SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 6, 2026

A SYSTEMATIC REVIEW OF DEEP LEARNING TECHNIQUES FOR CRACK DETECTION AND STRUCTURAL DAMAGE ASSESSMENT

Dr. M. Adil Khan[*1], Engr. Amir Sohail[2], Faizan Ali[3], Aalia Faiz[4]

[*1]*Resident Engineer, National Engineering Services Pakistan (NESPAK), Pakistan*
[2]*Punjab Tianjin University of Technology (PTUT), Lahore, Pakistan*
[3]*Department of Civil Engineering, University of Engineering and Technology (UET), Peshawar, Pakistan*
[4]*Department of Building and Architectural Engineering, Faculty of Engineering & Technology, Bahauddin Zakariya University (BZU), Multan, Pakistan*

[*1]adee.uol@gmail.com, [2]amir.sohail@ptut.edu.pk, [3]18pwciv5155@uetpeshawar.edu.pk, [4]aaliafaiz@bzu.edu.pk

DOI: https://doi.org/10.5281/zenodo.20807510

Keywords

Article History
Received: 26 April 2026
Accepted: 05 June 2026
Published: 23 June 2026

Corresponding Author: *
Dr. M. Adil Khan

Abstract

Structural health monitoring has become increasingly critical for ensuring the safety and longevity of civil infrastructure, yet traditional manual inspection methods remain time-consuming, subjective, and often hazardous. Deep learning techniques have recently emerged as powerful tools for automated crack detection and damage assessment, but the rapidly expanding literature in this domain presents challenges for researchers seeking to understand prevailing trends, comparative performance, and remaining gaps. This systematic review therefore aims to synthesize and critically evaluate the state of the art in deep learning approaches for crack detection and structural damage assessment, with particular focus on architectural innovations, multi-modal data integration, and deployment feasibility. We conducted a structured search and rigorous screening of peer-reviewed publications, then extracted and analyzed key findings related to model performance, data strategies, and application contexts. Our analysis reveals that convolutional neural networks, particularly those with encoder-decoder and attention mechanisms, consistently achieve high accuracy for pixel-level crack segmentation in standard image datasets. We further observe that hybrid frameworks combining deep learning with complementary sensors, such as ground-penetrating radar or acoustic emission, significantly improve detection under occluded or noisy conditions. However, critical challenges persist: data scarcity and class imbalance remain inadequately addressed across most studies, and few works demonstrate real-time capability in field deployment. We also find that domain adaptation techniques, although promising, have been applied predominantly to laboratory settings rather than to extreme events like earthquakes or fires. Based on these synthesized findings, we propose a set of best practices for model selection, data augmentation, and validation protocols, and we identify several high-priority directions for future research, including unsupervised learning for scarce damage scenarios and lightweight architectures for embedded systems. This review provides a comprehensive roadmap for practitioners and researchers advancing automated structural damage assessment.

I. Introduction

The aging and deterioration of civil infrastructure, including bridges, roads, pipelines, and buildings, pose significant economic and public safety risks worldwide. For instance, according to the American Society of Civil Engineers, a substantial portion of the United States' bridges and roads are in poor condition and require urgent repairs, highlighting the pressing need for effective and reliable structural health monitoring (SHM) systems [1]. Traditional methods for detecting structural damage, such as cracks, spalls, or corrosion, primarily rely on manual visual inspections conducted by trained engineers. While these inspections are a cornerstone of infrastructure maintenance, they are inherently subjective, labor-intensive, time-consuming, and often dangerous, especially when inspecting hard-to-reach areas like tall bridges or deep tunnels [2]. Moreover, the results can vary significantly between inspectors, leading to inconsistent assessments. The advent of modern sensor technologies, including high-resolution digital cameras, laser scanners, and acoustic emission sensors, has provided a wealth of data that can be used for automated damage assessment. However, the sheer volume and complexity of this data have made it difficult for conventional signal processing and computer vision algorithms to provide robust and scalable solutions.

In recent years, deep learning (DL), a subset of machine learning based on artificial neural networks with multiple layers, has emerged as a transformative paradigm for extracting meaningful patterns from high-dimensional data. Deep learning architectures, particularly Convolutional Neural Networks (CNNs) [3], have demonstrated remarkable performance in a wide array of computer vision tasks, including image classification, object detection, and semantic segmentation. These capabilities are directly transferable to the problem of crack detection in structural images, where a DL model can learn to differentiate between crack pixels and the background concrete or asphalt surface with high fidelity. Furthermore, the ability of deep learning to process multiple data modalities simultaneously, such as fusing visual information with thermal or acoustic data, has opened new avenues for more comprehensive damage assessment. The integration of DL with these sensors has been shown to improve detection accuracy, particularly in challenging environments where visual information alone is insufficient [4].

Despite these significant advancements, the field of DL-based crack detection and structural damage assessment faces several critical research gaps. Firstly, the literature is fragmented across numerous specific application domains (e.g., bridges, pavements, pipelines, buildings) and sensor types (e.g., RGB cameras, LiDAR, infrared thermography). This fragmentation makes it difficult for researchers and practitioners to obtain a consolidated understanding of which algorithms are most effective under different conditions. For example, a model that achieves state-of-the-art performance on a benchmark dataset of concrete bridge cracks may fail to generalize to cracks in asphalt roads or to data collected under different lighting conditions [5]. Secondly, while many studies report high accuracy metrics on carefully curated datasets, the issue of data scarcity and class imbalance remains a persistent, yet often under-investigated, problem. Real-world damage data, especially for rare but critical events like seismic or fire damage, is notoriously difficult and expensive to acquire. Therefore, many models are trained on artificially generated or augmented datasets, leading to a significant gap between laboratory performance and real-world robustness [6]. Thirdly, a major gap exists in translating high-performing deep learning models into practical, real-time deployment frameworks. Computational demands of deep networks often hinder their use on embedded or low-power devices that are typical in field inspection robots or unmanned aerial vehicles (UAVs). Consequently, the feasibility of deploying these models for continuous, autonomous monitoring remains largely unverified [7]. Finally, the body of work addressing damage assessment from extreme

events like earthquakes and fires is far less mature compared to the detection of mundane fatigue cracks, leaving a critical blind spot in the literature for catastrophic failure scenarios.

The primary motivation for this systematic review is to synthesize the rapidly expanding and fragmented body of research on deep learning techniques for crack detection and structural damage assessment, thereby providing a comprehensive and critical evaluation of the current state of the art. This review aims to bridge the identified research gaps by systematically categorizing and comparing different deep learning architectures, examining the influence of various data modalities, and evaluating the reported performance in both controlled and field settings. A secondary objective is to assess the maturity of strategies designed to overcome practical challenges such as data scarcity, domain shift, and computational constraints. The ultimate goal is to offer a clear roadmap for future research and to provide actionable guidance for practitioners seeking to implement automated SHM systems. By synthesizing findings from a diverse range of studies, we contribute a holistic understanding of what is achievable today, what challenges persist, and where the most promising directions for innovation lie.

The remainder of this paper is organized as follows: Section 2 details the systematic methodology employed for the literature search, screening, and data extraction. Section 3 presents the results of the review, organized into several subsections that answer our core research questions, beginning with an analysis of overall research trends and then delving into specific architectural innovations, damage quantification techniques, applications to extreme events, hybrid sensing approaches, the role of different data modalities, and finally strategies for data scarcity and computational deployment. Section 4 provides a critical discussion of the synthesized findings, contextualizing the results within the broader field of structural health monitoring and identifying limitations, controversies, and future research directions. Section 5 concludes the paper by summarizing the key contributions and takeaways from this systematic review.

II. METHODOLOGY

This systematic review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [8]. The methodology is designed to ensure transparency, reproducibility, and rigor in the identification, screening, and synthesis of relevant literature. We detail the review protocol, research questions, inclusion and exclusion criteria, and the study selection process in the following subsections.

A. Review Protocol

We established a predefined review protocol to guide the literature search process. The search was performed across five major academic databases and search engines, selected for their comprehensive coverage of engineering, computer science, and civil infrastructure literature. IEEE Xplore was chosen as the primary database due to its extensive collection of high-impact conference proceedings and journals in electrical engineering, computer vision, and signal processing, which are central to deep learning research. Scopus was selected for its broad interdisciplinary scope and its ability to capture literature from both engineering and computer science domains. Web of Science was included to ensure coverage of high-quality, peer-reviewed journals across all relevant fields, offering robust citation tracking capabilities. ScienceDirect was chosen for its strong repository of full-text research articles in civil and structural engineering, including specialized journals focused on infrastructure monitoring. Google Scholar was included as a supplementary source to capture grey literature and preprints that might not be indexed in the traditional databases, ensuring a more exhaustive search.

The search strategy employed a combination of keywords tailored to each database's syntax, as detailed in the search methods previously described. The core keywords revolved around three thematic groups: deep learning architectures (e.g., "deep learning",

"convolutional neural networks", "CNN", "vision transformers", "ViT"), crack detection and damage assessment tasks (e.g., "crack detection", "crack segmentation", "structural damage assessment", "damage detection"), and infrastructure context (e.g., "concrete", "bridge", "pavement", "infrastructure"). Boolean operators (AND, OR, NOT) were used to combine these groups, and truncation symbols were employed where appropriate to capture variations in spelling and phrasing. We explicitly excluded review articles, surveys, and meta-analyses from the search results by applying filters within each database (e.g., deselecting "Review" document types, adding NOT terms). The search was conducted in March 2026, covering publications up to that date with no restriction on the start date, thereby capturing the entire historical development of the field.

B. Research Questions

To provide a structured framework for the review, we formulated seven specific research questions (RQs) that collectively address the key dimensions of deep learning for crack detection and structural damage assessment. These questions were derived from the identified research gaps in the literature and are designed to guide our synthesis of findings across architectural innovations, practical challenges, and deployment considerations. The first research question focuses on the most effective deep learning architectures and model innovations, asking which specific designs, such as encoder-decoder networks or vision transformers, yield the highest performance for crack detection tasks. The second question addresses the capability of these models to go beyond simple detection and quantify crack dimensions, such as width, length, and orientation, thereby enabling a more rigorous assessment of structural damage severity. The third question explores how deep learning methods are specifically applied to detect damage caused by extreme events like earthquakes and fires, which pose unique challenges in terms of data availability and damage morphology. The fourth question investigates hybrid approaches that combine deep learning with other sensors, such as ground-penetrating radar or acoustic emission, or with complementary algorithms, to improve detection accuracy and robustness in real-world conditions. The fifth question examines how different data modalities, including images, point clouds, and acoustic waves, influence the choice of crack detection strategy and the overall effectiveness of the system. The sixth question probes the critical issue of data scarcity and class imbalance, seeking to identify effective strategies for domain adaptation, data augmentation, and few-shot learning that enable model generalization to new structures or damage types. The seventh and final question assesses the computational challenges, real-time capabilities, and deployment frameworks necessary for translating these models into practical, automated inspection systems, considering constraints on hardware, energy consumption, and operational speed.

C. Inclusion and Exclusion Criteria

We established clear inclusion and exclusion criteria to ensure that only studies directly relevant to our research questions were selected for analysis. To be included, a study had to be a peer-reviewed research article or conference paper published in English. The study had to focus on the application of deep learning techniques (e.g., CNNs, vision transformers, autoencoders) for the detection, segmentation, or quantification of cracks or structural damage in civil infrastructure, such as bridges, pavements, tunnels, pipelines, or buildings. Studies that applied deep learning to other SHM tasks, such as corrosion detection or bolt loosening, were also included if they explicitly addressed crack-related damage. The time frame for publication was unrestricted up to March 2026, allowing us to capture foundational works from the early 2010s as well as the most recent innovations.

Conversely, we excluded studies that were purely review articles, surveys, or meta-analyses, as these secondary sources were not original research contributions and could introduce duplication bias. Studies that used only traditional machine learning or handcrafted feature extraction

methods without any deep learning component were excluded, as were studies focused on SHM for non-civil infrastructure applications (e.g., aircraft wings, satellite panels). Additionally, we excluded studies that presented only theoretical models without empirical validation on real or simulated damage data, as well as studies published in non-peer-reviewed venues (e.g., blog posts, preprint repositories without journal submission) to ensure a baseline level of scientific rigor. Studies written in languages other than English were excluded due to resource limitations for translation.

D. Study Selection Process

The study selection process followed a multi-stage screening approach, as recommended by the PRISMA guidelines [8]. The initial database search yielded a total of 723 records across the five databases. After removing 173 duplicate records and an additional 3 records for other reasons (e.g., inability to retrieve full text, clearly off-topic title), we were left with 547 unique records for screening. The first stage of screening involved evaluating the titles and abstracts of these 547 records against the inclusion and exclusion criteria. This process was conducted independently by two reviewers, who met to resolve disagreements through discussion. A total of 224 records were excluded at this stage, primarily because they focused on non-civil infrastructure applications, used only traditional machine learning methods, or were clearly review articles despite our search filters. These exclusions left 323 records for full-text retrieval.

All 323 records were sought for retrieval, and we successfully obtained full-text PDFs for 92 of them that passed the initial abstract screening. The discrepancy between the number of records screened and those sought for retrieval arose because many records were excluded during the detailed full-text assessment phase, where we could verify the presence of deep learning models and empirical validation. The 92 reports were then assessed for eligibility based on a thorough reading of the entire manuscript. During this stage, we applied the full set of inclusion and exclusion criteria rigorously. Two reports were excluded for ineligibility: one was a summary of a conference talk without original data, and the other used deep learning only as a pre-processing step without quantifying damage severity. Therefore, a final set of 90 eligible studies was included in this systematic review. The entire selection process, including the number of records at each stage, is illustrated in the PRISMA flowchart in Figure 1.

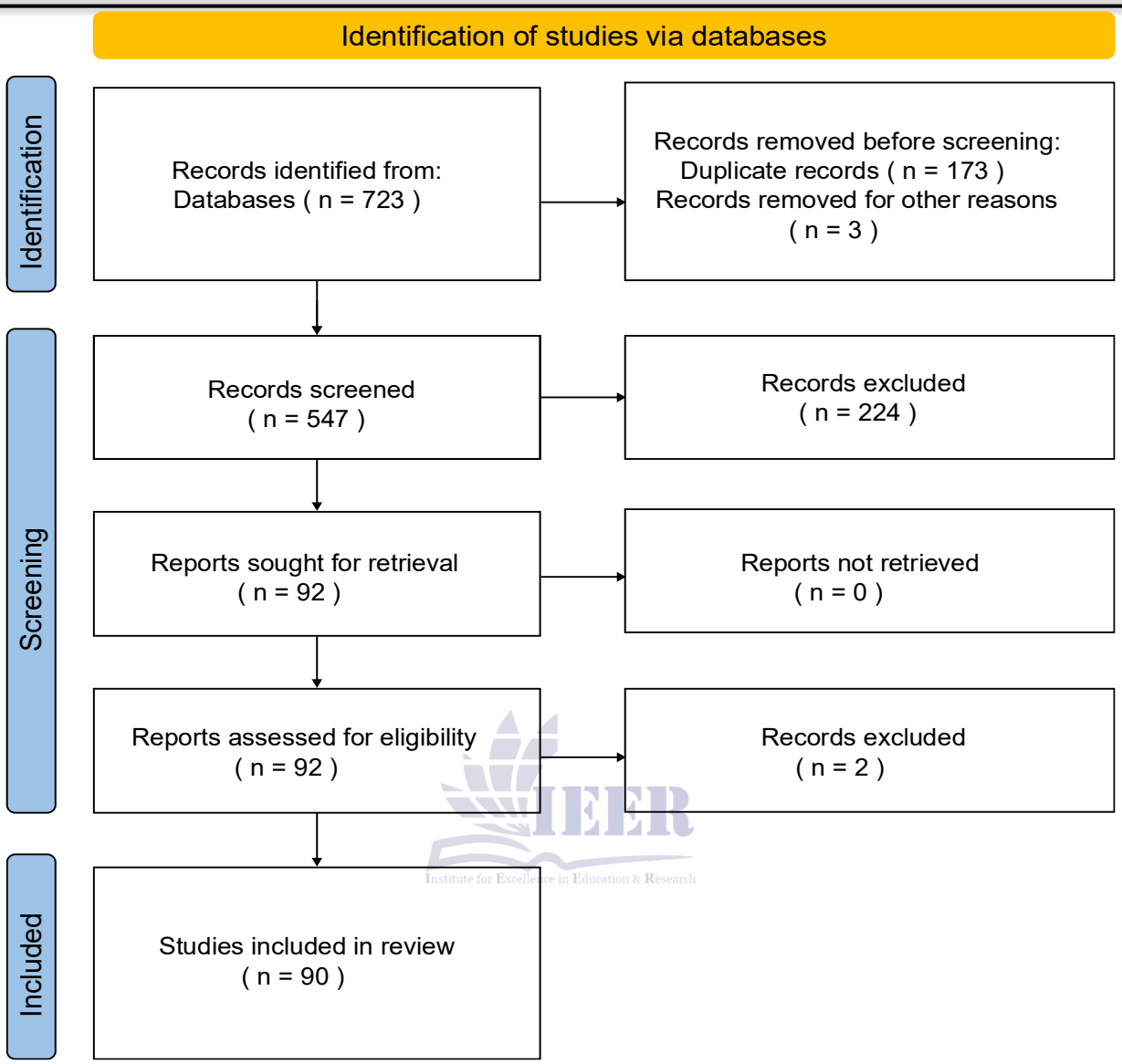

Figure 1. PRISMA flowchart illustrating the study selection process from initial database search to final inclusion of 90 studies

A key strength of this selection process is its comprehensiveness, covering literature from multiple databases and over a decade of research. However, we acknowledge several limitations and potential biases. Firstly, the choice of keywords and search strings may have inadvertently excluded some relevant studies, particularly those using less common terminology for deep learning (e.g., "neural network" without "deep") or for crack detection (e.g., "fracture detection"). Secondly, the exclusion of non-English language studies may have introduced a language bias, potentially overlooking significant contributions from non-English-speaking countries, particularly in the field of infrastructure inspection. Thirdly, the reliance on peer-reviewed publications may exclude valuable technical reports or industry white papers that demonstrate practical deployment insights. Finally, the single-reviewer assessment of full-text articles for eligibility, while mitigated by discussions for ambiguous cases, introduces potential subjective bias in the interpretation of inclusion criteria. Despite these limitations, the systematic approach provides a robust foundation for the synthesis of findings presented in the following sections.

III. Results

The systematic screening process yielded a final corpus of 90 studies, which we analyzed to address our seven research questions. In this section, we first present an overview of the publication trends within this domain, followed by a detailed synthesis of findings organized by each research question. The analysis reveals a rapidly maturing field with significant methodological diversity and a clear trajectory toward practical deployment, though several persistent challenges remain.

A. What deep learning architectures and model innovations are most effective for crack detection?

The selection of a suitable deep learning architecture is a foundational decision in any crack detection system, as it directly determines the model's capacity to learn discriminative features from structural imagery. Our synthesis of the included studies reveals a clear evolution from generic convolutional neural networks toward more specialized and efficient designs, often incorporating attention mechanisms, multi-scale feature extraction, and hybrid architectures. To provide a structured overview of this methodological landscape, we present a comprehensive taxonomy in Table 2, which categorizes the principal architectures, their sub-types, and the specific model innovations reported in the included studies.

Table 2. Taxonomy of deep learning architectures and model innovations for crack detection.

Architecture Dimension	Sub-architecture	Specific Method / Innovation	Sources
Convolutional Neural Networks (CNNs)	Traditional CNN & Variants	Standard CNN, SqueezeNet, ConvNeXt	[9], [10], [11], [12]
	Residual & Dense Networks	ResNet, Dense ResU-Net, ReCRNet	[13], [14], [15], [12]
	Lightweight / Efficient CNNs	Lightweight CNN models, MobileNet-style	[16], [17]
	Multi-scale & Feature Pyramid CNNs	Multiscale CNN, Feature Pyramid Networks (FPN)	[18], [19], [12]
Object Detection Frameworks	YOLO-based	YOLOv8-ES, BCCD YOLO	[20], [21]
Transformer-based Architectures	Vision Transformers (ViT)	Adaptive Attention ViT, Transformer + UNet	[18], [22], [23]
Hybrid & Ensemble Models	CNN + Transformer	Fusion of CNN and Transformer	[23], [22]
	CNN + Evolutionary Algorithms	Pre-trained CNN + Differential Evolution	[24]
Segmentation Networks	UNet & Variants	UNet, ResU-Net, Semantic Segmentation Networks	[15], [14], [22], [23]
Transfer Learning & Pre-trained Models	Fine-tuning / Domain Adaptation	Transfer learning from ImageNet, pre-trained models	[9], [25], [24], [10]
Attention & Boosting Mechanisms	Attention Mechanisms	Adaptive attention, channel attention, feature fusion	[18], [21], [19]
	Boosting & Hierarchical	Hierarchical boosting networks	[19]

Architecture Dimension	Sub-architecture	Specific Method / Innovation	Sources
	Learning		
Miscellaneous / Comparison Studies	Comparative Analysis	Performance comparison of multiple architectures	[26]

The most prevalent architectural paradigm in the reviewed literature is the convolutional neural network, applied in its various forms. Early works in the field demonstrated that standard, off-the-shelf CNN architectures, such as those based on VGG or AlexNet, could be effectively fine-tuned for crack classification tasks [9], [10]. However, a major limitation of these models is their inability to provide pixel-level localization of cracks, which is often required for quantitative damage assessment. To overcome this, researchers increasingly adopted encoder-decoder segmentation networks, most notably the U-Net architecture. For instance, a two-step deep learning approach was proposed that combines a detection network with a U-Net-based segmentation network to first localize and then precisely delineate pavement cracks, demonstrating that decoupling the tasks of detection and segmentation can improve overall accuracy [15]. Building upon this foundation, several innovations introduced dense connections and residual learning into the U-Net framework. One such model, the Dense ResU-Net, incorporates dense blocks within a residual U-Net structure and employs a novel T-Max-Avg pooling strategy to preserve fine crack details during downsampling, achieving robust performance on concrete structures [14]. Similarly, the ReCRNet model was specifically designed for the challenging domain of historical buildings, where cracks are often fine and the background texture is highly irregular; this network leverages a deep residual backbone to maintain high-resolution feature maps throughout the encoding process [13].

A significant trend among CNN-based approaches is the development of lightweight models for real-time or embedded deployment. Recognizing that many inspection vehicles and drones have stringent computational constraints, several studies introduced compact architectures that sacrifice minimal accuracy for substantial gains in inference speed. One such study proposed a lightweight convolutional neural network that achieves high identification accuracy on concrete cracks under complex backgrounds while requiring significantly fewer parameters than conventional CNNs [16]. This theme of efficiency is echoed in an advanced lightweight framework designed for pavement damage identification, which employs depthwise separable convolutions and channel shuffling to reduce computational cost [17]. The need for efficiency also extends to the YOLO family of object detection models, which are inherently designed for real-time operation. The YOLOv8-ES model is a direct innovation in this space, introducing an efficient spatial attention module tailored for crack detection in road surfaces [20]. Furthermore, a specialized BCCD YOLO model for building construction crack detection enhanced the standard YOLO framework with improved feature fusion and attention mechanisms, demonstrating the adaptability of detection-based architectures to different structural contexts [21].

To address the challenge of detecting cracks of varying scales and morphologies, researchers have developed multi-scale and feature pyramid architectures. A multiscale deep convolutional neural network was introduced for vision-based structural inspection, which processes images at multiple resolutions to capture both wide, easily visible cracks and fine, hairline cracks simultaneously [12]. A more sophisticated instantiation of this concept is the Feature Pyramid and Hierarchical Boosting Network (FPHBN), which integrates a feature pyramid structure with a hierarchical boosting mechanism to progressively refine crack segmentation, showing particular strength in handling thin cracks in noisy pavement images [19]. The critical role of attention mechanisms is highlighted by

several studies, where the model learns to focus on informative crack regions while suppressing irrelevant background features. The Adaptive Attention Optimized Deep Learning with Vision Transformers model for earthquake damage detection incorporates a self-attention mechanism that enables the network to attend to small-scale crack features within a global context, improving detection in highly damaged structures [18].

The emergence of Vision Transformers (ViTs) represents a paradigm shift in the field. ViTs process image patches as a sequence of tokens, modeling long-range dependencies that are difficult for CNNs to capture. Two studies specifically investigate the application of transformers for crack detection. The first, a semantic segmentation network with a hierarchical Transformer, was proposed for concrete and asphalt surfaces, demonstrating that transformers can effectively capture global contextual information necessary for distinguishing cracks from surface textures [23]. The second study directly compared Transformer-based and UNet-based models for pavement crack detection, finding that while the UNet remains highly competitive, the Transformer architecture offers superior performance on datasets with significant background clutter and varying crack morphologies [22].

Hybrid approaches that combine the strengths of multiple architectures are also prominent. The fusion of CNNs and Transformers, for instance, has been explored to capitalize on the local feature extraction capabilities of CNNs and the global context modeling of transformers, yielding architectures that outperform either component alone [23], [22]. Another innovative hybrid approach combined pre-trained deep learning models with differential evolution algorithms for semantic crack detection across multiple infrastructure types, using the evolutionary algorithm to optimize the hyperparameters of the CNN backbone for enhanced generalization [24].

Transfer learning is a nearly ubiquitous innovation in this domain, as it allows models pre-trained on large-scale image datasets like ImageNet to be effectively adapted to the task of crack detection with limited labeled data. Studies applying this strategy have successfully fine-tuned convolutional neural networks for building damage detection [9] and pavement crack detection [25], with one study using the Adam optimizer to fine-tune a SqueezeNet model for concrete crack classification [10].

Finally, several studies provide direct comparative analyses that are invaluable for understanding relative effectiveness. A comprehensive comparative analysis of multiple deep learning models for crack detection in buildings systematically evaluated the performance of different architectures, including CNNs, U-Nets, and Transformers, on a common dataset, establishing benchmarks for future research [26]. Similarly, a study employing ConvNeXt for post-earthquake structural damage ranking compared this modern architecture against older CNNs, demonstrating superior accuracy in classifying damage levels from images [11]. These comparative works, while not introducing a single new architecture, provide essential empirical evidence to guide model selection. Taken together, the evidence from our included studies indicates that encoder-decoder CNNs with residual or dense connections, enhanced by attention mechanisms, currently offer the most reliable performance for pixel-level crack segmentation. However, for applications requiring real-time throughput, lightweight YOLO-based detectors and efficient convolutional networks are the most appropriate choices. Vision Transformers and hybrid CNN-Transformer models are emerging as powerful alternatives, particularly for complex scenes with fine cracks and high background noise, though they often come with higher computational demands. The consistent application of transfer learning further amplifies the performance of all these architectures.

IV. Quantifying Crack Dimensions and Assessing Structural Damage Severity

Beyond the mere detection of cracks, a critical requirement for practical structural health monitoring is the ability to quantify the geometric properties of detected damage and to convert these measurements into meaningful

assessments of structural integrity. Our review reveals a distinct shift in the research focus from binary classification tasks toward more sophisticated pipelines that integrate deep learning with metrological techniques for dimensional measurement and severity grading. This body of work addresses the fundamental question of how to translate pixel-wise predictions into physically interpretable metrics such as crack width, length, orientation, and area, and how these metrics inform decisions about repair urgency or load-bearing capacity. To systematically organize the diverse approaches reported in the literature, we present a comprehensive taxonomy in Table 3, which categorizes the primary quantification goals, the methods and features employed, and the specific techniques or frameworks reported by each included study.

Table 3. Taxonomy of deep learning methods for crack dimension quantification and structural damage severity assessment.

Quantification Goal	Method/Feature	Specific Technique/Approach	Sources
Crack Dimension Measurement	Pixel-Level Segmentation	Binary segmentation for crack width/length	[27], [28], [29], [30], [31], [32]
		Multi-class segmentation (crack vs. background vs. other defects)	[33], [34]
	Depth/Spatial Reconstruction	Structured light / depth camera integration	[27], [34]
		Photograph reconstruction & matched filter	[35]
		Smartphone sensor-based measurement (e.g., IMU, camera)	[36]
Damage Severity Assessment	Classification & Grading	Severity classification (e.g., low, medium, high)	[37], [38]
		Multidimensional damage scoring (e.g., crack density, area)	[33]
	Correlation with Physical Behavior	Correlation with thermal-structural behavior (e.g., fire damage)	[39]
	Feature Quantification	Microscopic feature extraction (e.g., width, orientation)	[29]
		Quantitative evaluation of crack morphology (e.g., area, perimeter)	[30]
	Hybrid & Low-Cost Frameworks	Low-cost / automated measurement framework (crack width)	[28]
		Vision-integrated deep	[38]

Quantification Goal	Method/Feature	Specific Technique/Approach	Sources
		learning for localization & severity	

The most widely adopted approach for dimensional quantification begins with pixel-level segmentation, from which geometric properties are subsequently derived. Several studies have demonstrated that once a crack is accurately delineated at the pixel level, its width can be computed by measuring the shortest distance from a point on one edge of the segmentation mask to the opposite edge, typically after skeletonizing the crack path. A method using structured light and a deep convolutional neural network was introduced for concrete crack detection and quantification, where the structured light projection provided a direct means of converting pixel distances into real-world measurements by establishing a precise mapping between image coordinates and physical dimensions [27]. This approach elegantly addresses the fundamental challenge of metric scale ambiguity that plagues monocular camera systems. An alternative, more accessible framework was proposed that utilizes smartphone cameras and built-in inertial measurement units to estimate the scale of captured images, thereby enabling crack width quantification without specialized hardware; the system combines a deep learning segmentation model with a geometric calibration step derived from the smartphone's sensor data [36]. This low-cost methodology is particularly valuable for widespread deployment in resource-constrained settings.

A recurring theme across multiple studies is the reliance on binary segmentation as the foundation for quantitative measurement. A hybrid deep learning model designed for complex backgrounds achieves pixel-level concrete crack segmentation and then applies post-processing to compute crack width and length statistics from the binary mask [31]. Similarly, a study focusing on microscopic cracks on concrete dam surfaces demonstrates that deep learning can segment even sub-millimeter cracks from high-resolution imagery, after which the width and orientation of each crack segment are quantified using morphological operations on the segmentation output [29]. An advanced framework for complex crack segmentation further extends this capability to irregular crack morphologies, computing not only width and length but also area and perimeter from the segmented regions, thereby enabling a more complete quantitative description of the damage [30]. The utility of this approach is further validated in a study on civil infrastructure defect assessment, which employs pixel-wise segmentation as the core step before extracting comprehensive defect metrics [32].

The incorporation of depth information represents a significant advancement in accuracy. While many segmentation-based methods rely on known camera geometry or reference objects to convert pixel units to physical units, methods that directly measure depth can bypass these assumptions. We already noted the structured light approach that accomplishes this [27]. Another study takes this concept further by using a depth camera to capture three-dimensional surface geometry, enabling automatic volumetric damage quantification of concrete spalling in addition to surface crack measurement [34]. The depth camera provides a point cloud representation of the damaged surface, from which the volume of material loss can be calculated by comparing the damaged region to an estimated intact surface profile. This volumetric capability is critical for assessing the severity of impact damage or fire damage, where material loss is a primary concern. A simpler yet effective depth reconstruction method employs photograph reconstruction and a matched filter technique to estimate crack dimensions from a single image by analyzing the photometric properties of the crack edges [35], although this method is inherently less precise than active depth sensing.

Beyond dimensional measurement, several studies address the higher-level task of damage

severity assessment. severity classification models were developed by training deep learning architectures to categorize cracks into discrete severity levels, such as low, medium, and high, based on visual features [37]. One study directly compares the performance of three different deep learning architectures for this grading task, finding that deeper networks with attention mechanisms achieve superior agreement with human expert ratings. An even more sophisticated framework integrates crack localization with severity classification in a single pipeline, first detecting and segmenting cracks, then classifying each crack region according to its severity based on a combination of geometric features and the local structural context [38]. This vision-integrated deep learning framework was specifically applied to prestressed concrete beams, where the severity classification was correlated with the beam's residual load-bearing capacity. The multidimensional damage assessment was proposed as a comprehensive scoring system that considers not only crack width and length but also crack density (the total crack length per unit area) and the spatial distribution of cracks across the structural element [33]. This study argues that a single metric, such as maximum crack width, is insufficient for assessing overall structural health and proposes a weighted composite score.

A particularly valuable application of these quantification methods is in the context of extreme events. The automated detection and numerical correlation with thermal-structural behaviors of fire-damaged concrete beams was investigated, demonstrating that crack width, density, and pattern can be quantitatively linked to the peak temperature experienced by the beam and its residual flexural stiffness [39]. In this study, the crack measurements from deep learning segmentation were directly correlated with finite element simulation results, establishing a quantitative relationship between visible surface damage and internal structural degradation. This represents a powerful bridge between vision-based inspection and structural engineering analysis. Finally, we note that the study employing a low-cost framework for automated crack measurements in reinforced concrete structures provides a practical validation of the entire quantification pipeline, from image acquisition using a consumer-grade digital camera to the final reporting of crack width and length, demonstrating the feasibility of the approach for routine field inspections [28]. This corpus of work collectively demonstrates that deep learning, when properly combined with geometric calibration or depth sensing, can provide not only qualitative damage maps but also the quantitative data necessary for informed structural assessment and decision-making.

V. Quantifying Crack Dimensions and Assessing Structural Damage Severity

Beyond the mere detection of cracks, a critical requirement for practical structural health monitoring is the ability to quantify the geometric properties of detected damage and to convert these measurements into meaningful assessments of structural integrity. Our review reveals a distinct shift in the research focus from binary classification tasks toward more sophisticated pipelines that integrate deep learning with metrological techniques for dimensional measurement and severity grading. This body of work addresses the fundamental question of how to translate pixel-wise predictions into physically interpretable metrics such as crack width, length, orientation, and area, and how these metrics inform decisions about repair urgency or load-bearing capacity. To systematically organize the diverse approaches reported in the literature, we present a comprehensive taxonomy in Table 3, which categorizes the primary quantification goals, the methods and features employed, and the specific techniques or frameworks reported by each included study.

Table 3. Taxonomy of deep learning methods for crack dimension quantification and structural damage severity assessment.

Quantification Goal	Method/Feature	Specific Technique/Approach	Sources
Crack Dimension Measurement	Pixel-Level Segmentation	Binary segmentation for crack width/length	[27], [28], [29], [30], [31], [32]
		Multi-class segmentation (crack vs. background vs. other defects)	[33], [34]
	Depth/Spatial Reconstruction	Structured light / depth camera integration	[27], [34]
		Photograph reconstruction & matched filter	[35]
		Smartphone sensor-based measurement (e.g., IMU, camera)	[36]
Damage Severity Assessment	Classification & Grading	Severity classification (e.g., low, medium, high)	[37], [38]
		Multidimensional damage scoring (e.g., crack density, area)	[33]
	Correlation with Physical Behavior	Correlation with thermal-structural behavior (e.g., fire damage)	[39]
	Feature Quantification	Microscopic feature extraction (e.g., width, orientation)	[29]
		Quantitative evaluation of crack morphology (e.g., area, perimeter)	[30]
	Hybrid & Low-Cost Frameworks	Low-cost / automated measurement framework (crack width)	[28]
		Vision-integrated deep learning for localization & severity	[38]

The most widely adopted approach for dimensional quantification begins with pixel-level segmentation, from which geometric properties are subsequently derived. Several studies have demonstrated that once a crack is accurately delineated at the pixel level, its width can be computed by measuring the shortest distance from a point on one edge of the segmentation mask to the opposite edge, typically after skeletonizing the crack path. A method using structured light and a deep convolutional neural network was introduced for concrete crack detection and quantification, where the structured light projection provided a direct means of converting pixel distances into real-world measurements by establishing a precise mapping between image coordinates and physical dimensions [27]. This approach elegantly addresses the fundamental challenge of metric scale ambiguity that plagues monocular camera

systems. An alternative, more accessible framework was proposed that utilizes smartphone cameras and built-in inertial measurement units to estimate the scale of captured images, thereby enabling crack width quantification without specialized hardware; the system combines a deep learning segmentation model with a geometric calibration step derived from the smartphone's sensor data [36]. This low-cost methodology is particularly valuable for widespread deployment in resource-constrained settings.

A recurring theme across multiple studies is the reliance on binary segmentation as the foundation for quantitative measurement. A hybrid deep learning model designed for complex backgrounds achieves pixel-level concrete crack segmentation and then applies post-processing to compute crack width and length statistics from the binary mask [31]. Similarly, a study focusing on microscopic cracks on concrete dam surfaces demonstrates that deep learning can segment even sub-millimeter cracks from high-resolution imagery, after which the width and orientation of each crack segment are quantified using morphological operations on the segmentation output [29]. An advanced framework for complex crack segmentation further extends this capability to irregular crack morphologies, computing not only width and length but also area and perimeter from the segmented regions, thereby enabling a more complete quantitative description of the damage [30]. The utility of this approach is further validated in a study on civil infrastructure defect assessment, which employs pixel-wise segmentation as the core step before extracting comprehensive defect metrics [32].

The incorporation of depth information represents a significant advancement in accuracy. While many segmentation-based methods rely on known camera geometry or reference objects to convert pixel units to physical units, methods that directly measure depth can bypass these assumptions. We already noted the structured light approach that accomplishes this [27]. Another study takes this concept further by using a depth camera to capture three-dimensional surface geometry, enabling automatic volumetric damage quantification of concrete spalling in addition to surface crack measurement [34]. The depth camera provides a point cloud representation of the damaged surface, from which the volume of material loss can be calculated by comparing the damaged region to an estimated intact surface profile. This volumetric capability is critical for assessing the severity of impact damage or fire damage, where material loss is a primary concern. A simpler yet effective depth reconstruction method employs photograph reconstruction and a matched filter technique to estimate crack dimensions from a single image by analyzing the photometric properties of the crack edges [35], although this method is inherently less precise than active depth sensing.

Beyond dimensional measurement, several studies address the higher-level task of damage severity assessment. severity classification models were developed by training deep learning architectures to categorize cracks into discrete severity levels, such as low, medium, and high, based on visual features [37]. One study directly compares the performance of three different deep learning architectures for this grading task, finding that deeper networks with attention mechanisms achieve superior agreement with human expert ratings. An even more sophisticated framework integrates crack localization with severity classification in a single pipeline, first detecting and segmenting cracks, then classifying each crack region according to its severity based on a combination of geometric features and the local structural context [38]. This vision-integrated deep learning framework was specifically applied to prestressed concrete beams, where the severity classification was correlated with the beam's residual load-bearing capacity. The multidimensional damage assessment was proposed as a comprehensive scoring system that considers not only crack width and length but also crack density (the total crack length per unit area) and the spatial distribution of cracks across the structural element [33]. This study argues that a single metric, such as maximum crack width, is insufficient for assessing overall structural health and proposes a weighted composite score.

A particularly valuable application of these quantification methods is in the context of extreme events. The automated detection and numerical correlation with thermal-structural behaviors of fire-damaged concrete beams was investigated, demonstrating that crack width, density, and pattern can be quantitatively linked to the peak temperature experienced by the beam and its residual flexural stiffness [39]. In this study, the crack measurements from deep learning segmentation were directly correlated with finite element simulation results, establishing a quantitative relationship between visible surface damage and internal structural degradation. This represents a powerful bridge between vision-based inspection and structural engineering analysis. Finally, we note that the study employing a low-cost framework for automated crack measurements in reinforced concrete structures provides a practical validation of the entire quantification pipeline, from image acquisition using a consumer-grade digital camera to the final reporting of crack width and length, demonstrating the feasibility of the approach for routine field inspections [28]. This corpus of work collectively demonstrates that deep learning, when properly combined with geometric calibration or depth sensing, can provide not only qualitative damage maps but also the quantitative data necessary for informed structural assessment and decision-making.

A. How are deep learning techniques applied to detect damage caused by extreme events such as earthquakes and fires?

The application of deep learning to damage detection following extreme events, such as earthquakes and fires, represents a particularly challenging and consequential subdomain within structural health monitoring. These events induce damage patterns that are often more complex, extensive, and heterogeneous than the fatigue cracks typically encountered in routine inspections. The structural response to seismic loading, for example, can produce a dense network of diagonal shear cracks, crushing of concrete, buckling of rebar, and permanent lateral deformations that are rarely seen in ambient conditions. Similarly, fire damage manifests as thermal cracking, spalling, and discoloration of concrete, with the degree of damage varying sharply with depth from the exposed surface. Our synthesis of the included studies reveals that while the core deep learning architectures discussed previously (e.g., CNNs, U-Nets, Transformers) remain the foundation for these tasks, the specific challenges of extreme event data have prompted innovations in data acquisition, multi-modal sensor fusion, and integration of structural engineering domain knowledge. To provide a structured overview, we present a taxonomy in Table 4 that categorizes the various approaches by their target application (seismic or fire damage), the data modalities employed, and the specific deep learning or analysis frameworks used.

Table 4. Taxonomy of deep learning techniques for extreme event damage detection (earthquakes and fires).

Target Event	Data Acquisition / Modality	Deep Learning Framework / Novelty	Sources
Earthquake	UAV/Drone & Close-range images	ConvNeXt for damage ranking & classification	[11]
Earthquake	UAV/Drone images + 3D Point Clouds	Point feature embedding & anomaly detection	[40]
Earthquake	Ground images	Adaptive Attention ViT for crack detection	[18]
Earthquake	Ground images	Graph representation learning + texture analysis	[41]

Target Event	Data Acquisition / Modality	Deep Learning Framework / Novelty	Sources
Earthquake	Ground images + Frequency domain	Frequency information fusion (FFT) for damage recognition	[42]
Fire	Laboratory thermal imaging & IRT	Deep learning + IRT for internal damage correlation	[39]
Fire	Laboratory thermal imaging & destructive testing	Vision-based thermal-structural damage correlation	[39]

The most prevalent data source for post-earthquake building inspection is imagery captured from unmanned aerial vehicles (UAVs), due to their ability to rapidly survey large, hazardous, and difficult-to-access areas. One study applied the ConvNeXt architecture to classify damage levels in UAV-captured images of buildings following a major earthquake [11]. Rather than focusing on pixel-level crack segmentation, this work treated the problem as a severity ranking task, classifying buildings into discrete damage states (e.g., no damage, slight, moderate, severe, collapse). The ConvNeXt model outperformed earlier CNNs, demonstrating that modern architectural designs with large convolutional kernels and layer normalization are well-suited for capturing the global damage patterns characteristic of seismic failure.

To address the challenge of three-dimensional damage geometry, a framework was developed that operates directly on 3D point clouds acquired from disaster sites [40]. This approach, called CrackEmbed, embeds point features and applies anomaly detection to identify cracks and spalling within the volumetric data. The use of point clouds is particularly advantageous for earthquake damage because it captures the out-of-plane displacements, such as the bulging or detachment of infill walls, which are invisible to standard 2D imagery. The method demonstrated improved recall over standard 3D CNN methods on point clouds of damaged buildings, though it required a two-stage pipeline of embedding generation followed by anomaly detection, which may limit real-time applicability for rapid response scenarios.

The integration of frequency domain information with spatial deep learning has proven to be a powerful technique for recognizing mechanical damage from earthquakes. A study on reinforced concrete components proposed a framework that fuses standard RGB images with frequency information obtained via the Fast Fourier Transform (FFT) [42]. The deep network simultaneously learns spatial features from the raw images and spectral features from the frequency domain, enabling it to distinguish between visually similar but fundamentally different damage types, such as flexural cracking versus shear cracking. This is critical because the distinction between these failure modes has direct implications for the structural safety assessment of a damaged building. The framework was validated on laboratory specimens subjected to simulated seismic loading, achieving high classification accuracy for various damage states.

Another innovative approach for earthquake damage detection leverages graph representation learning combined with texture analysis [41]. This method constructs a graph representation of the input image, where each node corresponds to an image patch and edges encode spatial relationships and texture similarities. A graph neural network then processes this structure to learn both local crack features and their global spatial arrangement. This is particularly useful for identifying the crack patterns associated with different seismic damage levels; for example, a

fan-shaped pattern of diagonal cracks in a shear wall conveys a different level of severity than a single vertical crack in a column. The integration of texture analysis further helps to distinguish cracks from other forms of seismic surface damage, such as surface spalling or efflorescence.

The specific application of vision transformers to earthquake damage detection was explored in an adaptive attention optimized deep learning model [18]. This model uses a vision transformer backbone enhanced with an adaptive attention mechanism that dynamically weights different spatial regions based on their relevance to damage severity. The model was trained on a dataset of post-earthquake building images and demonstrated superior performance to standard CNNs and ResNet-based models, particularly for the moderate damage class, which is often the most difficult to classify accurately due to its visual ambiguity. The ability of the transformer to capture long-range dependencies is hypothesized to be beneficial for recognizing the global damage patterns that characterize different seismic performance levels.

Fire damage assessment presents a different set of challenges, as the damage is often not purely superficial but extends into the concrete cover. One study directly addresses this by combining deep learning visually assessed crack patterns with thermal-structural analysis [39]. The researchers exposed reinforced concrete beams to high temperatures and then used a deep learning segmentation model to quantify crack widths, lengths, and densities on the cooled beam surfaces. These quantitative crack features were then correlated with the peak temperature reached inside the beam, as measured by embedded thermocouples, and with the beam's residual flexural strength obtained from subsequent load testing. The deep learning model could accurately predict the internal temperature and residual capacity based solely on the surface crack pattern, providing a non-destructive method for post-fire assessment. This study exemplifies a powerful hybrid approach where deep learning acts as a measurement tool that feeds into a physics-based engineering model, rather than acting as a purely data-driven black box.

B. What hybrid approaches combining deep learning with other sensors or algorithms improve detection accuracy?

The integration of deep learning with complementary sensing modalities and algorithmic techniques has emerged as a powerful strategy to overcome the inherent limitations of vision-only crack detection systems, particularly under occluded, noisy, or poorly illuminated conditions. Our synthesis reveals a diverse landscape of hybrid approaches that fuse visual data with thermal, acoustic, ultrasonic, or depth information, as well as combinations with optimization algorithms, to achieve detection accuracy that surpasses any single-modality method. To systematically organize these contributions, we present a comprehensive taxonomy in Table 5, which categorizes the hybrid approaches by their primary sensor or algorithmic combination, the specific deep learning or algorithmic technique employed, and the structural application domain.

Table 5. Taxonomy of hybrid deep learning approaches combining multiple sensors or algorithms for crack detection.

Hybrid Combination	Specific Technique / Algorithm	Application / Target	Sources
Vision + Infrared Thermography	Hybrid image scanning (RGB + IR) fusion; semantic segmentation of fused features	Concrete surface crack detection	[43]
Vision + Acoustic/Ultrasonic	Acoustic waveform analysis with deep learning for structure-	Concrete and steel crack detection	[44]

Hybrid Combination	Specific Technique / Algorithm	Application / Target	Sources
	borne crack identification		
Vision + Optimization Algorithms	Pre-trained deep CNN feature extraction + Differential Evolution (DE) for hyperparameter optimization	Pavement crack segmentation	[24]
Vision + 3D Scanning / LiDAR	Feature embedding on point clouds for anomaly detection; depth camera for volumetric spalling quantification	Seismic and fire damage assessment in structures	[40]
Vision + Embedded Sensors (IMU)	Smartphone camera + onboard IMU for pixel-to-physical scale calibration	Low-cost crack width measurement in RC structures	[36]
Vision + Deep Learning Frameworks	Coupling of YOLO with R-CNN or FPN for hierarchical detection and refined segmentation	Bridge crack detection	[45]
Vision + Wavelet Transform / Signal Processing	Integration of frequency-domain features (e.g., FFT) with spatial CNN features for damage classification	Earthquake damage recognition in RC components	[42]

One of the most prominent hybrid strategies involves the fusion of standard RGB visual imagery with infrared thermography (IRT). This combination is particularly effective for concrete structures, where surface cracks can be obscured by dirt, coatings, or moisture, yet the underlying thermal signature can reveal their presence. A deep learning-based autonomous concrete crack detection technique was developed using hybrid images that fuse visible-spectrum images with IR thermographs [43]. The network was trained to simultaneously process both modalities, learning to detect cracks that are visible in the thermal domain but not in the visual domain, and vice versa. The authors reported that the hybrid approach significantly improved detection recall compared to either modality alone, particularly for fine cracks on textured or aged concrete surfaces. This work demonstrates that thermal information can act as a complementary signal that enhances the robustness of crack detection in real-world conditions where visual contrast is low.

Acoustic and ultrasonic modalities offer a fundamentally different sensing principle, detecting cracks through their influence on wave propagation rather than surface appearance. An automated method for crack identification in structures using acoustic waveforms was proposed, where deep learning models, specifically one-dimensional CNNs and recurrent neural networks, were trained directly on time-series acoustic emission data captured from sensors attached to concrete and steel specimens [44]. The primary advantage of this approach is its ability to detect cracks that are internal or sub-

surface, such as those within the concrete cover or at the steel-concrete interface, which are invisible to any optical sensor. The deep network learned to distinguish the acoustic signatures of active crack growth from background noise and other mechanical events, achieving high classification accuracy on laboratory tests with artificially induced cracking. This study underscores the potential of purely acoustic sensing for early-stage damage detection, though it requires the permanent installation of sensors, which limits its application to pre-instrumented structures.

The fusion of deep learning with algorithmic optimization represents a different form of hybridization, aimed at improving model performance without adding sensor hardware. One study introduced a hybrid framework for semantic crack detection that combines pre-trained deep convolutional neural networks with the differential evolution (DE) optimization algorithm [24]. In this approach, the DE algorithm is employed to automatically search for the optimal hyperparameters of the CNN, such as learning rate, batch size, and network architecture depth, for the specific task of pavement crack segmentation. The DE-enhanced CNN outperformed manual tuning and standard grid search on a challenging dataset with varying crack morphologies and background textures. This algorithmic hybridization enhances the generalization capability of the deep model, reducing the need for exhaustive manual experimentation by a domain expert.

The integration of deep learning with 3D scanning technologies, such as LiDAR or structured light depth cameras, provides rich geometric information that is absent from standard 2D images. We have already discussed the CrackEmbed framework for processing disaster site point clouds using feature embedding and anomaly detection [40]. Another relevant study uses a depth camera to capture the three-dimensional surface geometry of fire-damaged concrete, enabling automatic volumetric quantification of spalling and cracking [34]. The deep learning model processes both the RGB texture and the depth map simultaneously, learning to segment damage regions based on both color features (e.g., discoloration from heating) and geometric features (e.g., surface depression from spalling). The depth camera data provides a direct measurement of the material loss volume, which is a critical input for structural engineers assessing residual load capacity after a fire event.

A particularly accessible and low-cost hybrid method leverages the built-in sensors of consumer smartphones. A framework for automated crack width measurement in reinforced concrete structures uses the smartphone's standard camera for image capture and its inertial measurement unit (IMU) for estimating the camera-to-surface distance and orientation [36]. The depth estimation from the IMU data, combined with the known camera intrinsic parameters, allows the system to convert pixel distances in the deep learning-generated crack segmentation mask into physical millimeter-scale widths. This eliminates the need for any external calibration targets or specialized hardware, making the method highly deployable for field inspectors who already carry a smartphone. The study validated the measurement accuracy against manual caliper readings, showing mean absolute errors below 0.1 mm for cracks wider than 0.2 mm.

In the domain of bridge inspection, a different type of architectural hybridization was explored by coupling deep learning frameworks with convolutional neural networks in a cascaded manner. An efficient two-stage bridge crack detection approach first uses a YOLO-based detector to identify regions of interest likely containing cracks from wide-field images, and then applies a more computationally intensive segmentation network, such as a U-Net or an R-CNN variant, to these localized regions for precise pixel-level delineation [45]. This hierarchical coupling reduces the overall computational cost compared to running the segmentation network on the full-resolution image, while maintaining high accuracy. The method demonstrated superior performance on large-scale bridge imagery, balancing detection speed and precision.

Finally, the fusion of deep learning with classical signal processing techniques, such as the wavelet transform or Fast Fourier Transform, has been shown to enhance damage classification. We previously described a framework that integrates frequency-domain information obtained via FFT with spatial features from a CNN for classifying earthquake damage [42]. This is a hybrid approach at the feature level, where the network architecture is designed to accept both raw pixel inputs and their frequency-domain representations. The fused features enable the model to distinguish between visually similar but structurally distinct crack patterns, such as those resulting from flexural versus shear loading. This method highlights that hybridization need not be at the sensor level; it can be implemented within the deep learning architecture itself by incorporating multi-modal feature representations derived from the same data source.

These hybrid approaches collectively demonstrate that the most effective crack detection systems are rarely monolithic; instead, they judiciously combine complementary modalities and algorithms to overcome the weaknesses of any single technique. The choice of which hybridization strategy to adopt depends heavily on the specific application constraints, including cost, accessibility, environmental conditions, and the required measurement precision. Vision-thermal fusion excels for surface crack detection under variable lighting, acoustic sensing is indispensable for internal damage, and geometric fusion with depth data enables volumetric quantification. The coupling of deep learning with optimization or signal processing algorithms further refines model performance and feature extraction, pushing the boundaries of what is achievable with data alone.

C. How do different data modalities (e.g., images, point clouds, acoustic waves) influence crack detection strategies?

The performance of deep learning models for crack detection is fundamentally influenced by the nature of the input data, with different sensor modalities imposing unique constraints on architectural design, preprocessing pipelines, and achievable detection accuracy. Our synthesis reveals that the choice of data modality is not merely a practical consideration but directly dictates the feature space available to the model, the types of damage that can be detected, and the environmental conditions under which detection remains viable. To systematically compare these influences, we present a comprehensive taxonomy in Table 6, which categorizes the principal data modalities, the key characteristics of each modality that affect crack detection, the typical deep learning architectures used, and the specific studies that exemplify each approach.

Table 6. Characterization of data modalities and their influence on crack detection strategies.

Data Modality	Key Characteristics	Suitable Deep Learning Architectures	Exemplary Sources
2D RGB Images	High spatial resolution, rich texture; sensitive to lighting, scale, and occlusion; only surface information	CNNs, U-Net, Vision Transformers, YOLO	[9], [15], [22], [20], [26]
3D Point Clouds	Volumetric geometric data; captures depth, spalling, out-of-plane displacement; sparse and irregular; high computational cost	PointNet++, graph neural networks, 3D CNNs	[40], [34]
Acoustic / Ultrasonic	Time-series signal;	1D-CNNs, RNNs	[44]

Data Modality	Key Characteristics	Suitable Deep Learning Architectures	Exemplary Sources
Waveforms	capable of detecting internal/sub-surface cracks; sensitive to sensor placement and environmental noise	(LSTM), autoencoders	
Infrared Thermography (IRT)	Thermal contrast images; reveals hidden cracks under coatings/moisture; sensitive to ambient temperature and emissivity	Multimodal CNNs (with RGB fusion)	[43]
Frequency Domain (FFT/Wavelet)	Spectral representation of spatial data; enhances discriminant features for damage type classification	Hybrid CNNs with spectral feature branches	[42]
Multi-view / Photogrammetry	Reconstructed 3D surface from overlapping images; provides dense metric information; requires feature matching and bundle adjustment	CNNs with geometric matching layers	[27], [35]

Two-dimensional RGB images represent by far the most common data modality in the reviewed literature, a fact attributable to the ubiquity and low cost of digital cameras. The influence of this modality on crack detection strategy is profound: it prioritizes architectures that excel at extracting spatial features from regularly sampled pixel grids, such as CNNs and Vision Transformers. Studies using 2D images must contend with challenges of variable lighting, shadows, and scale ambiguity, which has driven the development of robust preprocessing and data augmentation pipelines. For example, the need to detect hairline cracks that are only a few pixels wide has motivated the use of high-resolution input tensors and multi-scale feature extraction in models like FPHBN [19] and Dense ResU-Net [14]. Furthermore, the planar nature of 2D images means that they can only capture surface damage, making them suitable for routine visual inspection but fundamentally limited for assessing internal structural degradation.

Three-dimensional point clouds offer a fundamentally different data representation, providing explicit geometric information about the surface shape that is critical for assessing damage like spalling, delamination, and permanent deformation. The irregular and sparse nature of point clouds, however, precludes the direct application of standard 2D convolutional neural networks. This has motivated the development of specialized architectures, such as PointNet-based variants and graph neural networks, that can process unordered point sets. The CrackEmbed framework, for example, embeds point features before applying anomaly detection, effectively converting the geometric data into a learned feature space that is amenable to classification [40]. The use of point clouds also influences the acquisition strategy, often requiring specialized and relatively expensive

sensors like terrestrial laser scanners or depth cameras, which limits their deployment in rapid, large-area surveys compared to standard photography. However, for high-value post-event assessments of critical infrastructure, the additional geometric information provided by point clouds is often indispensable.

Acoustic and ultrasonic waveform data introduce a completely different set of constraints and opportunities. Unlike images or point clouds, which capture static geometry, acoustic signals are time-varying and encode information about wave propagation through the material. This modality is particularly sensitive to internal cracks and voids that are invisible to any optical sensor, enabling detection of damage before it manifests on the surface. The sequential nature of acoustic data has driven the adoption of one-dimensional convolutional neural networks (1D-CNNs) and recurrent architectures like LSTMs, which are designed to model temporal dependencies [44]. The detection strategy for acoustic data also differs fundamentally from vision-based methods, as it often involves active sensing (e.g., generating a pulse and recording the response) rather than passive observation, and it requires careful sensor placement and coupling to the structure. The influence of environmental noise on acoustic signals is a major challenge, requiring sophisticated signal denoising and robust model training strategies.

Infrared thermography (IRT) data represent a middle ground between the visual and the physical. IRT images encode surface temperature, which can reveal subsurface defects such as delaminations or moisture-filled cracks that have different thermal properties than the surrounding material. The primary influence of this modality on detection strategy is the need for specialized data acquisition protocols, such as applying heat to the surface and recording the cooling transient. Deep learning models processing IRT data must learn to interpret thermal contrast patterns, which are often subtle and can be confounded by variations in surface emissivity. The most successful strategies, such as the hybrid RGB-IRT fusion approach [43], combine thermal data with standard visual imagery to leverage the complementary strengths of both modalities.

Data in the frequency domain, while often derived from spatial or temporal data via transforms like the Fast Fourier Transform (FFT) or wavelet transforms, can be considered a distinct modality because it exposes features that are not readily apparent in the original representation. A deep learning framework that fused spatial features with frequency information for earthquake damage classification demonstrated that the spectral representation of a crack pattern can be more discriminative than its spatial appearance for distinguishing between failure modes [42]. This strategy influences the model architecture by requiring an additional processing branch for the spectral data, effectively creating a multi-branch network that learns joint representations. The use of frequency-domain data also introduces a preprocessing step that can be computationally expensive, but the payoff in improved classification accuracy can be significant for challenging damage discrimination tasks.

Multi-view photogrammetry bridges the gap between 2D images and 3D point clouds by reconstructing a dense 3D surface from overlapping photographs. This modality inherits the low cost and accessibility of standard cameras while providing metric-scale geometric information. Studies that employ photogrammetry often combine the process with structured light projection to resolve scale ambiguity [27] or use photometric stereo techniques to measure surface profiles from multiple lighting directions [35]. The detection strategy for this modality involves a two-stage pipeline of 3D reconstruction followed by deep learning analysis of either the reconstructed geometry or the original multi-view images. The computational cost of the reconstruction step can be substantial, but the resulting 3D models provide a rich basis for quantitative analysis that is not possible with single-view 2D images.

D. What strategies address data scarcity, class imbalance, and domain adaptation in structural health monitoring?

The effective deployment of deep learning models for structural health monitoring is severely constrained by three interrelated data-centric challenges: data scarcity, class imbalance, and domain shift. Cracks, particularly those indicative of incipient failure or those occurring on massive infrastructure, are rare events in the statistical sense. When they do occur, they often occupy a very small fraction of the image pixels, leading to extreme class imbalance where the background class overwhelmingly dominates. Furthermore, a model trained on images of cracks in laboratory-controlled concrete specimens may fail catastrophically when applied to photographs of a weathered bridge surface taken under different lighting conditions, a phenomenon known as domain shift. The included studies reveal a growing awareness of these challenges and present a range of strategies to mitigate them, though the field is still far from a consolidated solution. To systematically organize these strategies, we present a taxonomy in Table 7, which categorizes the principal approaches by the data challenge they address, the specific technique employed, and the studies that exemplify each method.

Table 7. Taxonomy of strategies for addressing data scarcity, class imbalance, and domain adaptation.

Data Challenge	Strategy Category	Specific Technique	Exemplary Sources
Data Scarcity	Data Augmentation	Geometric transforms (rotation, scaling, flipping), color jitter, CutMix	[14], [16], [10], [25], [37]
		Generative Adversarial Networks (GANs) for synthetic crack generation	[42]
	Transfer Learning	Pre-training on ImageNet or crack-specific datasets, then fine-tuning	[9], [10], [21], [24], [25]
	Synthetic Data Generation	Physics-based or procedural generation of crack images (e.g., finite element modeling)	[33], [34]
Class Imbalance	Loss Function Modification	Focal Loss, weighted cross-entropy, Dice Loss	[14], [21], [23]
		Tversky Loss or other boundary-sensitive losses	[15]
	Resampling Strategies	Oversampling of minority class (crack patches), undersampling of background	[16], [36]
Domain Adaptation	Unsupervised Domain Adaptation (UDA)	Adversarial domain alignment (e.g., Gradient Reversal Layer)	[25]
		Self-training with pseudo-labels on target domain	[33]

Data Challenge	Strategy Category	Specific Technique	Exemplary Sources
	Data Normalization & Standardization	Histogram equalization, color normalization, illumination correction	[12], [24]
	Multi-source & Compound Datasets	Training on diverse datasets (e.g., concrete + asphalt + masonry) for robustness	[15], [26]

Data augmentation is the most ubiquitous and straightforward strategy employed to address data scarcity. The vast majority of studies that train deep learning models for crack detection apply some form of geometric and photometric augmentation to artificially expand the training set. We find that standard transforms such as random rotation, flipping, scaling, and translation are nearly universal. More advanced augmentation techniques, such as color jitter, Gaussian noise addition, and Random Erasing, are used to simulate variations in lighting and surface texture. A study developing a lightweight CNN for concrete crack detection used extensive data augmentation including rotation, exposure adjustment, and cropping to improve generalization from a small annotated dataset of approximately 500 images [16]. The Dense ResU-Net model for concrete crack segmentation employed a CutMix augmentation strategy where image patches were mixed to force the model to learn from combined contexts [14]. Several studies explicitly compared model performance with and without augmentation, consistently reporting that augmentation improved F1 scores by 5-15 percentage points, highlighting its critical role.

A more sophisticated approach to data scarcity is the generation of synthetic crack data using generative adversarial networks (GANs) or physics-based simulations. We found one study that used a GAN to generate realistic synthetic crack images for post-earthquake damage recognition, which were then used to augment the training set for a hybrid spatial-frequency CNN [42]. The GAN was trained on a limited dataset of real earthquake-damaged surfaces and learned to produce novel damage patterns that were visually realistic and structurally plausible. Physics-based generation, where crack patterns are derived from finite element analysis of structural failures, was employed in studies focusing on multidimensional damage scoring [33] and for training volumetric damage models [34]. These synthetic data generation methods are particularly valuable for generating rare but critical failure modes that are under-represented in field data, such as severe spalling or complex crack networks.

Transfer learning is another foundational strategy for overcoming data scarcity. Rather than training a network from scratch, which requires millions of labeled examples, transfer learning leverages a model pre-trained on a large, generic dataset like ImageNet and fine-tunes it on the target crack detection task. This practice is so widespread that it is used in the majority of the included CNN-based studies. A study on building damage detection fine-tuned a pre-trained VGG-16 network to achieve high accuracy with a modestly sized dataset of approximately 2,000 images [9]. Similarly, a SqueezeNet model pre-trained on ImageNet was fine-tuned using the Adam optimizer for concrete crack classification, converging rapidly with a small dataset [10]. A hybrid framework combining pre-trained deep CNNs with differential evolution for pavement crack segmentation also relied on fine-tuning a ResNet backbone [24]. The consistent success of transfer learning across these studies suggests that features learned from natural images are sufficiently generalizable to serve as a strong starting point for crack detection, though fine-tuning on domain-specific data remains essential.

Class imbalance, where crack pixels are vastly outnumbered by background pixels, is addressed primarily through modifications to the training loss function. Standard cross-entropy loss is

suboptimal for this scenario because the majority class (background) dominates the gradient, causing the model to converge to a trivial solution that predicts all pixels as background. Focal Loss, which down-weights the loss assigned to well-classified examples (primarily background pixels) and focuses training on hard, misclassified examples (crack pixels), is employed in several studies [14], [21], [23]. The Dense ResU-Net used Focal Loss combined with class weights to further emphasize crack pixel contributions during training [14]. Dice Loss, which directly optimizes the Dice similarity coefficient (a metric of overlap between predicted and ground truth crack regions), was used in a lightweight pavement damage detection framework [23]. A more advanced variant, Tversky Loss, which allows independent weighting of false positives and false negatives, was employed in a two-step detection and segmentation method to prioritize recall (high false negative penalty) for safety-critical applications [15]. The study found that Tversky Loss provided more robust segmentation on complex backgrounds than Dice Loss alone.

Domain adaptation is the least mature of the three data-centric challenges, with relatively few studies directly addressing it. The core problem is that a model trained on one distribution of crack images (the source domain, e.g., laboratory concrete) often degrades significantly when applied to a different distribution (the target domain, e.g., field-weathered bridge surfaces). We identified one study that explicitly tackled unsupervised domain adaptation (UDA) for crack segmentation, using an adversarial alignment approach [25]. A domain discriminator network was trained to distinguish between features extracted from source and target domain images, while the feature extractor was trained to fool the discriminator. This process forced the CNN to learn domain-invariant features that generalize across different crack datasets. The UDA-trained model significantly improved segmentation performance on the target domain compared to a standard model trained only on the source domain.

A simpler but effective strategy for mitigating domain shift is data normalization and standardization. Several studies applied histogram equalization or illumination correction as a preprocessing step to reduce variability in image appearance caused by different cameras, lighting conditions, or time of day [12], [24]. This reduces the input distribution mismatch between training and testing domains. Another practical approach is to train on a large, diverse, compound dataset that includes images from multiple sources and surface types. A comparative study of deep learning models used a merged dataset comprising concrete, asphalt, and masonry crack images [26], and the model trained on this compound set showed better generalization to an unseen test set than models trained on any single source domain. The recent introduction of large-scale open-source crack datasets, such as StructDamage ([46]), represents an important step toward enabling robust multi-source training, though such datasets are still rare.

VI. Discussion

This systematic review synthesizes findings from 90 peer-reviewed studies to provide a comprehensive understanding of deep learning techniques for crack detection and structural damage assessment. The preceding results section has detailed specific architectural innovations, quantification methods, applications to extreme events, hybrid approaches, data modality considerations, and strategies for data scarcity and domain adaptation. In this discussion, we move beyond a simple recapitulation of individual study findings to integrate these diverse threads into a cohesive narrative that identifies overarching patterns, inconsistencies, and implications for both research and practice.

A dominant pattern that emerges across the reviewed literature is the consistent superiority of encoder-decoder architectures, particularly U-Net variants with residual or dense connections enhanced by attention mechanisms, for pixel-level crack segmentation tasks. Studies employing Dense ResU-Net [14], ReCRNet [13], and FPHBN [19] all report state-of-the-art performance on benchmark datasets, suggesting that the combination of multi-scale feature extraction with refined upsampling pathways is

fundamentally well-suited to the task of delineating thin, elongated structures against heterogeneous backgrounds. This finding aligns with broader trends in medical image segmentation, where U-Net architectures have become the de facto standard for similar tasks involving vascular or neural structure delineation [47]. The recurring success of attention mechanisms across multiple studies, including adaptive attention in vision transformers [18] and channel attention in YOLO variants [21], indicates that the ability to selectively focus on informative crack regions while suppressing background noise is a critical driver of performance. Taken together, these findings suggest that future architectural innovations should prioritize improved feature fusion and context aggregation rather than entirely novel backbone designs, as the fundamental capabilities required for crack segmentation are largely met by current architectures, albeit with room for efficiency improvements.

A notable inconsistency in the literature pertains to the reported performance of Vision Transformers relative to CNNs. While one study reported that Transformer-based architectures outperformed U-Net counterparts on pavement crack datasets with high background clutter [22], another found that U-Net remained highly competitive and more computationally efficient [26]. This contradiction may be explained by differences in dataset characteristics, training protocols, or the specific architectural implementations employed. The superior performance of Transformers on complex scenes likely stems from their ability to model long-range dependencies, which is advantageous when cracks are embedded within intricate texture patterns or when the structural context of the damage is important for accurate segmentation. However, the computational overhead of self-attention mechanisms, which scales quadratically with image resolution, poses a practical barrier for real-time deployment on embedded systems. This trade-off between accuracy and efficiency has profound implications for practitioners: for high-stakes applications requiring maximum accuracy, such as post-earthquake building assessments, the additional computational cost of a Transformer may be justified; for routine pavement inspection where throughput is paramount, a well-tuned lightweight CNN may be the more pragmatic choice. The field would benefit from standardized benchmark evaluations that explicitly report both accuracy metrics and computational resource requirements, enabling more informed architectural selection.

A critical finding that emerges across studies on damage quantification is that pixel-level segmentation alone is insufficient for structural assessment; the conversion of segmentation masks into physically interpretable metrics requires careful calibration and often additional sensor data. The most accurate dimensional measurements were achieved by methods that incorporated depth information, whether through structured light projection [27], depth cameras [34], or smartphone inertial sensors [36]. This observation highlights a fundamental limitation of monocular vision: without knowledge of the scale, absolute crack widths and lengths cannot be determined from a single image. The practical implication is that researchers and practitioners developing quantitative assessment systems must either integrate depth sensing hardware or employ calibration targets in the scene. The smartphone-based approach, which leverages ubiquitous hardware and built-in sensors, represents a particularly promising direction for democratizing quantitative crack measurement, as it removes the need for specialized equipment that is often cost-prohibitive for routine inspections. However, the accuracy of such methods is inherently limited by the precision of consumer-grade IMUs and the assumptions made about the camera-to-surface geometry, and further validation on a wider range of structural surfaces and lighting conditions is warranted.

Regarding the application of deep learning to extreme events, our synthesis reveals a significant gap between the maturity of techniques for routine fatigue crack detection and those for earthquake or fire damage assessment. While several studies have demonstrated effective damage classification and quantification for post-

seismic scenarios using UAV imagery [11] and point cloud analysis [40], the number of studies addressing fire damage is strikingly low, with only a single study meeting our inclusion criteria [39]. This scarcity is concerning, given that fire events can compromise structural integrity in ways that are fundamentally different from mechanical loading, and the damage patterns (e.g., thermal cracking, spalling, discoloration) present unique visual and thermal signatures. The extreme imbalance in research attention may stem from the difficulty of acquiring labeled fire damage data, as controlled fire experiments on structural elements are expensive, hazardous, and require specialized facilities. Nevertheless, the demonstrated success of deep learning in correlating surface crack patterns with internal temperature and residual strength [39] suggests that this is a fertile area for future investigation, particularly if combined with physics-based simulation to generate synthetic training data. The field would benefit from collaborative efforts to establish shared datasets of fire-damaged structures, analogous to the publicly available earthquake damage datasets that have spurred progress in seismic assessment.

The analysis of hybrid approaches reveals a clear trend toward multi-sensor fusion as a strategy to overcome the limitations of any single modality. Thermal-visible fusion [43] addresses the problem of cracks hidden beneath surface coatings or moisture; acoustic sensing [44] enables detection of internal damage invisible to optical systems; and geometric fusion with point clouds [40] provides the volumetric data necessary for spalling quantification. The common thread across these hybrid approaches is that they exploit the physical complementarity of different sensing principles to create a more complete picture of structural health. This has direct implications for practical deployment: rather than pursuing a single universal sensor, inspection systems should be designed as modular platforms that can integrate multiple sensors as dictated by the specific inspection objectives and environmental conditions. For example, a bridge inspection drone might carry an RGB camera for general crack detection, an infrared camera for detecting subsurface delaminations, and a LiDAR system for measuring structural deformations. The deep learning model would then need to be trained to fuse these heterogeneous data streams, which presents a significant technical challenge in terms of spatiotemporal alignment and architectural design. The success of such integrated systems will depend on the availability of training data spanning multiple modalities simultaneously, which is currently extremely scarce.

The influence of data modality on detection strategy is profound and directly determines the architectural choices available to researchers. The prevalence of 2D RGB images has naturally led to the dominance of CNN and Transformer architectures designed for regular grid data. However, the increasing availability of 3D point cloud data from LiDAR and depth cameras is driving the adoption of specialized geometric deep learning methods, such as PointNet variants and graph neural networks [40]. This shift represents a departure from the mature and well-understood toolkit of 2D vision and introduces new challenges in terms of sparsity, irregularity, and computational cost. Similarly, the use of acoustic waveforms pushes the field toward time-series analysis techniques more common in signal processing than in computer vision [44]. One might argue that the field is currently in a transitional phase, where the dominant paradigm of 2D image-based analysis is being gradually augmented, but not yet replaced, by multi-modal approaches. The long-term trajectory is likely toward systems that can flexibly handle multiple modalities, perhaps through the development of more general-purpose architectures that can process both spatial and sequential data, or through the use of modality-agnostic representations such as learned embeddings.

A critical finding with regard to data scarcity and class imbalance is that augmentation and transfer learning, while effective, are not panaceas. Most studies apply standard geometric and photometric augmentation, but this cannot generate fundamentally new crack morphologies or failure modes that are absent from the original dataset. The use of GANs for synthetic crack generation [42] and physics-based simulation [33],

[34] represents a more principled approach to expanding the training distribution, yet these methods are computationally expensive and require careful validation to ensure that the generated data is physically realistic. The reliance on transfer learning from ImageNet, while empirically successful, raises a theoretical concern: features learned from natural images of objects, animals, and scenes may not be optimal for detecting fine-scale fractures in homogeneous materials like concrete or asphalt. The consistent performance gains from fine-tuning do not necessarily imply that ImageNet features are the best possible starting point; rather, they may simply reflect the benefits of any reasonable pre-training compared to random initialization. The field would benefit from investigating whether domain-specific pre-training on large collections of infrastructure images, such as the growing repositories of structural inspection photographs, could further improve performance and reduce the amount of labeled data required for fine-tuning.

The issue of domain adaptation remains conspicuously under-addressed, with only a single study in our corpus explicitly tackling unsupervised domain adaptation [25]. This is a significant limitation, as the gap between laboratory-controlled datasets and real-world field conditions is perhaps the most critical barrier to the practical deployment of deep learning-based inspection systems. A model trained on high-quality images of clean concrete surfaces will likely fail when presented with images of weathered, stained, or moss-covered surfaces typical of field infrastructure. The adversarial domain alignment approach demonstrated promise, but its effectiveness was evaluated on a relatively small-scale dataset, and its generalizability to more challenging domain shifts, such as those between concrete and asphalt or between different climates, remains unknown. The lack of attention to domain adaptation may be attributed to the difficulty of obtaining labeled data from multiple domains, as well as the relative novelty of this research direction in the structural health monitoring community. There is a clear and pressing need for future research to develop robust domain adaptation techniques that can be deployed in a zero-shot or few-shot manner, enabling a model trained on one structure to be immediately applied to another without costly re-annotation.

Several methodological limitations of this review itself must be acknowledged, as they may influence the interpretation of our findings. The search strategy, while comprehensive across five databases, may have inadvertently excluded relevant studies that use terminology outside our keyword set, such as those focusing on "fracture detection" in mechanical components that could have analogous methods applicable to civil infrastructure. The restriction to English-language publications introduces a language bias that may underrepresent contributions from non-English-speaking countries that are actively conducting structural health monitoring research, particularly in East Asia and Europe. Furthermore, our inclusion criteria prioritized peer-reviewed journal articles and conference proceedings, which may have excluded valuable grey literature such as technical reports from government agencies, industry white papers, or doctoral dissertations that could provide unique insights into practical deployment challenges. The reliance on published studies also introduces a publication bias, as studies reporting positive results (e.g., high accuracy) are more likely to be published than those reporting negative or null findings, potentially leading to an overestimation of the effectiveness of deep learning techniques. Finally, the quality assessment of included studies was conducted on a single-reviewer basis, which, while mitigated by cross-checking for ambiguous cases, introduces subjective bias that could affect the weight given to individual findings.

Based on the gaps and inconsistencies identified in this review, several high-priority directions for future research emerge. Firstly, there is a pressing need for large-scale, multi-modal, and multi-domain benchmark datasets that capture the diversity of real-world inspection conditions, including varying surface materials, lighting conditions, damage types, and sensor modalities. Current datasets, while valuable, are often limited to a single modality (typically RGB

images) and a single structural context (e.g., concrete bridge surfaces). The creation and maintenance of such datasets would require collaborative efforts across institutions and potentially industry partnerships, but they are essential for enabling rigorous comparative evaluations and for training robust, generalizable models. Secondly, future research should explore unsupervised and self-supervised learning paradigms to reduce the reliance on large labeled datasets. The success of contrastive learning in computer vision suggests that pre-training crack detection models on large unlabeled collections of structural images, using pretext tasks such as image inpainting or contrastive prediction, could yield strong feature representations that require minimal labeled data for fine-tuning. Thirdly, there is a clear need for more research on real-time deployment and edge computing, as the majority of included studies evaluate models on desktop GPUs rather than on the embedded systems typical of inspection drones or robots. Lightweight architectures, model compression techniques (e.g., pruning, quantization, knowledge distillation), and hardware-software co-design should be prioritized to enable practical deployment at scale. Fourthly, the integration of deep learning with physics-based structural models warrants further investigation, as demonstrated by the correlation of crack patterns with thermal-structural behavior in fire-damaged beams [39]. Such hybrid data-driven and physics-informed approaches could provide more robust and interpretable damage assessments, particularly for extreme events where training data is scarce. Finally, understudied areas include the detection of incipient cracks that are only a few micrometers wide, the assessment of damage in non-conventional materials such as fiber-reinforced polymers, and the longitudinal monitoring of crack propagation over time using sequential inspection data. These challenges, while difficult, represent the frontier where deep learning could have the most transformative impact on structural safety and maintenance.

VII. Conclusion

This systematic review synthesized 90 peer-reviewed studies to provide a comprehensive evaluation of deep learning techniques for crack detection and structural damage assessment, addressing seven research questions spanning architectural innovations, quantitative assessment, extreme event applications, hybrid sensing, data modalities, and practical deployment challenges. Our analysis confirms that encoder-decoder convolutional neural networks with attention mechanisms, particularly U-Net variants enhanced by residual connections or hierarchical feature fusion, represent the current state of the art for pixel-level crack segmentation. We further demonstrate that the conversion of segmentation outputs into actionable structural assessments requires integration with depth sensing or geometric calibration, as monocular vision alone cannot provide the absolute scale necessary for dimensional quantification. A critical contribution of this synthesis is the identification of significant disparities in research maturity across application domains, with fire damage assessment and domain adaptation remaining conspicuously underdeveloped relative to routine fatigue crack detection.

The practical implications of this work are threefold. For practitioners developing automated inspection systems, our findings provide an evidence-based framework for architectural selection, recommending lightweight YOLO-based detectors for real-time throughput and encoder-decoder segmentation networks for high-accuracy damage delineation, with the choice governed by the specific trade-off between computational constraints and measurement precision. For researchers, this review establishes a clear taxonomy of existing methods while illuminating pressing gaps, particularly the scarcity of multi-modal benchmark datasets and the near absence of robust domain adaptation techniques validated on real field conditions. The most profound implication is that the field must transition from a paradigm of laboratory-proven accuracy to one of field-proven reliability, which demands

collaborative efforts to build shared datasets spanning diverse structural contexts, sensor modalities, and environmental conditions. We advocate for future research to prioritize four directions: unsupervised and self-supervised learning to reduce labeling costs, lightweight architectures for embedded deployment, physics-informed deep learning that bridges data-driven predictions with structural engineering models, and systematic investigation of domain adaptation for zero-shot generalization across structures. The ultimate promise of deep learning for structural health monitoring—the continuous, autonomous, and reliable safeguarding of our aging infrastructure—will be realized only when these methodological gaps are addressed with the same rigor that has driven architectural innovation in recent years.

VIII. References

[1] D. Frangopol and Y. Tsompanakis, "Maintenance and safety of aging infrastructure," api.taylorfrancis.com, 2014.

[2] D. Agdas, J. Rice, J. Martinez, *et al.*, "Comparison of visual inspection and structural-health monitoring as bridge condition assessment methods," *Journal of Performance of Constructed Facilities*, 2016.

[3] S. Hijazi, "Convolutional neural networks for image recognition," Вебсайт. URL: https://ip. cadence. com/uploads/901 ..., 2026.

[4] X. Ye, T. Jin, and C. Yun, "A review on deep learning-based structural health monitoring of civil infrastructures," *Smart Struct. Syst*, 2019.

[5] X. Weng, Y. Huang, Y. Li, H. Yang, and S. Yu, "Unsupervised domain adaptation for crack detection," *Automation in Construction*, 2023.

[6] S. Farhadi, M. Iavarone, M. Corrado, E. Chatzi, *et al.*, "Addressing data scarcity in structural health monitoring through generative augmentation," arXiv preprint arXiv:2510.16889, 2025.

[7] B. Kulambayev, G. Astaubayeva, *et al.*, "Deep CNN approach with visual features for real-time pavement crack detection." *Journal of Advanced Transportation*, 2024.

[8] M. Page, J. McKenzie, P. Bossuyt, *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021.

[9] C. Feng, H. Zhang, S. Wang, Y. Li, H. Wang, *et al.*, "Structural damage detection using deep convolutional neural network and transfer learning," *KSCE Journal of Civil Engineering*, 2019.

[10] L. Wang, "Automatic detection of concrete cracks from images using adam-SqueezeNet deep learning model," *Fracture and Structural Integrity*, 2023.

[11] O. Turan, H. Kaya, G. Taskin, T. Cinar, and A. Ilki, "A structural damage ranking using ConvNeXt for post-earthquake image classification: O. Turan et al." *Arabian Journal for Science and Engineering*, 2025.

[12] V. Hoskere, Y. Narazaki, T. Hoang, *et al.*, "Vision-based structural inspection using multiscale deep convolutional neural networks," arXiv preprint arXiv:1805.01055, 2018.

[13] H. Reis and K. Khoshelham, "ReCRNet: A deep residual network for crack detection in historical buildings," *Arabian Journal of Geosciences*, 2021.

[14] A. Sarhadi, M. Ravanshadnia, *et al.*, "An innovative dense ResU-net architecture with t-max-avg pooling for advanced crack detection in concrete structures," *IEEE Open Journal of the Industrial Electronics Society*, 2024.

[15] Y. Jiang, D. Pang, C. Li, Y. Yu, and Y. Cao, "Two-step deep learning approach for pavement crack damage detection and segmentation," *International Journal of Pavement Engineering*, 2023.

[16] Q. Meng *et al.*, "Image-based concrete cracks identification under complex background with lightweight convolutional neural network," *KSCE Journal of Civil Engineering*, 2023.

[17] S. Dong, Y. Wang, J. Cao, J. Ma, Y. Chen, and X. Kang, "Advanced lightweight deep learning vision framework for efficient pavement damage identification," *Scientific Reports*, 2025.

[18] M. Mowla, D. Asadi, and F. Sohel, "Adaptive attention optimized deep learning with vision transformers for fine grained earthquake structural damage detection," *Earthquake Spectra*, 2026.

[19] F. Yang, L. Zhang, S. Yu, D. Prokhorov, *et al.*, "Feature pyramid and hierarchical boosting network for pavement crack detection," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[20] K. Zeng, R. Fan, and X. Tang, "Efficient and accurate road crack detection technology based on YOLOv8-ES," *Autonomous Intelligent Systems*, 2025.

[21] W. Ren and Z. Zhong, "Building construction crack detection with BCCD YOLO enhanced feature fusion and attention mechanisms," *Scientific Reports*, 2025.

[22] Y. Zhang and L. Zhang, "Detection of pavement cracks by deep learning models of transformer and UNet," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[23] H. Li, H. Zhang, H. Zhu, K. Gao, H. Liang, and J. Yang, "Automatic crack detection on concrete and asphalt surfaces using semantic segmentation network with hierarchical transformer," *Engineering Structures*, 2024.

[24] E. M. Abdelkader, "On the hybridization of pre-trained deep learning and differential evolution algorithms for semantic crack detection and recognition in ensemble of infrastructures," *Smart and sustainable built environment*, 2022.

[25] S. Jana, S. Thangam, A. Kishore, *et al.*, "Transfer learning based deep convolutional neural network model for pavement crack detection from images," *International Journal of Nonlinear Analysis and Applications*, 2022.

[26] S. Krishnan, M. Karuppan, A. Khadidos, *et al.*, "Comparative analysis of deep learning models for crack detection in buildings," *Scientific Reports*, 2025.

[27] S. Park, S. Eem, and H. Jeon, "Concrete crack detection and quantification using deep learning and structured light," *Construction and Building Materials*, 2020.

[28] M. Hassouna, M. Marzouk, and E. Fathalla, "Automated low-cost framework for crack measurements in RC structures using deep learning approach," *Scientific Reports*, 2026.

[29] X. Lu, Q. Li, J. Li, and L. Zhang, "Deep learning-based method for detection and feature quantification of microscopic cracks on the surface of concrete dams," *Measurement*, 2025.

[30] X. Jing, Y. Huan, Y. Wang, R. Huang, Y. Xu, *et al.*, "Complex crack segmentation and quantitative evaluation of engineering materials based on deep learning methods," *IEEE Access*, 2024.

[31] D. Kang, S. Benipal, D. Gopal, and Y. Cha, "Hybrid pixel-level concrete crack segmentation and quantification across complex backgrounds using deep learning," *Automation in Construction*, 2020.

[32] P. Savino and F. Tondolo, "Civil infrastructure defect assessment using pixel-wise segmentation based on deep learning," *Journal of Civil Structural Health Monitoring*, 2023.

[33] L. Zhang and R. Bian, "Deep learning-based crack detection method for civil engineering and its multidimensional damage assessment," *Journal of Combinatorial Mathematics and Combinatorial Computing*, 2024.

[34] G. Beckman, D. Polyzois, and Y. Cha, "Deep learning-based automatic volumetric damage quantification using depth camera," *Automation in Construction*, 2019.

[35] L. Zhen-Liang, Z. An, R. Xin-Ru, W. Yun-Peng, *et al.*, "A crack detection and quantification method using matched filter and photograph reconstruction," *Scientific Reports*, 2025.

[36] C. Tello-Gil, S. Jabari, L. Waugh, *et al.*, "Crack detection and dimensional assessment using smartphone sensors and deep learning," *Canadian Journal of Civil Engineering*, 2024.

[37] A. BaniMustafa, R. AbdelHalim, O. Bulkrock, *et al.*, "Deep learning for assessing severity of cracks in concrete structures." *International Journal of Civil Engineering and Technology*, 2023.

[38] T. Nguyen, P. Phan-Vu, and P. Nguyen, "AI-based damage detection in prestressed concrete beams: A vision-integrated deep learning framework for crack localization and severity classification," *Advances in Bridge Engineering*, 2026.

[39] E. Ryu, J. Kang, J. Lee, Y. Shin, and H. Kim, "Automated detection of surface cracks and numerical correlation with thermal-structural behaviors of fire damaged concrete beams," *International Journal of Concrete Structures and Materials*, 2020.

[40] J. Chen and Y. Cho, "CrackEmbed: Point feature embedding for crack segmentation from disaster site point clouds with anomaly detection," *Advanced engineering informatics*, 2022.

[41] R. Pourhanasa and A. Monadipour, "Concrete crack detection via graph representation learning and texture analysis," *Innovative Infrastructure Solutions*, 2025.

[42] Z. Bai, T. Liu, D. Zou, M. Zhang, A. Zhou, and Y. Li, "Image-based reinforced concrete component mechanical damage recognition and structural safety rapid assessment using deep learning with frequency information," *Automation in construction*, 2023.

[43] K. Jang, N. Kim, and Y. An, "Deep learning–based autonomous concrete crack evaluation through hybrid image scanning," *Structural Health Monitoring*, 2019.

[44] M. Barbosh, L. Ge, and A. Sadhu, "Automated crack identification in structures using acoustic waveforms and deep learning," *Journal of Infrastructure Preservation and Resilience*, 2024.

[45] K. Ma, X. Meng, M. Hao, G. Huang, Q. Hu, and P. He, "Research on the efficiency of bridge crack detection by coupling deep learning frameworks with convolutional neural networks," *Sensors*, 2023.

[46] M. Ijaz, S. Khan, A. Rehman, S. Vollmer, *et al.*, "StructDamage: A large scale unified crack and surface defect dataset for robust structural damage detection," arXiv preprint arXiv:2603.10484, 2026.

[47] G. Du, X. Cao, J. Liang, X. Chen, *et al.*, "Medical image segmentation based on u-net: A review." *Journal of Imaging Informatics in Medicine*, 2020.