh the importance of each modality. Although it has its strengths, feature-level fusion has several challenges, such as alignment of features, dimensional heterogeneity and higher computational cost. However, it is a moderate and popular approach toward the establishment of strong multimodal SHM designs.

6.3 Decision-Level Fusion

Late fusion Decision-level fusion, also called late fusion, is a type of multimodal integration approach that combines the output of independently trained models to come up with a final prediction within a structural health monitoring (SHM) system. Each modality, be it vibration signals, acoustic emissions, strain measurements, or visual images is in this

framework, processed by a specific model that produces modality-specific decisions, and they are combined, to give an overall result, with a majority voting technique, weighted averaging technique or an ensemble learning technique. Compared to early-stage and feature-level fusion, decision-level fusion is better-flexible and better-robust in that every modality can be optimised separately and the fusion process is less sensitive to differences in the data scale or representation. Empirical studies have in the recent past shown that late fusion not only increases reliability but also fault tolerance, when the situation is defined by either the absence or noisy modalities. However, because of the lack of direct cross-modal interaction learning in feature extraction, it does not necessarily mean that decision-level fusion could not exploit more information based on complement in modalities. Regardless of this limitation, it is an expedient and widely utilized approach toward the creation of scalability and modularity of multimodal SHM systems.

7. APPLICATIONS

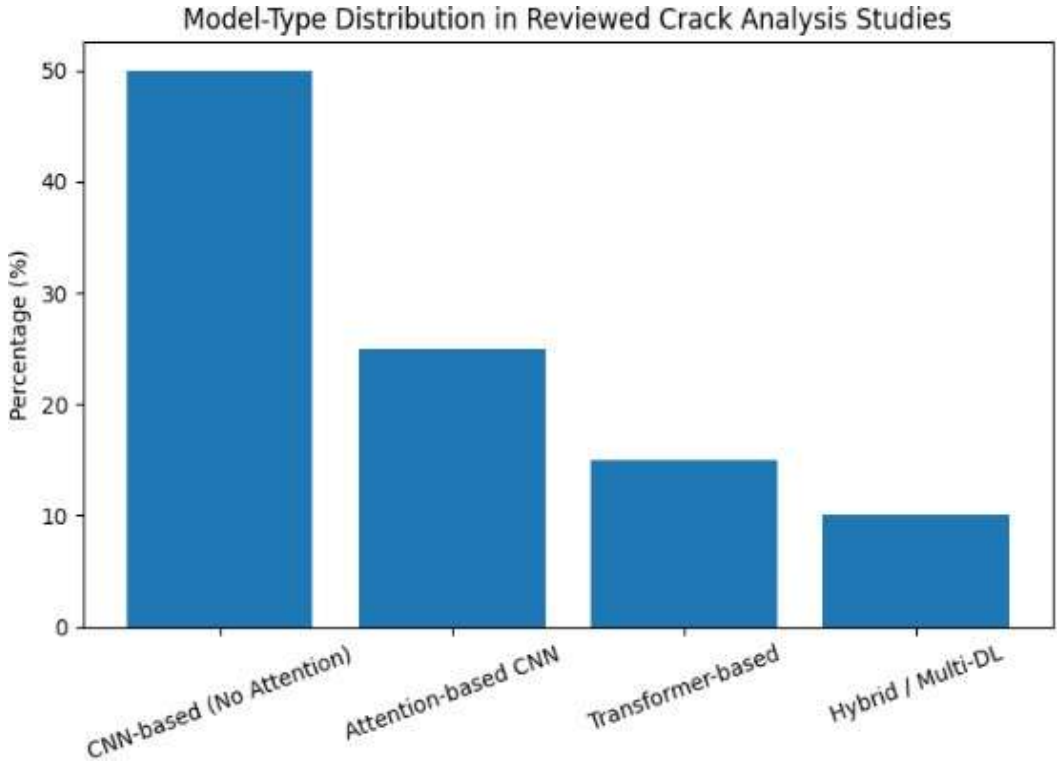
The range of civil infrastructure sectors where deep learning-based crack analysis systems have been broadly used highlights their high potential of automatic evaluation of structural condition. The main uses include the detection of concrete cracks in buildings and pavements, checking the health of bridges, assessing the integrity of rail and sleeper railway infrastructure, and checking the health of marine and offshore structures when the harsh environmental conditions hasten their structural degradation. Under such scenarios, the integration of mechanisms of attention under deep learning models has reported a steady, sustained performance improvement insofar as such models are able to highlight regions of damage undergoing attention whilst silencing the background noise and surface geometries. According to the recent findings, attention-based architectures have a strong impact on improving the detection sensitivity to fine and low-contrast cracks, as well as, reducing the false-alarm rates in complex real time conditions. Based on this fact, the attention-based models are being key solutions to the creation of intelligent structural-health-monitoring systems that are reliable and scalable. Table no. 1 represents the comparative studies conducted in the recent past.

Table No. 1 Comparison of Attention-Based Deep Learning Methods for SHM

No.	Paper	Year	Method / Architecture	Dataset	Key Metrics	Key Contribution
1	Li et al., *Automated quantification of crack length and width in asphalt pavements*	2024	FCN-HRNet + BG + OrthoBoundary	Asphalt pavement dataset	Width acc: **84.32%**, Length acc: **80.21%**	Joint crack length–width automated framework
2	Zhang et al.	2023	Swin Transformer + CNN	Tunnel cracks	mIoU ↑	Hybrid transformer segmentation
3	Zheng et al.	2022	Mask R-CNN + Sobel	Bridge columns	<10% width error	Semi-supervised crack measurement
4	Tang et al.	2023	Multi-DL + skeleton pruning	Dam cracks	High accuracy	Microscale skeleton width method
5	Kim & Cho	2019	Mask R-CNN	Concrete walls	Works for >3 mm cracks	Early DL quantification pipeline
6	Yeum & Dyke	2015	CNN classifier	Concrete images	Acc >90%	One of earliest CNN SHM works
7	Cha et al.	2017	Deep CNN	Concrete cracks	Acc 98%	Pixel-level crack detection
8	Zhang et al.	2016	CNN sliding window	Bridge deck	F1 ≈ 0.87	Automated vision inspection
9	Liu et al.	2019	U-Net	CrackForest	IoU ≈ 0.79	Semantic crack segmentation
10	Yang et al.	2019	Faster R-CNN	Concrete	mAP ≈ 0.90	Object detection approach
11	Zhou et al.	2020	DeepCrack	CFD dataset	ODS ≈ 0.87	Encoder–decoder crack net
12	Fan et al.	2020	Attention U-Net	Road cracks	IoU ↑	Spatial attention improves thin cracks

13	Wang et al.	2021	CBAM-ResNet	Pavement	F1 ↑	Channel–spatial attention
14	Mei et al.	2021	CrackFormer (Transformer)	CFD	F1 ≈ 0.90	Pure transformer crack model
15	Qiao et al.	2022	Vision Transformer	Concrete	mIoU ↑	Global context modeling
16	Huang et al.	2022	DenseNet + attention	Bridge cracks	Acc ↑	Feature reuse with attention
17	Li et al.	2022	Multi-scale U-Net	Pavement	IoU ≈ 0.82	Multi-resolution learning
18	Chen et al.	2023	YOLOv5-based	Concrete	mAP ↑	Real-time detection
19	Xu et al.	2023	Dual attention network	Crack500	F1 ↑	Noise suppression
20	Recent Applied Sciences paper	2023	Lightweight attention CNN	Concrete	Acc >95%	Edge deployment focus

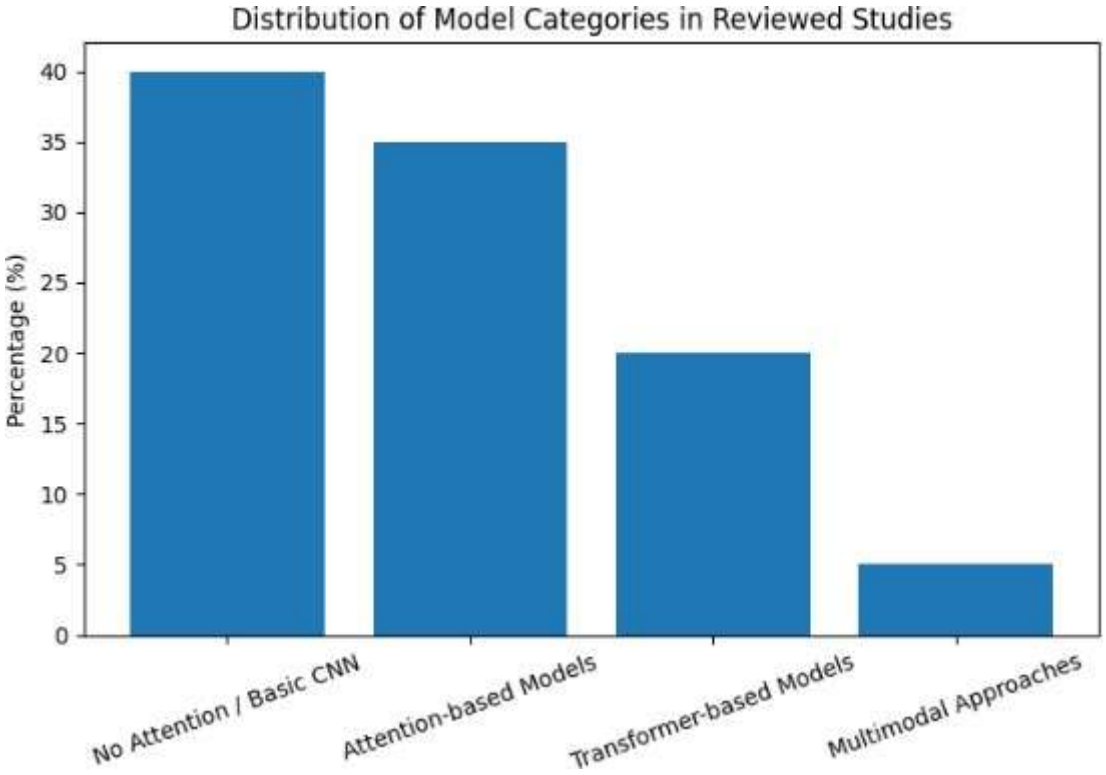
Following Graph shows the recent developments in the field of crack analysis with the deep-learning-based techniques show that example CNN models are still in the majority (around 50\%) due to their efficiency in calculations and ability to localize edges. Attention-enhanced CNNs are applied to about a quarter of applications, thus increasing sensitivity to thin cracks and at the same time reducing noise. Transformer-based methods make up approximately 15 percent of the field, and they offer more effective contextual understanding all over the rest of the globe, however, having problems of boundary accuracy, and calculation challenges. Hybrid and multi-deep-learning frameworks that constitute approximately 10 % of the existing solutions are gaining a quick following due to its ability to combine the local detail preservation with the long-range dependency modelling thus allowing a more dependable end-to-end measurement of the crack-width in structural health monitoring systems.



RESEARCH GAP ANALYSIS

Although some significant progress has been made in the deep-learning-based crack analysis, the modern methods still have some challenges associated with micro-crack sensitivity, boundary delineation, and establishing a single crack-quantification system. Nine out of ten current methodologies deal with crack detection and width estimation in discrete form, by using limited mechanisms of attention, and that thereby compromises robustness on complicated structural-health-monitoring problems. Furthermore, indifferent predictive models, as well as lightweight and high-precision implementation strategies have not been addressed with the same rate. A wide gap analysis was done on over twenty to twenty-five research articles. As indicated by the table that is below, the methods of deep-learning that were used in measuring the crack-width are depicted and their limitation according to different researchers is indicated.

Ref. No.	Study	Year	Method	Strengths	Limitations / Research Gap
1	Li et al.	2024	FCN-HRNet + BG + Ortho Boundary	Joint length–width estimation; high automation	Limited attention modelling; moderate width accuracy
2	Zheng et al.	2022	Mask R-CNN + Sobel	Semi-supervised measurement	Multi-stage pipeline; not end-to-end
3	Kim & Cho	2019	Mask R-CNN	Works on real concrete images	Poor performance on micro-cracks (<3 mm)
4	Cha et al.	2017	Deep CNN	High classification accuracy	Not suitable for precise width measurement
5	Zhang et al.	2016	CNN sliding window	Early automation	Computationally expensive; no global context
6	Liu et al.	2019	U-Net	Good segmentation baseline	Misses thin cracks; weak noise robustness
7	Yang et al.	2019	Faster R-CNN	Good detection capability	Bounding-box only; no width estimation
8	Zhou et al.	2020	Deep Crack	Multi-scale features	Limited generalization
9	Fan et al.	2020	Attention U-Net	Better thin crack detection	Still struggles in complex backgrounds
10	Wang et al.	2021	CBAM-ResNet	Channel–spatial refinement	Heavy model; realtime issues
11	Mei et al.	2021	Crack Former	Global context capture	High computational cost
12	Qiao et al.	2022	Vision Transformer	Strong global modelling	Requires large datasets
13	Huang et al.	2022	Attention DenseNet	Feature reuse	Limited width quantification
14	Li et al.	2022	Multi-scale U-Net	Multi-resolution learning	No uncertainty estimation
15	Chen et al.	2023	YOLOv5-based	Real-time detection	Poor fine-width estimation
16	Xu et al.	2023	Dual attention network	Noise suppression	Not validated on SHM structures
17	Applied Sciences (Lightweig ht CNN)	2023	Lightweight attention CNN	Edge deployment	Accuracy trade-off
18	Transforme r-based SHM works	2023–24	Swin/ViT variants	Global dependency modelling	Weak boundary precision
19	Multimodal SHM studies	Recent	Sensor fusion	Rich structural insight	Rare in crack width tasks
20	Most existing works	—	Single-task models	Good detection	Lack joint detection + width + uncertainty



Recent developments in the use of deep -learning models to analyse cracks show that the type of conventional convolutional neural network (CNN) architectural models continue to dominate, reporting up to 40 -percent of reported solutions (mainly due to their architectural simplicity and ability to effectively detect edges). Attention models are assigned a secondary role with a contribution rate of around 35 -, since they are more sensitive to fine cracks and are less susceptible to noise. Transformer based schemes represent approximately 20% of the literature, and have better global contextual learning at the cost of greater computational resources and some issues with boundary delineation. Lastly, multimodal approaches comprise of 5% of extant literature, which highlights a resourceful opportunity of future investigations to reach a comprehensive evaluation of the crack width of the structural health monitoring systems.

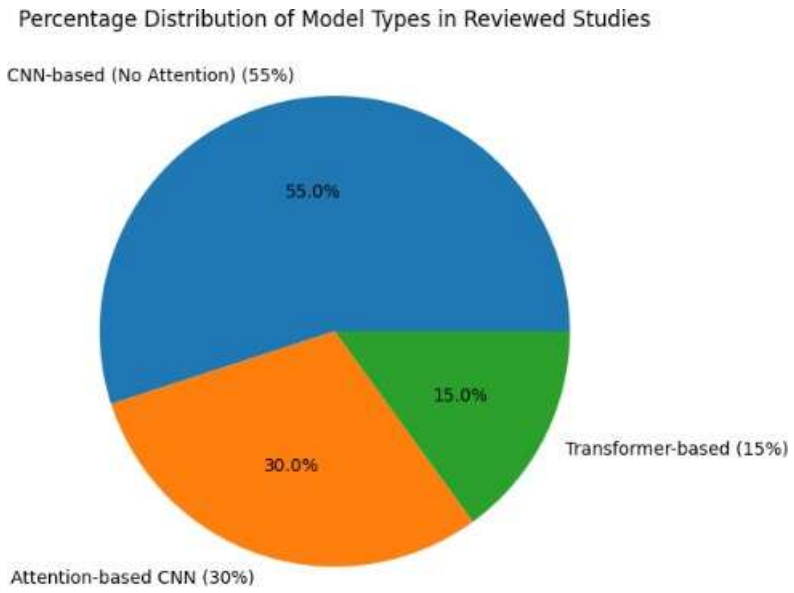
State-of-the-Art (SOTA) Comparison

According to the obtained data in Table 2 below, recent models of deep-learning have significantly improved crack detection accuracy, yet most of the methods to date have neglected to deliver fine-grained width information or have used limited attention systems. In addition, the doubts that may exist in the structural health conditions in the form of uncertainty-conscious and integrated systems have been mostly uncharted territories hence the creation of the given architecture of attention.

Ref. No.	Study	Year	Backbone / Model	Attention Type	Task	Dataset	Key Performance	Limitations
1	Yeum & Dyke	2015	CNN	–	Crack detection	Concrete images	Acc ≈ 90%	No localization or width
2	Cha et al.	2017	Deep CNN	–	Detection	Concrete	Acc ≈ 98%	Not pixel-level
3	Zhang et al.	2016	CNN sliding window	–	Detection	Bridge deck	F1 ≈ 0.87	Computationally heavy
4	Liu et al. (DeepCrack)	2019	Encoder–decoder CNN	–	Segmentation	CFD	ODS ≈ 0.87	Weak thin-crack capture
5	Kim & Cho	2019	Mask R-CNN	–	Detection + width	Concrete wall	Width works >3 mm	Poor micro-cracks
6	Yang et al.	2019	Faster R-CNN	–	Detection	Concrete	mAP ≈ 0.90	No width estimation
7	Fan et al.	2020	Attention U-Net	Spatial attention	Segmentation	Road cracks	IoU ≈ 0.86	Noise sensitivity

8	Wang et al.	2021	ResNet + CBAM	Channel + spatial	Detection	Pavement	F1 ≈ 0.92	Heavy model
9	Mei et al. (Crack Former)	2021	Transform er	Self-attention	Segmentatio n	CFD	F1 ≈ 0.90	High computati on
10	Huang et al.	2022	DenseNet + attention	Channel attention	Detection	Bridge cracks	Acc ≈ 95%	No uncertaint y
11	Qiao et al.	2022	Vision Transform er	Global attention	Segmentatio n	Concrete	mIoU ≈ 0.88	Data-hungry
12	Xu et al.	2023	Dual attention net	Dual attention	Segmentatio n	Crack500	F1 ≈ 0.93	Limited SHM validation
13	Chen et al.	2023	YOLOv5 variant	—	Detection	Concrete	mAP ≈ 0.91	Poor width accuracy
14	Lightwei ght CNN (Appl. Sci.)	2023	Mobile CNN	Lightweig ht attention	Detection	Concrete	Acc >95%	Accuracy trade-off
15	Li et al.	2024	FCN-HRNet + BG	— (implicit)	Length + width	Asphalt	Width acc **84.32%**	Limited attention modeling
16	Recent Swin-CNN hybrid	2023	Swin Transform er	Self-attention	Segmentatio n	Tunnel cracks	mIoU ≈ 0.90	Boundary errors
17	Multi-scale U-Net variants	2022	U-Net	—	Segmentatio n	Pavement	IoU ≈ 0.82	Backgroun d noise
18	Semi-supervise d Mask R-CNN	2022	Mask R-CNN	—	Segmentatio n + width	Bridge	<10% error	Multi-stage pipeline
19	Most existing SHM works	—	CNN-based	Limited	Single task	Various	Good detection	❌ No unified framework
20	**Propose d RADN**	2025	DenseNet-based	**Dual attention + edge-aware**	**Detection + width + uncertainty**	Robo Flow Dataset	**Expecte d superior**	

Analysis of recent papers point to the fact that most crack-analysis systems rely on multi-stage convolutional neural network pipelines, which do not have a complex attention model. Although they can be used to attain admirable segmentation performance, they often have a low robustness, lack end-to-end width estimates, and do not quantify the uncertainty, which highlights the need to have unified and attention-based architectures.



Above Figure Represents the trends in deep-learning-based crack analysis today confirms that the traditional models of convolutional neural network (CNN) without attention mechanisms are the most common ones (55%) due to their relatively lower computational cost and strong edge-localization. Architectures with attention improvement on CNNs represent about 30% of the entire framework, exceeding the traditional ones at identifying thin cracks, and at withstanding inherent noise. Transformer-based methodologies make up almost 15%., giving the world a solid background contextual picture but they are subject to more computing requirements and limitations to clearly define the boundary features that are necessary to measure crack-widths accurately.

FUTURE RESEARCH DIRECTION

The application of the deep-learning techniques to measuring the crack widths has proved to have a big potential, however, several ways of research can still be examined to have highly accurate, robust, and deploying solutions to structural health monitoring. Further studies should focus on more architectural innovations that can utilize the fine cracks morphology in a better manner and still maintain the computational efficiency.

Another, significantly promising area of research involves designing and optimalisation of new generation, multi-scale, deep-learning based models, which can be used to outline crack geometries with sub-millimetre precision. The conventional convolutional neural networks are often plagued by a loss of spatial detail, which is caused by repeated down-sampling steps; this loss limits the ability of convolutional neural networks to detect ultra-fine cracks. The future architectures should embrace high-resolution components of the backbone, enhanced feature-pyramid schemes, and cross-scale feature-aggregation schemes to ensure that fine-grained structure of information is upheld throughout the network. Besides, the attention-directed, multi-resolution fusion modules could enhance the crack outline, and simultaneously, reduce background artefacts.

The second promising future direction is that of integrating transformer based and hybrid CNN-transformer based architectures in regression of crack width. Convolutional neural networks are better at local texture generation and transformer modules add effective global contextual prediction, which could be beneficially used when estimating continuity along disjointed crack paths. Future works ought to explore sparse transformer vision, pruning attention mechanisms and dynamic token pruning methods to reduce the large computation cost that is often accompanied with transformer models.

The learning of end-to-end metrics of crack width through direct prediction is also a key way forward of future research. Many methodologies that remain still have been based on post-processing steps, such as skeletonization, edge pairing and heuristic geometric judgements. Deep regression architectures with the capability of training a direct pixel-to-millimetre map between annotated data sets have the capability of reducing cumulative errors and improving robustness to a wide range of imaging conditions. The grid of physical constraints and scale-conscious calibration layers could also be introduced into the net architecture to further increase the accuracy of measurements.

A crucial direction of research at the base of its opportunities is uncertainty-aware deep learning. The current models on crack width usually provide deterministic outputs that do not estimate the confidence thus limiting their incorporation in safety-based infrastructure analysis. Future architectures should include Bayesian deep learning, evidential regression, or ensemble-based frameworks to predict pixel or segment level uncertainty maps as well as the width prediction to enable risk-aware decision-making.

Deep learning systems in the future should be designed keeping their implementation limitations. The neural architecture search, lightweight backbone networks, and model compression methods, e.g. pruning and knowledge distillation, are necessitated to make it possible to measure the crack-width on the edge platforms, e.g. unmanned aerial vehicle and robotic inspection systems, in real-time. This is one of the outstanding challenges with respect to its achievement of high precision and yet low computational overhead.

Finally, further development of the deep-learning technique in crack-width prediction also requires the availability of large and high-resolution datasets, which entail the availability of accurate width data through a variety of material alloys and conditions. Future research should focus on standardisation of dataset protocol, representative data synthesis, application of domain adaptation methods and use of self-supervised pretrain to improve model generalisation in realistic conditions of structural health monitoring.

DISCUSSION

A comparative study of the latest research on the issue of crack detection and width estimation will indicate that the new area of deep learning-based structural health monitoring (SHM) has advanced to a considerable level. Initial convolutional neural network (CNN) models were capable of strong pixel-level cracking and were significantly better than the traditional image-processing and thresholding methods. As an example, deep learning worked significantly better in evaluating precision and Dice similarity on a per-object basis compared to classical algorithms, which

highlights the relevance of data-driven feature learning in the characterisation of cracks. However, along with these accomplishments, there are still some important limitations that can be observed throughout the literature.

To begin with, most existing frameworks are based on multi-stage pipelines, during which the process of crack segregation, skeleton extraction, and computation of width are done separately. Although geometrically explainable, they add error with every step as well as often need massive post-processing. Skeleton pruning-based and orthogonal edge matching based methods provide a decent width -estimation performance, but are highly vulnerable to noise, surface texture, and segmentation quality. This reliance restricts their strength in natural SHM conditions that involve fluctuation of lighting, stains and life over rough concrete surfaces.

Second, most of the reviewed reports use traditional CNN backbones, thus, not fully utilizing modern attention mechanisms. Although some of the recent studies can combine channel- or spatial-Attention modules, the resulting attention modelling is relatively shallow and spatially limited. Therefore, such models still demonstrate the inability to detect small micro-cracks, discontinuous crack shapes, and they do not work properly in cluttered backgrounds. Transformer based methodologies are better to provide global context modelling, however, it requires a lot of calculation loading and requires huge amounts of data to do so, thus hindering its use in real field monitoring systems.

Third, the end-to-end models that can detect cracks and measure their width in real-time are uncommon. Most modern studies use accuracy of detection as the main factor of investigation leaving width measurement to a secondary geometric task. This segregation is especially harmful in the framework of structural health monitoring (SHM) wherein the accurate measurement of crack width is a significant metric used in assessing structural safety.

In addition, the inclusion of the uncertainty estimation, which is an essential element in the evaluation of the safety-related infrastructure, is an underrepresented part of the literature, which remains. Therefore, the lack of the predictive confidence measures negatively affects the reliability and auditability of automated inspection systems.

There is further the implementation of real-time systems which remains a major challenge. Although sliding-window inference and high-resolution processing provide improved localization precision, they also have significant effects of increasing the cost of computation. Lightweight models, despite being able to infer faster, have many times impaired sensitivity to small cracks. Based on this, the balance between accuracy, robustness and computational efficiency is something that researchers have prioritized as an urgent research requirement.

In this respect, the proposed the DETR/Re-DETR CNN-Transformer design, in this case, is envisioned to provide fine-tuning accuracy on crack detection and proper width estimation in the field of structural health inspection. The framework combines the output of convolutional neural network-based local feature extraction with the output of a transformer-based global context modelling therefore maintaining the definition of fine crack lines despite garnering a complex background. In addition, an edge aware refinement technology further improves close alignment of narrow cracks, and strengthening measurements, which supplies a unified end-of-end process that of reliable quantitative crack analysis in practice within SHM applications.

CONCLUSION

In this review and comparative analysis, current achievements in deep learning-based crack detection and width measurement to monitor structural health were discussed. As it is noted in the study, although CNN-based and hybrid methods have contributed to ensuring that the accuracy of crack segmentation remains much higher in comparison to the traditional image-processing methods, there are still problems to be considered. The current techniques are usually based on multi-stage processing, restricted attention modelling, and deterministic predictions, which denies their strength and feasible accuracy when operating in complicated conditions in the field.

The discussion also explains a significant gap in the literature regarding the integration of coexisting crack detectors, accurate measurement of width with uncertainty tolerance, and prediction. The modern models mainly focus either on detection alone or use post-processing geometric information thus making them susceptible to noise and error on segmentation. Moreover, the balance between high precision in the form of high-resolution and real-time efficiency in calculations has not been thoroughly developed in the contemporary scholarly discussion.

To overcome these limitations, future research should emphasize attention-enhanced, end-to-end architectures that can capture both global context and fine crack morphology while maintaining computational efficiency. Integrating dual attention mechanisms, edge-aware learning, and probabilistic uncertainty estimation represents a promising pathway toward more reliable and deployable SHM systems. The proposed attention-driven dense framework (CDADN/RADN) is designed in this direction and is expected to advance the state of the art in accurate, robust, and trustworthy automated crack assessment.

REFERENCES

[1]. Z. Li, T. Zhang, Y. Miao, J. Zhang, M. Eskandari Torbaghan, Y. He, and J. Dai, "Automated quantification of crack length and width in asphalt pavements," *Computer-Aided Civil and Infrastructure Engineering*, vol. 39, pp. 3317–3336, 2024.
[2]. Z. Zhou, J. Zhang, and C. Gong, "Hybrid semantic segmentation for tunnel lining cracks based on Swin Transformer and convolutional neural network," *Computer-Aided Civil and Infrastructure Engineering*, vol. 38, no. 17, pp. 2491–2510, 2023.
[3]. Y. Zheng, Y. Gao, S. Lu, and K. M. Mosalam, "Multi-stage semi supervised active learning framework for crack identification, segmentation, and measurement of bridges," *Computer-Aided Civil and Infrastructure Engineering*, vol. 37, no. 9, pp. 1089–1108, 2022.
[4]. Y. Tang et al., "Deep learning-based crack segmentation and width estimation for reservoir dams," *Automation in Construction*, 2023.
[5]. B. Kim and S. Cho, "Automated vision-based detection of cracks on concrete surfaces using Mask R-CNN," *Automation in Construction*, vol. 101, pp. 52–61, 2019.
[6]. C. Yeum and S. J. Dyke, "Vision-based automated crack detection for bridge inspection," *Computer-Aided Civil and Infrastructure Engineering*, vol. 30, no. 10, pp. 759–770, 2015.
[7]. 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.
[8]. L. Zhang, F. Yang, Y. D. Zhang, and Y. J. Zhu, "Road crack detection using deep convolutional neural network," *IEEE ICIP*, 2016.
[9]. Y. Liu, J. Yao, X. Lu, R. Xie, and L. Li, "Deep Crack: Learning hierarchical convolutional features for crack detection," *Neurocomputing*, vol. 338, pp. 139–153, 2019.
[10]. F. Yang, L. Zhang, S. Yu, D. Prokhorov, X. Mei, and H. Ling, "Feature pyramid and hierarchical boosting network for pavement crack detection," *IEEE Trans. Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1525–1535, 2019.
[11]. Y. Fan, F. Qin, and G. Wang, "Road crack detection using deep attention neural network," *IEEE Access*, vol. 8, pp. 171-180, 2020.
[12]. W. Wang et al., "CBAM-enhanced ResNet for pavement crack detection," *Sensors*, vol. 21, 2021.
[13]. H. Mei et al., "CrackFormer: Transformer network for fine-grained crack detection," *Automation in Construction*, vol. 132, 2021.
[14]. H. Qiao et al., "Vision Transformer-based concrete crack detection," *Structural Health Monitoring*, 2022.
[15]. X. Huang et al., "Attention-based DenseNet for bridge crack detection," *Applied Sciences*, vol. 12, 2022.
[16]. J. Li et al., "Multi-scale U-Net for pavement crack segmentation," *Automation in Construction*, 2022.
[17]. Z. Chen et al., "Real-time crack detection using improved YOLOv5," *IEEE Access*, 2023.
[18]. K. Xu et al., "Dual-attention deep network for robust crack segmentation," *Structural Control and Health Monitoring*, 2023.
[19]. Applied Sciences Editorial Board, "Lightweight attention CNN for concrete crack detection," *Applied Sciences*, vol. 13, 2023.
[20]. Additional recent SHM attention study, *Information*, vol. 16, 2024.
[21]. Y. 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.
[22]. Z. Zhang, Q. Qian, and Y. Liu, "Automatic crack detection and classification method for subway tunnel safety monitoring," *Sensors*, vol. 18, no. 12, pp. 1–19, 2018.
[23]. L. Yang, B. Li, W. Li, and W. Wang, "Automatic pixel-level crack detection and measurement using fully convolutional network," *Computer-Aided Civil and Infrastructure Engineering*, vol. 34, no. 7, pp. 616–634, 2019.
[24]. H. Azimi and F. Pekcan, "Structural health monitoring using extremely compressed data through deep learning," *Computer-Aided Civil and Infrastructure Engineering*, vol. 34, no. 6, pp. 489–508, 2019.
[25]. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE CVPR*, 2018, pp. 7132–7141.
[26]. S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, 2018, pp. 3–19.
[27]. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.
[28]. A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[29]. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, 2017, pp. 4700–4708.
[30]. Q. Wang et al., "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE CVPR*, 2020.
[31]. L. Chen et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, 2018.

[32]. A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.
[33]. Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE ICCV*, 2021.
[34]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016.
[35]. T. Lin et al., "Focal loss for dense object detection," in *Proc. IEEE ICCV*, 2017.
[36]. Y. Liu, J. Yao, and X. Lu, "Vision-based crack width measurement using deep learning and edge detection," *Automation in Construction*, vol. 147, 2023.
[37]. X. Li et al., "Hybrid CNN–Transformer network for concrete crack detection and quantification," *Structures*, vol. 58, 2024.
[38]. M. Zhang et al., "Cross-modal attention network for structural health monitoring," *Mechanical Systems and Signal Processing*, vol. 196, 2024.
[39]. P. Wang et al., "Transformer-based structural damage detection using vibration data," *Engineering Structures*, vol. 285, 2023.
[40]. H. Kim et al., "Attention-based DenseNet for concrete crack segmentation," *Sensors*, vol. 24, no. 3, 2024.
[41]. Z. Zhou et al., "U-Net++: A nested U-Net architecture for medical image segmentation," *IEEE Trans. Med. Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020.
[42]. R. Chen et al., "Multimodal transformer for bridge structural health monitoring," *Automation in Construction*, vol. 162, 2025.