

AI-driven Defect Detection and Multi-Agent Disaster Response for Civil Infrastructure

By

Zhaohui Chen

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

In the

Faculty of Engineering

Supervisor: **A/Prof. Daniel Dias-da-Costa**

Co-Supervisor: **Prof. Luming Shen**

The University of Sydney, NSW 2006

Australia

2026

Originality statement

I hereby certify that this submission is my own work and, to the best of my knowledge, has not been submitted for any other degree or diploma. I declare that the intellectual content of this thesis is the product of my own work. I acknowledge that some parts of this thesis have been published as journal articles, as detailed in the section 'List of Publications'. All assistance received in the preparation of this thesis and all sources used have been duly acknowledged.

Zhaohui Chen

Generative AI statement

During the preparation of this thesis, the author used Microsoft Copilot for the purposes of text refinement, including paraphrasing, sentence structure improvement, and clarity enhancement. The author confirms that where text was modified by generative AI, the content was carefully reviewed for possible errors, inaccuracies, and bias. The author takes full responsibility for the submitted thesis and confirms that all research ideas, methodologies, experiments, and conclusions are their own. The use of generative AI was limited to editorial support and was conducted in accordance with the University of Sydney generative AI guidelines.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, A/Prof. Daniel Dias-da-Costa and Prof. Luming Shen, for their guidance, support, and encouragement throughout the course of this doctoral research. Their expertise, critical insights, and commitment to academic rigour have been helpful in shaping both the direction and quality of this thesis. I am particularly grateful for the freedom to explore interdisciplinary research at the intersection of civil engineering and artificial intelligence, which enabled the development of the ideas presented in this work.

I would also like to thank my collaborators, Elyas Asadi Shamsabadi and Sheng Jiang, for their valuable discussions, technical contributions, and constructive feedback across the studies included in this thesis. Their perspectives and expertise greatly enriched the research and helped refine both the methodology and presentation of the results.

My appreciation extends to the academic and technical staff of the University of Sydney for providing a supportive research environment and essential infrastructure. I am grateful for access to computational resources and experimental facilities that made this research possible.

I would like to acknowledge the financial support provided by the joint scholarships from the University of Sydney and China Scholarship Council, which enabled me to pursue this research full-time and to disseminate the outcomes through international publications.

Finally, I would like to thank my family and friends for their patience, understanding, and unwavering support throughout this journey. Their encouragement and belief in me provided constant motivation during challenging periods of the PhD. This thesis would not have been possible without their support.

List of publications

1. **Zhaohui Chen**, Elyas Asadi Shamsabadi, Sheng Jiang, Luming Shen, Daniel Dias-da-Costa. (2024). An average pooling designed Transformer for robust crack segmentation. *Automation in Construction*, 162, 105367.
2. **Zhaohui Chen**, Elyas Asadi Shamsabadi, Sheng Jiang, Luming Shen, Daniel Dias-da-Costa. (2024). Robust feature knowledge distillation for enhanced performance of lightweight crack segmentation models. *arXiv:2404.06258* (under review, preprint available).
3. **Zhaohui Chen**, Elyas Asadi Shamsabadi, Sheng Jiang, Luming Shen, Daniel Dias-da-Costa. (2024). Vision Mamba-based autonomous crack segmentation on concrete, asphalt, and masonry surfaces. *arXiv:2406.16518* (under review, preprint available).
4. **Zhaohui Chen**, Elyas Asadi Shamsabadi, Sheng Jiang, Luming Shen, Daniel Dias-da-Costa. (2026). Integration of large vision language models for efficient post-disaster damage assessment and reporting. *Nature Communications*.

Authorship statement

The author of this thesis is the primary contributor to the relevant publications and led model and framework designs, experimental developments, implementations, evaluations, and manuscript preparation. The statement of my authorship has been confirmed by all co-authors:

Daniel Dias-da-Costa

School of Civil Engineering, The University of Sydney, NSW 2006, Australia

Luming Shen

School of Civil Engineering, The University of Sydney, NSW 2006, Australia

Elyas Asadi Shamsabadi

School of Civil Engineering, The University of Sydney, NSW 2006, Australia

Sheng Jiang

College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing, 210024, China

Abstract

Rapid and reliable assessment of civil infrastructure condition is essential for effective disaster response and recovery. Traditional post-disaster inspection workflows rely heavily on manual, expert-driven field surveys, which are time-consuming, resource-intensive, and difficult to scale under conditions of limited accessibility, incomplete information, and evolving hazards. Although recent advances in computer vision have enabled automated damage detection from visual data, existing approaches often focus on isolated perception tasks and provide limited support for system-level interpretation and decision-making.

This thesis addresses these limitations by developing an integrated AI-driven framework that connects robust local-scale visual perception with structured global-scale reasoning for post-disaster infrastructure assessment. At the local scale, efficient and deployable computer vision models are developed for infrastructure defect detection under realistic inspection conditions. A Transformer-based crack segmentation architecture with average pooling-based feature aggregation is proposed to improve robustness and accuracy in noisy and heterogeneous visual environments. A Robust Feature Knowledge Distillation framework is further introduced to enhance noise resilience and deployment feasibility by transferring robust representations from complex teacher models to lightweight student models. Vision Mamba-based architectures are also investigated for high-resolution defect inspection, addressing challenges associated with large-scale image processing and long-range dependency modelling. At the global scale, an AI agent-based framework for disaster perception and assessment is proposed. The framework integrates local perceptual outputs with a multi-agent reasoning architecture that performs role-based task decomposition, information integration, and collaborative reasoning. This design enables the transformation of local observations into coherent global situational awareness while supporting interpretability and human-in-the-loop decision-making. Experimental evaluations demonstrate improved robustness, scalability, and practical relevance compared with conventional approaches.

Despite these contributions, challenges remain in handling extreme real-world uncertainty, ensuring long-horizon reasoning reliability, and supporting adaptive learning under evolving disaster conditions. Future research should therefore focus on uncertainty-aware perception, continual adaptation, and tighter integration between AI systems, sensing infrastructure, and decision-support environments, enabling more trustworthy and resilient intelligent infrastructure assessment systems.

Contents

Originality statement.....	i
Generative AI statement.....	ii
Acknowledgements.....	iii
List of publications	iv
Authorship statement	v
Abstract.....	vi
Contents	vii
Acronyms.....	xi
1. Introduction.....	1
1.1 Motivation and societal challenges.....	1
1.2 Traditional infrastructure inspection and disaster response	3
1.3 Computer vision for local-scale infrastructure inspection.....	7
1.4 AI-based global-scale disaster perception and assessment	9
1.5 Thesis outline.....	11
2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection.....	22
2.1 Introduction.....	22
2.2 Methodology	25
2.2.1 Convolutional patch embedding	26
2.2.2 Pooling blocks.....	27
2.2.3 Feature fusion misalignment.....	28
2.2.4 Feature selection modules.....	28
2.2.5 Feature alignment modules	29
2.3 Study design.....	30

2.3.1 Model selection.....	30
2.3.2 Evaluation metrics	30
2.3.3 Image datasets.....	31
2.3.4 Training configuration	33
2.3.5 Assessment framework	33
2.4 Results and discussion	34
2.4.1 Accuracy comparison.....	34
2.4.2 Efficiency comparison	36
2.4.3 Robustness study.....	37
2.5 Conclusion	43
Appendix 2.A.....	43
2.A.1 Ablation study results.....	43
2.A.2 Accuracy comparison with models in other publications	45
2.A.3 Application in colonoscopy segmentation	49
3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception.....	57
3.1 Introduction.....	57
3.2 Methodology	60
3.2.1 RFKD framework	60
3.2.2 PCT architecture	63
3.3 Study design.....	66
3.3.1 KD strategy and lightweight model selection.....	66
3.3.2 Evaluation metrics	66
3.3.3 Image datasets.....	67
3.3.4 Training configuration	68
3.3.5 Study framework.....	68
3.4 Results and discussion	69

3.4.1 Accuracy assessment of the RFKD framework	70
3.4.2 Robustness assessment of the RFKD framework	72
3.4.3 Accuracy and efficiency assessment of the PCT model.....	74
3.4.4 Robustness assessment of the PCT model	76
3.5 Conclusions.....	78
Appendix 3.A.....	80
3.A.1 Additional validation.....	80
3.A.2 Sensitivity analysis of hyperparameters.....	85
3.A.3 The architecture of the teacher model.....	88
3.A.4 Ablation study	89
4. Vision Mamba-based architectures for high-resolution local-scale defect inspection.....	98
4.1 Introduction.....	98
4.2 Study design.....	101
4.2.1 The architecture of Mamba	101
4.2.2 Vision Mamba based encoder-decoder network	104
4.2.3 Evaluation metrics	105
4.2.4 Datasets.....	106
4.2.5 Training configuration	107
4.3 Results and discussions.....	107
4.3.1 Accuracy comparison.....	108
4.3.2 Efficiency comparison	111
4.4 Conclusion	114
Appendix 4.A.....	114
4.A.1 Convolutional Neural Networks (CNNs).....	114
4.A.2 Vision Transformers.....	115
5. AI agent framework for global-scale defect perception.....	122
5.1 Introduction.....	122

5.2 Results and discussions.....	127
5.2.1 Autonomous disaster response.....	127
5.2.2 Agent output results evaluation.....	129
5.2.3 Limitations and deployment considerations	133
5.2.4 Summary	136
5.3 Methods.....	137
5.3.1 Large Vision Language Models (LVLMs) with augmented generation	137
5.3.2 Custom multi-LVLM agent framework design.....	138
5.3.3 Tools development.....	138
5.3.4 Agents definition for post-disaster management.....	140
Appendix 5.A.....	141
5.A.1 Intermediate process tasks	141
5.A.2 Output reports evaluation.....	144
5.A.3 Error propagation study	149
5.A.4 Allocation under scarce resources.....	153
5.A.5 Flood scenario preliminary study.....	155
5.A.6 Bushfire scenario preliminary study	157
6. Conclusions and Future Research.....	162
6.1 Concluding remarks.....	162
6.2 Limitations and future research	164

Acronyms

AI	Artificial Intelligence
API	Application Programming Interface
CNN	Convolutional Neural Network
FLOP	Floating Point Operation
GIS	Geographic Information System
KD	Knowledge Distillation
LLM	Large Language Model
LVLM	Large Vision-Language Model
mDS	Mean Dice Score
mIoU	Mean Intersection over Union
ML	Machine Learning
MLP	Multi-Layer Perceptron
RFKD	Robust Feature Knowledge Distillation
RNN	Recurrent Neural Network
S4	Structured State Space for Sequences
S6	Select-Scan Structured State Space for Sequences
SiLU	Sigmoid Linear Unit
SOTA	State of the Art
SS2D	Two-Dimensional Select-Scan
VSS	Visual State Space

1. Introduction

1.1 Motivation and societal challenges

Natural disasters such as earthquakes, floods, and hurricanes, have become increasingly frequent and destructive under the combined influence of climate change, rapid urban expansion, and the progressive aging of civil infrastructure systems (Akinsanola, et al., 2025). These disaster events bring substantial social and economic burdens on modern societies, not only through immediate casualties and property losses, but also through long-term disruptions to transportation networks, energy supply, water systems, healthcare accessibility, and overall community resilience (Rezvani, et al., 2023). Civil infrastructure, as the physical backbone supporting societal functions, is inherently vulnerable to such hazards, and its performance during and after extreme events directly determines the effectiveness of emergency response and recovery operations (Boeing and Ha, 2024; Oughton, et al., 2023).

In disaster scenarios, rapid and reliable assessment of infrastructure condition is a fundamental prerequisite for informed, trustworthy decision-making (Kopiika, et al., 2025). Damage to bridges, buildings, tunnels, dams, and infrastructure networks may pose structural safety hazards, restrict accessibility, and endanger rescue personnel and affected populations (Ouyang, 2014; Vlašić, et al., 2022). Consequently, authorities must quickly identify hazardous components, prioritise inspection resources, and determine appropriate intervention strategies under severe time pressure. However, the spatial extent, structural complexity, and functional interdependency of modern infrastructure networks frequently exceed the capacity of manual inspection and conventional reporting (Ouyang, 2014). This limitation is particularly evident in scenarios characterised by incomplete information, restricted accessibility, and evolving hazard conditions (Khajwal and Noshadravan, 2021). In practice, decision-makers must often act on fragmented observations and partial evidence, and such reliance on uncertain estimates of damage extent and severity increases the risk of overly conservative or insufficient responses (Khajwal and Noshadravan, 2021).

From a societal perspective, deficiencies in post-disaster assessment propagate into cascading consequences. Insufficient situational awareness may lead to inappropriate evacuation strategies, inefficient deployment of emergency resources, and prolonged disruption of essential services, thus amplifying both direct economic losses and indirect social impacts

1. Introduction

(Pescaroli and Alexander, 2016). As infrastructure systems become more interconnected, a single localised failure can swiftly trigger a network-wide crisis that is often beyond control under traditional fragmented management (Brunner, et al., 2024). For instance, a blocked transport route can delay emergency services, while power outages usually disrupt hospitals and communication networks. Additionally, if drainage or flood defences fail, the resulting secondary hazards pose an even greater threat to life and property. These challenges are especially pronounced in densely populated urban regions, where exposure is high, infrastructure redundancy is limited, and disruptions affect large populations within a short time window (Haggag, et al., 2022).

Beyond operational efficiency, post-disaster assessment also plays a critical role in public trust, governance, and risk communication (Bonfanti, et al., 2023). Road closures, building evacuations, and service suspensions are examples of regulatory interventions that authorities must make and justify when the safety of buildings, transport corridors, and lifeline facilities is uncertain (Fathollahzadeh, et al., 2024). These regulatory interventions may undermine trust in emergency management and make community compliance more difficult if they are thought to be inconsistent, delayed, or poorly supported (Mizrahi, et al., 2023). Additionally, prompt, transparent, and evidence-based assessment procedures promote better communication, lessen social tension, and enable coordinated initial recovery activities (Hu, et al., 2024). Post-disaster assessment is therefore crucial for institutional credibility and societal resilience rather than being a technical requirement.

The global distribution of disaster impacts is highly uneven. Regions with limited technical capacity, economic resources, and institutional support often experience disproportionate losses and prolonged recovery periods (Aka, et al., 2017). In such contexts, the availability of scalable, reliable, and cost-effective assessment technologies becomes critical for reducing vulnerability and improving equity in disaster management (Baraldo and Di Franco, 2024). Enhancing the efficiency and accessibility of infrastructure evaluation is thus directly linked to broader objectives of sustainable development, disaster risk reduction, and long-term social stability (Cai, et al., 2018).

Advances in sensing technologies, remote imaging, and artificial intelligence (AI) provide new opportunities to improve disaster preparedness and response (Al Shafian and Hu, 2024). The widespread deployment of unmanned aerial vehicles, mobile platforms, and satellite systems enables rapid acquisition of high-resolution visual information across large and heterogeneous areas (Zhang, et al., 2025). Modern machine learning and computer vision techniques further

1. Introduction

offer the capability to extract semantic and structural information from these data streams with a level of speed and consistency that exceeds manual analysis (Kazanskiy, et al., 2025). These developments create the possibility of shifting disaster assessment from a predominantly manual, expert-limited process to a more scalable, systematic, and data-driven paradigm. Yet, transforming these technological capabilities into operationally reliable and decision-relevant systems remains a significant research challenge. Disaster imagery is frequently affected by adverse illumination, occlusion, motion blur, debris, weathering, and viewpoint variation (Gupta, et al., 2019). The resulting noise and domain shift can substantially degrade model performance if algorithms are developed primarily on curated benchmarks (Ovadia, et al., 2019). Moreover, post-disaster decision-making requires more than isolated recognition of damage patterns: it demands structured reasoning about safety, accessibility, prioritisation, interdependency, and uncertainty (Pescaroli and Alexander, 2016). Bridging fine-grained visual perception with high-level situational understanding in a manner that is robust, interpretable, and deployable under field constraints therefore remains an open problem.

Beyond technical challenges, post-disaster assessment is inherently a socio-technical problem. Real-world disaster response involves coordination among multiple stakeholders, fragmented and evolving information sources, organisational constraints, and time-critical decision-making under uncertainty. Information must be interpreted, communicated, and prioritised across different roles, rather than processed as a purely computational task. These socio-technical constraints limit the effectiveness of isolated perception or analysis tools and highlight the need for structured frameworks that integrate perception, reasoning, and decision support.

This thesis is motivated by the need to address these gaps. It seeks to develop AI-driven methodologies that enhance both local and global understanding of infrastructure condition in disaster scenarios. By enabling robust, efficient, and interpretable analysis of visual information, the proposed approaches aim to support more timely and consistent decision-making, ultimately contributing to safer, more resilient, and more sustainable civil infrastructure systems.

1.2 Traditional infrastructure inspection and disaster response

Traditional infrastructure inspection and disaster response practices are mainly organised around expert-driven, field-based workflows (Tehler, et al., 2024). Following a disaster event, responsible agencies typically mobilise multidisciplinary teams, including structural engineers,

1. Introduction

geotechnical specialists, emergency responders, and local authorities, to conduct on-site evaluations (Kapucu, 2006; Perry, et al., 2001). Damage evidence is collected through visual inspection, manual measurements, photographic documentation, and reference to established assessment guidelines or rating systems (Graybeal, et al., 2002). These observations are subsequently transformed into situation reports that inform emergency actions, access control, and recovery planning (Kates, et al., 2006). Yet, when applied to large-scale, rapidly evolving disaster scenarios, traditional workflows exhibit fundamental limitations that constrain their effectiveness (Alexander, 2015).

One of the most significant challenges is the reliance on labour-intensive, time-consuming manual inspection. Inspection capacity is rapidly exceeded in large-scale disasters affecting thousands of assets, resulting in limited spatial coverage and delayed identification of critical hazards (Kopiika, et al., 2025). Decision-makers are therefore frequently forced to prioritise inspection targets in the absence of complete information, increasing uncertainty in early response actions. Accessibility and safety constraints further enhance these limitations. Post-disaster environments are usually characterised by debris, unstable structures, aftershocks, flooding, or hazardous materials, which restrict physical access to damaged assets (Al Shafian and Hu, 2024). Inspectors may be unable to reach critical locations for extended periods, particularly in dense urban areas or remote regions (Shi, et al., 2023). Even when access is possible, ensuring personnel safety often requires conservative procedures that slow inspection progress (Marshall, et al., 2013). These constraints delay the assessment of infrastructure condition precisely when timely information is most critical (Zhou, et al., 2018).

Another key limitation of traditional workflows lies in the fragmentation of data collection and integration. Information is typically gathered by multiple teams using heterogeneous tools and formats, ranging from handwritten notes and photographs to isolated measurements and verbal reports (Lozano and Tien, 2023). The subsequent aggregation of these observations relies heavily on manual compilation and subjective interpretation, which is prone to inconsistency, data loss, and miscommunication, particularly when information must be exchanged across organisational boundaries (Adams, et al., 2024). Forming a coherent, system-level understanding of infrastructure condition over a large geographic area, therefore, becomes difficult.

Variability in expert judgement also introduces uncertainty into traditional assessment processes. Although standardised guidelines and rating systems exist, the interpretation of damage severity, safety level, and serviceability often depends on individual experience,

1. Introduction

training background, and contextual judgement (Jayawardene, et al., 2021). Under emergency conditions, factors such as human fatigue, time pressure, and limited information further influence decision-making (Doyle, et al., 2019). This variability can lead to inconsistent assessments across teams and locations, complicating prioritisation and resource allocation (Sköld Gustafsson, et al., 2024). In safety-critical contexts, such an inconsistency directly affects risk management and public interest (Horita, et al., 2018).

Existing research has increasingly sought to enhance infrastructure inspection and disaster response through the integration of digital technologies and computational methods (Hlal, et al., 2025). Remote sensing platforms, unmanned aerial systems, and mobile sensing devices have been widely adopted to improve spatial coverage and reduce dependence on direct field access (Kim, et al., 2022). Vision-based inspection systems have been developed to automate damage detection and condition assessment, enabling rapid processing of large volumes of visual data (Spencer Jr, et al., 2019). At a higher level, geographic information system (GIS)-based platforms, digital twins, and decision-support systems have been proposed to facilitate information integration, spatial analysis, and situational awareness at regional scales (Cheng, et al., 2023). These research efforts can be mainly categorised into several dominant methodological paradigms. One major direction focuses on automated visual perception, including image-based damage classification, defect segmentation, and structural condition recognition, aiming to replace or augment manual inspection through machine learning and computer vision techniques (Guo, et al., 2024). A second paradigm emphasises data integration and spatial analytics, where heterogeneous data sources are aggregated using GIS frameworks, databases, and spatial models to support regional-level assessment and planning (Hlal, et al., 2025). A third line of research concentrates on decision-support and digital platforms, aiming to assist coordination, reporting, and information management through computational tools and interactive systems (Lagap and Ghaffarian, 2024).

Despite these advances, current approaches remain functionally fragmented (Hristidis, et al., 2010). Most studies address isolated technical components, such as perception, mapping, or reporting, without providing a framework that connects local observations to a global understanding of the condition (Hristidis, et al., 2010). Automated perception and high-level response planning are often treated as independent processes, limiting their cooperation in real disaster scenarios (Yu, et al., 2018). As a result, the transition from local-scale assessment to global-scale coordination remains largely manual and organisational rather than computational and structural (Yu, et al., 2018). Additionally, existing studies are frequently based on

1. Introduction

simplified assumptions, and models are often developed and validated using curated datasets acquired under controlled or semi-controlled environments, which do not fully reflect the noise, incompleteness, and heterogeneity characteristic of post-disaster imagery (Nguyen, et al., 2017; Opabola and Galasso, 2024; Torralba and Efros, 2011). At a more fundamental level, the predominance of single-model and single-task paradigms constitutes a structural limitation. While such approaches may achieve strong performance in specific analytical tasks, they lack the organisational and architectural structure required to reflect the multi-stage, multi-role nature of real disaster response workflows (Comfort, 2007). Human disaster management relies on coordinated expert teams performing distinct but interdependent functions, including assessment, interpretation, communication, prioritisation, and planning (Kapucu and Garayev, 2011). Most existing AI-based pipelines, however, are not designed to mirror this organisational structure, resulting in limited capacity for structured collaboration, reasoning, and coordination across tasks (Jiang, et al., 2026; Xu, et al., 2025). Despite the substantial progress in automating individual components of inspection and assessment, a fundamental gap remains in the development of coherent, system-level frameworks that integrate perception, reasoning, and decision support (Xu, et al., 2025). Addressing this gap requires moving beyond isolated technical solutions towards structured architectures that can support both local-scale analysis and global-scale coordination in disaster response.

Current limitations highlight a fundamental mismatch between the complexity of disaster environments and the capacity of traditional infrastructure inspection and response workflows. While expert judgement remains essential, there is a clear need for methodologies that can augment human decision-making by improving efficiency, consistency, and integration of information. Automated and data-driven approaches have the potential to address these gaps by enabling rapid processing of large volumes of observations, reducing subjectivity, and supporting more coherent system-level understanding.

This thesis positions AI-driven perception and structured reasoning as enabling technologies to complement traditional practices. Rather than replacing engineers and emergency managers, the proposed approaches aim to enhance situational awareness, provide consistent and interpretable assessment outputs, and support more informed decision-making. By addressing the limitations of traditional workflows, such systems can contribute to more effective, resilient, and scalable disaster response for civil infrastructure.

1.3 Computer vision for local-scale infrastructure inspection

The inspection and condition assessment of civil infrastructure have traditionally relied on direct visual observation by trained engineers and inspectors (Farahzadi, et al., 2025). Through experience-based judgement, inspectors identify surface defects, material degradation, and structural anomalies, inferring potential safety risks and serviceability issues. While such practices are grounded in engineering expertise and remain indispensable for complex diagnosis, they are inherently constrained by subjectivity, scalability, accessibility, and operational efficiency (Li, et al., 2021).

The limitations of purely manual inspection have become more apparent as infrastructure networks age and expand quickly (Spencer Jr, et al., 2019). Inspection resources are under significant strain due to geographically dispersed systems, dense urban environments, and extensive infrastructures (Karaaslan, et al., 2019; Koch, et al., 2015). Large amounts of visual data must be processed for routine monitoring and post-disaster evaluation, frequently under time constraints and safety restrictions (Yeum and Dyke, 2015). In this situation, maintaining coverage and consistency becomes challenging when relying solely on human observation. Advances in computer vision and deep learning have transformed the paradigm of visual inspection (Cha, et al., 2017; Spencer Jr, et al., 2019). Modern convolutional neural network (CNN) and Transformer-based vision models can learn hierarchical representations from large collections of annotated images, enabling automated identification, localisation, and quantification of surface defects (Li, et al., 2025; Savino and Tondolo, 2023; Yang, et al., 2020). In civil engineering applications, such techniques have been widely explored for the detection and segmentation of cracks, spalling, corrosion, efflorescence, and other forms of material degradation across a broad range of structural components (Cha, et al., 2024; Chen, et al., 2023; Hühthwohl, et al., 2019; Yeum and Dyke, 2015).

From a local-scale perspective, automated defect detection provides several critical advantages over traditional inspection practices. First, it enables the standardisation of visual assessment criteria, reducing inter-observer and intra-observer variability and supporting more objective condition evaluation (Dorafshan, et al., 2018). Second, it significantly accelerates the processing of large image datasets, allowing rapid screening of extensive infrastructures that would otherwise require substantial manual effort (Zhang, et al., 2016). Third, it enhances safety and accessibility by enabling remote inspection through unmanned aerial vehicles, mobile platforms, and fixed sensing systems, thereby reducing the exposure of personnel to

1. Introduction

hazardous environments (Ellenberg, et al., 2016; Metni and Hamel, 2007). Modern satellites and drones provide high-quality images with far greater detail and frequency than previously possible, enabling continuous or on-demand monitoring of infrastructures and supporting both routine maintenance and emergency assessment (Panigati, et al., 2025). In post-disaster scenarios, such remote sensing technologies allow rapid visual assessment of affected areas, providing early situational awareness even before physical access becomes possible (Fernandez Galarreta, et al., 2015). Despite these advances, the practical deployment of computer vision systems in real-world infrastructure inspection remains challenging. While high-resolution images are essential for identifying fine structural defects, processing them puts a heavy strain on computer memory and hardware (Chu and Chun, 2024). This high demand for computational resources conflicts with the limited resources of edge devices commonly used in field environments, such as unmanned aerial vehicles, mobile platforms, and embedded sensing systems (Meng, et al., 2023). As a result, models must balance accuracy with efficiency, requiring careful architectural design and optimisation. Additionally, real-world inspection data exhibit significant variability and noise (Spencer Jr, et al., 2019). Visual information acquired in operational environments is frequently affected by changes in illumination, viewpoint, surface texture, occlusion, weathering, debris, and environmental contamination (Shamsabadi, et al., 2022; Shamsabadi, et al., 2022). In post-disaster contexts, these factors are further compounded by motion blur, structural deformation, and scene complexity (Fernandez Galarreta, et al., 2015; Vetrivel, et al., 2018). Such conditions introduce domain shifts that can substantially degrade model performance if algorithms are developed and evaluated primarily on curated or controlled datasets (Torralba and Efros, 2011). These challenges highlight the need for vision models that are not only accurate but also robust and deployable. Local-scale inspection systems must maintain stable performance under noisy, heterogeneous, and resource-constrained conditions. Robustness to domain variation, resilience to degraded inputs, and efficiency in computation and memory usage are therefore as critical as the raw predictive accuracy.

From a methodological perspective, the evolution from manual inspection to automated defect detection represents a broader transformation in the role of visual perception within infrastructure management. Visual data are no longer treated merely as documentation but as structured sources of information that can be systematically analysed, integrated, and interpreted. Automated perception enables the transformation of raw image inputs into structured representations of local damage states, defect distributions, and condition indicators,

1. Introduction

providing a computational foundation for higher-level reasoning and decision-making processes. Yet, local-scale perception alone is insufficient for comprehensive disaster response and infrastructure management. Automated defect detection provides detailed information at the component level, but does not resolve questions of prioritisation, coordination, and strategic planning. The value of local-scale perception lies in its role as a foundational layer of a broader decision-support architecture, supplying reliable and interpretable inputs for system-level assessment.

This thesis uses computer vision-based defect detection as the starting point for infrastructure perception at the local scale. By focusing on efficient, robust, and deployable vision models, it seeks to establish a reliable perceptual foundation where higher-level reasoning and coordination mechanisms can be constructed. The subsequent chapters develop and evaluate a series of architectures and learning strategies designed to address the practical constraints of real-world deployment, forming the local-scale perceptual core of the proposed AI-driven disaster response framework.

1.4 AI-based global-scale disaster perception and assessment

While local-scale computer vision systems enable automated defect detection and detailed component-level assessment, effective disaster response and infrastructure management require a fundamentally broader level of understanding. Decision-making in disaster scenarios is not driven by isolated observations, but by the integration and interpretation of large volumes of heterogeneous information across spatial, functional, and organisational scales (Hlal, et al., 2025). Transforming local observations into coherent, global situational awareness therefore represents a critical challenge that extends beyond perception alone. Global-scale perception is about more than just gathering data. It involves combining visual inputs with historical records, geographic data, and expert knowledge, and needs to interpret the implications of damages for safety, accessibility, functionality, and risk propagation across interconnected infrastructure systems (Ouyang, 2014). This process involves a shift from separate physical damage reports to an understanding of whole-network impacts and strategic response priorities.

Traditional disaster management relies on human experts and organisational structures to perform the information integration for global-scale perception (Aka, et al., 2017). Multidisciplinary teams work through hierarchical, collaborative workflows to evaluate damage and share their insights, ensuring that response actions are effectively coordinated. Such workflows reflect a form of collaborative problem-solving, where different roles

1. Introduction

contribute specialised knowledge to a shared situational understanding. Yet, as disaster scenarios grow in scale and complexity, the cognitive and organisational burden of coordinating them becomes increasingly difficult to sustain.

Recent advances in artificial intelligence have created new possibilities for supporting global-scale disaster perception and assessment by emulating key aspects of human cognitive processes in computational systems (Xu, et al., 2025). The emergence of large vision-language models (LVLMs) represents a significant step toward multimodal intelligence, extending language-based reasoning capabilities to joint processing of visual and textual information (Alayrac, et al., 2022; Radford, et al., 2021). By learning from large-scale image-text corpora, such models can transform raw visual inputs into structured semantic representations, enabling the generation of descriptive, summarised, and interpretable outputs that are directly usable within decision-making processes (Li, et al., 2022). While these models lack genuine human-like cognition, their ability to model linguistic and visual patterns at scale allows them to function as powerful tools for large-scale information processing and semantic abstraction (Schick, et al., 2023). Beyond individual models, the development of agent-based AI systems introduces a new paradigm for modelling complex reasoning workflows (Wu, et al., 2024). Rather than relying on monolithic architectures, multi-agent systems decompose complex tasks into specialised functional roles, enabling different agents to perform distinct but interdependent operations such as perception, interpretation, validation, communication, coordination, and planning (Wu, et al., 2024). This organisational structure provides a scalable computational framework to mimic human disaster response teams, where distributed expertise and role differentiation support collaborative reasoning and coordinated decision-making. Within such systems, intelligence emerges not from a single model, but from structured interaction between specialised components. From a system perspective, global-scale disaster perception can therefore be understood as a process of structured information transformation across multiple cognitive layers. Local observations are first converted into interpretable representations of damage and condition through perceptual models. These representations are then integrated with contextual, spatial, and historical knowledge to form a coherent situational model of the disaster environment. Finally, this situational model supports higher-level reasoning processes, including prioritisation, risk interpretation, resource allocation, and strategic planning. Each stage of this transformation requires different forms of intelligence, ranging from visual perception and semantic understanding to structured inference and decision logic.

1. Introduction

Despite these advances, most existing AI-based approaches for disaster perception and assessment remain focused on isolated analytical functions (Xu, et al., 2025). Functions such as damage classification or hazard mapping are typically developed as standalone tools rather than integral parts of a unified decision-support framework (Nguyen, et al., 2017). Consequently, the connection between local perception and global decision-making remains structurally weak, and the transition from data to action is frequently mediated by manual interpretation and organisational processes rather than computational reasoning. Additionally, global-scale assessment requires interpretability and transparency for AI-based approaches (Doshi-Velez and Kim, 2017). In safety-critical contexts such as disaster response, decision-support systems must not only produce outputs but also provide explanations that can be understood and evaluated by human operators. To ensure accountability and trust, such systems must be designed to augment human judgment rather than hide critical logic behind opaque computational layers. These challenges indicate that global-scale disaster perception cannot be addressed through perception models alone. It requires structured reasoning frameworks that integrate perception, knowledge representation, and decision logic within a coherent system architecture. Such frameworks must support information fusion across scales, role-based task decomposition, collaborative reasoning, and human-AI interaction.

This thesis develops an AI agent-based framework for global-scale disaster assessment. By integrating local-scale perceptual models with a multi-agent reasoning architecture, it seeks to bridge the gap between detailed visual observations and high-level decision support. Local defect detection is treated as the perceptual foundation, while structured agent collaboration provides the mechanism for interpretation, integration, and coordination. In this framework, AI systems do not replace human expertise but act as structured cognitive support tools that assist in information processing, reasoning, and communication. The subsequent chapters operationalise this framework through the design of efficient and robust local-scale perception models and a multi-agent global reasoning framework for post-disaster assessment and response.

1.5 Thesis outline

Based on the preceding analysis, two interconnected research gaps are identified:

- **Local-scale gap:** The lack of efficient, robust, and deployable defect detection models that can operate reliably under resource constraints and severe noise, domain variability, and real-world environmental uncertainty.

1. Introduction

- Global-scale gap: The absence of autonomous, interpretable, and collaborative disaster assessment frameworks that extend beyond single-model perception and support structured reasoning, coordination, and decision-making at the system level.

These gaps indicate the need for a framework that connects local-scale perception with global-scale reasoning within a coherent system architecture. In response to these gaps, the primary objectives of this thesis are:

- To develop robust and accurate computer vision architectures for local-scale infrastructure defect detection, with a focus on reliability and generalisation under realistic inspection conditions.
- To design robust training frameworks for noise-resilient and deployable defect detection, enhancing model stability, anti-noise robustness, and deployment feasibility under heterogeneous and degraded visual environments.
- To investigate new visual architectures for high-resolution defect inspection, addressing the challenges of large-scale image processing, long-range dependency modelling, and fine-grained defect representation.
- To develop an AI agent-based framework for global-scale defect perception and disaster assessment that integrates local-scale perceptual outputs with structured multi-agent reasoning for interpretation, coordination, and decision support.

The thesis is organised into six chapters, and the overall structure is shown in Figure 1.1.

1. Introduction

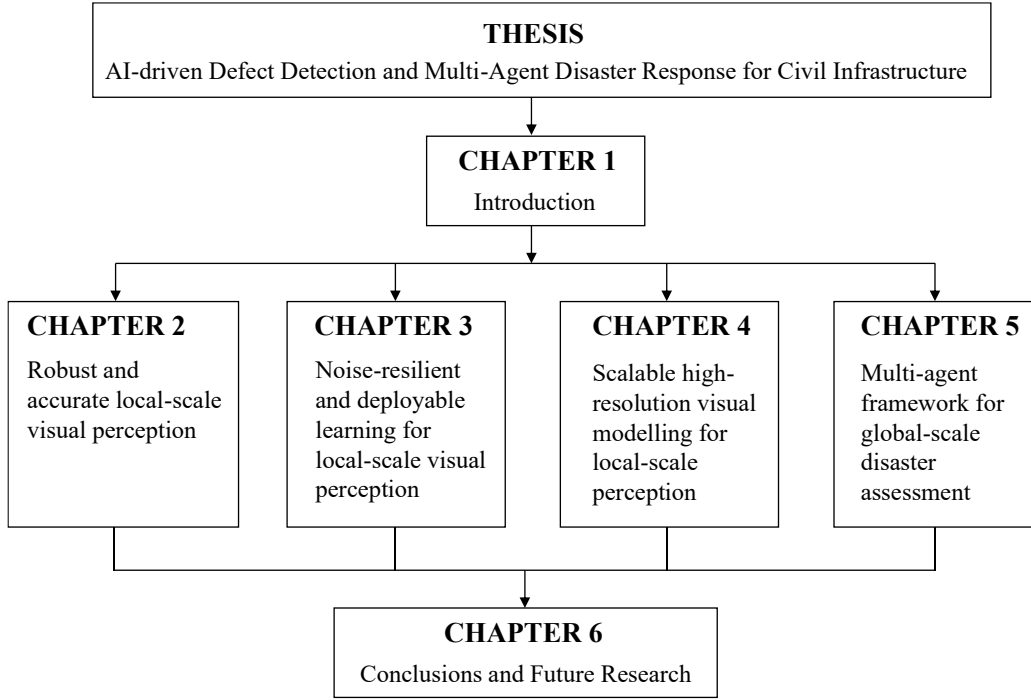


Figure 1.1 Overview of the thesis structure and main research components

Chapter 1 introduces the research background, societal motivation, and methodological context of AI-driven infrastructure inspection and disaster response, outlining the challenges of both local-scale defect perception and global-scale disaster assessment, and establishing the conceptual framework of the thesis.

In Chapter 2, robust and accurate computer vision models for local-scale infrastructure defect detection are investigated. This chapter presents the proposed PoolingCrack model (Chen, et al., 2024), which focuses on improving the robustness and reliability of crack segmentation through architectural optimisation and feature aggregation strategies. Comprehensive evaluations are conducted to analyse detection accuracy, robustness, and generalisation performance under realistic inspection conditions, establishing a reliable perceptual foundation for local-scale infrastructure inspection.

Chapter 3 proposes a Robust Feature Knowledge Distillation (RFKD) framework (Chen, et al., 2024) for noise-resilient and deployable defect detection. This chapter systematically studies the challenges of model robustness and generalisation under noisy, heterogeneous, and degraded visual environments, and introduces the RFKD framework that enhances anti-noise robustness, stability, and deployment while reducing model complexity. Extensive experiments are conducted to evaluate robustness, performance consistency, and deployment feasibility under realistic operational constraints.

1. Introduction

In Chapter 4, Vision Mamba-based architectures for high-resolution defect inspection are developed and analysed (Chen, et al., 2024). This chapter investigates sequence-model-based visual representations for high-resolution infrastructure images, addressing the challenges of large-scale image processing, long-range dependency modelling, and fine-grained defect representation. The proposed architectures are evaluated in terms of accuracy, efficiency, and scalability for high-resolution local inspection scenarios.

Chapter 5 proposes an AI agent-based framework, DisasTeller (Chen, et al., 2026), for global-scale defect perception and disaster assessment. This chapter presents a multi-agent reasoning architecture that integrates local-scale perceptual outputs with structured agent collaboration for interpretation, coordination, and decision support. The framework enables the transformation of local observations into coherent global situational awareness through role-based task decomposition, collaborative reasoning, and interpretable decision-making.

Finally, Chapter 6 concludes the thesis by summarising the main findings and contributions, discussing the limitations of the proposed approaches, and outlining directions for future research in AI-driven infrastructure inspection and disaster response.

References

- Adams R.M., C.M. Evans, L. Peek, Defining, collecting, and sharing perishable disaster data, *Disasters* 48 (1) (2024), pp. e12592.
- Aka F.T., G.W. Buh, W.Y. Fantong, Issa, I.T. Zouh, S.L.B. Djomou, R.T. Ghogomu, T. Gibson, M.-A. Marmol del, L.N. Sigha, Disaster prevention, disaster preparedness and local community resilience within the context of disaster risk management in Cameroon, *Natural hazards* 86 (1) (2017), pp. 57-88.
- Akinsanola A., P. Singhai, T. Taguela, A. Folorunsho, O. Adeyeri, T. Morakinyo, A. Adebisi, A review of urban resilience to weather and climate extremes, *City and Built Environment* 3 (1) (2025), pp. 24.
- Al Shafian S., D. Hu, Integrating machine learning and remote sensing in disaster management: A decadal review of post-disaster building damage assessment, *Buildings* 14 (8) (2024), pp. 2344.
- Alayrac J.-B., J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, Flamingo: a visual language model for few-shot learning, *Advances in neural information processing systems* 35 (2022), pp. 23716-23736.

1. Introduction

Alexander D.E., Disaster and emergency planning for preparedness, response, and recovery, Oxford University Press 2015.

Baraldo M., P.D.G. Di Franco, Place-centred emerging technologies for disaster management: A scoping review, *International Journal of Disaster Risk Reduction* 112 (2024), pp. 104782.

Boeing G., J. Ha, Resilient by design: Simulating street network disruptions across every urban area in the world, *Transportation research part A: policy and practice* 182 (2024), pp. 104016.

Bonfanti R.C., B. Oberti, E. Ravazzoli, A. Rinaldi, S. Ruggieri, A. Schimmenti, The role of trust in disaster risk reduction: a critical review, *International journal of environmental research and public health* 21 (1) (2023), pp. 29.

Brunner L.G., R. Peer, C. Zorn, R. Paulik, T. Logan, Understanding cascading risks through real-world interdependent urban infrastructure, *Reliability Engineering & System Safety* 241 (2024), pp. 109653.

Cai H., N.S. Lam, Y. Qiang, L. Zou, R.M. Correll, V. Mihunov, A synthesis of disaster resilience measurement methods and indices, *International journal of disaster risk reduction* 31 (2018), pp. 844-855.

Cha Y.-J., W. Choi, O. Büyüköztürk, Deep Learning-Based Crack Damage Detection Using Convolutional Neural Networks, *Computer-Aided Civil and Infrastructure Engineering* 32 (5) (2017), pp. 361-378.

Cha Y.-J., R. Ali, J. Lewis, O. Büyüköztürk, Deep learning-based structural health monitoring, *Automation in Construction* 161 (2024), pp. 105328.

Chen L., W. Chen, L. Wang, C. Zhai, X. Hu, L. Sun, Y. Tian, X. Huang, L. Jiang, Convolutional neural networks (CNNs)-based multi-category damage detection and recognition of high-speed rail (HSR) reinforced concrete (RC) bridges using test images, *Engineering Structures* 276 (2023), pp. 115306.

Chen Z., E.A. Shamsabadi, S. Jiang, L. Shen, D. Dias-da-Costa, Vision Mamba-based autonomous crack segmentation on concrete, asphalt, and masonry surfaces, *arXiv preprint arXiv:2406.16518* (2024).

Chen Z., E.A. Shamsabadi, S. Jiang, L. Shen, D. Dias-da-Costa, Robust feature knowledge distillation for enhanced performance of lightweight crack segmentation models, *arXiv preprint arXiv:2404.06258* (2024).

Chen Z., E.A. Shamsabadi, S. Jiang, L. Shen, D. Dias-da-Costa, An average pooling designed transformer for robust crack segmentation, *Automation in Construction* 162 (2024), pp. 105367.

1. Introduction

- Chen Z., E. Asadi Shamsabadi, S. Jiang, L. Shen, D. Dias-da-Costa, Integration of large vision language models for efficient post-disaster damage assessment and reporting, *Nature Communications* (2026).
- Cheng R., L. Hou, S. Xu, A review of digital twin applications in civil and infrastructure emergency management, *Buildings* 13 (5) (2023), pp. 1143.
- Chu H., P.j. Chun, Fine-grained crack segmentation for high-resolution images via a multiscale cascaded network, *Computer-Aided Civil and Infrastructure Engineering* 39 (4) (2024), pp. 575-594.
- Comfort L.K., Crisis management in hindsight: Cognition, communication, coordination, and control, *Public administration review* 67 (2007), pp. 189-197.
- Dorafshan S., R.J. Thomas, M. Maguire, Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete, *Construction and Building Materials* 186 (2018), pp. 1031-1045.
- Doshi-Velez F., B. Kim, Towards a rigorous science of interpretable machine learning, *arXiv preprint arXiv:1702.08608* (2017).
- Doyle E.E., D.M. Johnston, R. Smith, D. Paton, Communicating model uncertainty for natural hazards: A qualitative systematic thematic review, *International journal of disaster risk reduction* 33 (2019), pp. 449-476.
- Ellenberg A., A. Kontsos, F. Moon, I. Bartoli, Bridge deck delamination identification from unmanned aerial vehicle infrared imagery, *Automation in Construction* 72 (2016), pp. 155-165.
- Farahzadi L., I. Odeh, M. Kioumars, B. Shafei, Automated Image-Based Condition Assessment of Built Environment: A State-of-the-Art Investigation of Damage Characteristics and Detection Requirements, *Results in Engineering* (2025), pp. 104978.
- Fathollahzadeh A., J. Babaie, I. Salmani, M.A. Morowatisharifabad, M.-R. Khajehaminian, Challenges of disaster risk communication from the perspectives of experts and affected people: A conventional content analysis, *International Journal of Disaster Risk Reduction* 108 (2024), pp. 104553.
- Fernandez Galarreta J., N. Kerle, M. Gerke, UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning, *Natural hazards and earth system sciences* 15 (6) (2015), pp. 1087-1101.
- Graybeal B.A., B.M. Phares, D.D. Rolander, M. Moore, G. Washer, Visual inspection of highway bridges, *Journal of nondestructive evaluation* 21 (3) (2002), pp. 67-83.

1. Introduction

- Guo J., P. Liu, B. Xiao, L. Deng, Q. Wang, Surface defect detection of civil structures using images: Review from data perspective, *Automation in Construction* 158 (2024), pp. 105186.
- Gupta R., B. Goodman, N. Patel, R. Hosfelt, S. Sajeev, E. Heim, J. Doshi, K. Lucas, H. Choset, M. Gaston, Creating xBD: A dataset for assessing building damage from satellite imagery, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, (2019), pp. 10-17.
- Haggag M., M. Ezzeldin, W. El-Dakhakhni, E. Hassini, Resilient cities critical infrastructure interdependence: a meta-research, *Sustainable and resilient infrastructure* 7 (4) (2022), pp. 291-312.
- Hlal M., J.-C. Baraka Munyaka, J. Chenal, R. Azmi, E.B. Diop, M. Bounabi, S.A. Ebnou Abdem, M.A.S. Almouctar, M. Adraoui, Digital twin technology for urban flood risk management: a systematic review of remote sensing applications and early warning systems, *Remote Sensing* 17 (17) (2025), pp. 3104.
- Horita F.E., J.P. de Albuquerque, V. Marchezini, Understanding the decision-making process in disaster risk monitoring and early-warning: A case study within a control room in Brazil, *International journal of disaster risk reduction* 28 (2018), pp. 22-31.
- Hristidis V., S.-C. Chen, T. Li, S. Luis, Y. Deng, Survey of data management and analysis in disaster situations, *Journal of Systems and Software* 83 (10) (2010), pp. 1701-1714.
- Hu S., J. Zhou, L. Tang, Y. Zhao, H. Xing, The impact of risk communication, trust, and coping appraisal on individual preparedness decisions in geological hazard-prone areas, *International Journal of Disaster Risk Reduction* 100 (2024), pp. 104166.
- Hüthwohl P., R. Lu, I. Brilakis, Multi-classifier for reinforced concrete bridge defects, *Automation in Construction* 105 (2019), pp. 102824.
- Jayawardene V., T.J. Huggins, R. Prasanna, B. Fakhruddin, The role of data and information quality during disaster response decision-making, *Progress in disaster science* 12 (2021), pp. 100202.
- Jiang Y., J. Wang, X. Shen, K. Dai, Q. Ge, Multitask unified large vision-language model for post-earthquake structural damage assessment of buildings, *Automation in Construction* 182 (2026), pp. 106720.
- Kapucu N., Interagency communication networks during emergencies: Boundary spanners in multiagency coordination, *The American review of public administration* 36 (2) (2006), pp. 207-225.
- Kapucu N., V. Garayev, Collaborative decision-making in emergency and disaster management, *International Journal of Public Administration* 34 (6) (2011), pp. 366-375.

1. Introduction

- Karaaslan E., U. Bagci, F.N. Catbas, Artificial intelligence assisted infrastructure assessment using mixed reality systems, *Transportation Research Record* 2673 (12) (2019), pp. 413-424.
- Kates R.W., C.E. Colten, S. Laska, S.P. Leatherman, Reconstruction of New Orleans after Hurricane Katrina: a research perspective, *Proceedings of the national Academy of Sciences* 103 (40) (2006), pp. 14653-14660.
- Kazanskiy N., R. Khabibullin, A. Nikonorov, S. Khonina, A comprehensive review of remote sensing and artificial intelligence integration: advances, applications, and challenges, *Sensors* 25 (19) (2025), pp. 5965.
- Khajwal A.B., A. Noshadravan, An uncertainty-aware framework for reliable disaster damage assessment via crowdsourcing, *International Journal of Disaster Risk Reduction* 55 (2021), pp. 102110.
- Kim D., J. Won, E. Lee, K.R. Park, J. Kim, S. Park, H. Yang, M. Cha, Disaster assessment using computer vision and satellite imagery: Applications in detecting water-related building damages, *Frontiers in Environmental Science* 10 (2022), pp. 969758.
- Koch C., K. Georgieva, V. Kasireddy, B. Akinci, P. Fieguth, A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure, *Advanced Engineering Informatics* 29 (2) (2015), pp. 196-210.
- Kopiika N., A. Karavias, P. Krassakis, Z. Ye, J. Ninic, N. Shakhovska, S. Argyroudis, S.-A. Mitoulis, Rapid post-disaster infrastructure damage characterisation using remote sensing and deep learning technologies: A tiered approach, *Automation in Construction* 170 (2025), pp. 105955.
- Lagap U., S. Ghaffarian, Digital post-disaster risk management twinning: A review and improved conceptual framework, *International Journal of Disaster Risk Reduction* 110 (2024), pp. 104629.
- Li J., D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, *International conference on machine learning*, PMLR, (2022), pp. 12888-12900.
- Li T., M. Alipour, D.K. Harris, Mapping textual descriptions to condition ratings to assist bridge inspection and condition assessment using hierarchical attention, *Automation in Construction* 129 (2021), pp. 103801.
- Li Y., H. Zhao, Y. Wei, T. Bao, T. Li, Q. Wang, N. Wang, M. Zhao, Vision-guided crack identification and size quantification framework for dam underwater concrete structures, *Structural Health Monitoring* 24 (4) (2025), pp. 2125-2148.

1. Introduction

Lozano J.-M., I. Tien, Data collection tools for post-disaster damage assessment of building and lifeline infrastructure systems, *International journal of disaster risk reduction* 94 (2023), pp. 103819.

Marshall J.D., K. Jaiswal, N. Gould, F. Turner, B. Lizundia, J.C. Barnes, Post-earthquake building safety inspection: Lessons from the Canterbury, New Zealand, earthquakes, *Earthquake Spectra* 29 (3) (2013), pp. 1091-1107.

Meng S., Z. Gao, Y. Zhou, B. He, A. Djerrad, Real-time automatic crack detection method based on drone, *Computer-Aided Civil and Infrastructure Engineering* 38 (7) (2023), pp. 849-872.

Metni N., T. Hamel, A UAV for bridge inspection: Visual servoing control law with orientation limits, *Automation in construction* 17 (1) (2007), pp. 3-10.

Mizrahi S., N. Cohen, E. Vigoda-Gadot, D.N. Krup, Compliance with government policies during emergencies: Trust, participation and protective actions, *Governance* 36 (4) (2023), pp. 1083-1102.

Nguyen D.T., F. Ofli, M. Imran, P. Mitra, Damage assessment from social media imagery data during disasters, *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, (2017), pp. 569-576.

Opabola E.A., C. Galasso, Informing disaster-risk management policies for education infrastructure using scenario-based recovery analyses, *Nature Communications* 15 (1) (2024), pp. 325.

Oughton E.J., T. Russell, J. Oh, S. Ballan, J.W. Hall, Global Vulnerability Assessment of Mobile Telecommunications Infrastructure to Climate Hazards using Crowdsourced Open Data, *arXiv preprint arXiv:2311.04392* (2023).

Ouyang M., Review on modeling and simulation of interdependent critical infrastructure systems, *Reliability engineering & System safety* 121 (2014), pp. 43-60.

Ovadia Y., E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, J. Snoek, Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, *Advances in neural information processing systems* 32 (2019).

Panigati T., M. Zini, D. Striccoli, P.F. Giordano, D. Tonelli, M.P. Limongelli, D. Zonta, Drone-based bridge inspections: Current practices and future directions, *Automation in Construction* 173 (2025), pp. 106101.

Perry R.W., M.K. Lindell, K.J. Tierney, Facing the unexpected: Disaster preparedness and response in the United States, Joseph Henry Press 2001.

1. Introduction

- Pescaroli G., D. Alexander, Critical infrastructure, panarchies and the vulnerability paths of cascading disasters, *Natural Hazards* 82 (1) (2016), pp. 175-192.
- Radford A., J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Aspell, P. Mishkin, J. Clark, Learning transferable visual models from natural language supervision, *International conference on machine learning*, PmLR, (2021), pp. 8748-8763.
- Rezvani S., M.J. Falcão, D. Komljenovic, N.M. de Almeida, A systematic literature review on urban resilience enabled with asset and disaster risk management approaches and GIS-based decision support tools, *Applied Sciences* 13 (4) (2023), pp. 2223.
- Savino P., F. Tondolo, Civil infrastructure defect assessment using pixel-wise segmentation based on deep learning, *Journal of Civil Structural Health Monitoring* 13 (1) (2023), pp. 35-48.
- Schick T., J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language models can teach themselves to use tools, *Advances in Neural Information Processing Systems* 36 (2023), pp. 68539-68551.
- Shamsabadi E.A., C. Xu, D. Dias-da-Costa, Robust crack detection in masonry structures with Transformers, *Measurement* 200 (2022), pp. 111590.
- Shamsabadi E.A., C. Xu, A.S. Rao, T. Nguyen, T. Ngo, D. Dias-da-Costa, Vision transformer-based autonomous crack detection on asphalt and concrete surfaces, *Automation in Construction* 140 (2022), pp. 104316.
- Shi Y., S. Liu, Z. Li, Y. Ding, Review on quick safety assessment of building structures in complex urban environment after extreme explosion events, *International Journal of Protective Structures* 14 (3) (2023), pp. 438-458.
- Sköld Gustafsson V., T. Andersson Granberg, S. Pilemalm, M. Waldemarsson, Identifying decision support needs for emergency response to multiple natural hazards: an activity theory approach, *Natural Hazards* 120 (3) (2024), pp. 2777-2802.
- Spencer Jr B.F., V. Hoskere, Y. Narazaki, Advances in computer vision-based civil infrastructure inspection and monitoring, *Engineering* 5 (2) (2019), pp. 199-222.
- Tehler H., A. Cedergren, M.d.G. de Herve, J. Gustavsson, H. Hassel, H. Lindbom, L. Nyberg, M. Wester, Evidence-based disaster risk management: A scoping review focusing on risk, resilience and vulnerability assessment, *Progress in Disaster Science* 23 (2024), pp. 100335.
- Torralba A., A.A. Efros, Unbiased look at dataset bias, *CVPR 2011*, IEEE, (2011), pp. 1521-1528.
- Vetrivel A., M. Gerke, N. Kerle, F. Nex, G. Vosselman, Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution

1. Introduction

oblique aerial images, and multiple-kernel-learning, *ISPRS journal of photogrammetry and remote sensing* 140 (2018), pp. 45-59.

Vlašić A., M. Srbić, D. Skokandić, A. Mandić Ivanković, Post-earthquake rapid damage assessment of road bridges in Glin County, *Buildings* 12 (1) (2022), pp. 42.

Wu Q., G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, Autogen: Enabling next-gen LLM applications via multi-agent conversations, *First Conference on Language Modeling*, (2024).

Xu F., J. Ma, N. Li, J.C. Cheng, Large language model applications in disaster management: an interdisciplinary review, *International Journal of Disaster Risk Reduction* (2025), pp. 105642.

Yang Q., W. Shi, J. Chen, W. Lin, Deep convolution neural network-based transfer learning method for civil infrastructure crack detection, *Automation in Construction* 116 (2020), pp. 103199.

Yeum C.M., S.J. Dyke, Vision-Based Automated Crack Detection for Bridge Inspection, *Computer-Aided Civil and Infrastructure Engineering* 30 (10) (2015), pp. 759-770.

Yeum C.M., S.J. Dyke, Vision-based automated crack detection for bridge inspection, *Computer-Aided Civil and Infrastructure Engineering* 30 (10) (2015), pp. 759-770.

Yu M., C. Yang, Y. Li, Big data in natural disaster management: a review, *Geosciences* 8 (5) (2018), pp. 165.

Zhang L., F. Yang, Y.D. Zhang, Y.J. Zhu, Road crack detection using deep convolutional neural network, *2016 IEEE international conference on image processing (ICIP)*, IEEE, (2016), pp. 3708-3712.

Zhang M., Z. Chen, J. Wang, B. Kar, M. Pierce, K. Tiampo, R. Eguchi, M. Glasscoe, Optical Remote Sensing for Global Flood Disaster Mapping: A Critical Review Towards Operational Readiness, *Remote Sensing* 17 (11) (2025), pp. 1886.

Zhou H., A. Taal, S. Koulouzis, J. Wang, Y. Hu, G. Suci Jr, V. Poenaru, C. de Laat, Z. Zhao, Dynamic real-time infrastructure planning and deployment for disaster early warning systems, *International conference on computational Science*, Springer, (2018), pp. 644-654.

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

This chapter presents the paper “Zhaohui Chen, Elyas Asadi Shamsabadi, Sheng Jiang, Luming Shen, Daniel Dias-da-Costa. (2024). An average pooling designed Transformer for robust crack segmentation. *Automation in Construction*, 162, 105367”. This chapter focuses on the development of a Transformer-based segmentation architecture for robust and accurate crack detection under realistic inspection conditions. The proposed model introduces an average pooling-based feature aggregation mechanism combined with convolutional projection to enhance feature stability and representation consistency. This design improves robustness against noise, surface texture interference, and visual degradation, while maintaining high segmentation accuracy and efficiency compared with current advanced models. Comprehensive experimental evaluations demonstrate strong robustness and generalisation performance across heterogeneous datasets and degraded visual environments. This chapter establishes a reliable local-scale perceptual foundation for subsequent developments in robustness-oriented learning frameworks and scalable visual modelling.

2.1 Introduction

Civil structures are prone to various deterioration forms that may cause loss of their anticipated performance due to external loads and environmental conditions. Cracks are one of the most common defects, with the potential to expand rapidly and, in some cases, leading to structural failures. In most cases, crack-related issues can be mitigated or even resolved if detected in their early stages where interventions are rather effective. In this domain, the implementation of regular inspection and condition evaluation schedules of civil infrastructure can contribute to the safety of structures and their long-term serviceability (Ai, et al., 2023).

Regular inspections typically require accurate detection of cracks, which can also prove important to the success of a damage or defect diagnosis system. Traditional crack inspection methods rely on the manual recording of structural damages, which is time-consuming, labour-intensive, sometimes hazardous, and subjective to the experience and skills of the inspector. As a result of these issues, computer vision-based methods are becoming a viable solution as machine learning (ML) techniques continue developing (Spencer, et al., 2019). Image processing methods have been a topic of research for years; however, their low generalisation

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

has impeded their application in practical terms. E.g., traditional threshold- (Jiang, et al., 2021; Kirschke and Velinsky, 1992; Xu, et al., 2011) and morphology-based (Canny, 1986; Tsai, et al., 2013; Yamaguchi and Hashimoto, 2010) methods cannot achieve satisfactory detection results when analysing crack images with complex textures. Compared to these, classic ML (Li, et al., 2017; Peng, et al., 2021; Rodriguez-Lozano, et al., 2020) deploys more generalisable features, although still relying on the manually designed crack features while being ineffective in capturing the dependencies in visual data. Addressing these issues, convolutional neural networks (CNNs) have gradually become a more common approach for accurate crack detection (Cha, et al., 2017; Cha, et al., 2018).

Leveraging the capabilities of CNNs, researchers have developed several patch-based (Ali, et al., 2021; Cha, et al., 2017; Park, et al., 2019) and pixel-wise-based (Ali, et al., 2022; Hsieh and Tsai Yichang, 2020; Kang, et al., 2020) methods to enable accurate crack detection. As each model has its shortcomings, new works have gradually emerged, building upon the successes of previous models. Various CNN-based networks, including fully convolutional networks (FCN) (Dung, 2019; Ji, et al., 2020; Yang, et al., 2018), U-Net variants (Ai, et al., 2020; Huyan, et al., 2020; Liu, et al., 2019; Ronneberger, et al., 2015; Zhang, et al., 2021), Feature Pyramid Networks (FPN) (Lin, et al., 2017), and DeepLabv3+ (Chen, et al., 2017; Chen, et al., 2018) were applied for crack segmentation over various materials, achieving high accuracy crack semantic segmentation in controlled settings (Bang, et al., 2019; Dais, et al., 2021). Despite the advantages, a broader application of CNN-based models in the industry was hampered by the existing limitations of convolutional operation. In fact, while the convolutional operation enables CNNs to achieve even higher accuracy than humans in controlled environments (He, et al., 2015), the use of small and immutable convolutional kernels directs the attention of CNNs towards local information (Zhu, et al., 2023). As a result, CNNs largely rely on local dependencies (e.g., the texture of semantic objects) to run vision-based tasks, while humans use a combination of global and local information for visual perception (Tuli, et al., 2021). To address this issue and build upon the success of self-attention in natural language processing, self-attention was recently introduced in computer vision through Transformers (Dosovitskiy, et al., 2020; Raghu, et al., 2021; Vaswani, et al., 2017). Self-attention equips deep learning architectures with the ability to access multi-range

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

information in input images, allowing them to capture a broader range of contextual cues and improve performance on vision-based tasks (Raghu, et al., 2021).

Transformers have shown remarkable performance in crack segmentation (Cao, et al., 2022; Chen, et al., 2021; Liu, et al., 2021; Strudel, et al., 2021; Xie, et al., 2021). Liu et al. (Liu, et al., 2021) proposed CrackFormer for pixel-level crack detection, with fewer parameters and higher optimal dataset scale values than classical CNN-based models. Pointing to the similarities between the human recognition system and Transformer-based models, Asadi Shamsabadi et al. (Shamsabadi, et al., 2022; Shamsabadi, et al., 2022) evaluated the performance of Transformer-based models (Chen, et al., 2021; Xie, et al., 2021) in concrete, asphalt and masonry crack segmentation, demonstrating significant Precision, robustness and generalisation. Meanwhile, Wang et al. (Wang and Su, 2022) developed SegCrack by using MMSegmentation (Contributors, 2020) and the OHEM strategy (Shrivastava, et al., 2016) to enhance the crack segmentation accuracy. Subsequently, Guo et al. (Guo, et al., 2023) proposed Crack Transformer (CT), a model that combines elements from the Swin Transformer (Liu, et al., 2021) and SegFormer (Xie, et al., 2021). This hybrid approach could provide enhanced accuracy compared to CNN-based models in crack segmentation, with fewer floating-point operations (FLOPs). Furthermore, several hybrid-designed models (Xiang, et al., 2023; Xu, et al., 2022; Yang, et al., 2023; Zhang, et al., 2023) that combine Transformers and CNNs were put forward recently, all to achieve higher accuracy in crack segmentation.

Despite the promising results of Transformer-based crack detection, current models still have limitations. Although they can achieve high accuracy and generalisation, practical deployment can be undermined by unnecessary redundancy in their design architecture. In most cases, the traditional self-attention module (Vaswani, et al., 2017) is deployed, which applies complex scaled dot-product operations and needs to concatenate Queries (Q), Keys (K), and Values (V) matrices for the attention score calculation. In addition, the self-attention module depends on position encoding, which requires extra linear projection and position information process, thus incrementing the computational complexity. As for developing a Transformer-CNN hybrid network (Chen, et al., 2021; Xiang, et al., 2023; Zhang, et al., 2023), the inclusion of the extra CNN part also leads to an unnecessary increase in computational complexity. Consequently, this complexity can negatively impair the model's efficiency. The recurrent down-sampling and up-sampling in Transformer-based models also bring about spatial

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

misalignments of feature map pixels in the fusion stages, which weakens the output feature representation and final prediction accuracy (Huang, et al., 2021).

For practical real-time inspection applications, crack segmentation algorithms must have stable performance in various texture surfaces and lighting conditions as well as outstanding efficiency. This is challenging as the effectiveness of robust models is currently reliant on the use of computationally expensive modules. Considering the trade-off between robustness and efficiency, we aim to break this balance and develop a light model for crack segmentation with lower computational cost but even stronger generalisation and robustness against various noise types than the representative models. By tailoring efficient patch embedding, token mixer and feature fusion correction design, we develop a novel model named PoolingCrack. Our main hypothesis is that the convolutional patch embedding and the average pooling design can enable a lighter design than traditional self-attention modules while maintaining the same performance, whereas feature alignment would contribute to the accurate feature fusion, resulting in the enhancement of generalisation performances.

The overall structure of the paper is as follows. Section 2.2 illustrates the detailed architecture of the proposed model. Section 2.3 introduces the performance study implementation, including crack datasets, training configuration, and assessment framework. In Section 2.4, the comprehensive experiment results are presented and discussed in detail, with a focus on explaining the higher performance of PoolingCrack. Finally, the conclusions are drawn in Section 2.5.

2.2 Methodology

Figure 2.1 shows the overall structure of the proposed model. To extract and fuse hierarchical features, the model employs four down-sampling and up-sampling stages: 1) convolutional patch embedding, 2) pooling blocks, 3) feature selection modules (FSM), and 4) feature alignment modules (FAM). The corresponding details are discussed in the next sections.

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

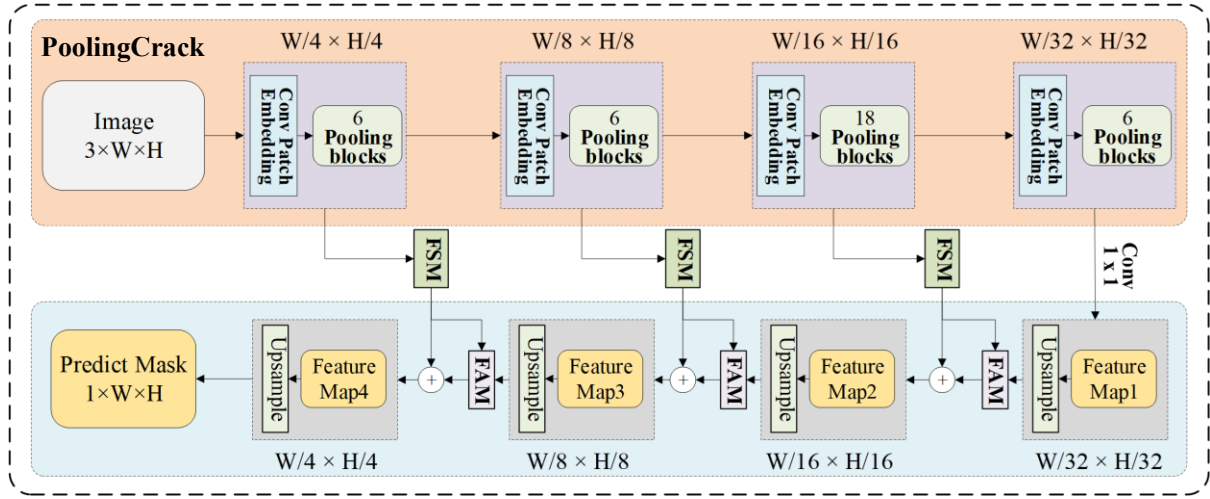


Figure 2.1 Framework of the proposed crack segmentation model (PoolingCrack).

2.2.1 Convolutional patch embedding

Convolutional patch embedding – see Figure 2.2-b – utilises zero-padding and non-overlap convolution projection to encode patch tokens (Wu, et al., 2021). Unlike the linear projection, which is shown in Figure 2.2-a, in the absolute position encoding (Dosovitskiy, et al., 2020) and relevant position encoding (Shaw, et al., 2018), zero-padding and non-overlap convolutions allow the model to construct local spatial relationships without the need for position embedding. It reduces the feature map size with increased feature dimensions after each stage of patch embedding. Removing position encoding contributes to its efficiency and fewer FLOPs.

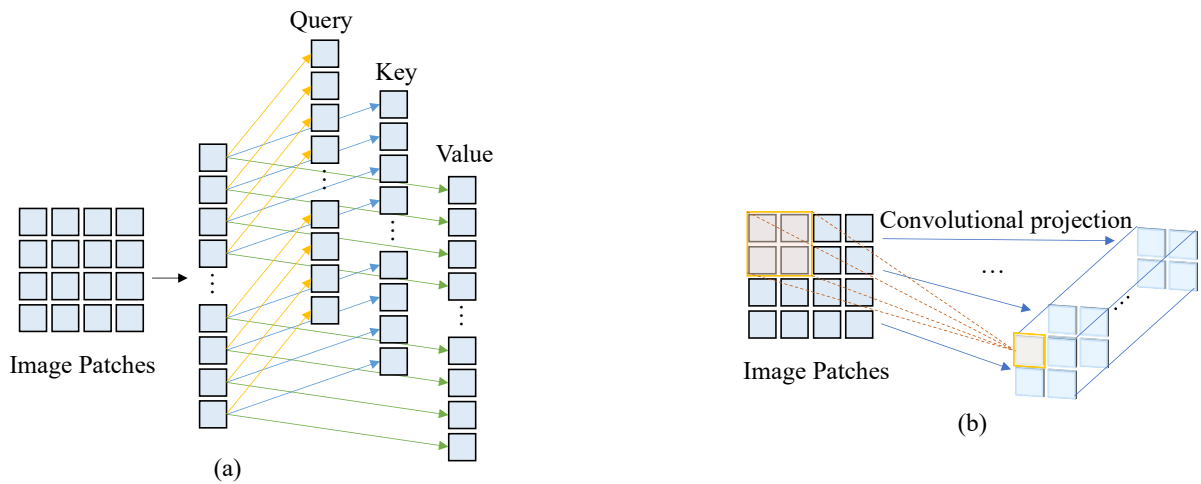


Figure 2.2 Different patch embedding methods: (a) linear projection in Transformers (Dosovitskiy, et al., 2020); and (b) convolutional projection (Wu, et al., 2021).

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

2.2.2 Pooling blocks

Since the traditional self-attention mechanism in Transformers concatenates Query, Key, and Value matrices and conducts extensive scaled dot-product operations, the computational complexity is relatively high. The average pooling design (Yu, et al., 2022) is leveraged here to construct a more straightforward model by replacing the traditional self-attention modules without extra learnable parameters and performance drop. Figure 2.3 illustrates the overall structure of the pooling block design.

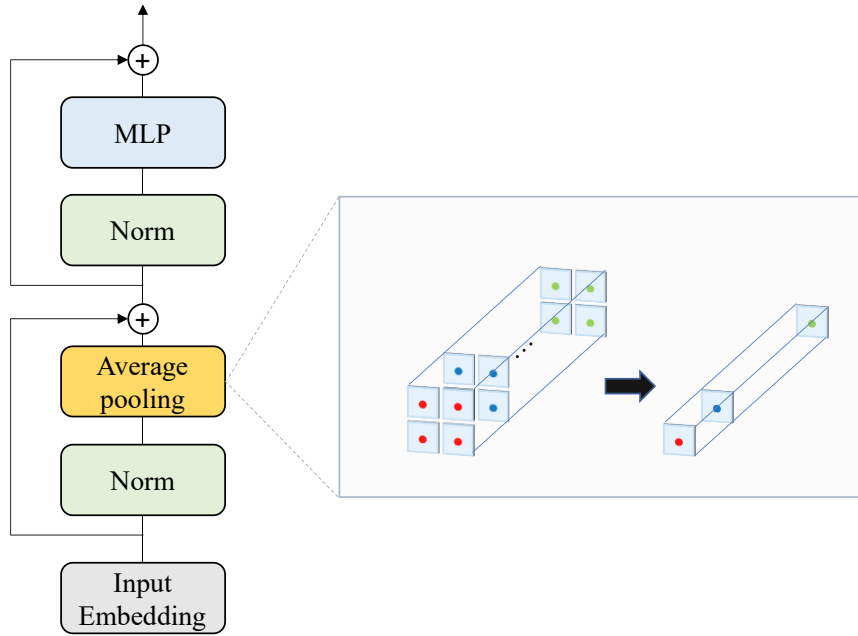


Figure 2.3 Architecture of the pooling block.

First, the image input tensor I is processed by convolutional patch embedding to output the embedded token X . This can be formulated as:

$$X = \text{PatchEmb}(I) . \quad (1)$$

Then, embedded tokens X are fed to the pooling block containing two residual sub-blocks. In the first sub-block, embedded tokens X are normalised through Group Normalisation (Wu and He, 2018), followed by the average pooling operation and residual addition. The output Y of the first sub- block is calculated by:

$$Y = \text{AvePooling}(\text{Norm}(X)) + X . \quad (2)$$

In the second sub-block, the output Y of the first sub-block is first normalised and then processed by a multilayer perceptron (MLP) (Gardner and Dorling, 1998) with the GELU (Hendrycks and Gimpel, 2016) non-linear activation function. The final output Z of the pooling block can be obtained by:

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

$$\mathbf{Z} = \text{GELU}(\text{Norm}(\mathbf{Y})\mathbf{W}_1)\mathbf{W}_2 + \mathbf{Y}, \quad (3)$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable weights of the MLP.

2.2.3 Feature fusion misalignment

As for the encoder-decoder design of the semantic segmentation models, the encoder employs down-sampling operations to obtain higher-level features with smaller pixel maps in each stage, and then these feature maps are fused in the corresponding up-sampling stage of the decoder. As shown in Figure 2.4, after several down-sampling operations, there are inevitable changes in the spatial location of the feature map pixels. As a result, spatial misalignment happens between the down-sampled feature maps and the corresponding up-sampled feature maps, which causes improper information representation. In this work, feature selection and alignment methods (Dai, et al., 2017; Hu, et al., 2018; Huang, et al., 2021) are applied to tackle this problem, which is explained in the subsequent sections.

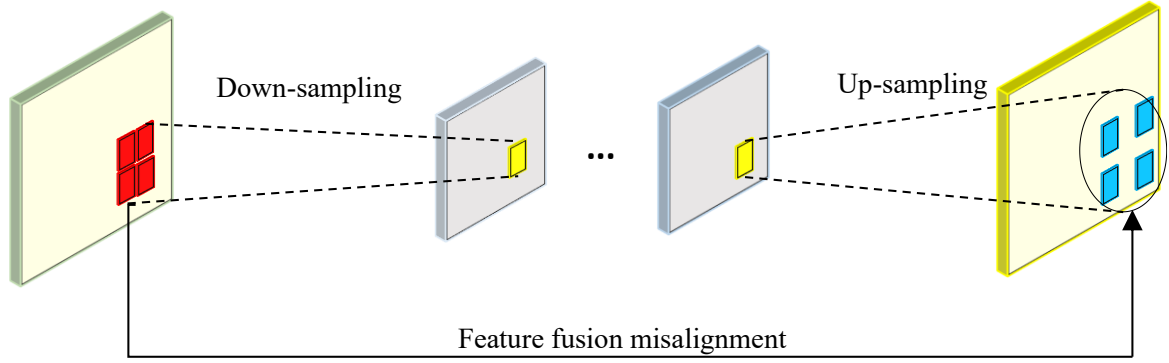


Figure 2.4 Schematic of the disparity in the positions of feature pixels.

2.2.4 Feature selection modules

Instead of applying a simple 1×1 convolutional layer to extract features from the encoder, a feature selection module (Huang, et al., 2021) is employed to reshape the feature maps according to the importance of features. The overview of the FSM is illustrated in Figure 2.5.

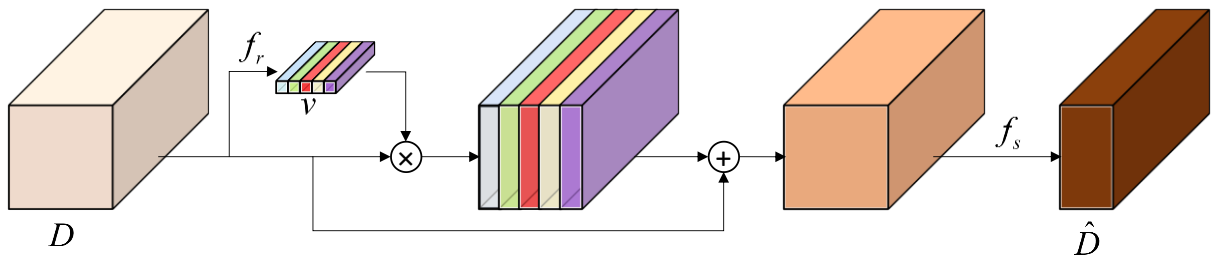


Figure 2.5 The feature selection module.

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

First, the feature map D from the encoder is followed by a feature importance representation layer f_r (a 1×1 convolutional layer with a sigmoid activation function) and outputs an importance vector $\mathbf{v}$. Next, the original feature map D is multiplied by the importance vector $\mathbf{v}$, and then added to the original feature map to obtain a reconstructed feature map. Finally, a feature selection layer f_s (a 1×1 convolutional layer) removes worthless features for channel shrinkage (Hu, et al., 2018). The procedure of FSM is formulated as follows:

$$\hat{D} = f_s(D + f_r(D) * D). \quad (4)$$

2.2.5 Feature alignment modules

As illustrated by Figure 2.6, the feature alignment module mainly applies the feature maps from FSM and up-sampling of the decoder to modify the spatial misalignment.

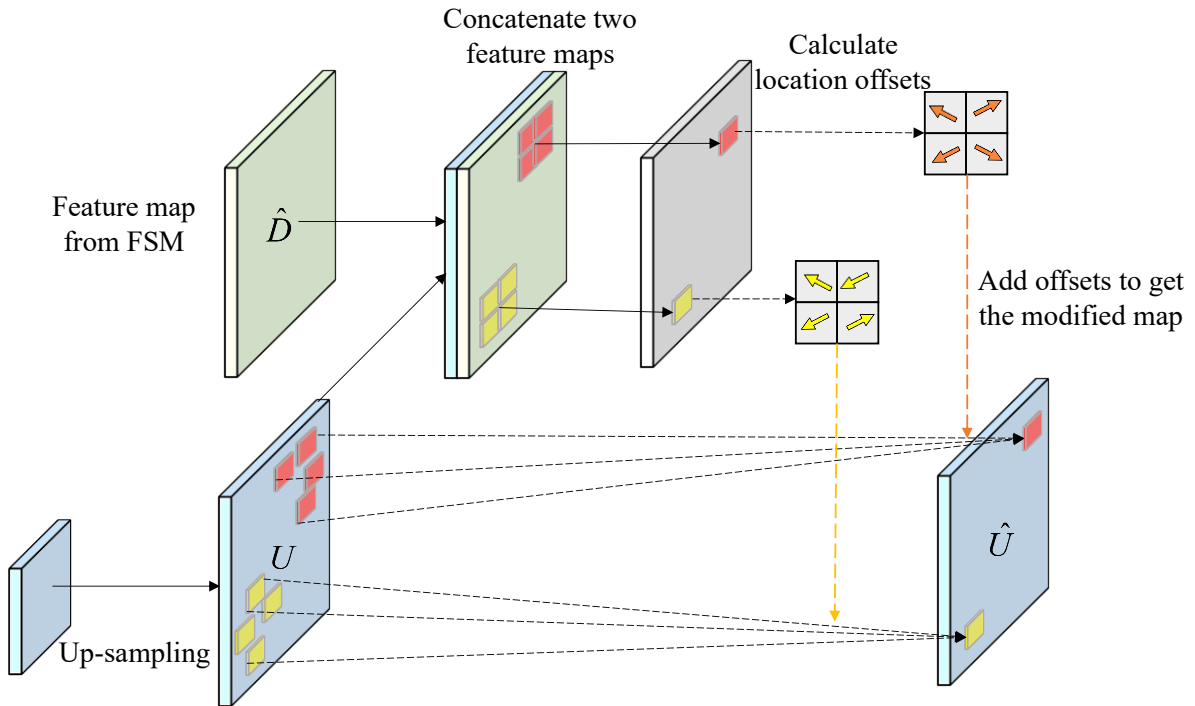


Figure 2.6 The feature alignment module.

To begin with, the feature map $\hat{D}$ from the FSM is concatenated to the feature map U from up-sampling. Afterwards, a deformable convolution (Dai, et al., 2017) is used as the feature alignment function to learn the offsets from the spatial differences. Lastly, the learned offsets are aligned to the up-sampling feature map U to obtain the modified feature map $\hat{U}$. The revised feature map is then added to the feature map $\hat{D}$ from the FSM as the input for the next up-sampling.

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

2.3 Study design

In this section, advanced crack segmentation models are selected for the performance comparison, followed by introductions for the evaluation metrics, image datasets, model training configuration, and assessment framework.

2.3.1 Model selection

To leverage the achievements of prior research and to better assess the capabilities of the designed model, a comparison is performed with the following representative large Transformer-based architectures: SegFormer-B5 with 110 million parameters (Xie, et al., 2021) and Crack Transformer with 50 million parameters (Guo, et al., 2023). These models have shown good performance in crack segmentation and can provide a benchmark for evaluation. Additionally, UNet with EfficientNet (Tan and Le, 2019) backbone (UNet-EB7) is selected as a robust CNN-based model due to its advanced performance in many image segmentation tasks (Shamsabadi, et al., 2022).

2.3.2 Evaluation metrics

The objective of a crack segmentation model is to separate the crack regions and the background, i.e., pixel-wise image classification. Herein, four common evaluation metrics, including Precision, Recall, mean Dice score (mDS), and mean intersection over union (mIoU) score, are used to comprehensively assess the performance of the different models. The definition of Precision and Recall are formulated as follows:

$$Precision = \frac{T_p}{T_p + F_p}, \quad (5)$$

$$Recall = \frac{T_p}{T_p + F_N}, \quad (6)$$

where T_p denotes the number of true positive pixels, F_p represents the false positives, and F_N shows the false negatives.

Although Precision and Recall are widely employed for segmentation performance evaluation, they face challenges concerning threshold dependence, class imbalance, boundary bias, and the trade-off between completeness and correctness (Euzenat, 2007). Different threshold values can lead to significantly different Precision and Recall scores due to the changes between false negatives and positives. In contrast, IoU and DS metrics evaluate the spatial overlap between

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

the predicted and ground truth regions, making them insensitive to the specific threshold used to binarise predictions (Terven, et al., 2023). Moreover, Precision and Recall do not inherently address class imbalance. As a result, a high Precision or Recall can be biased toward the majority class – the background – in crack segmentation (Choi and Cha, 2020; Kang and Cha, 2021), where background pixels significantly outnumber crack pixels. In addition, IoU and DS metrics assess the similarity of regions rather than raw pixel counts, focusing on the overlap of semantic regions. Following this method, both false positives and negatives are taken into measurement equally, providing a balanced evaluation of the model performance. Precision or Recall may not effectively reflect the limitations of models that exhibit boundary bias, making it harder to assess the quality of the predictions near object edges (Zhang, et al., 2021), while IoU and DS metrics evaluate the overlap of the entire region, including the boundary; models that tend to predict regions with boundary bias are penalised, as the overlap near the edges is crucial for a high IoU or DS score. Overall, Precision and Recall might not offer the same level of interpretability as IoU and DS metrics, which overcome these limitations by considering mask overlap, handling class imbalance effectively, and providing a more balanced assessment of segmentation accuracy (Li, et al., 2019). To ensure a comprehensive evaluation, this study also incorporates IoU and DS metrics (Eq. (7) and (8)).

$$DS = \frac{2|P \cap T|}{|P| + |T|}, \quad (7)$$

$$IoU = \frac{|P \cap T|}{|P \cup T|}, \quad (8)$$

where P denotes the predicted pixel map, and T represents the labelled pixel map. Although DS and IoU scores seem similar, they provide different results for model performance. DS tends to measure the average performance of the model, while the IoU score is inclined towards the worst-case performance (Morera, et al., 2020).

2.3.3 Image datasets

Four public datasets were used to comprehensively evaluate the performance of the proposed model on various backgrounds. These datasets have been extensively used in existing literature (Guo, et al., 2023; Jun, et al., 2022; Shamsabadi, et al., 2022; Shamsabadi, et al., 2022) and include cracks on various backgrounds, such as concrete with homogenous background and masonry surfaces with many crack-like features, which may correspond to most real-world challenges that a model is likely to encounter. This enabled the comprehensive assessment of

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

the performance of the proposed model in practical applications where a trained model may face changing environments. Details of the utilised datasets are presented below.

- Crack500 (Zhang, et al., 2016): this crack dataset was created from 500 high-resolution images (3264×2448) captured by cell phone cameras on the main campus of Temple University. This dataset primarily includes crack images of asphalt and concrete pavement surfaces. The developers of the dataset split each original image, and only the split images (640×360) containing crack pixels were retained. The released version of Crack500 comprises 1896 and 1124 images for training and testing. To prepare the dataset in this study, each image and its annotation were randomly cropped into 640×320 pixels in the data pre-processing stage to make the dataset compatible with the varying input requirements of all the studied models.
- GAPS384 (Yang, et al., 2019): this dataset contains 384 asphalt pavement crack images manually selected from the German Asphalt Pavement Distress (GAPs) dataset (Eisenbach, et al., 2017), with the original image resolution of 1920×1080 pixels. The released version of GAPS384 includes 509 640×540 images and their corresponding annotations. Due to the existence of images taken in low light conditions and thin crack semantics, this dataset can present detection challenges for pixel-level crack segmentation. After randomly cropping all images and masks into 448×448 pixels, 90% of the data was assigned for training and validation while the rest 10% for testing.
- Ozgenel (Özgenel, 2019): this dataset includes 458 high-resolution concrete crack images (4032×3024 pixels) with the corresponding annotation maps of crack semantics. The original images were collected from several buildings in the Middle East Technical University. To prepare the dataset for the evaluation purpose in this study, the images were split into patches with the resolution of 512×512 pixels. Two thousand images and their annotations were obtained from this dataset, with 80% and 20% of them for training and test purposes, respectively.
- MC448 (Shamsabadi, et al., 2022): this dataset consists of 3355 masonry crack images (448×448) and their corresponding masks. These images were extracted from 45 copyright-free images, 153 full-resolution photos taken manually, and 15 artificially generated images. This dataset presents significant challenges for crack segmentation, primarily due to the inclusion of non-crack images and the presence of crack-like background textures on the masonry surface. We allocated 90% of images and their masks

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

for the model training and evaluated the trained model with the rest 10% of images. Herein, the input image sizes and the image number of training and testing sets in these datasets are shown in Table 2.1.

Table 2.1 Details of four datasets

Dataset	Input image size	Total number	Train set	Test set
Crack500	640x320	3020	1896	1124
GAPS384	448x448	509	465	44
Ozgenel	512x512	2000	1600	400
MC448	448x448	3351	3000	351

2.3.4 Training configuration

The proposed model was developed in Python using the PyTorch (Paszke, et al., 2019) open-source framework. All the model training and evaluation were conducted on a workstation equipped with an AMD Ryzen 9 5900X 12-Core CPU, an NVIDIA GeForce RTX 3090 Ti GPU, and 32 gigabytes of RAM. To run a fair comparison, all models were trained, validated, and tested within the same framework. The learning rate was set to 5×10^{-5} , a batch size of 2 images was used, and the training process was continued for 200 epochs with AdamW optimiser and the Dice loss function – see Eq. (7). To avoid over-fitting, the training set was first shuffled randomly, and then colour-jittering and horizontal flipping were utilised for random image augmentation. The backbone was pretrained in ImageNet (Deng, et al., 2009) datasets, and the trained weights were used for model initialisation to accelerate training and improve the final model performance.

2.3.5 Assessment framework

Models were trained respectively on four different datasets with the same training settings. As illustrated in Figure 2.7, three types of model evaluations (accuracy, efficiency, and robustness) were conducted in the assessment stage. The ablation study, additional model accuracy comparisons and application in another segmentation task can be found in Appendix 2.A.1, Appendix 2.A.2, and Appendix 2.A.3. In addition to Precision, Recall, mIoU, and mDS metrics which were used to analyse the accuracy of crack pixel classification, three indices, including model parameters, FLOPs, and inference time, were taken into evaluation to assess the model efficiency. Moreover, the noise robustness of the proposed model against five common noises, including gaussian, salt, pepper, gaussian blur, and crack texture, were compared to the

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

representative models. Additionally, the performances of trained models were investigated under luminance changes.

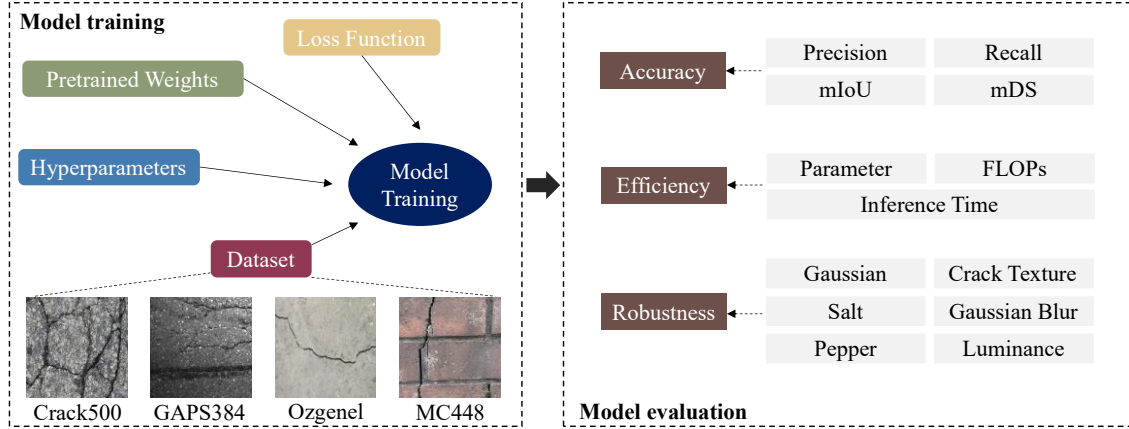


Figure 2.7 The schematic of the assessment framework.

2.4 Results and discussion

In this section, the accuracy of the proposed model is first evaluated and compared to the representative models, followed by an efficiency comparison and robustness study.

2.4.1 Accuracy comparison

Table 2.2 presents the performance of the proposed model on four datasets (representing concrete, asphalt, and masonry cracks) while being compared with the performance of the representative models trained with the same framework. Further performance comparison is also presented in Appendix 2.A.2, where the performance of PoolingCrack is compared with published results.

Overall, the results in Table 2.2 demonstrate the improved performance of the designed PoolingCrack model compared to UNet-EB7, CT, and SegFormer-B5 across all four datasets, as evidenced by its higher mIoU and mDS scores. Occasionally, the other three models can exhibit better Recall and Precision scores. The Recall and Precision can be subject to bias as they solely rely on the number of correctly and incorrectly classified pixels without considering the spatial extent and localisation of the predictions. This limitation may produce misleading evaluations, particularly when dealing with complex tasks like crack segmentation in which background pixels dominate the crack region (Euzenat, 2007). Considering this shortage, mIoU and mDS provide a more reliable measurement, as they account for the spatial overlap and localisation of segmentation masks and, therefore, a more comprehensive view of how well the model captures the actual boundaries of the segmented objects (Li, et al., 2019).

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

Table 2.2 Evaluation results of different models in four datasets (bold values are the best).

Datasets	Models	Precision (%)	Recall (%)	mDS (%)	mIoU (%)
Crack500	UNet-EB7	64.8	75.2	66.5	52.7
	CT	64.9	76.0	67.4	53.5
	SegFormer-B5	68.8	71.8	67.9	54.3
	PoolingCrack	67.2	75.7	68.6	54.8
GAPS384	UNet-EB7	59.4	57.5	54.8	40.2
	CT	63.1	58.3	54.6	40.1
	SegFormer-B5	65.5	58.6	57.5	42.1
	PoolingCrack	72.5	55.3	59.2	43.7
Ozgenel	UNet-EB7	91.7	85.6	87.7	79.3
	CT	87.0	87.8	86.7	77.8
	SegFormer-B5	90.1	86.6	87.4	79.0
	PoolingCrack	89.1	89.3	88.6	80.5
MC448	UNet-EB7	73.1	73.7	70.7	59.6
	CT	70.8	74.4	70.4	59.0
	SegFormer-B5	74.9	79.6	76.1	65.6
	PoolingCrack	75.7	79.5	76.5	65.8

UNet-EB7 and CT presented 2.0% - 6.8% lower mDS and mIoU scores than SegFormer-B5 and PoolingCrack in the two more challenging datasets, GAP384 and MC448, with crack images of poor light conditions and crack-like backgrounds. The dissimilarities between these two metric values are less pronounced in the other two datasets, Crack500 and Ozgenel, where PoolingCrack still shows 0.5% - 2.7% higher scores. As for Recall and Precision, PoolingCrack achieves the best Precision scores on the GAP384 and MC448 datasets and the maximum Recall value on the Ozgenel dataset, with 0.8% - 13.1% improvement over other models. But compared with all the topmost indices values, it experienced an incremental decrease of 0.1% - 3.3%. When the predicted profile of cracks becomes thicker, there would be an increase in Recall due to a higher rate of correctly identified true positives (cracks) and fewer false negatives. Yet, this may also lead to a decrease in Precision with the increased false positives.

Figure 2.8 visualises the randomly selected segmentation results from the four experimental datasets. In general, the proposed model presents a good performance in accurately segmenting cracks under different background textures. For the Crack500 and GAPS384 datasets, PoolingCrack could provide more details of the crack profiles compared to the other representative models. Over the Ozgenel dataset, except for the crack map produced by PoolingCrack, some visual image noises exist in the prediction masks, highlighting the possibility of texture bias in the other three models, although they could portray the complete crack semantic. As for the results in the MC448 dataset, the PoolingCrack created crack patterns most consistent with the ground truth.

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

Due to its design, PoolingCrack performs well on the datasets of GAPS384 and MC448 with low lighting conditions or crack-like background textures despite having the fewest model parameters. Unlike the other models compared against, the feature alignment module in PoolingCrack can correct the spatial position of feature pixels after recurrent down-sampling and up-sampling, contributing to an enhanced feature fusion and a better presentation of the crack prediction. Quite different from the max-pooling parts in the other models, PoolingCrack adopts an average pooling design. This design performs as a mean filter (Jiang, 2011). When the crack images have low lighting conditions or crack-like backgrounds, the probability of non-crack (background) pixels misleading crack segmentation models to identify them as crack pixels will enhance (Shamsabadi, et al., 2022). This average pooling design helps smooth random non-crack pixel values in crack images with these backgrounds, improving the robustness of the model.

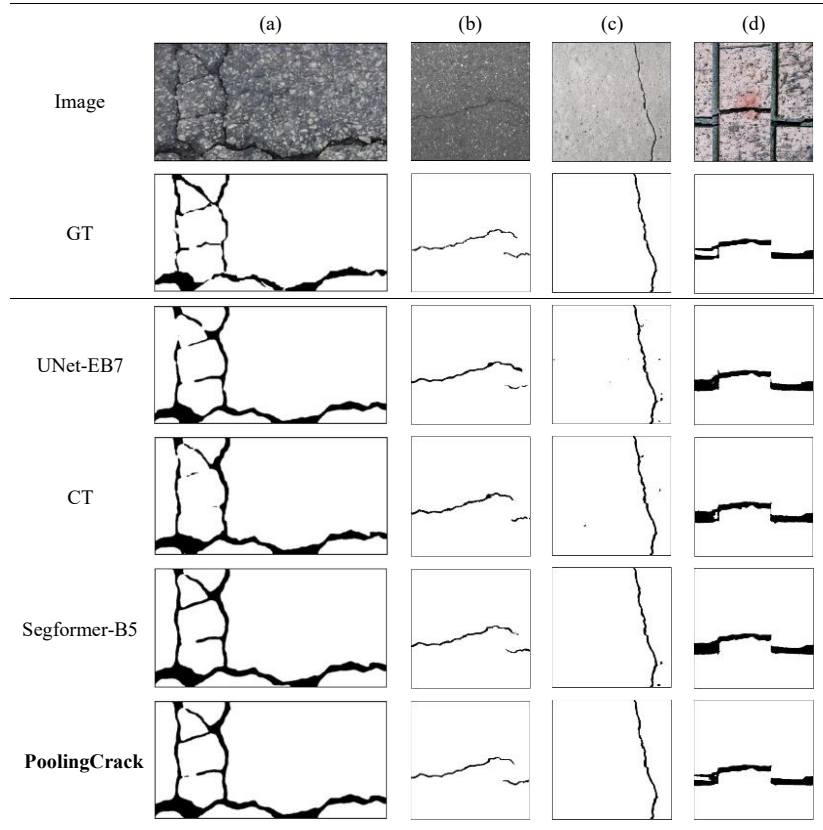


Figure 2.8 Randomly selected segmentation results of different models in four datasets: (a) Crack500, (b) GAPS384, (c) Ozgenel, and (d) MC448.

2.4.2 Efficiency comparison

To evaluate the efficiency of the proposed model and compare it with large representative models, three factors, including the inference time, the total number of parameters, and FLOPs of each model, are presented in Figure 2.9. PoolingCrack consists of 32 million training

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

parameters and executes 36 giga FLOPs for each 512×512 resolution image with 19 ms inference time (for one 512×512 image). Its training parameters are 52.2%, 36.0%, and 62.4%, respectively, fewer than UNet-EB7, CT, and SegFormer-B5. While the inference time of CT is shorter than the PoolingCrack by 6 ms, FLOPs of the PoolingCrack are 10.0%, 28.0%, and 67.3% lower than the other three models. For a stable overall performance, PoolingCrack stacks more blocks in each down-sampling of the decoder, causing a relatively slower inference speed than CT. However, its number of parameters and FLOPs are much lower than the other three models due to the efficient pooling block design, which reduces the dependence on high memory and performance GPUs and could contribute to its successful deployment on edge devices. Further efficiency evaluation and comparison to the representative models for crack detection in the literatures are presented in Appendix 2.A.2, further highlighting the potential suitability of PoolingCrack for practical deployments.

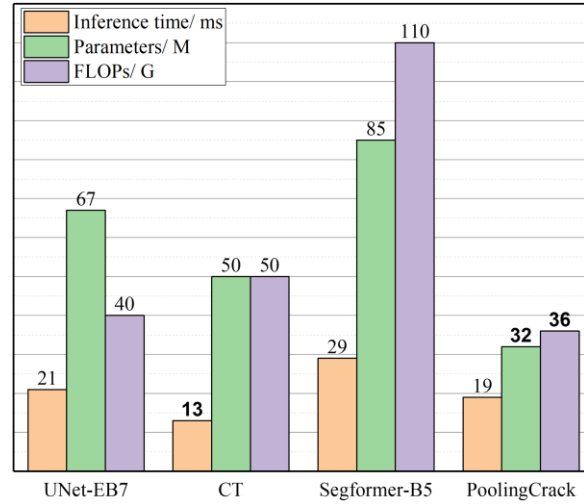


Figure 2.9 Efficiency comparison for processing one 512×512 resolution image.

In the above, PoolingCrack was compared with the similar size representative models. Additionally, unlike some other smaller-scale Transformer-based (Lewis, et al., 2023) and CNN-based (Ali and Cha, 2022; Choi and Cha, 2020; Kang and Cha, 2021) networks that may face the issues of inferior generalisation and anti-noise robustness, PoolingCrack could exhibit robust performance in different datasets, which would be validated in the next section.

2.4.3 Robustness study

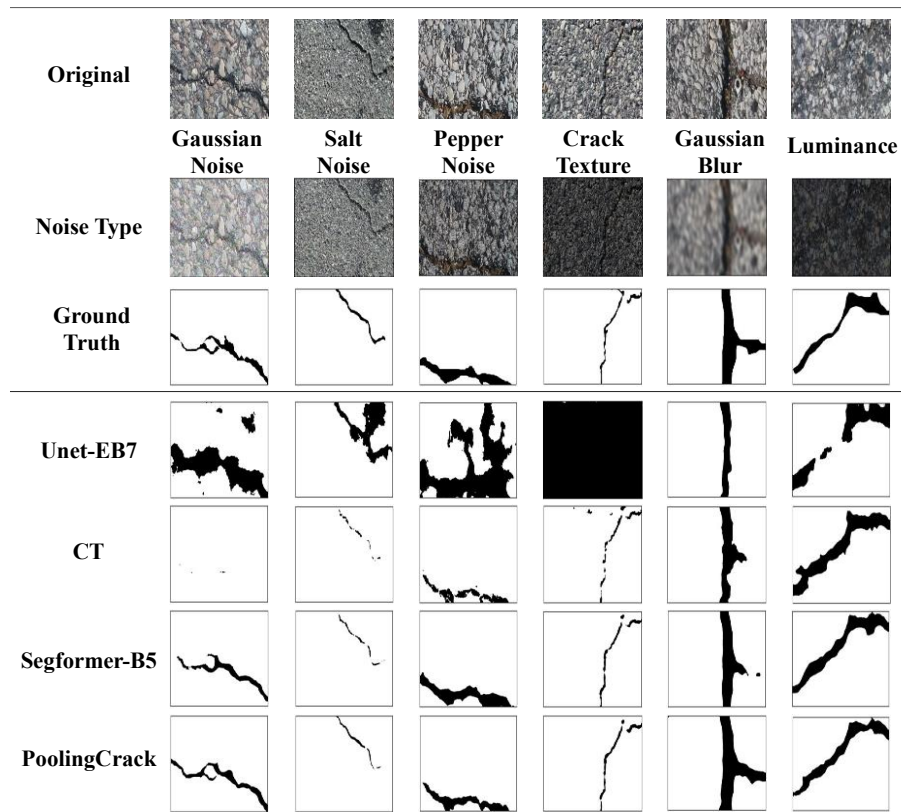
Structural health monitoring is a safety-critical task where maintaining a high level of accuracy in diverse conditions is of utmost importance. In real-world scenarios, the adaptive ability to handle noise disturbances is required for automated defect detection, thus ensuring the efficacy

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

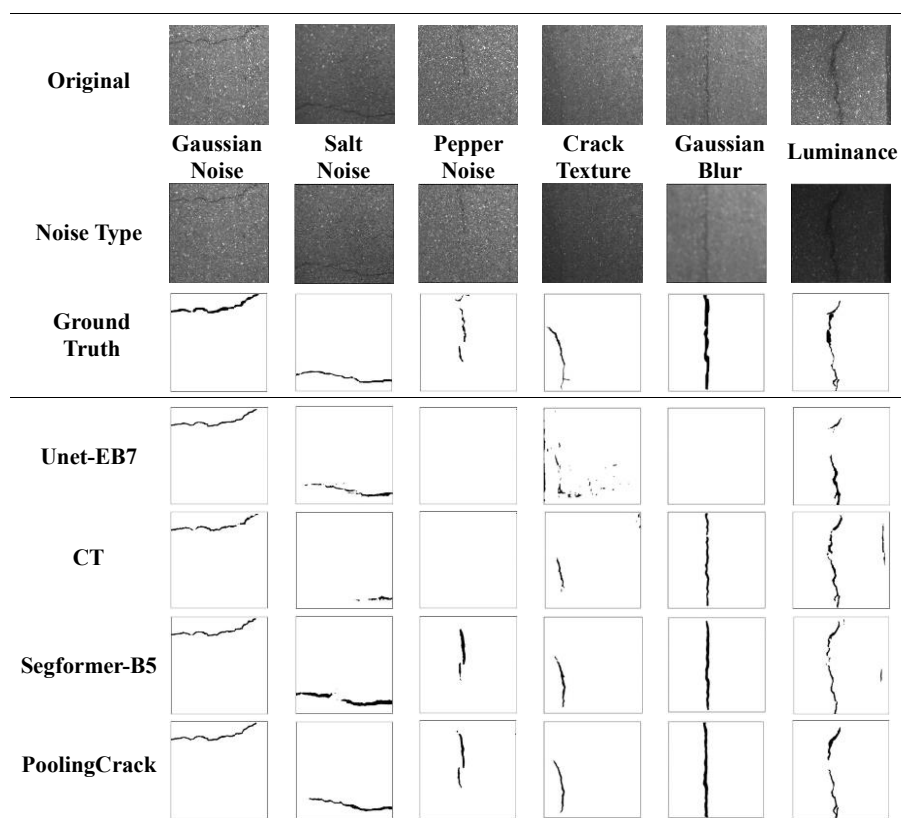
and practicality of the designed framework. Considering this, six image distortions, including Gaussian, salt, pepper, gaussian blur and crack texture noises, as well as luminance changes, were employed to test the performance of the proposed model in challenging and intricate environments. Figure 2.10 manifests the qualitative assessment results, and Figure 2.11 illustrates the quantitative assessment results on distorted images. Overall, in most cases, PoolingCrack demonstrates better robustness compared to the other three models, particularly against crack texture noise, thus highlighting its stability in detecting crack-like objects and true crack pixels. The observed performance of PoolingCrack can be linked to the FAM design. Transformers have shown better noise robustness and generalisation than CNNs because of their capacity to capture global information and reduced biases towards local clues (Tuli, et al., 2021); yet, the diverse designs in Transformers can also lead to robustness variations. Noise primarily distorts images by introducing random values to the original pixel values. The average pooling design in PoolingCrack helps to smooth out noise and reduce the impact of random variations in pixel values. Additionally, in the recurrent down-sampling and up-sampling stages, the locations of crack features are significantly disordered by the adjacent noise-caused mutant pixel values. Thus, without a specific average pooling design and feature location alignment, SegFormer-B5 and CT are less robust than PoolingCrack, although they have more parameters.

As illustrated in Figure 2.10 and Figure 2.11, overall, PoolingCrack shows robustness against image noises despite having fewer parameters than other studied models. Occasionally, however, other models can be found to have relatively better robustness, although without consistency. For example, when facing Gaussian and pepper noises in the GAP384 and MC448 datasets, SegFormer-B5 shows higher accuracy in crack map recovery as noise intensity increases. Visualised results in Figure 2.10-b and d indicate that the higher mDS of SegFormer-B5 in these cases result from a larger area of predicted crack pixels, which enhances the intersection with ground truth pixels. Nevertheless, more discontinuity and false positives can be discovered. As an example of the lack of consistency of other models, CT shows the best robustness against salt noise over concrete surface images (Figure 2.11-c, Ozgenel dataset) when the intensity of noise is more than $\sim 30\%$, while almost in all other cases, PoolingCrack could perform better.

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

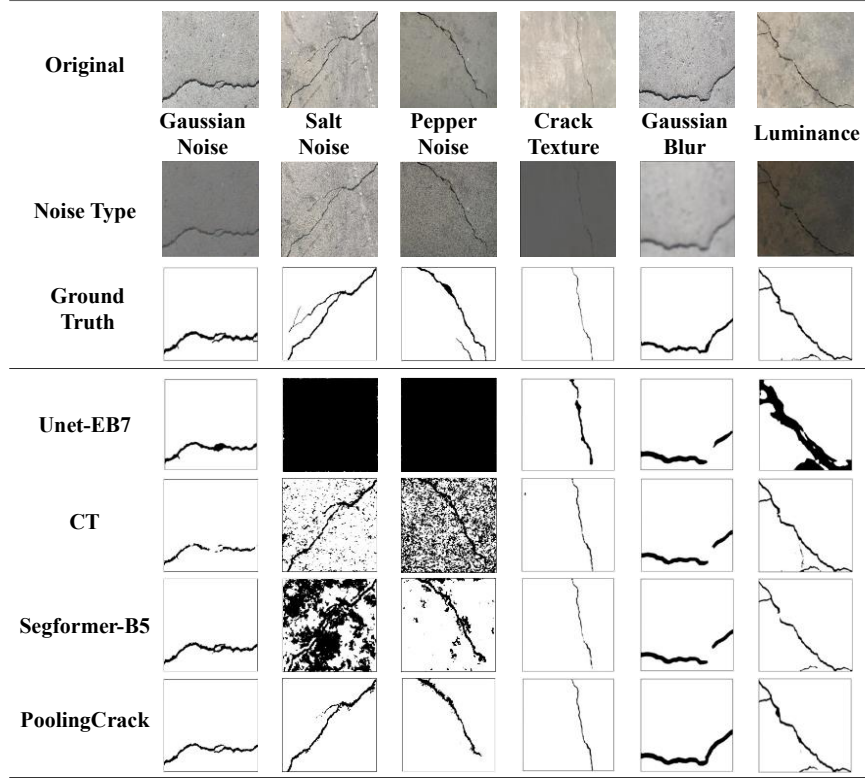


(a) Crack500 (concrete & asphalt)

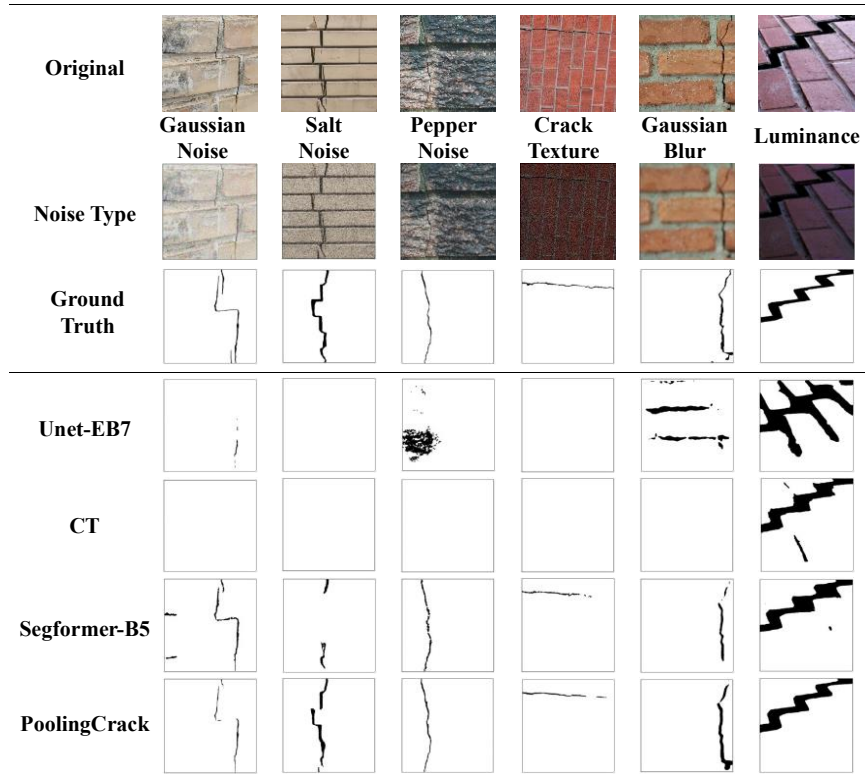


(b) GAPS384 (asphalt)

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection



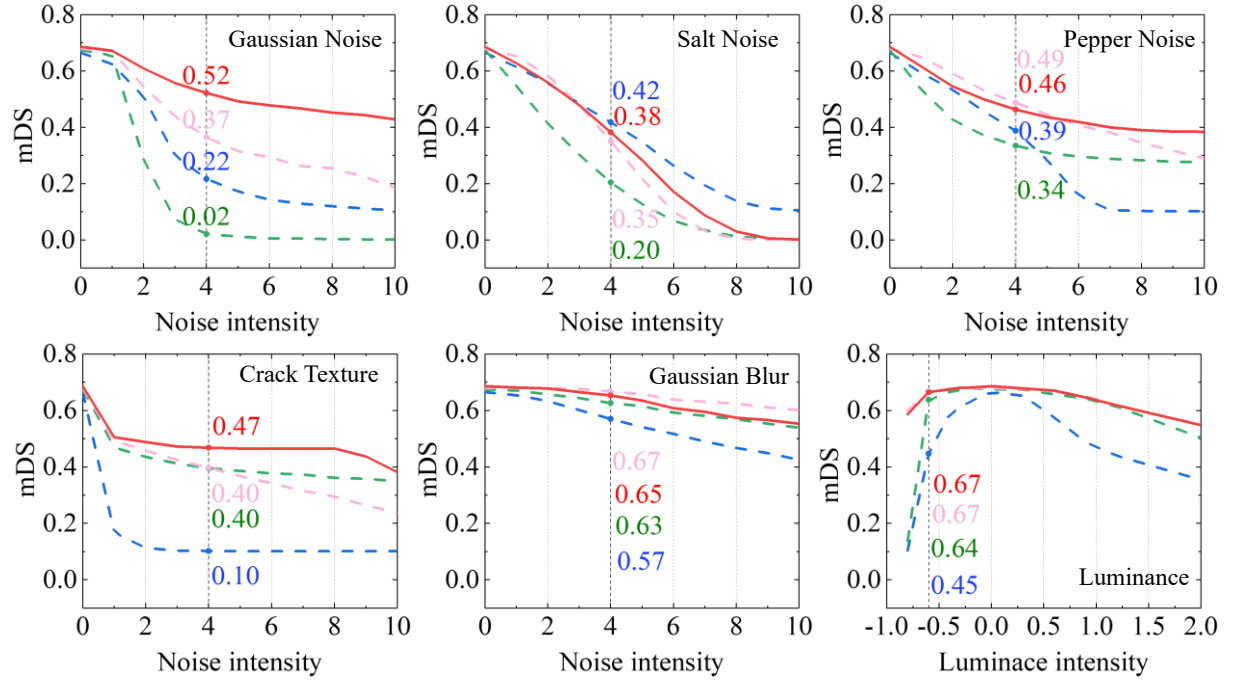
(c) Ozgenel (concrete)



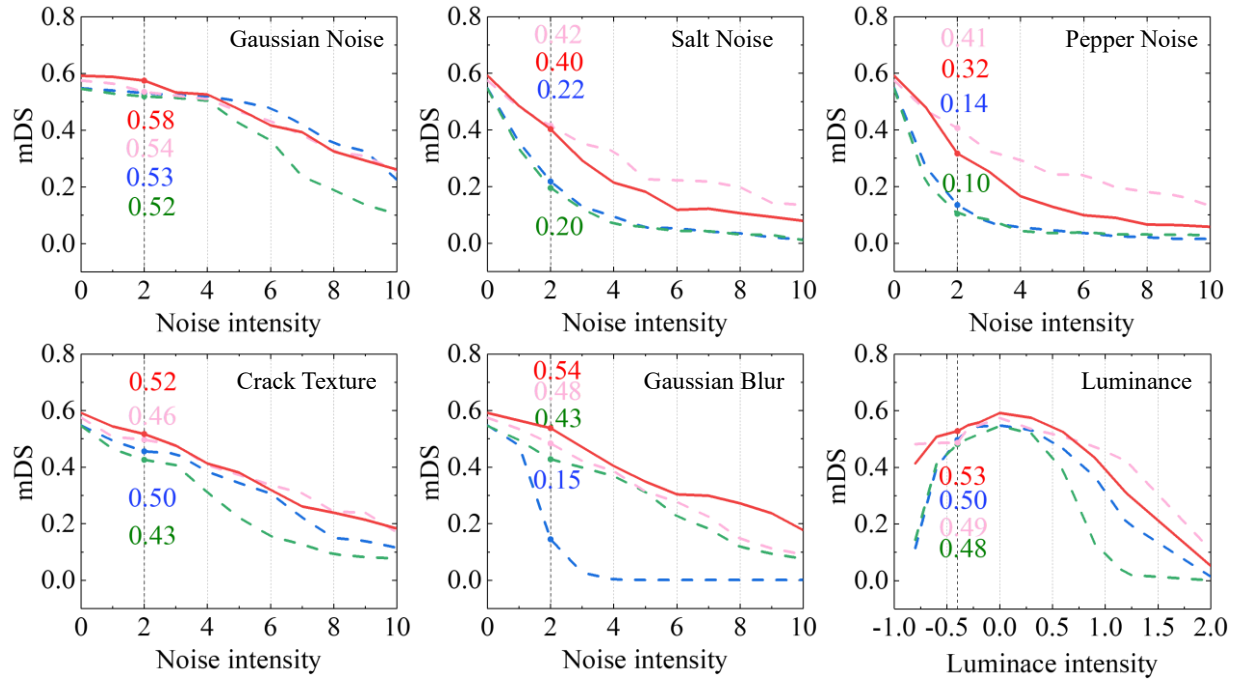
(d) MC448 (masonry)

Figure 2.10 Qualitative assessment of noise effects in four datasets: (a) Crack500; (b) GAPS384; (c) Ozgenel; and (d) MC448. The noise intensities are highlighted in the reference line of Fig.11 (i.e., Fig. 10-a illustrates the result in the Crack500 dataset with noise intensity 4 and luminance intensity -0.6, referred to as Fig. 11-a).

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

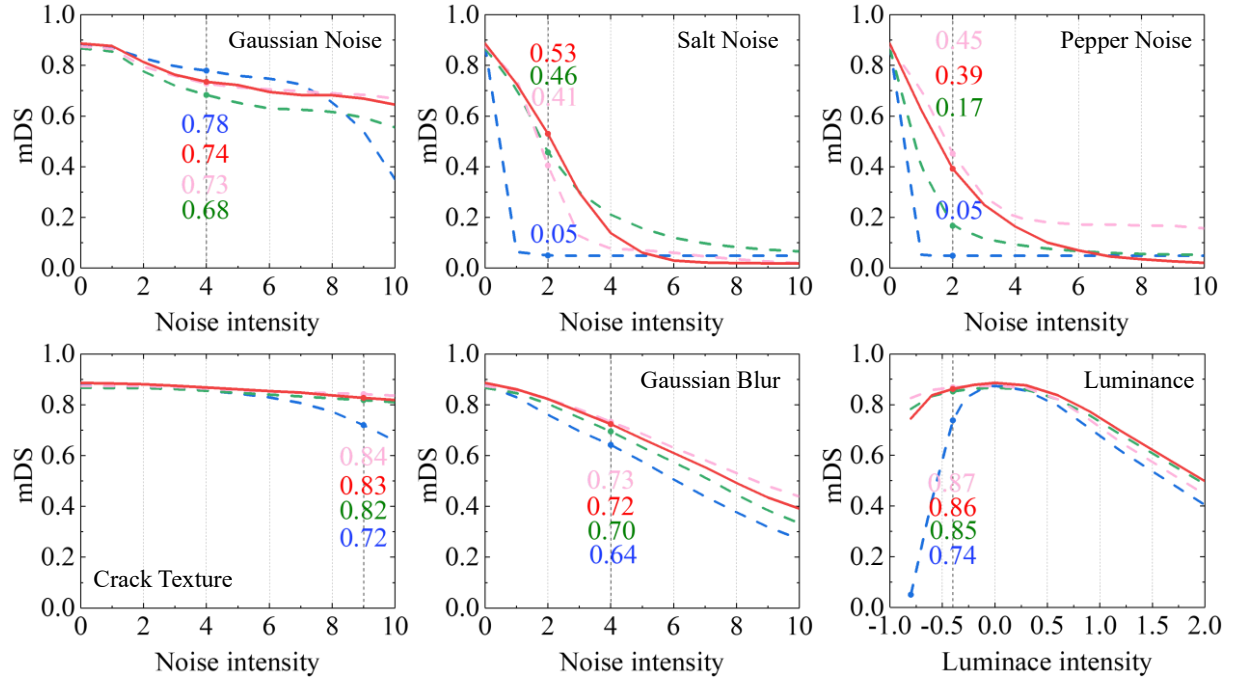


(a) Crack500 (concrete & asphalt)

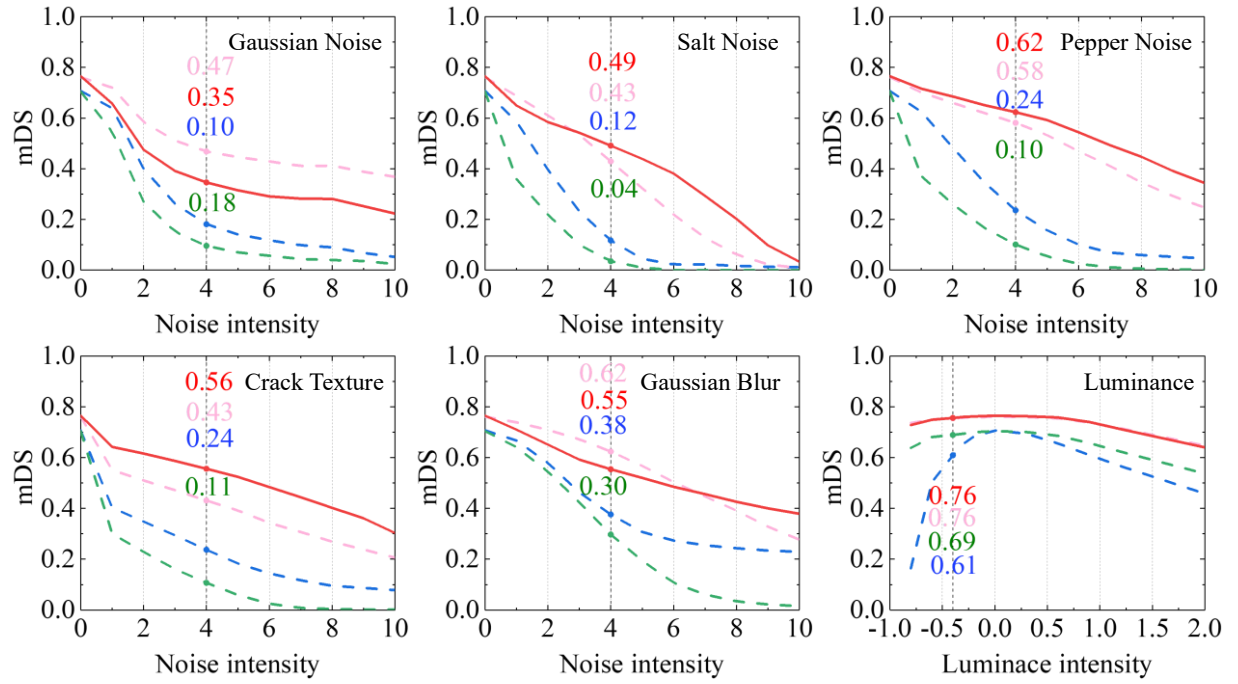


(b) GAPS384 (asphalt)

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection



(c) Ozgenel (concrete)



(d) MC448 (masonry)

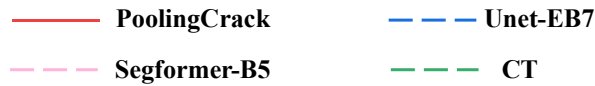


Figure 2.11 Quantitative assessment of noise effects in four datasets: (a) Crack500; (b) GAPS384; (c) Ozgenel; (d) MC448; The noise intensity (0-10) is defined according to the noise-distributed probability (0%-100%) in original images, and the luminance intensity is defined by varying brightness (i.e., -0.5 represents 50% brightness reduction).

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

2.5 Conclusion

Given the importance of developing a model with high accuracy, efficiency, and robustness for crack detection, a novel Transformer-based network named PoolingCrack is proposed in this study. This model leverages a hierarchical architecture to seize multi-scale feature information. It deploys a convolutional patch embedding to remove position encoding, an average pooling design for structural efficiency, and a spatial location correction module for exact feature fusion. Our main hypothesis is that this designed model could show robust crack segmentation capabilities under various challenging background textures and noise interference despite the fewer parameters.

To demonstrate the performance of the proposed model, a comprehensive assessment was implemented on four datasets with asphalt, concrete, and masonry backgrounds. The validation results on four datasets indicated that PoolingCrack could achieve 0.4% to 6.8% higher mDS and mIoU than UNet-EB7, CT, and SegFormer-B5, while it deployed only 33 million training parameters, 36% to 62% fewer than the other three representative models. It is worth mentioning that the enhanced performance of PoolingCrack was also confirmed by comparing its performance with published results in the literature, both in terms of accuracy and efficiency. Furthermore, five noise and luminance variations with different intensities were applied to investigate noise adaptability. Qualitative and quantitative noise robustness analysis showed the reliable performance of the PoolingCrack under various scenarios, with up to 52% higher mDS compared to the other studied models.

Since the designed framework was validated to achieve improved robustness, accuracy and efficiency, requiring only known subcomponents, it can be considered a new type of network structure that opens new possibilities for progressing crack segmentation model design.

Appendix 2.A

2.A.1 Ablation study results

In machine learning systems, it is typical to have multiple learning modules working together. An ablation study is one approach to comprehending the impact of each module. We systematically removed or changed one of the four primary modules (convolutional patch embedding dimensions, pooling blocks, FSM, FAM) in the proposed framework to investigate its contribution to the performance of the whole model.

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

This section implements four experiments on the GAPS384 dataset to examine the impacts of four modules (convolutional patch embedding, pooling blocks, FSM, and FAM) within the proposed model. We conducted the specific studies by reducing the pooling blocks, increasing the patch embedding dimensions, removing FAM, and deactivating FSM, respectively. The relative **improvements/decays** compared to the benchmark are shown in Table 2.3.

Due to the low light conditions of crack images in the GAPS384 dataset, significantly raising the crack segmentation performance is difficult. It is noticeable that the FSM slightly influences the holistic performance of the proposed model. Compared with pooling blocks and patch embedding dimensions, FAM improves similar performance with only a 0.9 million parameter increase. It would be inefficient to strengthen the proposed model by increasing the embedding dimensions due to an almost 80% parameter boost with around 1.0% accuracy improvement. Reducing the pooling block number brings a few parameters decrease but with a 0.8% precision drop. Overall, the number of pooling blocks, patch embedding dimension and FAM modules are the critical parts of the proposed model.

Table 2.3 Test results of models with different components

	Encoder blocks	PatchEmb dim	Decoder module	Parameters(M)	mDS(%)	mIoU(%)
Benchmark	[6,6,18,6]	[64,128,320,512]	FSM+FAM	31.95	59.2	43.7
Encoder	[4,4,12,4]	[64,128,320,512]	FSM+FAM	22.48(-9.47)	58.4(-0.8)	42.9(-0.8)
	[6,6,18,6]	[96,192,384,768]	FSM+FAM	57.13(+25.18)	60.2(+1.0)	44.6(+0.9)
Decoder	[6,6,18,6]	[64,128,320,512]	FSM	31.05(-0.90)	58.3(-0.9)	43.3(-0.4)
	[6,6,18,6]	[64,128,320,512]	FAM	31.83(-0.12)	59.1(-0.1)	43.6(-0.1)

Another ablation study was conducted to further explore the model performance contributions of the average pooling design in pooling blocks and FAM modules. As shown in Table 2.4, when changing average pooling to max-pooling or deactivating FAM modules, the performances of ablated models drop significantly. The performance improvement of Average Pooling over Max Pooling is consistently observed across multiple evaluation metrics (and datasets, if applicable), indicating stable effectiveness rather than random variation. Thus, compared with other modules in the designed model, these two parts primarily maintain the model performance in different datasets.

Table 2.4 Test results of ablated models

Datasets	Ablated Models	mDS (%)	mIoU (%)
	PoolingCrack (Original)	68.6	54.8

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

Crack500	PoolingCrack (max-pooling)	67.7 (-0.9)	54.0 (-0.8)
	PoolingCrack (max-pooling + no FAM)	67.2 (-1.4)	53.4 (-1.4)
GAPS384	PoolingCrack (Original)	59.2	43.7
	PoolingCrack (max-pooling)	56.6 (-2.6)	41.8 (-1.9)
	PoolingCrack (max-pooling + no FAM)	56.1 (-3.1)	41.1 (-2.6)
Ozgenel	PoolingCrack (Original)	88.6	80.5
	PoolingCrack (max-pooling)	87.0 (-1.6)	78.6 (-1.9)
	PoolingCrack (max-pooling + no FAM)	84.2 (-4.2)	76.1 (-4.4)
MC448	PoolingCrack (Original)	76.5	65.8
	PoolingCrack (max-pooling)	73.0 (-3.5)	61.8 (-4.0)
	PoolingCrack (max-pooling + no FAM)	72.7 (-3.8)	61.7 (-4.1)

2.A.2 Accuracy comparison with models in other publications

In addition to the validation presented in the paper, where all models were trained with the same framework (under the same training and evaluation strategy such as image preprocessing methods, hyperparameters, and identical library versions), another validation is shown in this Appendix. This adopts the same training settings in the available publications. Since other authors employed different training strategies and some training details could not be found, the already known training settings were utilised and other settings were kept identical. Then, the model comparisons are conducted one by one.

CT (2023) (Guo, et al., 2023): This model has been compared with PoolingCrack on the public Crack500 dataset, described in Section 2.4. The published training and evaluation settings (Guo, et al., 2023) were recreated, accordingly resizing the images of the Crack500 dataset to 320×640 pixels with batch size 2 and the accuracy of models was evaluated by considering crack and background as 2 classes with the tolerance margin as 2 pixels. The items not provided in the publication were set to “-”. The comparisons with published results [56] in Crack500 dataset are shown in Table 2.5 and Table 2.6. Overall, PoolingCrack shows better accuracy than other models, with fewer parameters, FLOPs and slightly higher inference time than CT (2023).

Table 2.5 Accuracy comparison with published results on Crack500 dataset

Networks	mF1 (%)	mIoU (%)	mPrecision (%)	mRecall (%)
PoolingCrack	90.50	84.85	89.00	95.45

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

CT (2023)	88.73	-	87.45	90.12
DeepLabv3	83.19	-	81.39	85.26
FCN-Unet	79.04	-	75.99	83.12
FCN	83.07	-	80.15	86.74
PSPNet	80.99	-	81.00	80.98

Table 2.6 Efficiency comparison on Crack500 dataset

Networks	Parameters(M)	FLOPs(G)	inference time(s)
PoolingCrack	32	28	0.015
CT (2023)	50	38	0.011

ACAU-Net (2022) (Jun, et al., 2022): This publication provided a comparison of the performance of the models in the public GAP384 dataset. The same training settings as in the publication were adopted: learning rate 0.0008, Adam optimiser, one class crack image labels, cross-entropy loss, 470/509 images as training set and 39/509 images as test set, image size of 512×512 pixels, and 80 training epochs. As illustrated in Table 2.7, PoolingCrack is compared with the presented results of the GAPS384 dataset. Overall, PoolingCrack has shown higher accuracy.

Table 2.7 Accuracy comparison with published results on GAPS384 dataset

Networks	mF1 (%)	mIoU(%)	mPrecision(%)	mRecall(%)
PoolingCrack	54.20	38.80	62.10	55.30
ACAU-Net (2022)	48.24	33.86	68.22	41.72
CPFNet	43.46	30.05	66.12	36.56
ACNet	38.15	26.77	73.97	30.90
CENet	39.41	26.85	69.42	31.69

TransUNet (Shamsabadi, et al., 2022): The same training settings as the previous publication (Shamsabadi, et al., 2022) were adopted. 2,200 images and their corresponding masks of 448×448 pixels were extracted from Ozgenel dataset, of which 85% and 15% were respectively used for training and testing with SGD optimiser (initial learning rate: 0.01–0.02; momentum: 0.9; damping: 0; learning rate scheduler: StepLR with step-size: 2–5, and gamma: 0.9). The comparisons with published results are provided in Table 2.8. As shown in Table 2.8, PoolingCrack has the same accuracy with TransUNet, but with less inference time, parameters and FLOPs.

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

Table 2.8 Performance comparison with published results on Ozgenel dataset

	mF1 (%)	mIoU(%)	inference time(s)	Parameters(M)	FLOPs(G)
PoolingCrack	88	82	0.015	32	27
TransUNet	88	82	0.055	105	126
DeepLabv3+	87	80	0.028	-	-
UNet	86	79	0.025	-	-

DTrC-Net (2023) (Xiang, et al., 2023): The authors of this publication conducted the model comparison in the open-source dataset Crack3238 (Xiang, et al., 2023). The Segmentation Models Pytorch metric (Iakubovskii, 2019) was adopted here with the same training strategies mentioned in the publication: 0.75Dice+0.25BCE loss function, image size 256 × 256 pixels, batch size 16, 200 training epochs, SGD optimiser, momentum value 0.9, poly scheduler control factor 3, and initial learning rate 0.1. Table 2.9 and Table 2.10 present the performance comparison with results in the publication. As illustrated in Table 2.9 and Table 2.10, PoolingCrack has similar accuracy with the state-of-the-art DTrC-Net (2023), but with fewer parameters, FLOPs and higher FPS.

Table 2.9 Accuracy comparison with published results on Crack3238 dataset

Networks	mF1 (%)	mIoU(%)	mPrecision(%)	mRecall(%)
PoolingCrack	76.50	63.40	73.80	79.50
DTrC-Net (2023)	76.44	64.30	75.60	78.86
Unet-RestNet34	71.21	58.78	72.32	73.14
Att-Unet	72.78	60.39	75.55	72.60
DeepLabv3+	71.55	58.37	67.83	78.74
CENet	73.32	60.61	74.75	74.70
TransFuse	71.12	58.26	73.18	72.56
UTNet	73.68	61.49	74.19	75.61
FAT-Net	73.47	61.31	75.07	74.33
DeepCrack	71.38	58.70	71.97	73.72
DcsNet	73.07	60.73	73.97	74.88

Table 2.10 Efficiency comparison on Crack3238 dataset

Networks	Parameters(M)	FLOPs(G)	FPS
PoolingCrack	32	9	125
DTrC-Net (2023)	42	41	80

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

Previous baseline models: Attention U-net (Oktay, et al., 2018), CrackSegNet (Ren, et al., 2020), Deeplab V3+ (Chen, et al., 2017), FPHBN (Yang, et al., 2019), and Unet++ (Zhou, et al., 2018) were also chosen for the performance comparison. Herein, the same training strategies in this paper were utilised for these networks. UNet++ and Deeplab V3+ are equipped with the ResNet101 backbone to have model sizes similar to PoolingCrack. As for the evaluation metrics of all the models below, we only consider the cracks as one class. Table 2.11 shows the comparison results between PoolingCrack and previous baseline models. PoolingCrack, with parameters and inference time comparable to other models, exhibits the best mF1 and mIoU in all four datasets. Although FPHBN, CrackSegNet and DeepLabV3+ have fewer parameters or less inference time, these models show inferior generalisation than PoolingCrack, with unstable performance in different datasets.

Overall, compared with other state-of-the-art and baseline similar-scale models in their respective datasets, PoolingCrack exhibits efficiency with a comparable if not higher accuracy, while it shows significant robustness against various types of noise, even compared to the large scale models. Compared to smaller scales models, PoolingCrack presents significantly better accuracy and generalisation with a comparable efficiency.

Table 2.11 Performance comparison with previous state-of-the-art models

Dataset	Model	mF1 (%)	mIoU(%)	mPrecision(%)	mRecall(%)	Parameters(M)	inference time(s)
Crack500	FPHBN	64.4	50.4	64.5	71.1	5.4	0.011
	UNet++	58.5	45.7	63.8	62.2	68.0	0.020
	AttenUNet	61.5	47.9	61.7	71.6	34.9	0.019
	DeepLabV3+	54.0	41.3	62.9	57.7	45.7	0.012
	CrackSegNet	63.6	49.6	62.2	74.4	15.7	0.015
	PoolingCrack	68.6	54.8	67.2	75.7	32.0	0.015
GAPS384	FPHBN	45.1	32.3	56.0	46.6	5.4	0.011
	UNet++	49.7	35.9	57.1	52.8	68.0	0.020
	AttenUNet	48.0	34.0	57.2	49.9	34.9	0.019
	DeepLabV3+	42.5	29.8	53.0	44.4	45.7	0.012
	CrackSegNet	52.0	37.7	62.3	52.8	15.7	0.015
	PoolingCrack	59.2	43.7	72.5	55.3	32.0	0.015
Ozgenel	FPHBN	87.9	79.5	90.6	86.8	5.4	0.013
	UNet++	87.8	79.4	89.7	87.5	68.0	0.025
	AttenUNet	87.6	79.2	91.2	85.7	34.9	0.024
	DeepLabV3+	86.1	77.2	88.9	85.2	45.7	0.014

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

	CrackSegNet	87.8	79.4	89.4	87.7	15.7	0.019
	PoolingCrack	88.6	80.5	89.1	89.3	32.0	0.019
MC448	FPHBN	60.0	48.7	72.1	57.9	5.4	0.011
	UNet++	62.5	51.2	66.2	67.9	68.0	0.020
	AttenUNet	60.4	49.3	69.6	60.4	34.9	0.019
	DeepLabV3+	60.7	49.2	65.1	64.3	45.7	0.012
	CrackSegNet	62.5	51.5	72.0	61.3	15.7	0.015
	PoolingCrack	76.5	65.8	75.7	79.5	32.0	0.015

2.A.3 Application in colonoscopy segmentation

Since PoolingCrack has shown considerable performance for crack segmentation, it is here applied to another task, i.e. colonoscopy segmentation. Herein, one of the representative models PSNet (Lewis, et al., 2023) is selected for comparison in the two datasets Kvasir-SEG and ClinicDB, and the results are show in Table 2.13.

Table 2.12 Details of the two datasets

Dataset	input image Size	Total	Train	Valid	Test
Kvasir-SEG	512 x 512	1000	900	90	100
ClinicDB	513 x 512	612	550	55	62

Table 2.13 Performance comparison with previous representative models

Model	Kvasir-SEG		ClinicDB	
	mF1 (%)	mIoU(%)	mF1 (%)	mIoU(%)
Unet++	82.1	74.3	79.4	72.9
ViT	92.4	87.4	90.9	86.4
TransFuse	92.0	87.0	94.2	89.7
PSNet	92.9	87.9	92.8	87.9
PoolingCrack	95.0	91.7	97.7	94.5

Although the semantics features in the colonoscopy image segmentation can be different from cracks, PoolingCrack shows a good precision among the above models.

References

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

Ai D., G. Jiang, S.-K. Lam, P. He, C. Li, Automatic pixel-wise detection of evolving cracks on rock surface in video data, *Automation in Construction* 119 (2020), pp. 103378.

Ai D., G. Jiang, S.-K. Lam, P. He, C. Li, Computer vision framework for crack detection of civil infrastructure—A review, *Engineering Applications of Artificial Intelligence* 117 (2023), pp. 105478.

Ali R., J.H. Chuah, M.S.A. Talip, N. Mokhtar, M.A. Shoaib, Automatic pixel-level crack segmentation in images using fully convolutional neural network based on residual blocks and pixel local weights, *Engineering Applications of Artificial Intelligence* 104 (2021), pp. 104391.

Ali R., Y.-J. Cha, Attention-based generative adversarial network with internal damage segmentation using thermography, *Automation in Construction* 141 (2022), pp. 104412.

Ali R., J.H. Chuah, M.S.A. Talip, N. Mokhtar, M.A. Shoaib, Structural crack detection using deep convolutional neural networks, *Automation in Construction* 133 (2022), pp. 103989.

Bang S., S. Park, H. Kim, H. Kim, Encoder–decoder network for pixel-level road crack detection in black-box images, *Computer-Aided Civil and Infrastructure Engineering* 34 (8) (2019), pp. 713-727.

Canny J., A computational approach to edge detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (6) (1986), pp. 679-698.

Cao H., Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, *European Conference on Computer Vision*, Springer, (2022), pp. 205-218.

Cha Y.-J., W. Choi, O. Büyüköztürk, Deep learning-based crack damage detection using convolutional neural networks, *Computer-Aided Civil and Infrastructure Engineering* 32 (5) (2017), pp. 361-378.

Cha Y.J., W. Choi, G. Suh, S. Mahmoudkhani, O. Büyüköztürk, Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types, *Computer-Aided Civil and Infrastructure Engineering* 33 (9) (2018), pp. 731-747.

Chen J., Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: transformers make strong encoders for medical image segmentation, *arXiv preprint arXiv:2102.04306* (2021).

Chen L.-C., G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (4) (2017), pp. 834-848.

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

Chen L.-C., Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, *Proceedings of the European Conference on Computer Vision*, (2018), pp. 801-818.

Choi W., Y.J. Cha, SDDNet: real-time crack segmentation, *IEEE Transactions on Industrial Electronics* 67 (9) (2020), pp. 8016-8025.

Contributors M., MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark, 2020.

Dai J., H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, *Proceedings of the IEEE International Conference on Computer Vision*, (2017), pp. 764-773.

Dais D., I.E. Bal, E. Smyrou, V. Sarhosis, Automatic crack classification and segmentation on masonry surfaces using convolutional neural networks and transfer learning, *Automation in Construction* 125 (2021), pp. 103606.

Deng J., W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), pp. 248-255.

Dosovitskiy A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint* (2020).

Dung C.V., Autonomous concrete crack detection using deep fully convolutional neural network, *Automation in Construction* 99 (2019), pp. 52-58.

Eisenbach M., R. Stricker, D. Seichter, K. Amende, K. Debes, M. Sesselmann, D. Ebersbach, U. Stoeckert, H.-M. Gross, How to get pavement distress detection ready for deep learning? A systematic approach, *2017 International Joint Conference on Neural Networks* (2017), pp. 2039-2047.

Euzenat J., Semantic precision and recall for ontology alignment evaluation, *20th International Joint Conference on Artificial Intelligence* (2007), pp. 348-353.

Gardner M.W., S. Dorling, Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences, *Atmospheric Environment* 32 (14-15) (1998), pp. 2627-2636.

Guo F., Y. Qian, J. Liu, H. Yu, Pavement crack detection based on transformer network, *Automation in Construction* 145 (2023), pp. 104646.

He K., X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, *Proceedings of the IEEE International Conference on Computer Vision*, (2015), pp. 1026-1034.

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

Hendrycks D., K. Gimpel, Gaussian error linear units (GELUs), arXiv preprint arXiv:1606.08415 (2016).

Hsieh Y.-A., J. Tsai Yichang, Machine learning for crack detection: review and model performance comparison, *Journal of Computing in Civil Engineering* 34 (5) (2020), pp. 04020038.

Hu J., L. Shen, G. Sun, Squeeze-and-excitation networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), pp. 7132-7141.

Huang S., Z. Lu, R. Cheng, C. He, FaPN: feature-aligned pyramid network for dense image prediction, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), pp. 864-873.

Huyan J., W. Li, S. Tighe, Z. Xu, J. Zhai, CrackU-net: a novel deep convolutional neural network for pixelwise pavement crack detection, *Structural Control and Health Monitoring* 27 (8) (2020), pp. e2551.

Iakubovskii P., Segmentation models pytorch, GitHub, 2019, p. GitHub repository.

Ji A., X. Xue, Y. Wang, X. Luo, W. Xue, An integrated approach to automatic pixel-level crack detection and quantification of asphalt pavement, *Automation in Construction* 114 (2020), pp. 103176.

Jiang S., L. Shen, F. Guillard, I. Einav, Characterisation of fracture evolution of a single cemented brittle grain using in-situ X-ray computed tomography, *International Journal of Rock Mechanics and Mining Sciences* 145 (2021), pp. 104835.

Jiang X., Iterative truncated arithmetic mean filter and its properties, *IEEE Transactions on Image Processing* 21 (4) (2011), pp. 1537-1547.

Jun F., L. Jiakuan, S. Yichen, Z. Ying, Z. Chenyang, ACAU-Net: atrous convolution and attention U-Net model for pavement crack segmentation, *2022 International Conference on Computer Engineering and Artificial Intelligence* (2022), pp. 561-565.

Kang D., S.S. Benipal, D.L. Gopal, Y.-J. Cha, Hybrid pixel-level concrete crack segmentation and quantification across complex backgrounds using deep learning, *Automation in Construction* 118 (2020), pp. 103291.

Kang D.H., Y.-J. Cha, Efficient attention-based deep encoder and decoder for automatic crack segmentation, *Structural Health Monitoring* 21 (5) (2021), pp. 2190-2205.

Kirschke K.R., S.A. Velinsky, Histogram-based approach for automated pavement-crack sensing, *Journal of Transportation Engineering* 118 (5) (1992), pp. 700-710.

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

Lewis J., Y.-J. Cha, J. Kim, Dual encoder–decoder-based deep polyp segmentation network for colonoscopy images, *Scientific Reports* 13 (1) (2023), pp. 1183.

Li G., X. Zhao, K. Du, F. Ru, Y. Zhang, Recognition and evaluation of bridge cracks with modified active contour model and greedy search-based support vector machine, *Automation in Construction* 78 (2017), pp. 51-61.

Li X., X. Sun, Y. Meng, J. Liang, F. Wu, J. Li, Dice loss for data-imbalanced NLP tasks, *arXiv preprint* (2019).

Lin T.-Y., P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, (2017), pp. 2117-2125.

Liu H., X. Miao, C. Mertz, C. Xu, H. Kong, Crackformer: transformer network for fine-grained crack detection, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), pp. 3783-3792.

Liu Z., Y. Cao, Y. Wang, W. Wang, Computer vision-based concrete crack detection using U-net fully convolutional networks, *Automation in Construction* 104 (2019), pp. 129-139.

Liu Z., Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), pp. 10012-10022.

Morera Á., Á. Sánchez, A.B. Moreno, Á.D. Sappa, J.F. Vélez, SSD vs. YOLO for detection of outdoor urban advertising panels under multiple variabilities, *Sensors* 20 (16) (2020), pp. 4587.

Oktay O., J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, Attention U-Net: learning where to look for the pancreas, *arXiv preprint arXiv:1804.03999* (2018).

Özgenel Ç.F., Concrete crack segmentation dataset, *Mendeley Data* 1 (2019), pp. 2019.

Park S., S. Bang, H. Kim, H. Kim, Patch-based crack detection in black box images using convolutional neural networks, *Journal of Computing in Civil Engineering* 33 (3) (2019), pp. 04019017.

Paszke A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimselshein, L. Antiga, Pytorch: an imperative style, high-performance deep learning library, *Advances in Neural Information Processing Systems* 32 (2019).

Peng X., X. Zhong, C. Zhao, A. Chen, T. Zhang, A UAV-based machine vision method for bridge crack recognition and width quantification through hybrid feature learning, *Construction and Building Materials* 299 (2021), pp. 123896.

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

Raghu M., T. Unterthiner, S. Kornblith, C. Zhang, A. Dosovitskiy, Do vision transformers see like convolutional neural networks?, *Advances in Neural Information Processing Systems* 34 (2021), pp. 12116-12128.

Ren Y., J. Huang, Z. Hong, W. Lu, J. Yin, L. Zou, X. Shen, Image-based concrete crack detection in tunnels using deep fully convolutional networks, *Construction and Building Materials* 234 (2020), pp. 117367.

Rodriguez-Lozano F.J., F. León-García, J.C. Gámez-Granados, J.M. Palomares, J. Olivares, Benefits of ensemble models in road pavement cracking classification, *Computer-Aided Civil and Infrastructure Engineering* 35 (11) (2020), pp. 1194-1208.

Ronneberger O., P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, *Medical Image Computing and Computer-Assisted Intervention—18th International Conference, Munich, Germany Springer*, (2015), pp. 234-241.

Shamsabadi E.A., C. Xu, D. Dias-da-Costa, Robust crack detection in masonry structures with Transformers, *Measurement* 200 (2022), pp. 111590.

Shamsabadi E.A., C. Xu, A.S. Rao, T. Nguyen, T. Ngo, D. Dias-da-Costa, Vision transformer-based autonomous crack detection on asphalt and concrete surfaces, *Automation in Construction* 140 (2022), pp. 104316.

Shaw P., J. Uszkoreit, A. Vaswani, Self-attention with relative position representations, *arXiv preprint arXiv:1803.02155* (2018).

Shrivastava A., A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), pp. 761-769.

Spencer B.F., V. Hoskere, Y. Narazaki, Advances in computer vision-based civil infrastructure inspection and monitoring, *Engineering* 5 (2) (2019), pp. 199-222.

Strudel R., R. Garcia, I. Laptev, C. Schmid, Segmenter: transformer for semantic segmentation, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), pp. 7262-7272.

Tan M., Q. Le, Efficientnet: rethinking model scaling for convolutional neural networks, *Proceedings of Machine Learning Research*, (2019), pp. 6105-6114.

Terven J., D.M. Cordova-Esparza, A. Ramirez-Pedraza, E.A. Chavez-Urbiola, Loss functions and metrics in deep learning, *arXiv preprint arXiv:2307.02694* (2023).

Tsai Y., V. Kaul, A. Yezzi, Automating the crack map detection process for machine operated crack sealer, *Automation in Construction* 31 (2013), pp. 10-18.

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

Tuli S., I. Dasgupta, E. Grant, T.L. Griffiths, Are convolutional neural networks or transformers more like human vision?, arXiv preprint arXiv:2105.07197 (2021).

Vaswani A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems* 30 (2017).

Wang W., C. Su, Automatic concrete crack segmentation model based on transformer, *Automation in Construction* 139 (2022), pp. 104275.

Wu H., B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, Cvt: introducing convolutions to vision transformers, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), pp. 22-31.

Wu Y., K. He, Group normalization, *Proceedings of the European Conference on Computer Vision* (2018), pp. 3-19.

Xiang C., J. Guo, R. Cao, L. Deng, A crack-segmentation algorithm fusing transformers and convolutional neural networks for complex detection scenarios, *Automation in Construction* 152 (2023), pp. 104894.

Xie E., W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, *Advances in Neural Information Processing Systems* 34 (2021), pp. 12077-12090.

Xu X., S. Xu, L. Jin, E. Song, Characteristic analysis of Otsu threshold and its applications, *Pattern Recognition Letters* 32 (7) (2011), pp. 956-961.

Xu Z., H. Guan, J. Kang, X. Lei, L. Ma, Y. Yu, Y. Chen, J. Li, Pavement crack detection from CCD images with a locally enhanced transformer network, *International Journal of Applied Earth Observation and Geoinformation* 110 (2022), pp. 102825.

Yamaguchi T., S. Hashimoto, Fast crack detection method for large-size concrete surface images using percolation-based image processing, *Machine Vision and Applications* 21 (5) (2010), pp. 797-809.

Yang F., L. Zhang, S. Yu, D. Prokhorov, X. Mei, H. Ling, Feature pyramid and hierarchical boosting network for pavement crack detection, *IEEE Transactions on Intelligent Transportation Systems* 21 (4) (2019), pp. 1525-1535.

Yang L., S. Bai, Y. Liu, H. Yu, Multi-scale triple-attention network for pixelwise crack segmentation, *Automation in Construction* 150 (2023), pp. 104853.

2. Robust and accurate computer vision models for automatic local-scale infrastructure defect detection

Yang X., H. Li, Y. Yu, X. Luo, T. Huang, X. Yang, Automatic pixel-level crack detection and measurement using fully convolutional network, *Computer-Aided Civil and Infrastructure Engineering* 33 (12) (2018), pp. 1090-1109.

Yu W., M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, S. Yan, Metaformer is actually what you need for vision, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), pp. 10819-10829.

Zhang E., L. Shao, Y. Wang, Unifying transformer and convolution for dam crack detection, *Automation in Construction* 147 (2023), pp. 104712.

Zhang L., F. Yang, Y.D. Zhang, Y.J. Zhu, Road crack detection using deep convolutional neural network, *2016 IEEE International Conference on Image Processing* (2016), pp. 3708-3712.

Zhang L., J. Shen, B. Zhu, A research on an improved Unet-based concrete crack detection algorithm, *Structural Health Monitoring* 20 (4) (2021), pp. 1864-1879.

Zhang Y., S. Mehta, A. Caspi, Rethinking semantic segmentation evaluation for explainability and model selection, *arXiv preprint arXiv:2101.08418* (2021).

Zhou Z., M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: a nested u-net architecture for medical image segmentation, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Granada, Spain, September 20, 2018, *Proceedings 4*, Springer, (2018), pp. 3-11.

Zhu H., B. Chen, C. Yang, Understanding why vit trains badly on small datasets: an intuitive perspective, *arXiv preprint arXiv:2302.03751* (2023).

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

This chapter presents the manuscript “Zhaohui Chen, Elyas Asadi Shamsabadi, Sheng Jiang, Luming Shen, Daniel Dias-da-Costa. (2024). Robust feature knowledge distillation for enhanced performance of lightweight crack segmentation models. arXiv:2404.06258”, which proposes a robustness-oriented knowledge distillation framework for noise-resilient and deployable defect detection.

The chapter introduces the Robust Feature Knowledge Distillation (RFKD) framework, which transfers not only predictive knowledge but also robust feature representations from teacher to student models through structured multi-level supervision. By integrating feature-level distillation with mixed clean-noisy training, RFKD enables lightweight models to learn noise-invariant and stable representations under degraded visual conditions.

A lightweight Transformer-based student model with only 0.5 million parameters is further designed to support robust distillation, enabling the construction of compact models that achieve strong robustness, generalisation, and deployment efficiency under severe resource constraints.

Comprehensive experiments demonstrate that the proposed framework significantly improves robustness, stability, and performance consistency in noisy and heterogeneous environments, establishing a principled foundation for noise-resilient and deployable local-scale perception.

3.1 Introduction

Crack detection is crucial in preserving the structural integrity of various infrastructures. The appearance of cracks may indicate aging, deterioration, or internal structural issues, which may cause major failures if not detected early (Wang, et al., 2021). Structural elements are prone to cracking under various combinations of loads, resulting in potential safety risks and reduced structural lifespan (Mohammadkhorasani, et al., 2023). Despite the importance of continuous monitoring and inspection of structures (Rubio, et al., 2019), traditional manual methods are discussed to have various drawbacks, mainly in the aspects of efficiency and safety. Computer vision-based methods provide a feasible solution that identifies cracks automatically through algorithm-based image processing. Yet, traditional threshold- (Kirschke and Velinsky, 1992;

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

Xu, et al., 2011) and morphology-based (Canny, 1986; Yamaguchi and Hashimoto, 2010) algorithms fail to achieve satisfactory performance when analysing crack images with complex textures. The emergence of deep learning algorithms (Jiang, et al., 2022; Karimi, et al.; Karimi, et al., 2024) provides a more common approach for remarkably enhanced automatic crack detection (Asadi Shamsabadi, et al., 2024).

Among deep learning models, convolutional neural networks (CNNs) have been widely adopted due to their strong local feature extraction ability (Jha and Babiceanu, 2023; Lan and Dong, 2022). However, CNNs often fail to capture long-range dependencies, which limits their robustness under complex conditions. Transformers, originally proposed in natural language processing, have recently been introduced to crack detection tasks due to their ability to model global context (Gao, et al., 2022; Shamsabadi, et al., 2022). Despite improved accuracy and generalisation, Transformer-based models are typically heavyweight and computationally demanding, making them impractical for edge-device deployment (Xu, et al., 2023). This attribute of Transformer-based models creates a fundamental contradiction between robustness and lightweight design, especially when deploying models on resource-constrained devices for real-time crack segmentation. Although several lightweight CNN-based models have been proposed (Ali and Cha, 2022; Choi and Cha, 2020; Dong, et al., 2024; Kang and Cha, 2021), they often struggle with performance degradation under noisy conditions or sacrifice performance for deployment requirements (Chu and Chun, 2023; Zhu, et al., 2023).

Knowledge distillation (KD), a process that knowledge is distilled from the heavyweight to smaller models to improve the patterns learnt during the model training, offers a promising solution to this problem by transferring the knowledge from a large, robust teacher model to a smaller student model (Hinton, et al., 2015; Wang, et al., 2020). However, existing KD approaches (Gou, et al., 2021; Li, et al., 2023; Wang and Yoon, 2021) for crack segmentation mainly focus on accuracy transfer and neglect robustness under real-world distortions such as noise and blur. Chen et al. (Chen, et al., 2023) proposed a CNN-based lightweight model with only 1.18 M parameters for crack segmentation and improved the model performance through hyperthermal and non-isothermal KD strategies.

This lightweight model was then demonstrated suitable for an edge device. Wang et al. (Wang, et al., 2023) designed a CNN-based light model with 5.92 M parameters and enhanced model accuracy through channel-wise KD strategies (Shu, et al., 2021). The applications of KD have

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

achieved comparable results for crack segmentation, however, conveying knowledge only from the logit layer (the final output layer that produces the raw scores or predictions by the network before applying the activation function) can only transfer accuracy without imparting anti-noise robustness (the ability to maintain accuracy under image noise disturbance) to the student models. Industry vision-based crack detection faces specific challenges due to the presence of distortions in input data. This occurs mainly because several factors can influence the quality of images captured by edge devices including noise. It is essential to ensure the ability of the model in generalising by improving the patterns learned during the training process. Given the limited computational power of edge devices, only small models can nonetheless be used for real-time crack segmentation. These models are often CNN-based and incapable of learning rich patterns when trained in a supervised manner, showcasing a significantly lower performance under image input distortions than Transformer-based models (Shamsabadi, et al., 2022; Shamsabadi, et al., 2022).

Considering the limitation of existing KD strategies and CNN-based small models, this work aims to achieve accurate and robust crack segmentation for possible practical industrial applications through developing the robust KD strategy and Transformer-based lightweight model. By distilling additional knowledge from intermediate feature layers with mixed clean and noise image input, we propose a novel framework, named robust feature knowledge distillation (RFKD), for the distillation of robust patterns from a teacher to a student. Considering the need for a capable and precise lightweight Transformer-based model which is missing in the literature, a Transformer-based lightweight crack segmentation model named PoolingCrackTiny (PCT) is also developed as part of this work. The core hypothesis is that the student model can learn robust crack features under the supervision of the teacher model when exposed to both clean and noisy data. Additional knowledge from the feature layers of the teacher model can help the student learn precise crack feature representations, resulting in the improvement of generalisation performance for the practical deployment. Our main contributions consist of two parts: the proposal of a novel knowledge distillation framework (RFKD) to enhance accuracy and robustness of lightweight crack segmentation models; the development of a lightweight Transformer-based model that integrates considerable accuracy, efficiency, and robustness after knowledge distillation.

This paper is organised as follows. Section 3.2 introduces the specific architectures of the proposed RFKD framework and PCT lightweight model. Section 3.3 illustrates the experiment implementation, including the dataset information, training settings, and assessment flowchart.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

In Section 3.4, the comprehensive results of the experiments are showcased and thoroughly discussed. Finally, the conclusions are drawn in Section 3.5.

3.2 Methodology

In this section, a novel robust knowledge distillation framework for crack segmentation was developed to improve both the accuracy and robustness of the distilled student model, followed by a designed student model for better leveraging the advance of the proposed framework and verifying its practical application.

3.2.1 RFKD framework

KD is a model compression technique where a smaller, efficient model (referred to as the student model) learns to replicate the behaviour of a larger, more accurate model (called the teacher model). The teacher model is typically a high-capacity network trained on labelled data, while the student model is trained using both the ground truth labels (also called hard labels) and the teacher's outputs (also called soft labels (Chen, et al., 2023)). This process enables the student to capture refined feature representations and generalization capabilities of the teacher, often achieving competitive performance with significantly fewer parameters and faster inference, making it suitable for deployment on resource-constrained devices. Currently, KD strategies for crack segmentation are based on transferring knowledge between the logit layers of teacher and student models (Chen, et al., 2023; Wang, et al., 2023). Yet, these logit layer-based KD methods have limited influence on the performance of the student model. Since feature representation learning was introduced by Bengio et al. (Bengio, et al., 2013), feature-based knowledge from the intermediate layers has become a reasonable extension of logit layer-based KD to supervise the training of the student model (Liu, et al., 2019; Yu, et al., 2019). Recent studies have further enriched feature-based distillation for dense prediction tasks. WaveNet (Zhou, et al., 2023) demonstrates that saliency-guided and wavelet-enhanced feature transfer can improve boundary fidelity in multimodal segmentation. SSRNet (Zhou, et al., 2025) combines hierarchical feature distillation with contrastive learning to stabilise cross-modality representation learning, while HKDNet (Zhou, et al., 2024) shows that aligning multi-scale features through hybrid convolution-attention modules enhances spatial consistency in crowd estimation. MDNet (Zhou, et al., 2024) extends this direction by integrating long-range contextual modelling with an efficient distillation framework for dense urban scene prediction. In addition, FCDENet (Zhou, et al., 2025) and MPMTNet (Zhou, et al., 2025) illustrate that

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

multi-path and multi-field feature fusion can guide student models toward more complete structural understanding under challenging visual conditions. These studies collectively suggest that intermediate feature transfer, rather than reliance solely on logit distillation, plays a crucial role in dense prediction tasks. While these methods substantially advance feature-based distillation and demonstrate the effectiveness of transferring richer intermediate representations, most existing studies are primarily evaluated under clean input conditions. Consequently, the question of how to systematically convey a teacher model’s robustness characteristics to a lightweight student has received comparatively less attention. Building upon these valuable contributions and inspired by the adversarial training with specific noises to enhance model robustness against these designated noises (Maroto, et al., 2022; Shao, et al., 2021), we designed a novel framework (see Figure 3.1) to explore an orthogonal direction by incorporating clean-noisy paired inputs and encouraging the student to learn noise-resilient feature patterns distilled from the teacher.

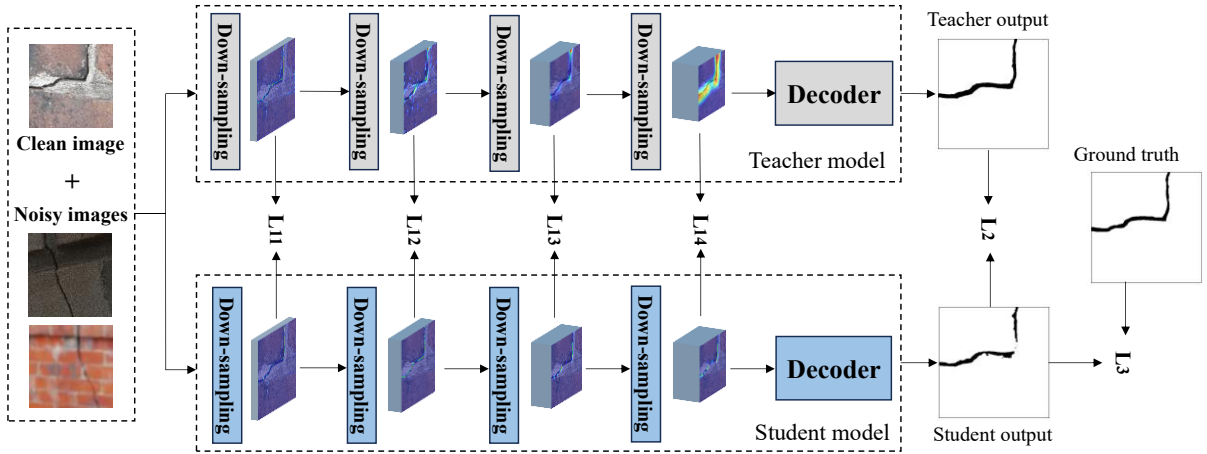


Figure 3.1 Framework of the proposed robust feature knowledge distillation (RFKD).

In this study, four common image noises, including pepper, salt, crack texture and gaussian blur, are utilised to distort parts of the clean images in the dataset. A study concerning hyperparameter tuning is presented in Appendix 3.A.2. The ratio of clean and noisy images is finally established as 2:1, with an equal quantity of different types of noisy images. The ground truth labels are first used to train the teacher model. In the next stage, as shown in Figure 3.1, the student model training is under the supervision of three parts: intermediate feature maps (the outputs of hidden layers after all down-sampling stages of the network) of the trained teacher model, the final output of the trained teacher model, and the ground truth labels. L_{1i} ($i = 1, 2, 3, 4$), L_2 , and L_3 are the respective losses for these three parts, and the total training loss is formulated as:

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

$$L = \sum \alpha_i \cdot L_{1i} + \beta \cdot L_2 + \gamma \cdot L_3, \quad (1)$$

where α_i ($i=1, 2, 3, 4$), β and γ are hyperparameters, representing the weight of each part, that are set before training. In the first two items of the total loss, to maximise the similarity of the student and teacher model outputs, the pair-wise distance loss (Yu, et al., 2019) is utilised to minimise the pixel-level Euclidean distance of their intermediate and final outputs. Unlike traditional Mean Squared Error loss (Adriana, et al., 2015) or Kullback-Leibler divergence losses (Hinton, et al., 2015) that typically focus on aggregated outputs or probability distributions, the pair-wise distance loss directly minimises the spatial discrepancy at each pixel location between intermediate representations, which is particularly beneficial for segmentation tasks, where spatial consistency and local structural features are critical. Obtaining feature map tensors from the intermediate and logits layer of the student and teacher models, the pair-wise distance loss L^{pw} is formulated as follows:

$$L^{pw} = \|F^S - F^T\|, \quad (2)$$

here F^S and F^T denote the feature map tensors of the student and the teacher models, and $\|\cdot\|$ is the Frobenius norm (Böttcher and Wenzel, 2008).

For the student model training with the ground truth labels in the last item of the total loss, Dice loss (Sudre, et al., 2017) was chosen due to its balanced performance in image segmentation. This loss function (L^{DS}) is derived from the Dice score (DS) (Sudre, et al., 2017), as shown in the following:

$$L^{DS} = 1 - DS, \quad (3)$$

$$DS = \frac{2|P \cap T|}{|P| + |T|}, \quad (4)$$

where P represents the predicted pixel map, and T denotes the labelled pixel map. According to Eq. (1), the total loss for the student model training is:

$$L = \sum \alpha_i \cdot L_{1i}^{pw} + \beta \cdot L_2^{pw} + \gamma \cdot L_3^{DS}, \quad (5)$$

where the pair-wise distance loss is applied for the soft label distillation from the teacher (L_{1i}^{pw} and L_2^{pw}), and Dice loss is selected for the hard label distillation from the ground truth (L_3^{DS}). To consider each loss item equally, the hyperparameters are set as: $\alpha_i = 0.25$ ($i=1, 2, 3, 4$), $\beta = 1$, $\gamma = 1$.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

Unlike previous KD strategies for crack segmentation (Chen, et al., 2023; Wang, et al., 2023) only considering L_2 and L_3 , RFKD enhances robustness by jointly constraining the student network at multiple representational levels under clean-noisy paired inputs. Noise primarily perturbs low- and mid-level features, and the intermediate-feature distillation loss encourages the student to project these perturbed activations back toward the teacher’s stable feature manifold. The pairwise distance loss further regularises the relational structure of feature embeddings, preserving spatial similarity patterns under perturbations. At the semantic level, logit distillation enforces consistency between the student’s predictions and the teacher’s noise-invariant outputs, while Dice loss maintains spatial fidelity in the segmentation mask. These complementary constraints operate across different abstraction levels, limiting the propagation of noise through the network and providing a theoretical rationale for the robustness improvements achieved by RFKD.

General-purpose robustness techniques for semantic segmentation commonly rely on online per-sample augmentations or adversarial training. While online augmentation broadens the input distribution, it does not regulate the batch-level composition of clean and noisy data and therefore cannot guarantee consistent representational alignment across noise conditions. Adversarial training, although effective for norm-bounded perturbations, incurs substantial computational overhead and targets a threat model distinct from the structured degradations considered in this study. RFKD instead adopts an offline partial-corruption strategy that deterministically forms clean-only, noisy-only, and mixed batches, enabling the student to stabilise feature alignment on clean batches while learning noise-resilient patterns from corrupted ones. The resulting robustness gain arises primarily from representation-level transfer: RFKD aligns the student to the teacher’s noise-invariant intermediate features and logits under a controlled exposure schedule, an effect that cannot be achieved through input-level augmentation alone. This structured design enables RFKD to provide an efficient and effective robustness solution for real-world degradations encountered in structural health monitoring.

3.2.2 PCT architecture

To demonstrate the advantage of the RFKD-based trained student model, we designed a novel Transformer-based student model with only 0.5 M parameters, specifically for robust and accurate crack segmentation in practical deployment. The overall structure of the proposed

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

lightweight PoolingCrackTiny (PCT) is illustrated in Figure 3.2, built upon the PoolingCrack model (Chen, et al., 2024) (see Appendix 3.A.3).

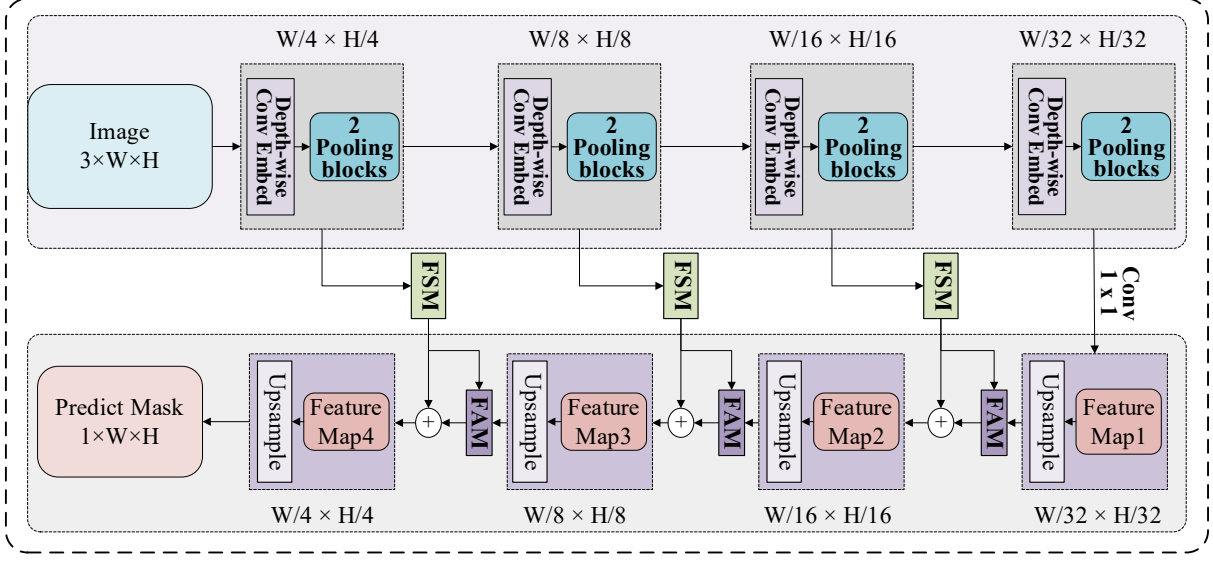


Figure 3.2 Architecture of the proposed lightweight model PCT.

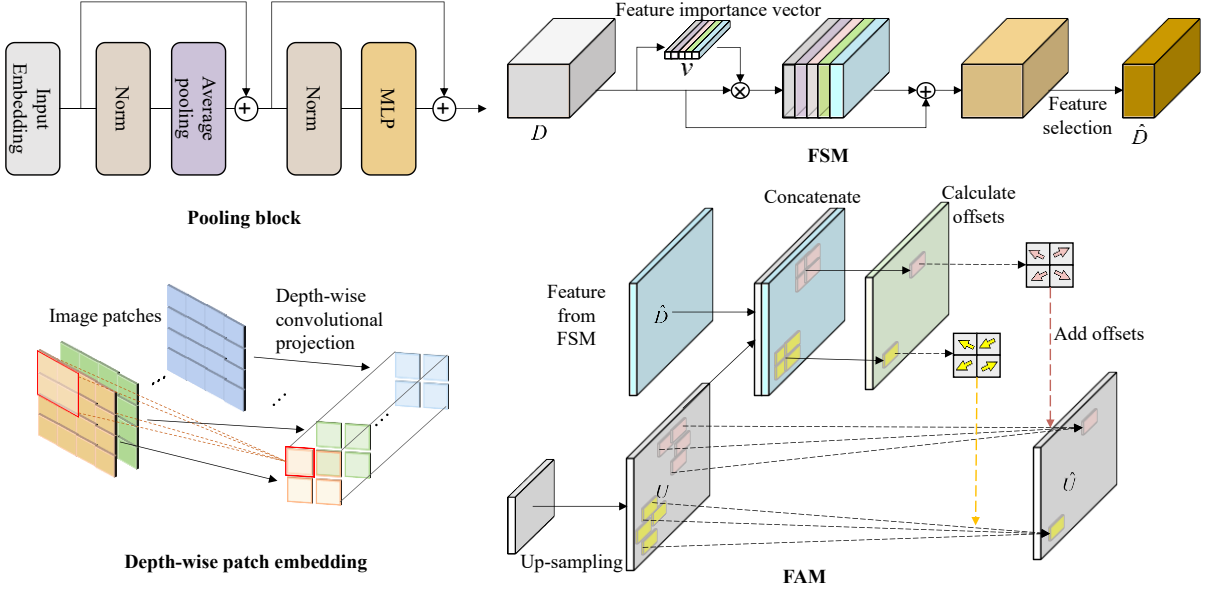


Figure 3.3 Details of each module for the proposed lightweight model PCT.

PCT consists of four encoder stages with increased channels and decreased sizes of feature maps. Each stage includes a depth-wise patch embedding module followed by two pooling blocks (see Figure 3.2). The image input is first transformed to embedded tokens by depth-wise patch embedding, followed with the embedded tokens normalisation, average pooling operation, additional normalisation, and a multilayer perceptron (MLP) processing (Gardner and Dorling, 1998), with two residual additions in the pooling blocks (see Figure 3.3). During decoding, multi-scale features are progressively upsampled and fused with high-level features

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

using the feature selection module (FSM) (Huang, et al., 2021) and feature alignment module (FAM) (Dai, et al., 2017). Since a depth-wise convolution applies one filter for an independent channel while a standard convolution applies a filter across all input channels simultaneously, the depth-wise convolution significantly decreases the convolutional operation, which is efficient and widely used in the model design for edge devices (Chollet, 2017). The feature selection module and feature alignment module can spatially align feature representations via learned offset corrections and improve boundary localisation in segmentation masks without introducing redundant parameters (Chen, et al., 2024). To maintain the efficiency and accuracy of PCT model, we applied the depth-wise convolution in the convolutional patch embedding to shrink the model scale and used the feature selection module and feature alignment module to increase the model performance without significant model expansion. Compared with PoolingCrack, the number of pooling blocks and patch embedding channels were respectively reduced to (2, 2, 2, 2) and (48, 96, 192, 384) in the PCT model to further improve model efficiency.

Table 3.1 Efficiency and accuracy comparison between PoolingCrack and RFKD distilled PCT

Model	Parameters (M)	FLOPs (G)	mDS (%)	mIoU (%)	Inference time (ms)	Processing time for test set (s)
PoolingCrack	31.9	27.9	68.6	54.8	15.0	16.9
PCT	0.5	2.7	67.4	53.4	5.6	6.3

Table 3.1 illustrates the efficiency and accuracy comparison between PoolingCrack and RFKD distilled PCT in the Crack500 dataset (Zhang, et al., 2016). Although PCT is derived from the PoolingCrack family, its novelty goes far beyond simple parameter reduction. The model restructures the pooling-based Transformer into a lightweight form specifically adapted for robust distillation and edge deployment, where each encoder stage is redesigned with depth-wise patch embeddings and cross-scale feature alignment to enable spatially matched teacher-student supervision, a capability not present in the original PoolingCrack architecture. This architectural-distillation co-design allows PCT to interact effectively with RFKD and forms a new design space for pooling-based Transformers. At only 0.5M parameters, PCT represents, to our knowledge, the first sub-million-parameter Transformer framework for crack segmentation that maintains competitive accuracy while achieving substantially improved robustness, including large gains in mean Dice Score (mDS) (Sudre, et al., 2017) compared with existing lightweight models, demonstrating that its performance cannot be achieved through naïve downscaling. The depth-wise embedding and feature alignment components

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

were newly formulated for efficiency and robustness rather than inherited from PoolingCrack, establishing PCT as a distinct and technically meaningful contribution rather than a simplified variant. In summary, RFKD integrates structured robustness distillation, deterministic exposure scheduling, relational feature alignment, and a co-designed lightweight PCT architecture, forming the core methodological contributions of this study.

3.3 Study design

In this section, the representative KD strategies and lightweight models were selected to run the comparison, thus validating the performance of the proposed RFKD framework and PCT model. The evaluation metrics, image dataset, training configuration, and study framework are also discussed in detail.

3.3.1 KD strategy and lightweight model selection

To verify the performance of the proposed RFKD method and evaluate the capabilities of the distilled PCT lightweight model, two comparisons were conducted respectively with the representative KD strategies and lightweight models. Due to their extensive applicability and superior performance, Naïve Knowledge Distillation (NKD) (Hinton, et al., 2015), Channel-wise Knowledge Distillation (CWD) (Shu, et al., 2021), and DIST (Huang, et al., 2022) were chosen for the KD strategy comparison. Current widely applied models on edge devices, CNN-based MobileNetV2 (Sandler, et al., 2018), MobileOne (Vasu, et al., 2023), EfficientNet-B0 (Tan and Le, 2019), and Transformer-based MobileViT (Mehta and Rastegari, 2021) were also selected as established and SOTA CNN- and Transformer-based models to analyse the performance of the lightweight PCT model, designed in this work.

3.3.2 Evaluation metrics

Since a crack segmentation model aims to distinguish the crack regions and the background in the crack images, two common evaluation metrics are frequently utilised to appraise the performance of the student model with KD strategies, including mean Dice score (Sudre, et al., 2017) and mean intersection over union (mIoU) score (Jaccard, 1912). DS metric is introduced in Eq. (4), and the IoU metric can be formulated as:

$$IoU = \frac{|P \cap T|}{|P \cup T|}, \quad (6)$$

similar to DS metric, P is the predicted pixel map, while T is the labelled pixel map. As these two metrics are insensitive to the specific threshold used to crack pixel classification and

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

evaluate the overlap of the entire region, including the boundary, they provide a more balanced evaluation of segmentation performance than Precision and Recall, alleviating the issues of threshold dependence and boundary bias (Euzenat, 2007; Li, et al., 2019).

3.3.3 Image datasets

To assess the student model performance with different KD strategies, a crack image dataset MC448 (Shamsabadi, et al., 2022) was selected for the validation experiment. This dataset was selected primarily due to the complexity of the masonry surface, which poses a greater challenge for computer vision-based frameworks compared to asphalt and concrete surfaces. The dataset poses significant challenges for crack segmentation due to the inclusion of non-crack images and the presence of crack-like textures in the background of masonry surfaces. The dataset comprises 3351 masonry crack images, each with a resolution of 448×448 pixels, along with their respective masks. These images were sourced from various origins, including 45 copyright-free images, 153 high-resolution manually captured photographs, and 15 artificially generated images. Since four image noises were chosen in this study for evaluating the proposed RFKD framework, 3000 images and their corresponding masks including 2000 clean and 1000 noisy images (250 images for each type of noise) were allocated for training, reserving the remaining 351 images for model evaluation. In most real-world scenarios, the number of each type of noisy image would be different, thus resulting in the dataset imbalance. The various kinds of noisy images can be adjusted to have the same quantity through oversampling (increasing the number of specific noisy images through data augmentation techniques such as rotation, flipping or cropping) and undersampling (randomly removing the specific noisy images to reduce the number) (Shelke, et al., 2017). In this study, we mainly adjusted and kept the numbers of different noisy images the same for performance evaluation; the relevant image samples with their masks are shown in Figure 3.4.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

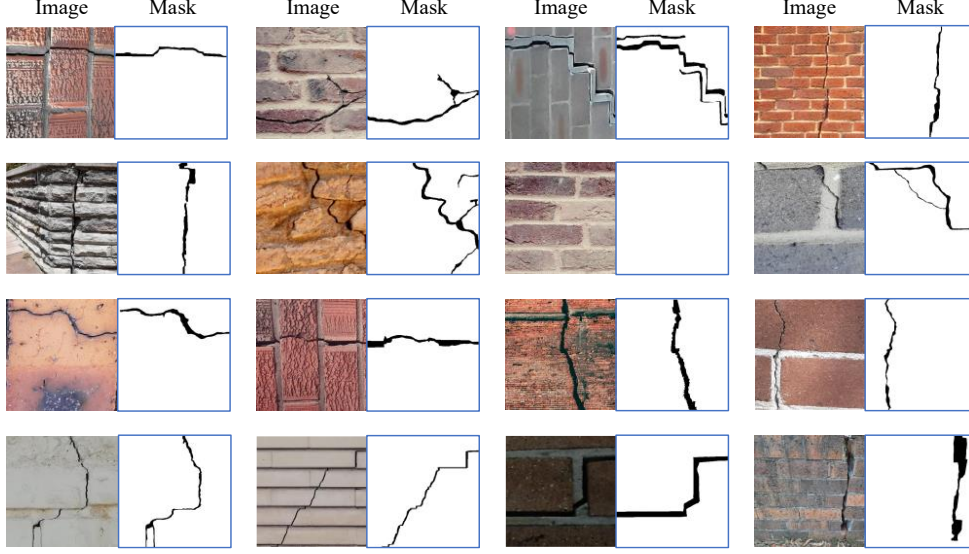


Figure 3.4 Image samples with their masks in dataset MC448 (Shamsabadi, et al., 2022).

3.3.4 Training configuration

The proposed framework was implemented in Python with the PyTorch (Paszke, et al., 2019) open-source platform. All the validation and evaluation experiments were performed on a workstation featuring an AMD Ryzen 9 5900X 12-Core CPU, an NVIDIA GeForce RTX 3090 Ti GPU, and 32 gigabytes of RAM. The learning rate was configured to 0.001, utilising a batch size of two images, and the training progress for knowledge distillation lasted for 500 epochs employing the AdamW optimiser. To enhance the generalisation of the training models and prevent overfitting, standard image augmentation techniques were applied during training. Based on common practice in computer vision literature for data augmentation, colour jittering with a probability of 20% was used to randomly perturb image brightness, contrast, saturation, and hue, introducing colour variation similar to real-world lighting conditions (Krizhevsky, et al., 2012). Additionally, horizontal flipping was applied with a 50% probability to mirror the image along the vertical axis, thus augmenting geometric diversity (Krizhevsky, et al., 2012).

3.3.5 Study framework

In this study, all the experiments are conducted on the MC448 masonry dataset, which includes non-crack images and crack-like background textures of masonry surfaces and showcases more challenges than asphalt and concrete surfaces to better verify and highlight the performance of RFKD framework. Given the previous studies utilised several common image noises (Patidar, et al., 2010) and a designed crack texture noise (a noise for assessing the bias of the networks towards the shape and texture of crack objects) (Shamsabadi, et al., 2022) to evaluate the anti-

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

noise robustness, we selected four of these noises (pepper, salt, crack texture and gaussian blur) to assess the model robustness against noises in this study. As illustrated in Figure 3.5, the proposed RFKD framework was first compared with the SOTA KD methods to verify its novel attributes of knowledge transferring accuracy and robustness. Then, the proposed lightweight model PCT with 0.5 M parameters was further compared with the existing representative lightweight models which have been deployed on edge devices, to evaluate the possibility of its practical deployment. In the first stage, three representative KD strategies (NKD, CWD, DIST) were chosen for the knowledge transferring accuracy and robustness comparison. In the second stage of the proposed PCT model assessment, all the representative model backbones were pretrained in the ImageNet (Deng, et al., 2009) dataset to improve their final performances. Two metrics, including mIoU and mDS, were utilised to evaluate the accuracy of crack segmentation. Key practical deployment metrics, including model parameters, floating point operations (FLOPs) and inference time, were considered to appraise the model efficiency (Chen, et al., 2023). Additional validation of RFKD framework, hyperparameter tuning, and ablation study were introduced in Appendix 3.A.

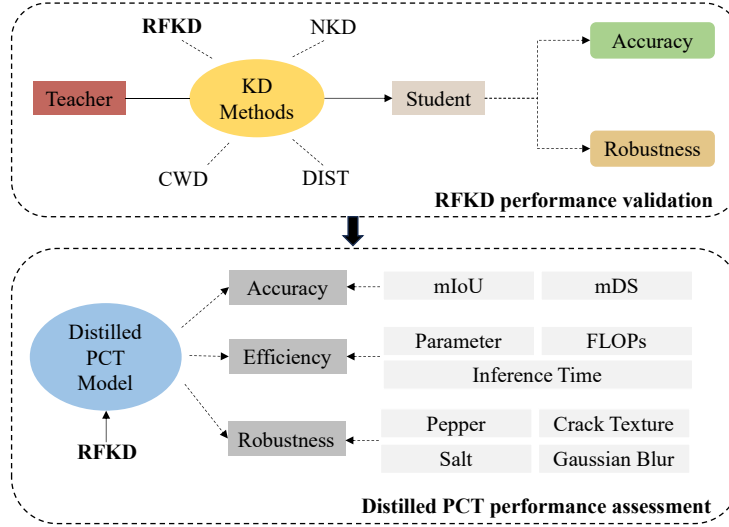


Figure 3.5 The schematic of the study framework.

3.4 Results and discussion

In this section, the proposed RFKD framework was first compared to the representative knowledge distillation methods and validated for its hypothetical performance. Then, the proposed PCT lightweight model trained with the help of RFKD, was compared to the commonly used lightweight model on edge devices and verified for its practical deployment. To fairly compare the performance of different knowledge distillation frameworks, the teacher model and student model were kept identical for each framework implementation, with the

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

same training configuration. In this study, PoolingCrack with 32 M parameters and PCT with 0.5 M parameters were chosen as the teacher and student models, respectively.

3.4.1 Accuracy assessment of the RFKD framework

As illustrated in Table 3.2, the distilled student models with different KD strategies are compared with the student model trained from scratch in terms of the model segmentation accuracy under clean and noisy (40% intensity) images. The proposed RFKD method presents the best accuracy improvement for the student model, which has 0.2% - 2.2% higher mDS and 1.1% - 2.9% higher mIoU than the other KD strategies. In addition, compared to training the student model from scratch, training the student model with different KD methods exhibits a significant improvement of 8.8% - 11.0% higher mDS and 8.6% - 11.5% mIoU. Since the students acquire knowledge from the teacher model, they also learn the improved feature representations and could produce more accurate segmentation results. Under most types of noisy images, the NKD, CWD, and DIST methods present significant accuracy drops. Since these KD methods primarily distil logit-level or shallow-layer knowledge under clean conditions, they lack robustness-aware feature alignment for the teacher and student, making the student vulnerable to local noise perturbations. By contrast, our proposed RFKD exhibits 14.4% - 41.2% mDS improvement under different noise disturbances. This difference would be attributed to RFKD’s incorporation of intermediate feature-level distillation under both clean and noisy inputs, allowing the student to learn robust, hierarchical representations that are less sensitive to local disruptions.

Table 3.2 Quantitative accuracy comparison for the different KD methods.

Knowledge distillation methods		Student from scratch	NKD	CWD	DIST	RFKD
Student model accuracy (clean)	mDS (%)	58.0	67.1 (+ 9.1)	66.8 (+ 8.8)	68.8 (+ 10.8)	69.0 (+ 11.0)
	mIoU (%)	46.2	55.5 (+ 9.3)	54.8 (+ 8.6)	56.6 (+ 10.4)	57.7 (+ 11.5)
Student model accuracy (crack texture)	mDS (%)	26.9	8.0 (- 18.9)	8.4 (- 18.5)	32.0 (+ 5.1)	57.0 (+ 30.1)
Student model accuracy (Gausssian blur)	mDS (%)	45.1	41.8 (- 3.3)	40.0 (- 5.1)	41.7 (- 3.4)	59.5 (+ 14.4)
Student model accuracy (Salt noise)	mDS (%)	22.4	1.6 (- 20.8)	8.4 (- 14.0)	17.7 (- 4.7)	63.6 (+ 41.2)
Student model accuracy (Pepper noise)	mDS (%)	27.4	6.1 (- 21.3)	10.2 (- 17.2)	25.0 (- 2.4)	63.1 (+ 35.7)

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

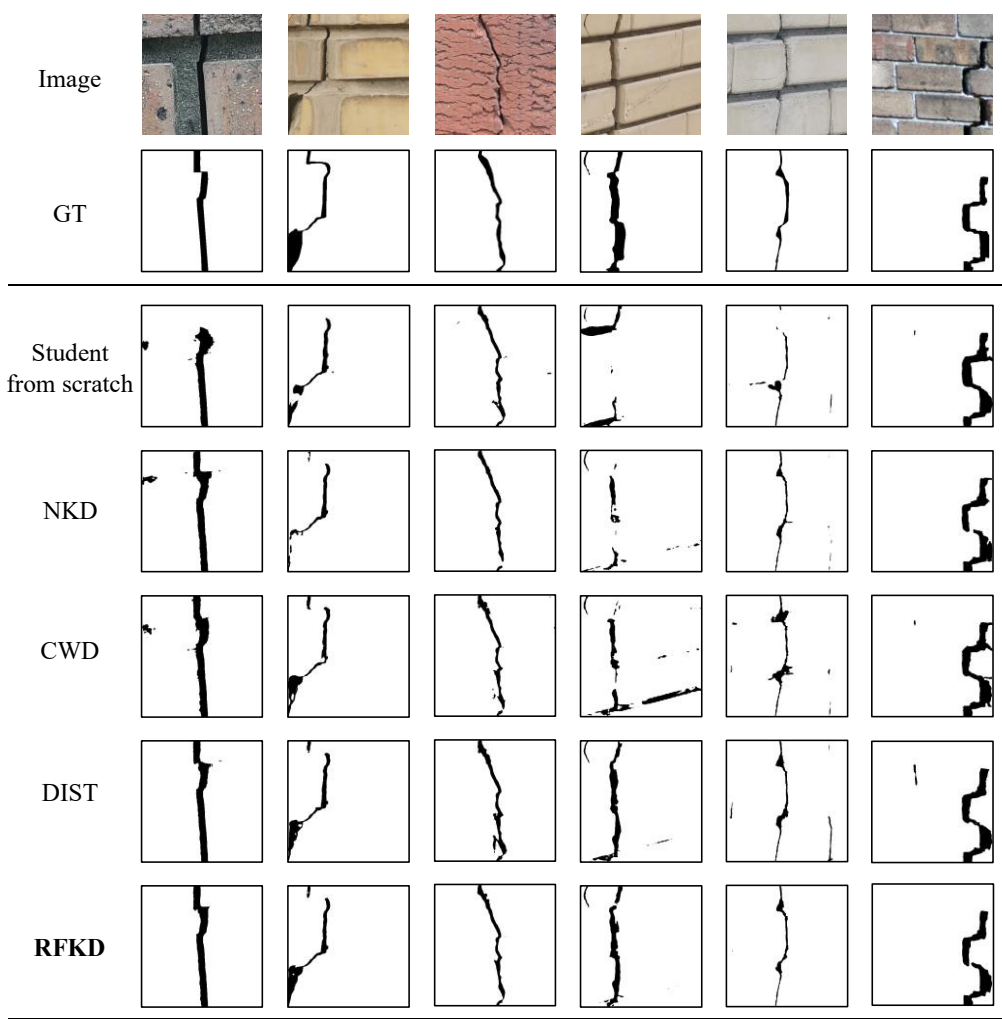


Figure 3.6 Qualitative accuracy comparison for the different KD methods. From top to bottom: original images, ground truth masks, and predictions from the student trained from scratch, NKD, CWD, DIST, and the proposed RFKD. Across the examples, RFKD provides noticeably finer crack-edge continuity (especially in thin and branching regions) and reduces false-positive artifacts around textured backgrounds. These improvements are particularly evident in (i) the narrow vertical cracks where other methods produce fragmented predictions, and (ii) complex brick or stone textures where competing methods mistakenly activate background edges.

Figure 3.6 shows the ground truth and segmentation results of the masks produced by the distilled student models trained with different KD methods. Overall, the distilled student with the RFKD method presents the best ability to segment cracks over the complex masonry background texture. Compared with the student model trained from scratch, KD methods improve the crack pixel identification accuracy with more correctly classified crack pixels. Yet, the distilled students with NKD, CWD and DIST methods produce some visual noises as they generate the crack maps, in contrast to the distilled student with RFKD methods. The main reason for the improved segmentation performance of the distilled student under the RFKD method would be attributable to the additional feature learning (extra loss functions for controlling the learning process) of intermediate layers from the teacher model (Adriana, et al., 2015). This additional feature learning allows the RFKD-trained model to better generalise

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

accurate multiple-scale features from the teacher model's knowledge. It helps the student model generate more exact predictions in the final feature fusion stage, resulting in enhanced segmentation quality, especially in complex visual contexts on the masonry dataset.

3.4.2 Robustness assessment of the RFKD framework

To achieve automatic crack detection in a practical environment with different kinds of challenging backgrounds and input image signal interference, the model should be robust against various noise disturbances and exhibit stable performance (Shamsabadi, et al., 2022). Since the previous KD methods (Chen, et al., 2023; Wang, et al., 2023) for crack segmentation are not able to improve the robustness of the student model, the proposed RFKD method is aimed at addressing this issue. To validate the robustness of the RFKD method, four types of image noises (Qin, et al., 2014; Shamsabadi, et al., 2022), including crack texture, gaussian blur, and salt and pepper noises, are selected for the performance evaluation. Figure 3.7 presents the robustness comparison of the distilled student models employing different KD strategies. The distilled student with the proposed RFKD method exhibits the best robustness, with up to a significant 62% higher mDS than the competing alternatives under these four noise disturbances. Notably, compared to the student model trained from scratch, while the distilled student models through the representative KD methods (NKD, CWD, DIST) can enhance mDS under low-intensity noise interference, their accuracy drops below that of the student from scratch as the noise intensity increases. However, the proposed RFKD method improves the mDS under all noise intensities, revealing its robust knowledge-transferring performance. The RFKD method utilises the mixed noisy and clean image inputs for training. During training, the student model is often exposed to inputs with added noise, encouraging it to learn to ignore such disturbances. This exposure helps the student model learn to filter out irrelevant information, focusing only on the essential features that contribute to stable performance under increasing noise intensity.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

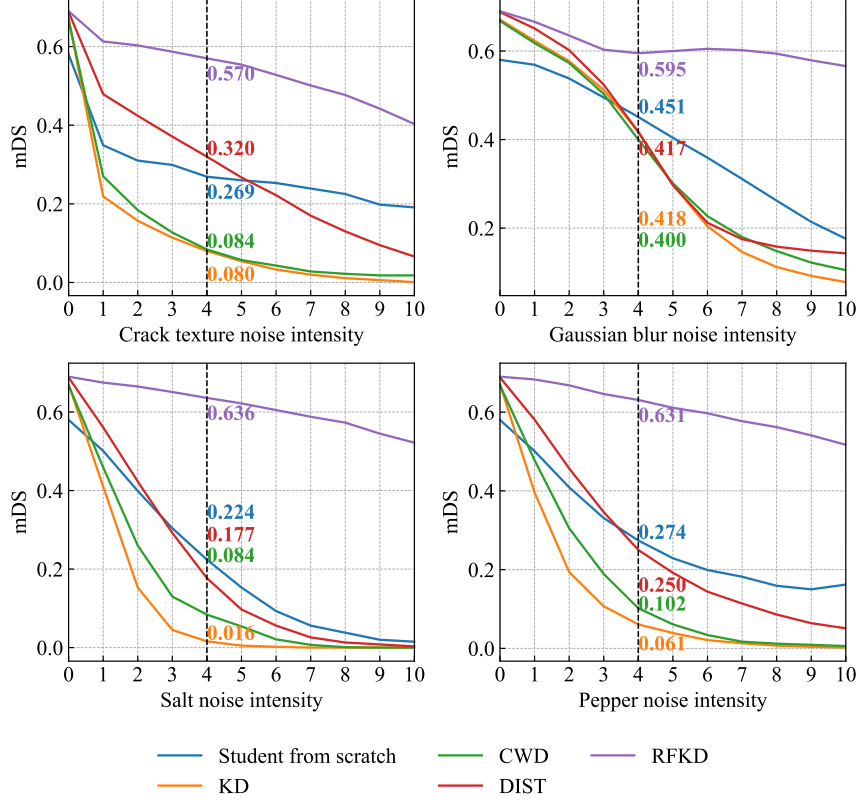


Figure 3.7 Quantitative robustness comparison for the different KD methods.

Figure 3.8 contains the segmentation results generated by the distilled students with different KD methods on crack images distorted with 40% noise intensity. The distilled student with the RFKD method portrays the most complete crack profile in contrast to the others. Compared to the student trained from scratch, the RFKD distilled student exhibits significant improvement in accurate crack mask generation, while the distilled students with other KD methods have unstable performances. Unlike other KD methods, RFKD employs noisy and clean images for training. The student model can learn the general features from the clean data and master the robust features from the noisy data. Thus, the RFKD distilled student model can maintain the segmentation quality in most cases in Figure 3.8, while other KD methods would fail to predict crack patterns under noise disturbance. Overall, it is evident that the proposed RFKD can effectively enhance the student model's accuracy and robustness.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

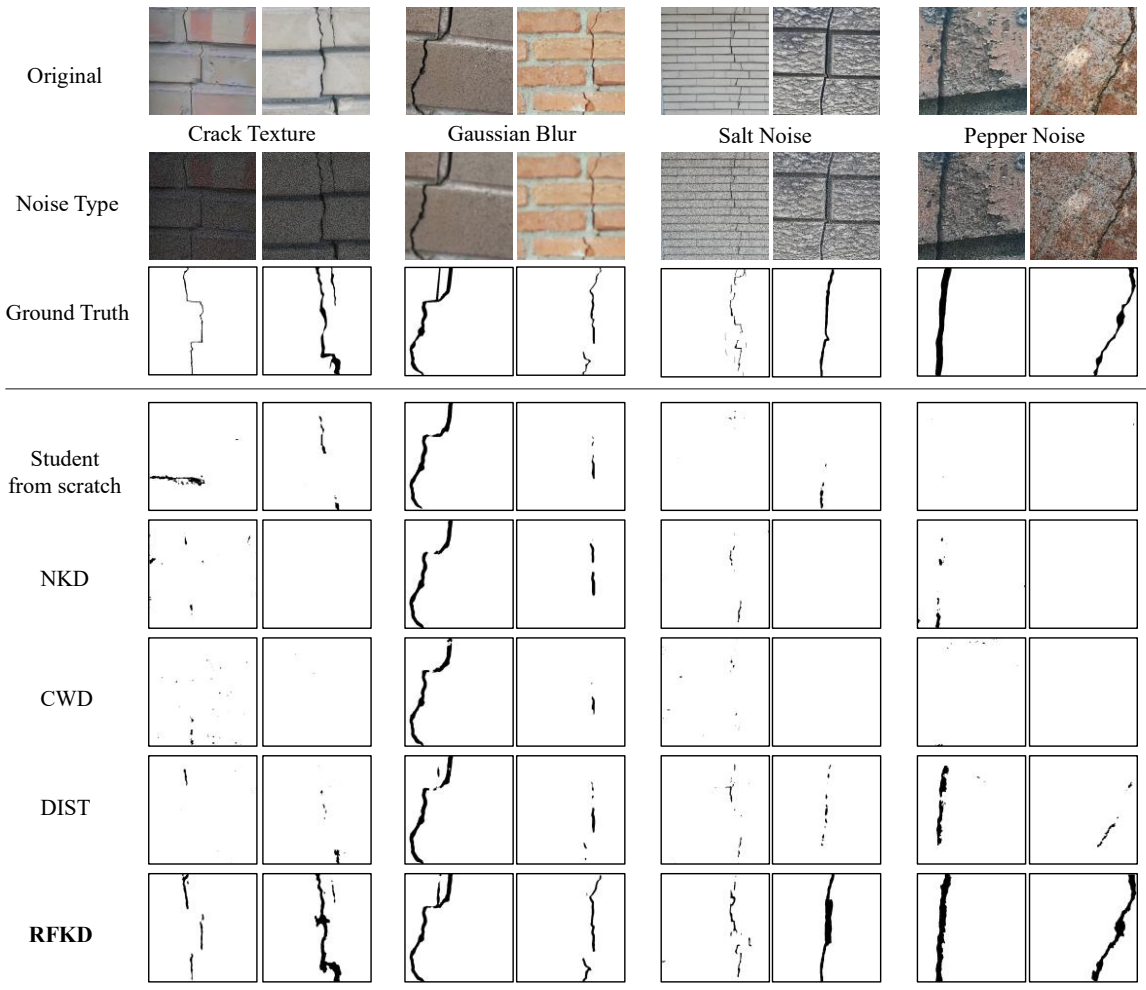


Figure 3.8 Qualitative robustness comparison for the different KD methods. From top to bottom: original images, noisy images, ground truth masks, and predictions from the student trained from scratch, NKD, CWD, DIST, and the proposed RFKD. Across all examples, RFKD preserves crack continuity more reliably under severe darkening, brightening, and blur. Improvements are particularly evident in regions where other methods either miss thin crack segments or produce noise-induced false edges on textured brick or stone surfaces. RFKD suppresses these spurious activations while maintaining clearer and more stable crack structures under challenging distortions.

3.4.3 Accuracy and efficiency assessment of the PCT model

Since the PCT model trained with the RFKD strategy is developed offline on high-performance GPUs rather than on edge devices, the computational considerations relevant to deployment primarily concern lightweight model parameters, inference latency, and floating-point operations. For completeness, we report that training the PCT model with RFKD for 500 epochs on the MC448 dataset required approximately 32 hours on a single NVIDIA RTX 3090 Ti (24 GB), with a peak memory usage below 18 GB. This training overhead is comparable to existing distillation frameworks such as CWD and DIST. As the training process is carried out only once and offline, these costs do not affect the feasibility of deploying the distilled model on resource-constrained edge devices. To evaluate the ability of the RFKD distilled PCT model for practical deployment, currently established and SOTA representative lightweight models,

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

widely applied on edge devices, were selected for comparison. Key practical deployment metrics, including model parameters, floating point operations and inference time, were used to evaluate the model efficiency (Chen, et al., 2023). The results of the accuracy and efficiency assessment are presented in Table 3.3 and Figure 3.9. Overall, MobileOne and the PCT model could show the best accuracy on clean data. However, although the distilled PCT model presents the same mDS and 0.1% lower mIoU than MobileOne, the PCT model has 74% - 92% fewer parameters, 4% - 75% fewer FLOPs, and 16% - 74% less inference time than all others including MobileOne. It is also worth mentioning that the robustness analysis presented in Figure 3.10 further reveals the significantly improved performance of the PCT model on noisy data. Since there is a trade-off between the model scale and the prediction accuracy, the main aim is not only to design a model with high accuracy on clean data but also to develop a light model with high generalisation robustness against noisy data for practical deployment, as real-world environments often involve unpredictable and challenging conditions.

Table 3.3 Quantitative accuracy comparison for the different lightweight models

Model	Parameters (M)	FLOPs (G)	mDS (%)	mIoU (%)	Inference time (ms)	Processing time for test set (s)
MobileOne	6.3	10.9	69.0	57.8	21.5	7.5
MobileNetV2	4.2	7.5	68.5	57.2	6.7	2.4
EfficientNet-B0	5.8	7.8	68.9	57.5	11.1	3.9
MobileViT	1.9	2.8	62.0	49.7	7.9	2.8
PCT	0.5	2.7	69.0	57.7	5.6	2.0

As shown in Figure 3.9, the segmentation results of different lightweight models in terms of the randomly selected crack images are visualised. Overall, the segmented crack profiles by the distilled PCT demonstrate comparable completeness in contrast to other lightweight models. Yet, when one crack image in the third column is under low luminance conditions, the distilled PCT maintains stable segmentation performance, while other lightweight models cannot provide accurate segmentation results. This difference would be attributed to the inferior robustness of those other lightweight models and the robust features learned by the distilled PCT.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

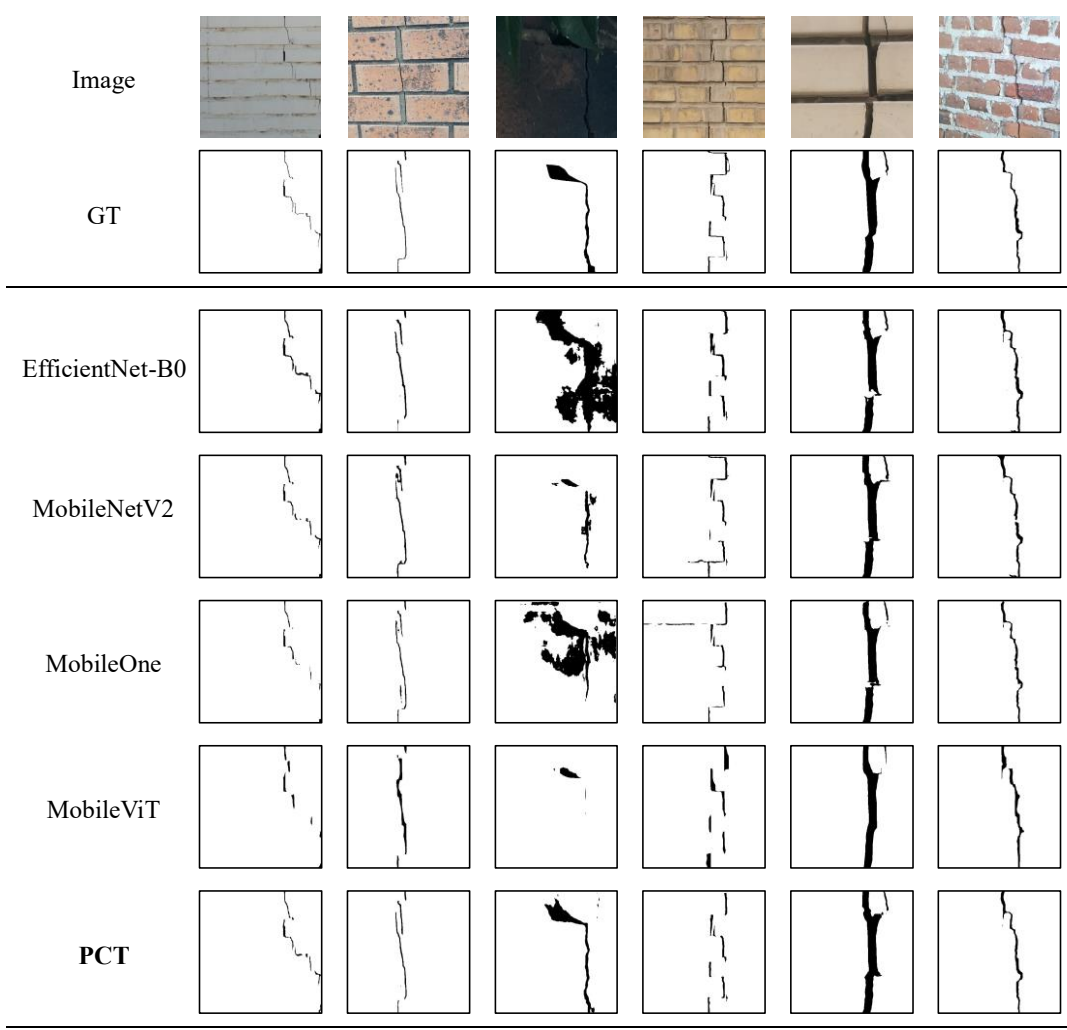


Figure 3.9 Qualitative accuracy comparison for the different lightweight models. From top to bottom: original images, ground truth masks, and predictions from several lightweight models. Across the examples, RFKD applied to the PCT model produces the most complete and coherent crack structures, particularly in regions where other models either miss thin crack segments or generate false responses along textured brick or stone surfaces. The improvements are especially visible in challenging cases with fine cracks and low contrast, where RFKD-PCT maintains clearer boundaries and fewer noise-induced artefacts.

3.4.4 Robustness assessment of the PCT model

As presented in Figure 3.10, the distilled PCT model presents the best crack segmentation precision under the interferences of all types of image noises, with up to 59.2% higher mDS than the other ones. While the other lightweight models can have considerable accuracy and provide consistent crack profile predictions without noise disturbances, their performances drop as the noise intensity increases. Figure 3.11 shows the segmented results of 40% noise intensity crack images for different lightweight models. The distilled PCT generates more precise and consistent crack masks than the competitive models. Due to the inconsistent performance under noise interferences, the other models fail to portray the complete crack profile or sometimes produce several false crack pixels. Since the RFKD distilled PCT model

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

learns more robust crack feature representations through the exposure to noisy inputs and the supervision of the teacher model, it can exhibit stable performance under noise interruptions.

Overall, given that current CNN-based lightweight crack segmentation models present unsatisfactory performance under the distortions of image inputs, and this weakness hinders their application to varying scenarios, the proposed RFKD framework is demonstrated to strengthen the performance and robustness of lightweight crack segmentation models, thus providing a possible solution to reinforce lightweight crack segmentation models for practical applications.

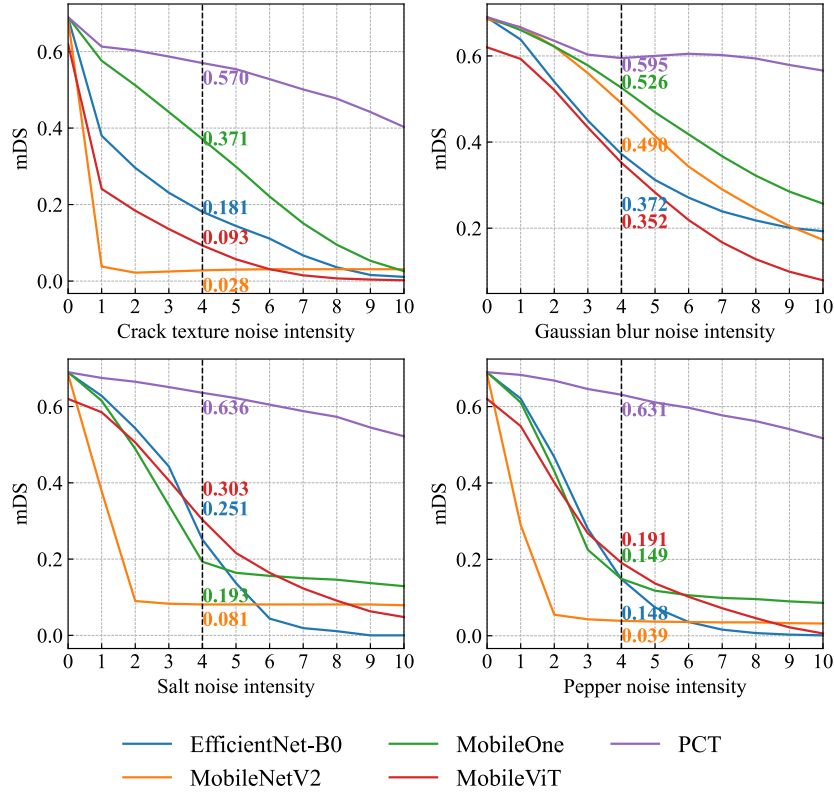


Figure 3.10 Quantitative robustness comparison for the different lightweight models.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

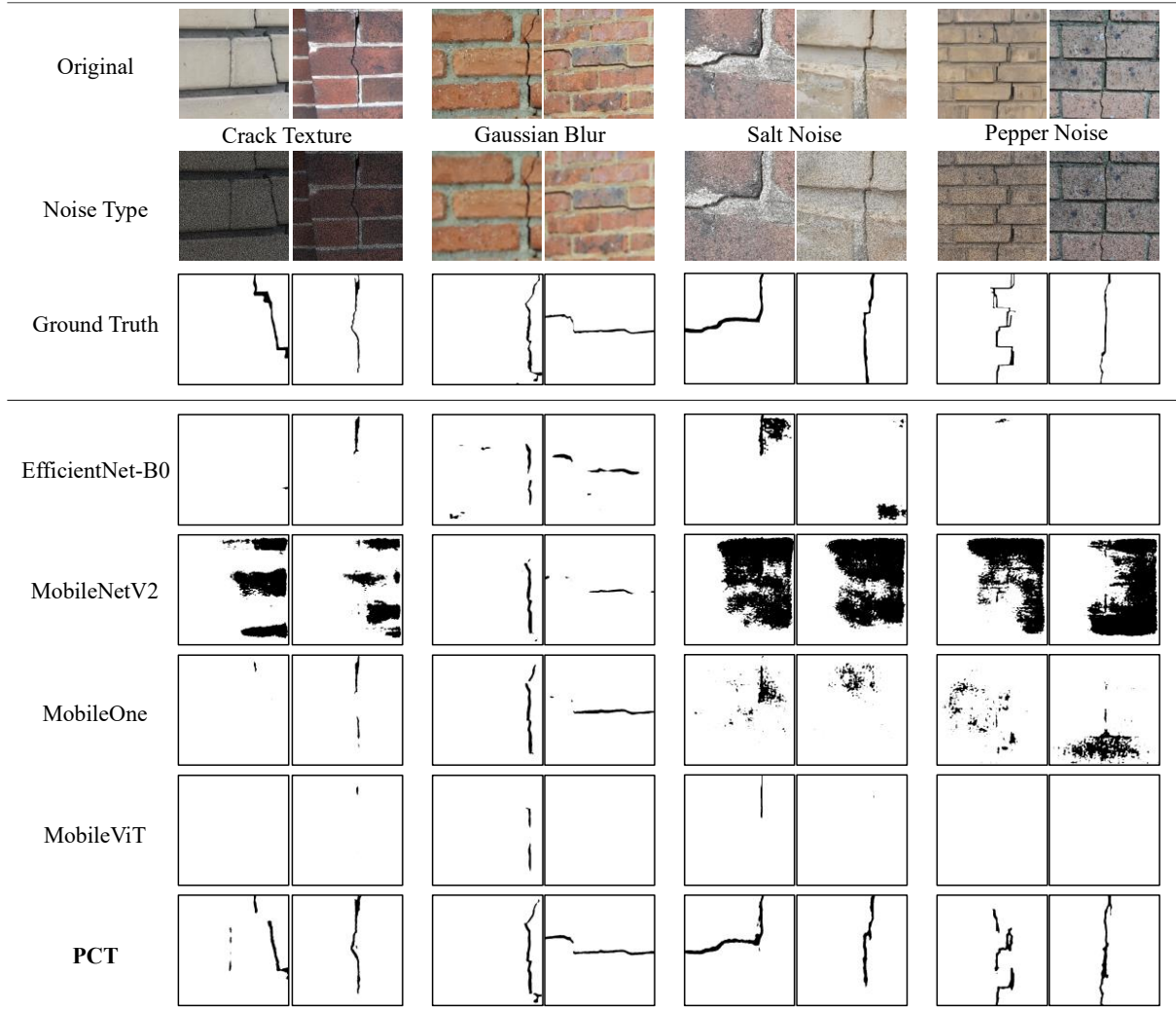


Figure 3.11 Qualitative robustness comparison for the different lightweight models. From top to bottom: original images, noisy images, ground truth masks, and predictions from several lightweight models. Under strong illumination changes and blur distortions, many lightweight models produce significant false-positive activations or lose large portions of thin crack structures, particularly in textured brick and stone regions. In contrast, the RFKD-enhanced PCT model maintains stable crack continuity and suppresses noise-induced artefacts across all examples. The improvements are especially visible in cases where competing models generate large spurious blobs or fail to recover faint crack segments.

3.5 Conclusions

Given the challenges for simultaneously lightweight, accurate, and anti-robust crack detection, current knowledge distillation methods for crack detection fail to improve both the accuracy and robustness of the lightweight models. To address this issue, a novel framework named RFKD is proposed, which enforces the student model to learn features of the teacher model in both intermediate and logit layers through the input of mixed noisy and clean training images. This proposed framework simultaneously improves accuracy and noise-resilience, providing a lightweight yet effective solution for practical crack segmentation in real-world industrial scenarios.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

To verify the performance of the proposed RFKD method, we conducted a validation comparison on the challenging MC448 dataset with non-crack and crack-like masonry background images, utilising three commonly used image noises (pepper, salt, and Gaussian blur) and a designed crack texture noise for evaluating the ability of models to distinguish the shape and texture of crack objects. The results showed that RFKD could assist the student model in obtaining 0.2% - 2.2% higher mDS and 1.1% - 2.9% higher mIoU than the previous NKD, CWD, and DIST methods on the clean data, and up to 62% higher mDS on the noisy data. In addition, to demonstrate the further application of the RFKD, a lightweight Transformer-based student model PCT was designed and compared with the current representative lightweight models. The comparative outcomes indicated that the PCT model distilled by RFKD could achieve identical mDS and slightly lower mIoU (0.1%) than the most accurate and largest MobileOne model on the clean data, while up to 59.2% higher mDS than other representative lightweight models on the noisy data, with 74% - 92% fewer parameters, 4% - 75% fewer FLOPs, and 16% - 74% less inference time.

However, RFKD would have some limitations: the knowledge distillation process requires extensive training epochs to achieve satisfactory results, thus becoming time-consuming; if the teacher exhibits uneven robustness across different distortion types, the student may overfit to the noise patterns most strongly represented in the teacher's outputs, resulting in weaker performance under certain degradations; the information from the teacher model may not be sufficient to guide the student model effectively at very high noise levels, which can lead to the student model inheriting and even amplifying the noise. These limitations can be potentially addressed in future research: the incorporation of attention mechanisms to guide the lightweight student model on where to focus, improving its ability to capture essential features from the teacher model, thus accelerating the knowledge distillation process; adopting multi-teacher distillation, the student can learn a more balanced and noise-agnostic representation that generalises better across diverse degradation types. the development of more robust teacher models or the ensemble of different robust teacher models handling different noises well to provide intense supervision for the lightweight student model since the performance of the lightweight student model relies on the teacher model's performance.

The deployment of lightweight and noise-robust crack segmentation models enables real-time structural health monitoring on resource-constrained platforms such as mobile devices, UAVs, and embedded sensors. This capability can reduce inspection costs, enhance the frequency and coverage of structural assessments, and support earlier detection of safety-critical deterioration.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

Such models may therefore contribute to improved public safety and more efficient asset-management workflows. At the same time, responsible deployment requires awareness of domain shift, environmental variability, and the need for human verification, so that automated predictions complement rather than replace expert judgement in real-world scenarios.

Appendix 3.A

3.A.1 Additional validation

Another validation of RFKD framework was conducted in the Crack500 (Zhang, et al., 2016) dataset, which included crack images on asphalt and concrete pavement surfaces. The released version of Crack500 comprised 1896 and 1124 images for training and testing. All the images with their annotations were cropped into 640×320 pixels to satisfy the requirements of all the studied models. As shown in Figure 3.12, RFKD improved both accuracy (noise intensity equal to zero) and robustness for the PCT model trained from scratch in the Crack500 dataset, thus demonstrating its generalisation for different crack datasets.

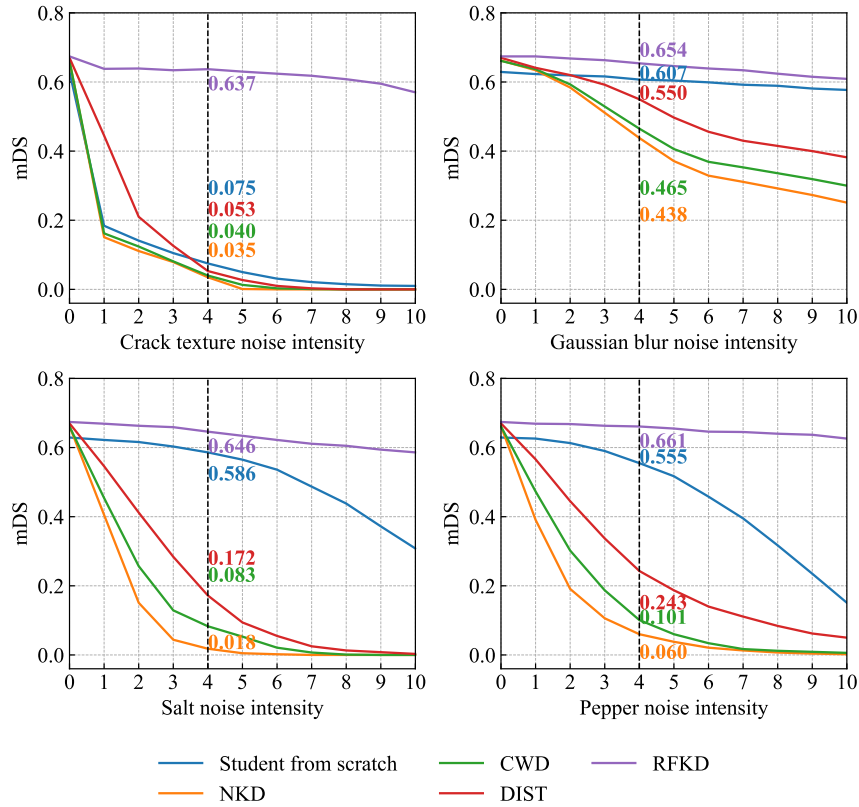


Figure 3.12 Performance validation of the proposed RFKD in the Crack500 dataset.

Additional recent lightweight models, MobileViTv3 (Wadekar and Chaurasia, 2022) and EdgeNeXt (Maaz, et al., 2022), were added for the distilled PCT model performance

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

comparison. In Table 3.4, Figure 3.13, and Figure 3.14, we compared the model efficiency, accuracy and robustness in both the Crack500 and MC448 datasets. The distilled PCT model exhibited the overall best accuracy, efficiency, and robustness performance in these two datasets.

Table 3.4 Quantitative efficiency and accuracy comparison for the different lightweight models in two datasets

Model	Parameters (M)	FLOPs (G)	Crack500		MC448		Inference time (ms)
			mDS (%)	mIoU (%)	mDS (%)	mIoU (%)	
MobileOne	6.3	10.9	66.3	52.4	69.0	57.8	21.5
MobileNetV2	4.2	7.5	66.9	52.7	68.5	57.2	6.7
EfficientNet-B0	5.8	7.8	67.1	53.2	68.9	57.5	11.1
MobileViT	1.9	2.8	66.9	52.9	62.0	49.7	7.9
MobileViTv3	3.1	3.9	62.9	49.3	60.6	49.2	5.7
EdgeNeXt	2.0	1.9	66.8	52.9	62.1	50.4	6.6
PCT	0.5	2.7	67.4	53.4	69.0	57.7	5.6

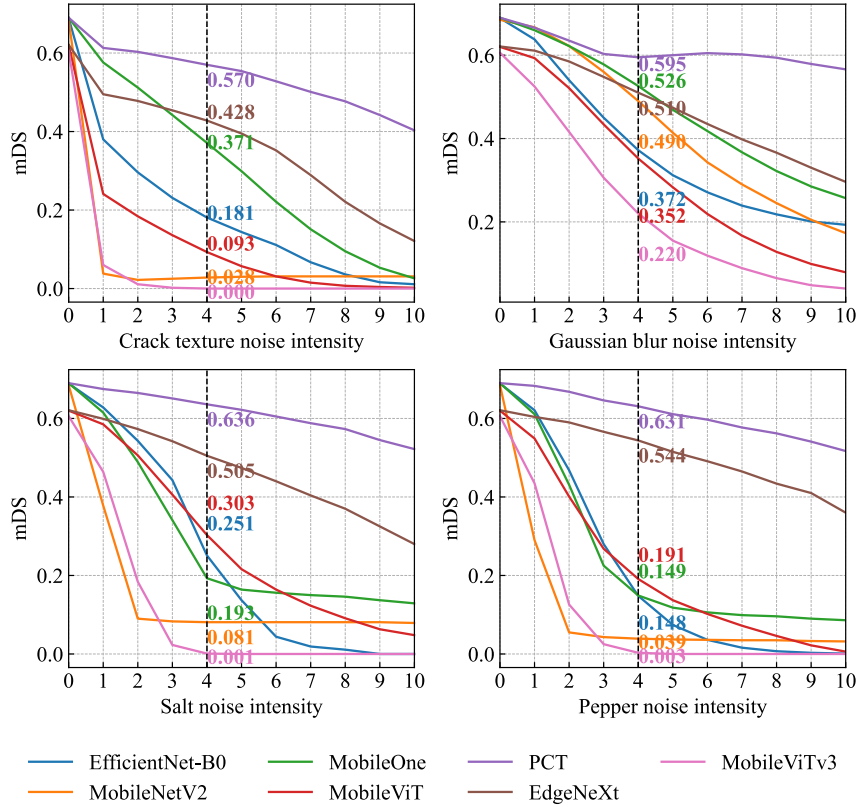


Figure 3.13 Model robustness comparison in the MC448 dataset.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

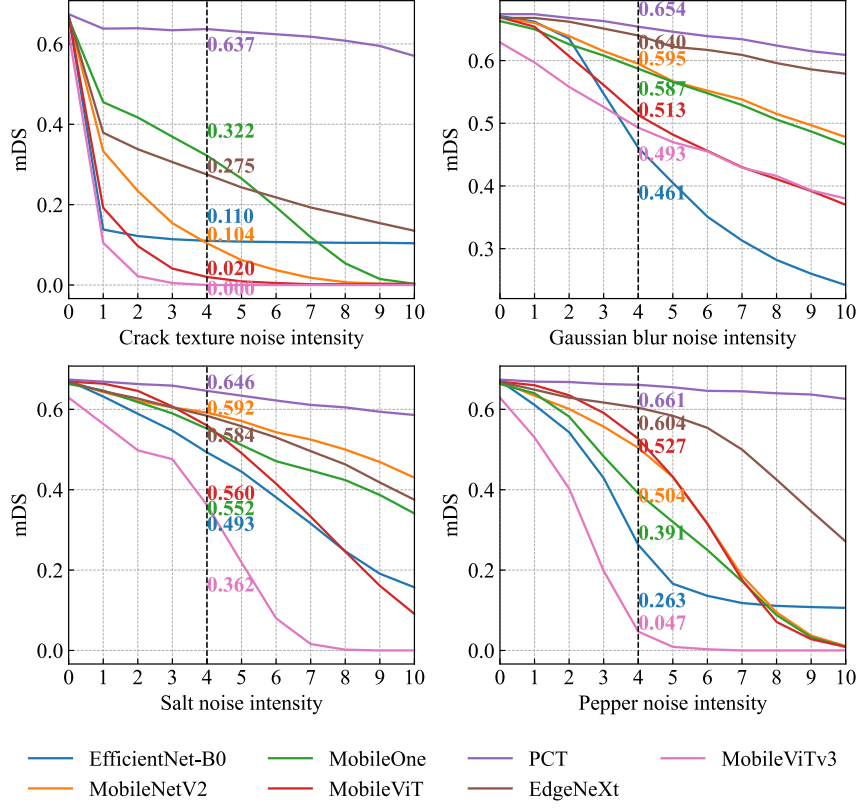


Figure 3.14 Model robustness comparison in the Crack500 dataset.

We also conducted the cross-dataset testing for the lightweight models (see Figure 3.15 and Figure 3.16). The distilled PCT model showed the improved accuracy and robustness compared to PCT model trained from scratch (see the item of ‘Student from scratch’ in Figure 3.15 and Figure 3.16) in all the cross-dataset testing, further validating the performance of RFKD. Yet, despite the distilled PCT presenting comparable robustness, its accuracy in clean data (noise intensity equal to zero) dropped in both cross-dataset testing. This accuracy drop would be because this cross-domain adaptation task did not provide out-of-domain data for the Transformer-based PCT to train its self-attention module, unlike CNN-based models, which could have comparable performance in clean out-of-domain data due to the inductive bias (Dosovitskiy, et al., 2020). Yet, the inductive bias of CNN could undermine its anti-noise robustness (Shamsabadi, et al., 2022). When trained and tested on the same domain, PCT could achieve the better overall performance compared to other lightweight models, as shown in Table 3.4, Figure 3.13, and Figure 3.14.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

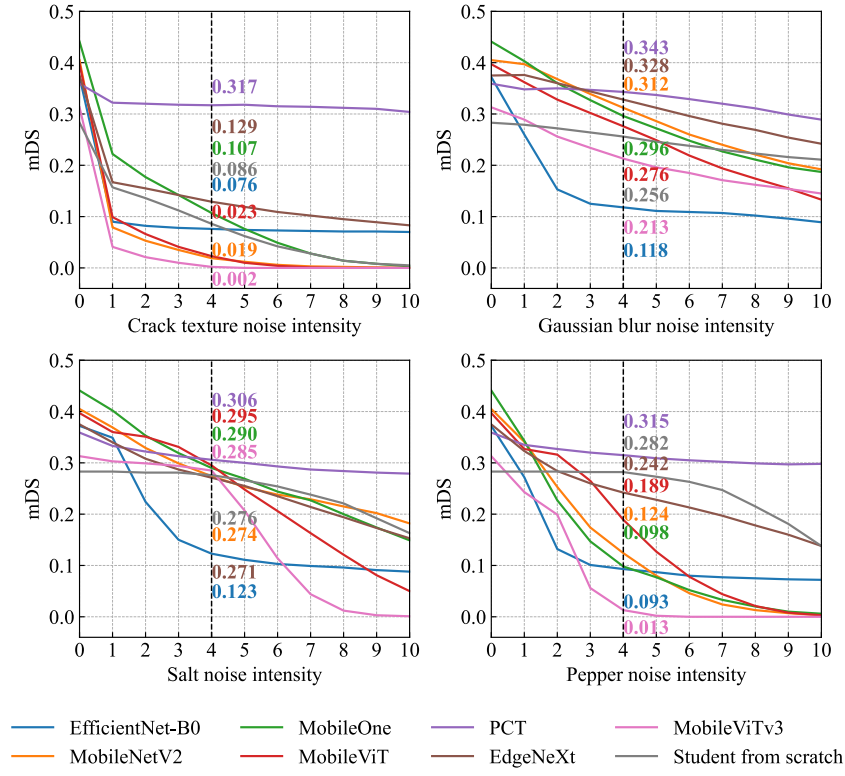


Figure 3.15 Robustness comparison for different models (training in MC448 and testing in Crack500).

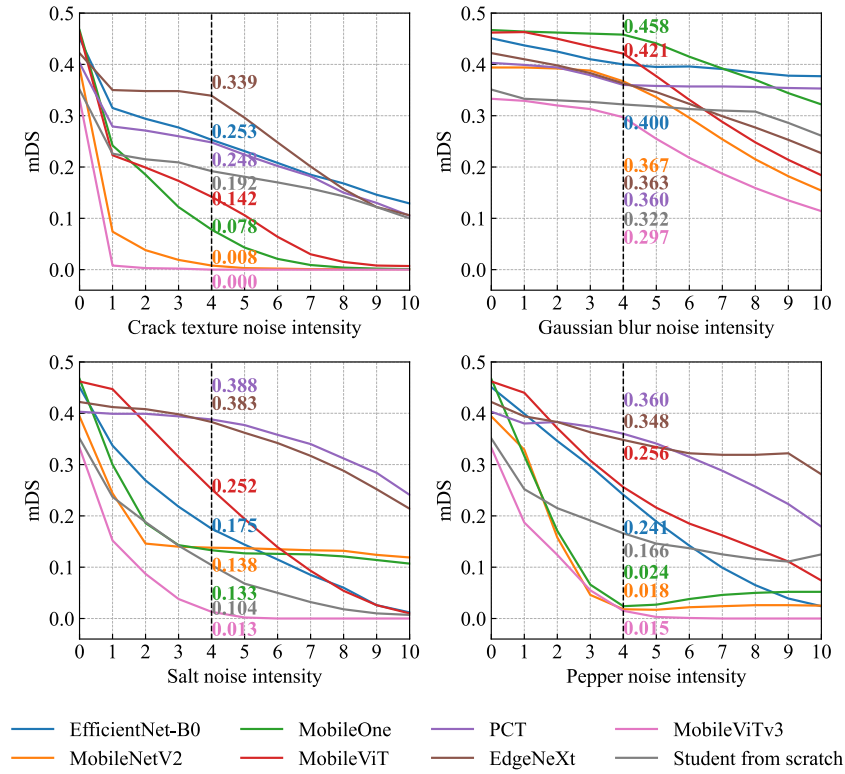


Figure 3.16 Robustness comparison for different models (training in Crack500 and testing in MC448).

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

Additional validation was performed on two challenging real-world distortions, namely motion blur and illumination change (ranging from darkening to brightening, intensity 1–10), to further assess the general applicability of RFKD. As illustrated in Figure 3.17, RFKD consistently improves the robustness of the PCT model across the entire distortion range, with particularly notable gains under severe motion blur where the baseline model exhibits substantial performance degradation. The qualitative examples in Figure 3.18 further demonstrate that the RFKD-distilled PCT retains crack structures even under strong luminance change and blur conditions, successfully identifying crack patterns that are visually ambiguous and, in some cases, difficult for human observers to discern. These results support the broader effectiveness of RFKD beyond synthetic noise types, highlighting its potential for deployment under diverse real-world imaging variations.

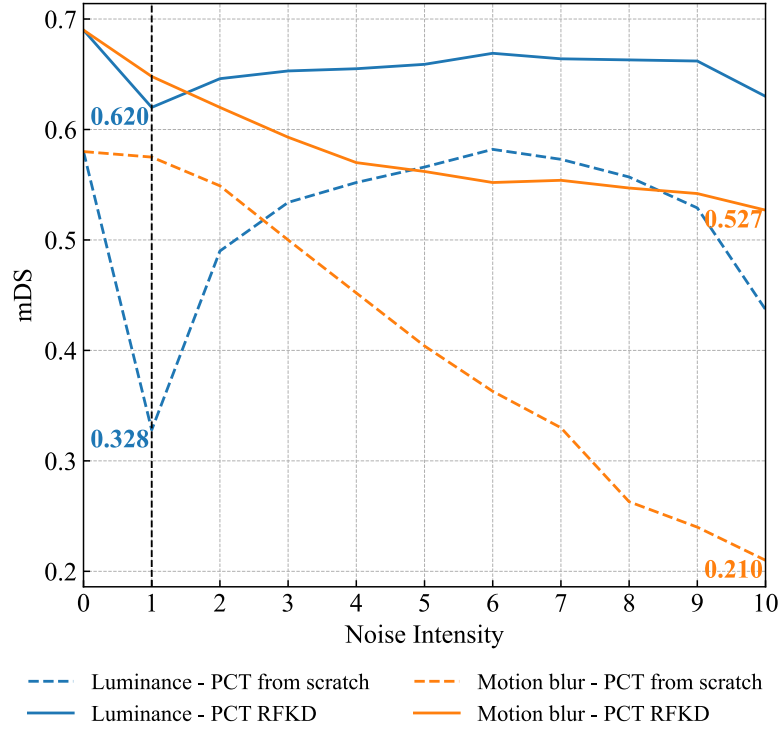


Figure 3.17 Quantitative robustness results of RFKD under motion blur and illumination change distortions.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

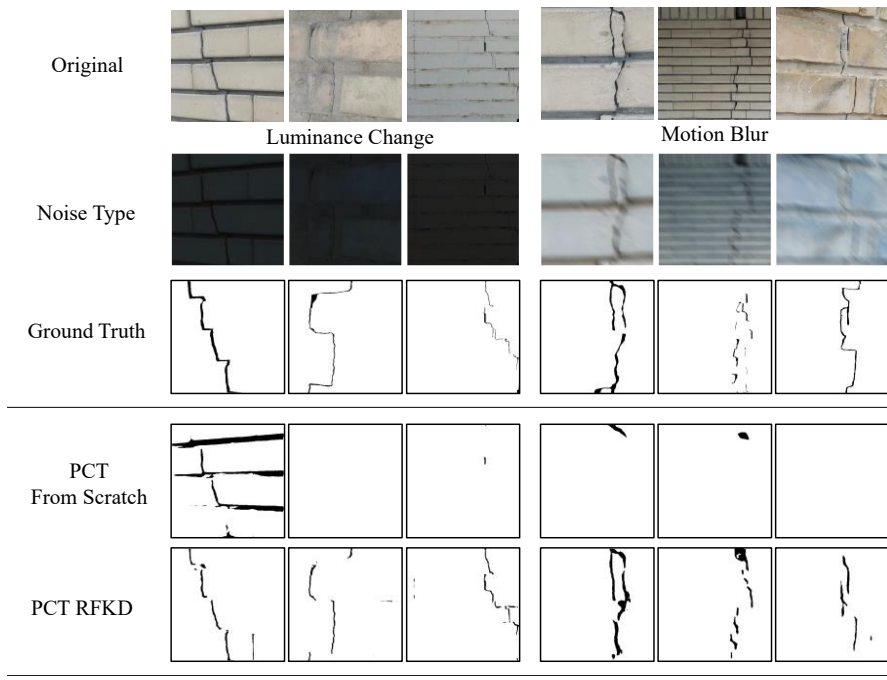


Figure 3.18 Qualitative robustness results of RFKD under motion blur and illumination change distortions

3.A.2 Sensitivity analysis of hyperparameters

Due to focusing on explaining the RFKD framework in the manuscript, we used the specific case of hyperparameter setting based on initial comparison experiments: the ratio of clean to noisy images (2:1); $\alpha_i = 0.25$ ($i = 1, 2, 3, 4$), $\beta = 1$, $\gamma = 1$, for the loss function (Eq. 1). This hyperparameter setting would not be the optimal choice but contributed to the comparable performance of RFKD framework. In this section, we fully explored the hyperparameter setting through the experiments in MC448 and Crack500 dataset.

For the ratio of clean to noisy images in the training set, four cases were selected:

- Case 1: the ratio of clean to noisy images 2:1
- Case 2: the ratio of clean to noisy images 1:1
- Case 3: the ratio of clean to noisy images 1:2
- Case 4: full noise

The comparison results for the different cases are shown in Figure 3.19 and Figure 3.20. Overall, the ratio settings of clean to noisy images in Cases 1 have comparable performance under four noise disturbances. Since the anti-noise robustness is not very sensitive to the ratio settings of clean to noisy images, and the high noise ratio would slightly undermine the model performance in clean data, the ratio in Case 1 makes a proper balance for model performance

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

in clean and noisy data. Given that the optimal ratio of clean to noisy images would vary in different datasets due to the various data distributions, tuning experiments are always required to confirm the best hyperparameter. Based on the results of our experiments, we suggest a rule-of-thumb ratio of clean to noisy image (2:1) for future studies.

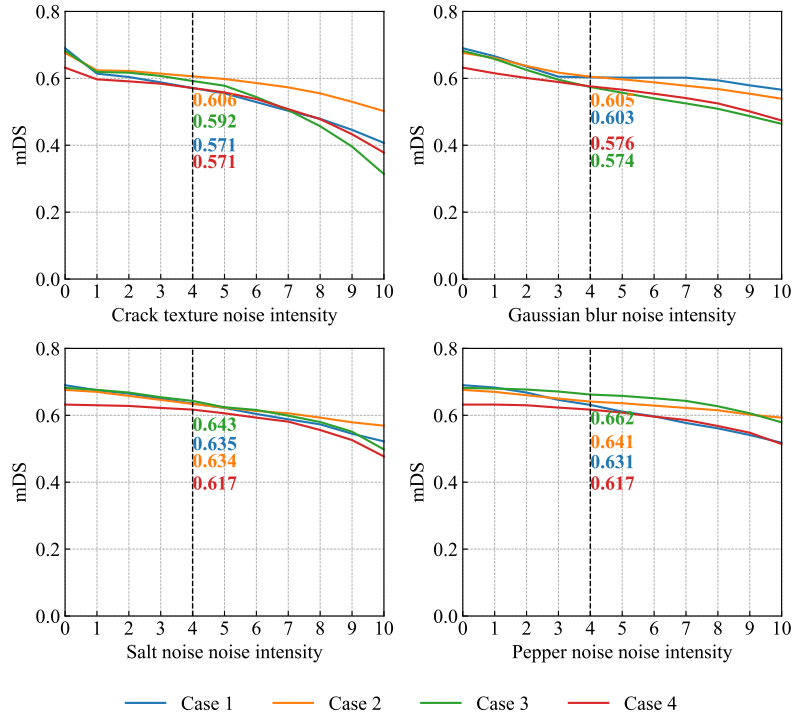


Figure 3.19 The comparison results of the different ratios of clean to noisy images in MC448 dataset.

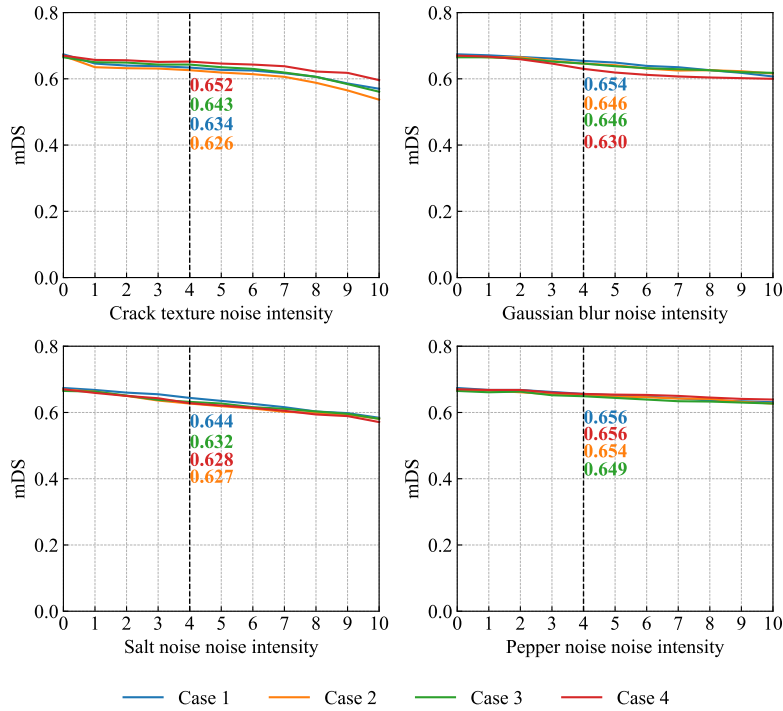


Figure 3.20 The comparison results of the different ratios of clean to noisy images in Crack500 dataset.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

For the hyperparameter setting in the loss function, four cases were selected:

- Case 1: $\alpha_1 = 0.25, \alpha_2 = 0.25, \alpha_3 = 0.25, \alpha_4 = 0.25, \beta = 1, \gamma = 1$
- Case 2: $\alpha_1 = 0.1, \alpha_2 = 0.2, \alpha_3 = 0.3, \alpha_4 = 0.4, \beta = 1, \gamma = 1$
- Case 3: $\alpha_1 = 0.25, \alpha_2 = 0.25, \alpha_3 = 0.25, \alpha_4 = 0.25, \beta = 10, \gamma = 1$
- Case 4: $\alpha_1 = 0.25, \alpha_2 = 0.25, \alpha_3 = 0.25, \alpha_4 = 0.25, \beta = 1, \gamma = 10$

The comparison result for the different cases is exhibited in Figure 3.21 and Figure 3.22. The hyperparameter setting in Case 1 shows the best overall performance in both datasets. As mentioned previously, the hyperparameter setting in Case 1 is recommended as the rule-of-thumb. Yet, the optimal hyperparameters still need to be confirmed through tuning experiments in future studies.

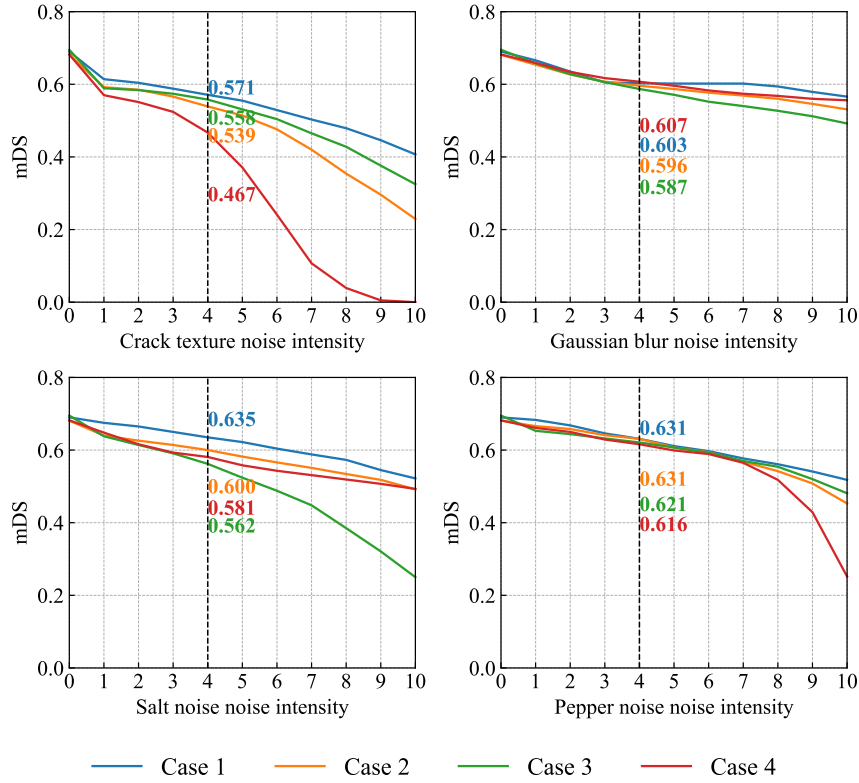


Figure 3.21 The comparison results of the different hyperparameter settings in the loss function in MC448 dataset.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

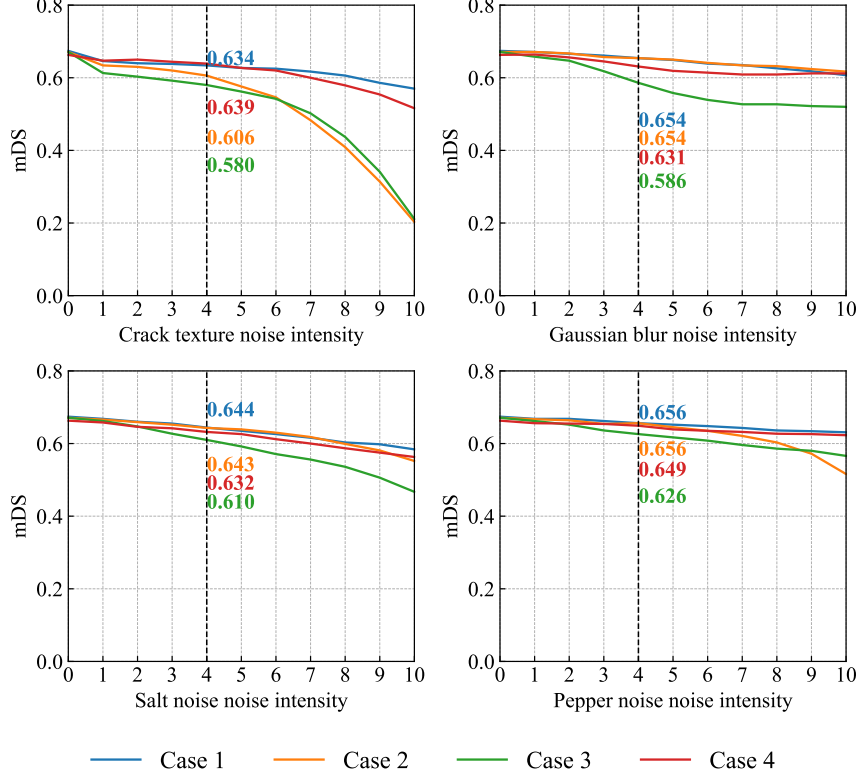


Figure 3.22 The comparison results of the different hyperparameter settings in the loss function in Crack500 dataset.

3.A.3 The architecture of the teacher model

Since the PoolingCrack model (Chen, et al., 2024) has demonstrated advanced performance in several crack image datasets, it was chosen as the teacher model in this study. Figure 3.23 and Figure 3.24 present the overall and detailed structure of the PoolingCrack model. This model consists of four components, including convolutional patch embedding (Wu, et al., 2021), pooling blocks (Yu, et al., 2022), feature selection modules (Huang, et al., 2021) and feature alignment modules (Dai, et al., 2017), with four down-sampling and up-sampling stages. The details of the pooling blocks, feature selection modules and feature alignment modules are introduced in Figure 3.24. Further details about the model can be found in (Chen, et al., 2024).

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

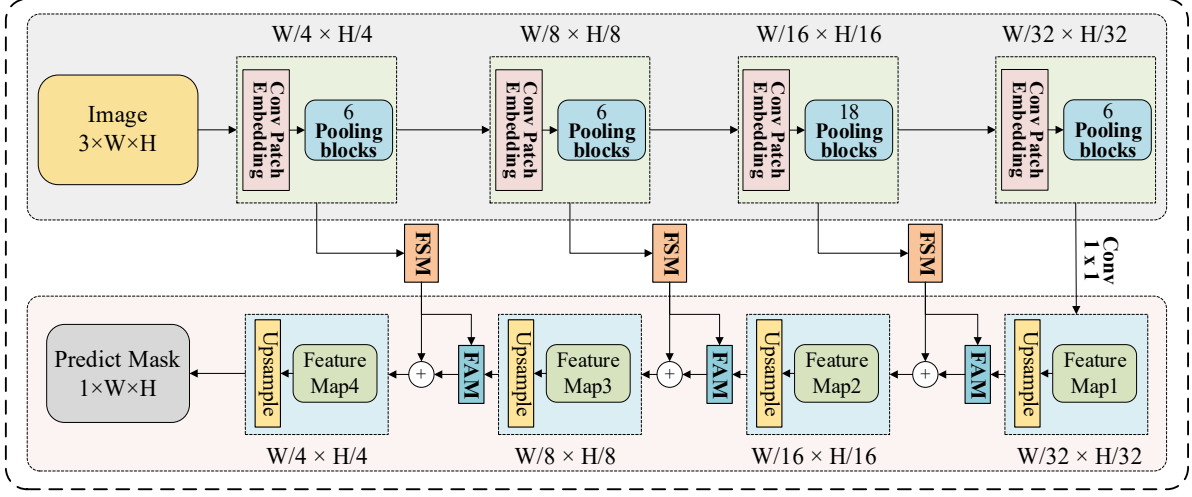


Figure 3.23 The overall architecture of the PoolingCrack model.

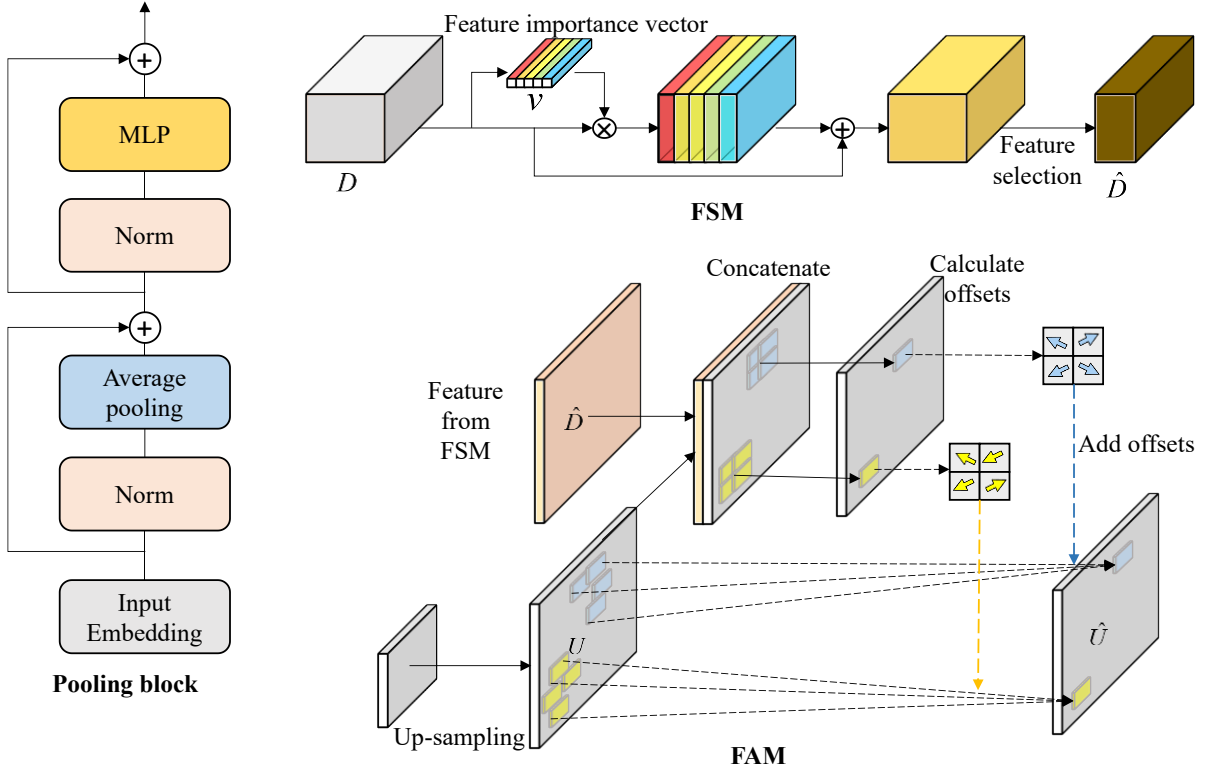


Figure 3.24 The details of the component parts in the PoolingCrack model.

3.A.4 Ablation study

To disentangle the respective contributions of knowledge distillation and noisy-data training, we introduced an ablated variant of our framework by disabling the mixed clean-noisy input path while retaining all distillation losses. This variant, referred to as FKD, isolates the effect of feature-based distillation alone. Figure 3.25 compares the robustness of FKD with that of RFKD and other KD strategies. Although FKD achieves higher robustness than NKD, CWD,

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

and DIST, indicating that feature-based knowledge distillation itself provides measurable robustness benefits, its performance still degrades rapidly as noise intensity increases. In contrast, the full RFKD framework consistently maintains superior mDS values across all four noise types. These findings demonstrate that feature-based knowledge distillation contributes to robustness but cannot fully account for the improvements, and that the mixed clean-noisy training strategy plays a crucial and complementary role in enabling the student model to learn noise-resilient feature patterns.

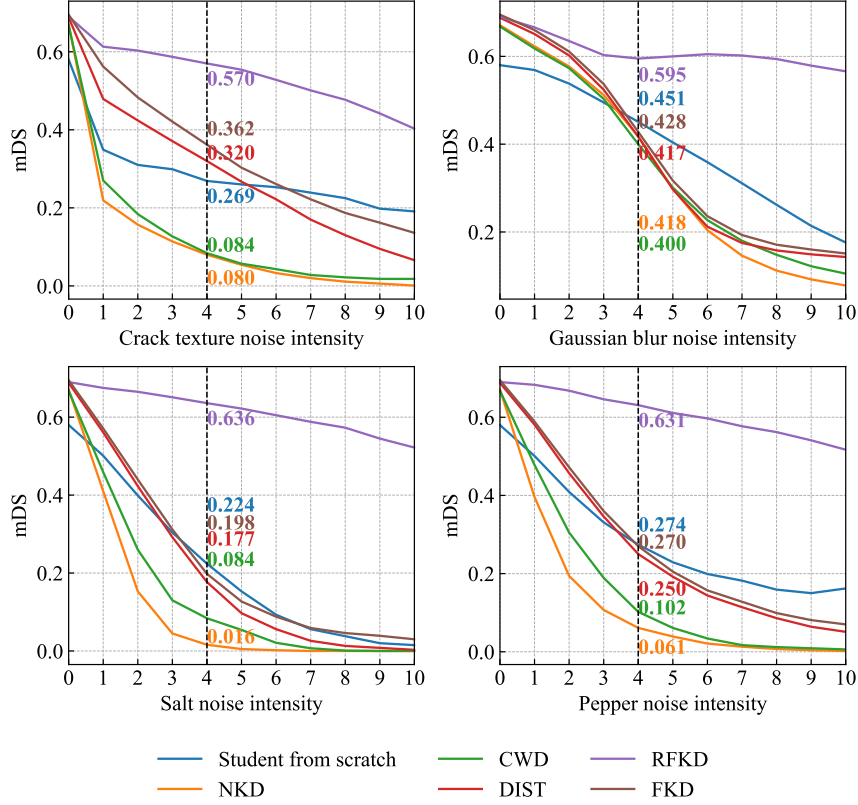


Figure 3.25 Robustness comparison for ablated FKD and other strategies.

Since the proposed RFKD has been demonstrated to be available for the designed Transformer-based PCT lightweight student model, another validation is conducted to assess RFKD's general performance for other student models. The CNN-based UNet-Resnet18 (Ronneberger, et al., 2015) and MobileOne (Vasu, et al., 2023) models were chosen for this verification, utilising the same training configuration and RFKD strategy as the PCT model. Table 3.5 and Figure 3.26 illustrate the accuracy and robustness comparison of the UNet-ResNet18, MobileOne and PCT students.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

Table 3.5 Accuracy comparison under the RFKD strategy.

Student Model	Parameters (M)	Training from scratch		RFKD	
		mDS (%)	mIoU (%)	mDS (%)	mIoU (%)
UNet-ResNet18	14.3	61.8	50.7	68.1 (+6.3)	57.2 (+6.5)
MobileOne	6.3	64.8	53.2	68.6 (+3.8)	57.9 (+4.7)
PCT	0.5	58.0	46.2	69.0 (+11.0)	57.7 (+11.5)

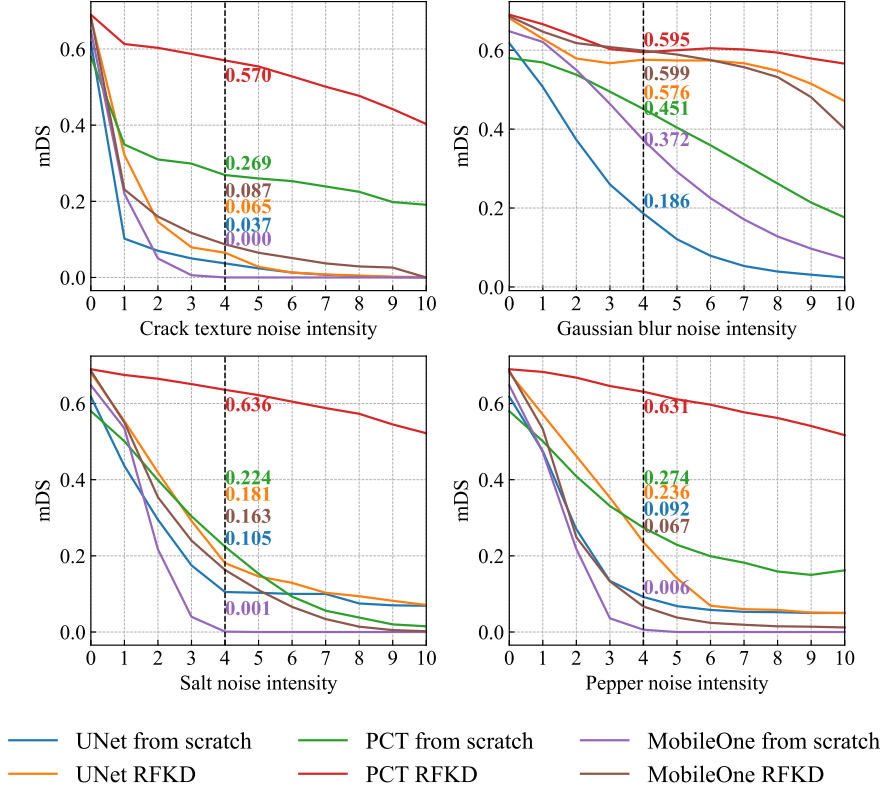


Figure 3.26 Robustness comparison under the RFKD strategy.

As presented in Table 3.5 and Figure 3.26, the accuracy (mDS/mIoU) of PCT is comparable to the other two models under clean test conditions, while PCT maintains significantly higher robustness and lower performance degradation than both UNet and MobileOne under noisy and perturbed conditions (e.g., Crack texture, salt, pepper noise). In terms of computational efficiency, PCT requires substantially fewer parameters and FLOPs while achieving consistent results. These findings demonstrate that PCT not only preserves competitive accuracy but also provides superior robustness and efficiency that the other models can typically achieve only partially (either accuracy-efficiency or accuracy-robustness, but not all three). Yet, the proposed RFKD method still enhances the robustness of UNet-ResNet18 and MobileOne (see Figure 3.26), yielding substantial improvements under Gaussian blur distortion and mild gains

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

across the remaining three noise types. Since noises disrupt the local details of images, and CNN-based models have the inherent limitation of fixed-size convolutional kernels and local attention, they are weaker than similar-scale Transformer-based models when processing noisy images (Dosovitskiy, et al., 2020). The RFKD framework helps the CNN-based UNet-ResNet18 and MobileOne learn noise-invariant features from the teacher model, enhancing the student's ability to handle noisy images. Yet, the instinctive shortcoming of CNNs restricts the further improvement of their robustness through the RFKD method.

References

- Adriana R., B. Nicolas, K.S. Ebrahimi, C. Antoine, G. Carlo, B. Yoshua, Fitnets: Hints for thin deep nets, *International Conference on Learning Representations* 2 (3) (2015), pp. 1.
- Ali R., Y.-J. Cha, Attention-based generative adversarial network with internal damage segmentation using thermography, *Automation in Construction* 141 (2022), pp. 104412.
- Asadi Shamsabadi E., S.M.H. Erfani, C. Xu, D. Dias-da-Costa, Efficient semi-supervised surface crack segmentation with small datasets based on consistency regularisation and pseudo-labelling, *Automation in Construction* 158 (2024), pp. 105181.
- Bengio Y., A. Courville, P. Vincent, Representation learning: A review and new perspectives, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (8) (2013), pp. 1798-1828.
- Böttcher A., D. Wenzel, The Frobenius norm and the commutator, *Linear Algebra and Its Applications* 429 (8-9) (2008), pp. 1864-1885.
- Canny J., A computational approach to edge detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (6) (1986), pp. 679-698.
- Chen J., Y. Liu, J.-a. Hou, A lightweight deep learning network based on knowledge distillation for applications of efficient crack segmentation on embedded devices, *Structural Health Monitoring* (2023), pp. 14759217221139730.
- Chen Z., E. Asadi Shamsabadi, S. Jiang, L. Shen, D. Dias-da-Costa, An average pooling designed Transformer for robust crack segmentation, *Automation in Construction* 162 (2024), pp. 105367.
- Choi W., Y.J. Cha, SDDNet: real-time crack segmentation, *IEEE Transactions on Industrial Electronics* 67 (9) (2020), pp. 8016-8025.
- Chollet F., Xception: Deep learning with depthwise separable convolutions, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), pp. 1251-1258.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

Chu H., P.j. Chun, Fine-grained crack segmentation for high-resolution images via a multiscale cascaded network, *Computer-Aided Civil and Infrastructure Engineering* (2023).

Dai J., H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, *Proceedings of the IEEE International Conference on Computer Vision*, (2017), pp. 764-773.

Deng J., W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), pp. 248-255.

Dong X., C. Zhang, J. Wang, Y. Chen, D. Wang, Real-time detection of surface cracking defects for large-sized stamped parts, *Computers in Industry* 159 (2024), pp. 104105.

Dosovitskiy A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint* (2020).

Euzenat J., Semantic precision and recall for ontology alignment evaluation, *20th International Joint Conference on Artificial Intelligence* (2007), pp. 348-353.

Gao L., J. Zhang, C. Yang, Y. Zhou, Cas-VSwin transformer: A variant swin transformer for surface-defect detection, *Computers in Industry* 140 (2022), pp. 103689.

Gardner M.W., S. Dorling, Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences, *Atmospheric Environment* 32 (14-15) (1998), pp. 2627-2636.

Gou J., B. Yu, S.J. Maybank, D. Tao, Knowledge distillation: A survey, *International Journal of Computer Vision* 129 (2021), pp. 1789-1819.

Hinton G., O. Vinyals, J. Dean, Distilling the knowledge in a neural network, *arXiv preprint* (2015).

Huang S., Z. Lu, R. Cheng, C. He, FaPN: feature-aligned pyramid network for dense image prediction, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), pp. 864-873.

Huang T., S. You, F. Wang, C. Qian, C. Xu, Knowledge distillation from a stronger teacher, *Advances in Neural Information Processing Systems* 35 (2022), pp. 33716-33727.

Jaccard P., The distribution of the flora in the alpine zone. 1, *New phytologist* 11 (2) (1912), pp. 37-50.

Jha S.B., R.F. Babiceanu, Deep CNN-based visual defect detection: Survey of current literature, *Computers in Industry* 148 (2023), pp. 103911.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

Jiang S., Z. Cheng, L. Yang, L. Shen, An auto-tuned hybrid deep learning approach for predicting fracture evolution, *Engineering with Computers* (2022).

Kang D.H., Y.-J. Cha, Efficient attention-based deep encoder and decoder for automatic crack segmentation, *Structural Health Monitoring* 21 (5) (2021), pp. 2190-2205.

Karimi N., M. Mishra, P.B. Lourenço, Automated Surface Crack Detection in Historical Constructions with Various Materials Using Deep Learning-Based YOLO Network, *International Journal of Architectural Heritage*, pp. 1-17.

Karimi N., M. Mishra, P.B. Lourenço, Deep learning-based automated tile defect detection system for Portuguese cultural heritage buildings, *Journal of Cultural Heritage* 68 (2024), pp. 86-98.

Kirschke K.R., S.A. Velinsky, Histogram-based approach for automated pavement-crack sensing, *Journal of Transportation Engineering* 118 (5) (1992), pp. 700-710.

Krizhevsky A., I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems* 25 (2012).

Lan Z.-X., X.-M. Dong, MiniCrack: A simple but efficient convolutional neural network for pixel-level narrow crack detection, *Computers in Industry* 141 (2022), pp. 103698.

Li X., X. Sun, Y. Meng, J. Liang, F. Wu, J. Li, Dice loss for data-imbalanced NLP tasks, *arXiv preprint* (2019).

Li Z., P. Xu, X. Chang, L. Yang, Y. Zhang, L. Yao, X. Chen, When object detection meets knowledge distillation: A survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

Liu Y., K. Chen, C. Liu, Z. Qin, Z. Luo, J. Wang, Structured knowledge distillation for semantic segmentation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), pp. 2604-2613.

Maaz M., A. Shaker, H. Cholakkal, S. Khan, S.W. Zamir, R.M. Anwer, F. Shahbaz Khan, Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications, *European conference on computer vision*, Springer, (2022), pp. 3-20.

Maroto J., G. Ortiz-Jiménez, P. Frossard, On the benefits of knowledge distillation for adversarial robustness, *arXiv preprint* (2022).

Mehta S., M. Rastegari, Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer, *arXiv preprint* (2021).

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

- Mohammadkhorasani A., K. Malek, R. Mojidra, J. Li, C. Bennett, W. Collins, F. Moreu, Augmented reality-computer vision combination for automatic fatigue crack detection and localization, *Computers in Industry* 149 (2023), pp. 103936.
- Paszke A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimeshein, L. Antiga, Pytorch: an imperative style, high-performance deep learning library, *Advances in Neural Information Processing Systems* 32 (2019).
- Patidar P., M. Gupta, S. Srivastava, A.K. Nagawat, Image de-noising by various filters for different noise, *International Journal of Computer Applications* 9 (4) (2010), pp. 45-50.
- Qin F., L. Zhu, L. Cao, W. Yang, Blind Single-Image Super Resolution Reconstruction with Gaussian Blur and Pepper & Salt Noise, *Journal of Computers* 9 (4) (2014), pp. 896-902.
- Ronneberger O., P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, *Medical Image Computing and Computer-Assisted Intervention–18th International Conference, Munich, Germany Springer*, (2015), pp. 234-241.
- Rubio J.J., T. Kashiwa, T. Laiteerapong, W. Deng, K. Nagai, S. Escalera, K. Nakayama, Y. Matsuo, H. Prendinger, Multi-class structural damage segmentation using fully convolutional networks, *Computers in Industry* 112 (2019), pp. 103121.
- Sandler M., A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), pp. 4510-4520.
- Shamsabadi E.A., C. Xu, D. Dias-da-Costa, Robust crack detection in masonry structures with Transformers, *Measurement* 200 (2022), pp. 111590.
- Shamsabadi E.A., C. Xu, A.S. Rao, T. Nguyen, T. Ngo, D. Dias-da-Costa, Vision transformer-based autonomous crack detection on asphalt and concrete surfaces, *Automation in Construction* 140 (2022), pp. 104316.
- Shao R., J. Yi, P.-Y. Chen, C.-J. Hsieh, How and when adversarial robustness transfers in knowledge distillation?, *arXiv preprint* (2021).
- Shelke M.S., P.R. Deshmukh, V.K. Shandilya, A review on imbalanced data handling using undersampling and oversampling technique, *International Journal of Recent Trends in Engineering & Research* 3 (4) (2017), pp. 444-449.
- Shu C., Y. Liu, J. Gao, Z. Yan, C. Shen, Channel-wise knowledge distillation for dense prediction, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), pp. 5311-5320.
- Sudre C.H., W. Li, T. Vercauteren, S. Ourselin, M. Jorge Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, *Deep Learning in Medical*

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, and 7th International Workshop, Québec City, QC, Canada, Springer, (2017), pp. 240-248.

Tan M., Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, International Conference on Machine Learning, PMLR, (2019), pp. 6105-6114.

Vasu P.K.A., J. Gabriel, J. Zhu, O. Tuzel, A. Ranjan, MobileOne: An Improved One Millisecond Mobile Backbone, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2023), pp. 7907-7917.

Wadekar S.N., A. Chaurasia, Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features, arXiv preprint arXiv:2209.15159 (2022).

Wang H., Y. Li, L.M. Dang, S. Lee, H. Moon, Pixel-level tunnel crack segmentation using a weakly supervised annotation approach, Computers in Industry 133 (2021), pp. 103545.

Wang L., K.-J. Yoon, Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (6) (2021), pp. 3048-3068.

Wang W., C. Su, G. Han, H. Zhang, A lightweight crack segmentation network based on knowledge distillation, Journal of Building Engineering 76 (2023), pp. 107200.

Wang Y., W. Zhou, T. Jiang, X. Bai, Y. Xu, Intra-class feature variation distillation for semantic segmentation, European Conference on Computer Vision, Springer, (2020), pp. 346-362.

Wu H., B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, Cvt: introducing convolutions to vision transformers, Proceedings of the IEEE/CVF International Conference on Computer Vision, (2021), pp. 22-31.

Xu X., S. Xu, L. Jin, E. Song, Characteristic analysis of Otsu threshold and its applications, Pattern Recognition Letters 32 (7) (2011), pp. 956-961.

Xu Y., Y. Fan, H. Li, Lightweight semantic segmentation of complex structural damage recognition for actual bridges, Structural Health Monitoring (2023), pp. 14759217221147015.

Yamaguchi T., S. Hashimoto, Fast crack detection method for large-size concrete surface images using percolation-based image processing, Machine Vision and Applications 21 (5) (2010), pp. 797-809.

Yu L., V.O. Yazici, X. Liu, J.v.d. Weijer, Y. Cheng, A. Ramisa, Learning metrics from teachers: Compact networks for image embedding, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2019), pp. 2907-2916.

3. Robust feature knowledge distillation for noise-resilient and deployable defect detection models in local-scale perception

Yu W., M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, S. Yan, Metaformer is actually what you need for vision, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), pp. 10819-10829.

Zhang L., F. Yang, Y.D. Zhang, Y.J. Zhu, Road crack detection using deep convolutional neural network, *2016 IEEE International Conference on Image Processing (2016)*, pp. 3708-3712.

Zhou W., F. Sun, Q. Jiang, R. Cong, J.-N. Hwang, WaveNet: Wavelet network with knowledge distillation for RGB-T salient object detection, *IEEE Transactions on Image Processing* 32 (2023), pp. 3027-3039.

Zhou W., H. Wu, Q. Jiang, MDNet: Mamba-effective diffusion-distillation network for RGB-thermal urban dense prediction, *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

Zhou W., X. Yang, W. Yan, Q. Jiang, Hybrid knowledge distillation for RGB-T crowd density estimation in smart surveillance systems, *IEEE Internet of Things Journal* (2024).

Zhou W., B. Jian, Y. Liu, Feature contrast difference and enhanced network for RGB-D indoor scene classification in internet of things, *IEEE Internet of Things Journal* (2025).

Zhou W., B. Jian, Y. Liu, Q. Jiang, Multiattentive Perception and Multilayer Transfer Network Using Knowledge Distillation for RGB-D Indoor Scene Parsing, *IEEE Transactions on Neural Networks and Learning Systems* (2025).

Zhou W., Y. Wang, X. Qian, Knowledge distillation and contrastive learning for detecting visible-infrared transmission lines using separated stagger registration network, *IEEE Transactions on Circuits and Systems I: Regular Papers* (2025).

Zhu G., J. Liu, Z. Fan, D. Yuan, P. Ma, M. Wang, W. Sheng, K.C. Wang, A lightweight encoder–decoder network for automatic pavement crack detection, *Computer-Aided Civil and Infrastructure Engineering* (2023).

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

This chapter presents the manuscript “Zhaohui Chen, Elyas Asadi Shamsabadi, Sheng Jiang, Luming Shen, Daniel Dias-da-Costa. (2024). Vision Mamba-based autonomous crack segmentation on concrete, asphalt, and masonry surfaces. arXiv:2406.16518”, which investigates the use of state-space-based visual architectures for efficient and scalable high-resolution defect inspection.

The chapter introduces a Vision Mamba (VMamba)-based encoder-decoder framework that leverages linear-complexity state-space modelling to address the computational limitations of CNN- and Transformer-based architectures in high-resolution visual inspection. By replacing quadratic-complexity self-attention with selective state-space scanning, the proposed architecture enables efficient long-range dependency modelling while maintaining strong segmentation accuracy and generalisation performance.

Comprehensive experimental evaluations demonstrate that the VMamba-based models achieve performance comparable to Transformer-based approaches with significantly reduced model parameters, computational cost, and floating-point operations, particularly under high-resolution input conditions. This chapter establishes a scalable and computationally efficient local-scale perception framework, supporting further practical deployment for high-resolution infrastructure inspection.

4.1 Introduction

A timely detection of defects in structural elements is crucial for ensuring the safety and longevity of infrastructure. The early identification of deficiencies can allow for an early intervention, thus reducing the likelihood of major failures. Nonetheless, the traditionally employed manual inspection methods have inherent drawbacks, including but not limited to being time-consuming, labour-intensive, and subject to human errors and biases (Sarkar, et al., 2024). Moreover, manual inspections can be hazardous, particularly when assessing hard-to-reach or unsafe areas. The subjective nature of human judgment in identifying defects may also lead to inconsistencies and inaccuracies in assessments (Yang, et al., 2022). As such, there is a growing need for automated and efficient defect detection techniques. Technologies such as computer vision (Ai, et al., 2023; Deng, et al., 2022), machine learning (ML) (Shiuly, et al.,

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

2022), and remotely operated vehicles can streamline inspection and enhance infrastructure safety (Tran, et al., 2023; Xiong, et al., 2024).

The capabilities of ML for defect detection over various surfaces with differing levels of complexity, including concrete (Han, et al., 2022), asphalt (Nguyen, et al., 2023), and masonry (Shamsabadi, et al., 2022), have been assessed with a series of works using classic and deep learning models (Shi, et al., 2024). Initially, classic machine learning models, such as random forest and multilayer perceptron, were deployed to develop frameworks mainly for crack classification and segmentation. Although the developed frameworks could have significant accuracy in controlled settings, their generalisation was limited as each pixel is treated individually, while cracks are continuous patterns over the surface. Because of this, such frameworks cannot use the spatial connectivity between defect pixels (Cha, et al., 2017). Convolutional neural networks (CNNs) enhanced crack detection by allowing models to use local crack pixel dependencies (Shamsabadi, et al., 2022). By building on convolutional operations, CNNs can extract the local pixel connectivity, thus enhancing the accuracy and applicability in more practical settings. Still, CNNs show a bias towards the texture. CNN-based models (Zhang, et al., 2023) use texture-based information more than shape-based cues, limiting their performance on unseen data. This contrasts with the human ability to maintain a balance between shape and texture, which can vary depending on the situation (Geirhos, et al., 2018; Shamsabadi, et al., 2022).

Transformers can address the bias towards the texture by incorporating self-attention mechanisms to capture multi-range interdependencies in diverse data types, including images (Dosovitskiy, et al., 2020). Self-attention calculates the importance of each element in a sequence by considering its relationship with all other elements, thus allowing the model to focus on different parts of the input as needed (Chen, et al., 2024). Transformers have demonstrated generalisation capabilities even in scenarios involving manipulated or noisy images (Chen, et al., 2024; Shamsabadi, et al., 2022; Shamsabadi, et al., 2022). Nonetheless, the self-attention mechanism of Transformers requires computing the pairwise attention scores between all patches, which leads to a significant computational cost with the long flattened patch embedding length (Mohammadi Farsani and Pazouki, 2020). This computational cost of the self-attention module poses deployment challenges on edge devices with limited processing power and memory, such as drones and mobile phones, thus limiting the development of efficient and effective crack detection solutions.

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

To address the computational challenges posed by the Transformer's self-attention mechanism, several architectures have been developed to decrease the computational cost. The linear Transformer reduces the complexity of self-attention from quadratic to linear by approximating the SoftMax function and using the kernel method but with the trade-off of performance (Katharopoulos, et al., 2020). Considering the low-order complexity of recurrent neural networks (RNNs) (Schuster and Paliwal, 1997), the Structured State Space for Sequences (S4) model (Gu, et al., 2021) integrated the benefits of the Hippo framework (Gu, et al., 2020) and RNNs and then fused linear attention (Katharopoulos, et al., 2020) form the Hungry Hungry Hippos (H3) model (Fu, et al., 2022) for more enhanced efficiency and effectiveness of sequence modelling. In addition, similar to the H3 model, the retentive network (RetNet) reduces the inner S4 layer with a state dimension equal to one, which can be regarded as a specific case of a linear state space model (Sun, et al., 2023). The Mamba (Gu and Dao, 2023) architecture represents a recent further evolution in sequence modelling, enabling the integration and enhancement of the strengths of previous S4/H3 models by introducing varying time steps to optimise performance. The new-generation architecture of Mamba has enabled the development of more efficient and effective visual models (Liu, et al., 2024).

Efficient vision models are essential for crack detection, especially for real-time processing, which necessitates the integration of generalisable models on devices with limited computational capabilities. However, while recent developments in the literature can achieve human-level adaptive generalisation across several material surface cracks, they are often computationally expensive (Chen, et al., 2024). Apart from model capabilities, another essential factor contributing to the generalisation and high accuracy of models is higher input image resolutions (Chu and Chun, 2023), as they contain more contextual information and more accurate semantic representations. Yet, higher resolution requires greater computational power. These considerations necessitate the development of efficient models so that the limited computational power of edge devices can be utilised for higher-resolution processing (Chen, et al., 2024). Given the optimised performance of the state-space model for large inputs compared to previous architectures, we propose a Mamba-based framework for crack segmentation and demonstrate that, with a smaller number of parameters and floating-point operations, the proposed framework can achieve high segmentation accuracy. Our assumption is that a VMamba-based crack segmentation framework can achieve the same stable and advanced performance as Transformer-based models across different material surface cracks yet with significantly fewer parameters and reduced computational power requirements.

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

The structure of this paper is outlined as follows: Section 4.2 presents the implementation of the performance study and discusses the differences among CNN, Transformer, and Mamba. Section 4.3 details the experimental results with a focus on the analysis of the high efficiency and accuracy of the Mamba-based network. Finally, Section 4.4 draws the main conclusions and suggests future developments.

4.2 Study design

This section illustrates and discusses how the Mamba architecture operates - particularly the 2D-Select-Scan module, which is the core of the VMamba framework proposed in this paper. Subsequently, the computational complexity of CNN, Transformer, and Mamba-based architectures are analysed, highlighting the great potential of Mamba for an efficient network design. The architectures of CNN and Transformer are introduced in Appendix 4.A. A baseline VMamba-based encoder-decoder network is deployed for crack segmentation. The evaluation metrics, datasets and training settings are presented in the next parts of this section. Our implemented framework, which consists of two primary stages (training and test), is outlined in Figure 4.1. The model is first trained with the help of the Dice loss function (see Figure 4.1-a) and pre-set hyperparameters and then evaluated in three commonly utilised surface crack datasets (concrete, asphalt, and masonry) with different background textures (see Figure 4.1-b).

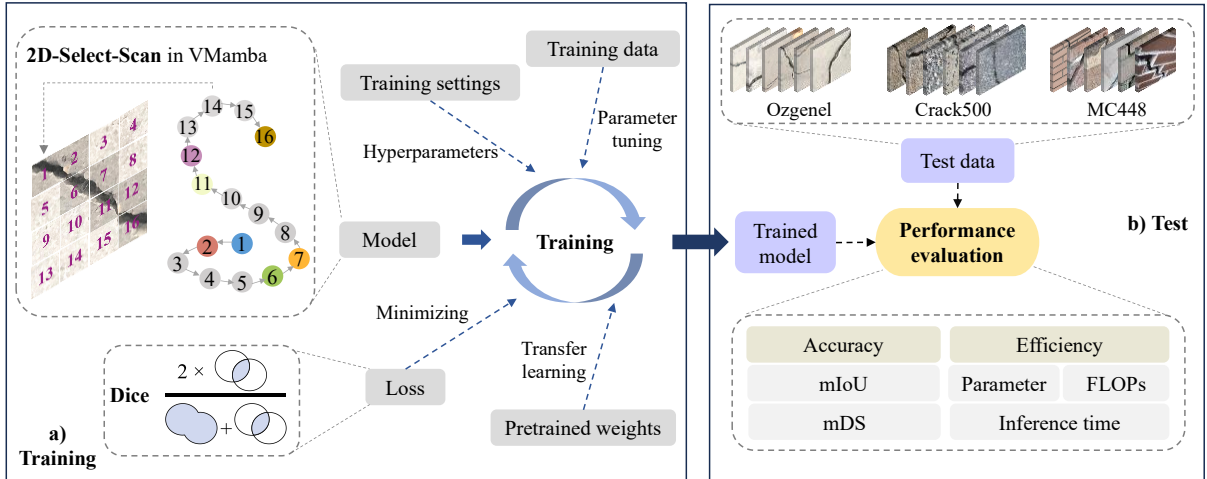


Figure 4.1 A diagram of the implemented framework.

4.2.1 The architecture of Mamba

Based on the recent Select-Scan Structured State Space for Sequences (S6) model, Mamba has been developed as a more efficient choice than CNNs and Transformers for natural language processing (Gu and Dao, 2023). Vision Mamba (VMamba), with a 2D-Select-Scan (SS2D)

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

module, has been introduced for vision tasks, showing high accuracy with a significantly enhanced efficiency (Liu, et al., 2024). As shown in Figure 4.2, the SS2D module starts with a scan expansion, including four routes of top-left to bottom-right, top-right to bottom-left, bottom-left to top-right, and bottom-right to top-left scanning, to obtain four sets of image patches. After patch embedding, each patch x is transformed into a sequence (sequence length: L , dimension: d) before going through the S6 block:

$$h_k = \bar{A} h_{k-1} + \bar{B} x_k, \quad (1)$$

$$y_k = C h_k + D x_k, \quad (2)$$

$$\bar{A} = e^{\Delta A}, \quad (3)$$

$$\bar{B} = (\Delta A)^{-1} (e^{\Delta A} - I) \cdot \Delta B, \quad (4)$$

where the matrices $B \in \mathbb{R}^{L \times H}$ and $C \in \mathbb{R}^{L \times H}$ with a hidden state size H , varying time step $\Delta \in \mathbb{R}^{L \times d}$, are obtained from the linear projection of the input sequence $x \in \mathbb{R}^{L \times d}$. The matrix $A \in \mathbb{R}^{d \times H}$ for hidden state space and $D \in \mathbb{R}^{L \times L}$ for the skip connection consist of trainable parameters that are updated during backpropagation. The outputs of the S6 block are then transformed into the original size of the input through the scan merging operation.

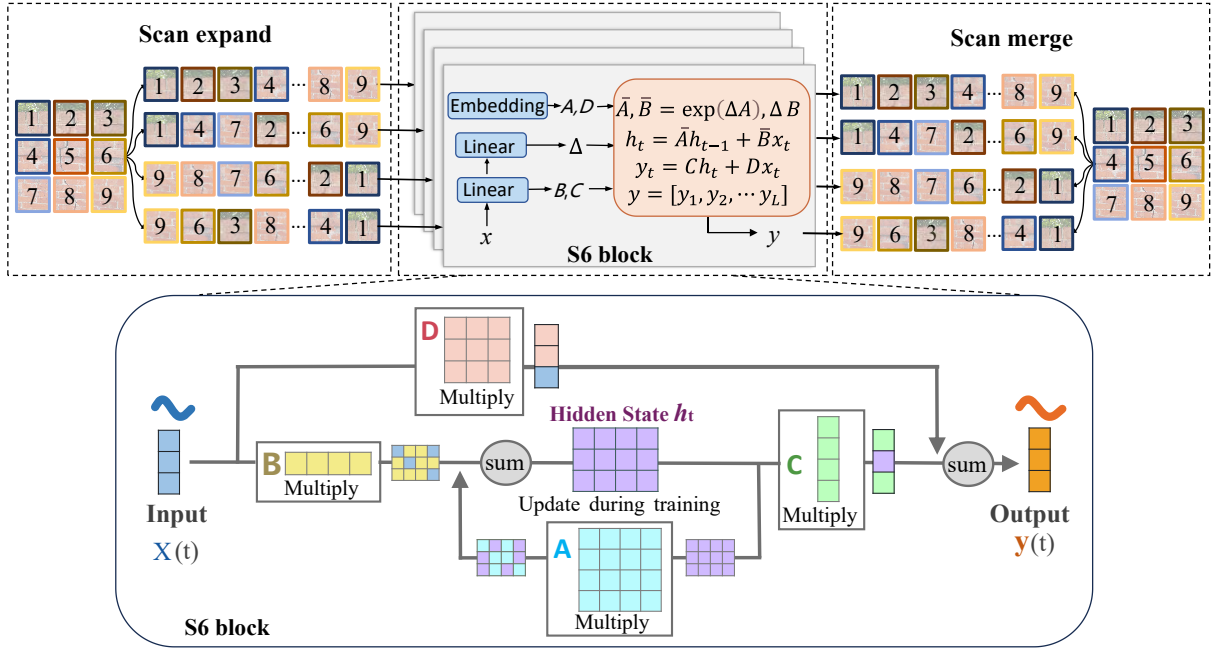


Figure 4.2 Illustration of the 2D-Select-Scan (SS2D) in VMamba.

The S6 block can be transformed to operate in another format of Q, K, and V matrices (Liu, et al., 2024). For the sequences $X := [x_a; \dots; x_b]$ with length L , the following variables are defined as:

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

$$\mathbf{V} := [\mathbf{V}_1; \dots; \mathbf{V}_L] \in \mathbb{R}^{L \times d_v}, \text{ where } \mathbf{V}_i = x_{a+i-1} \Delta_{a+i-1}; \quad (5)$$

$$\mathbf{K} := [\mathbf{K}_1; \dots; \mathbf{K}_L] \in \mathbb{R}^{L \times d_k}, \text{ where } \mathbf{K}_i = \mathbf{B}_{a+i-1}; \quad (6)$$

$$\mathbf{Q} := [\mathbf{Q}_1; \dots; \mathbf{Q}_L] \in \mathbb{R}^{L \times d_k}, \text{ where } \mathbf{Q}_i = \mathbf{C}_{a+i-1}; \quad (7)$$

$$\mathbf{w} := [\mathbf{w}_1; \dots; \mathbf{w}_L] \in \mathbb{R}^{L \times d_k \times d_v}, \text{ where } \mathbf{w}_i = \prod_{j=1}^i e^{A \Delta_{a-1+j}}; \quad (8)$$

$$\mathbf{H} := [\mathbf{h}_a; \dots; \mathbf{h}_b] \in \mathbb{R}^{L \times d_k \times d_v}; \quad (9)$$

$$\mathbf{Y} := [\mathbf{y}_a; \dots; \mathbf{y}_b] \in \mathbb{R}^{L \times d_v}; \quad (10)$$

based on the definitions, the time-varying hidden state can be written as:

$$\mathbf{h}_b = \mathbf{w}_L \odot \mathbf{h}_a + \sum_{i=1}^L \frac{\mathbf{w}_L}{\mathbf{w}_i} \odot (\mathbf{K}_i^T \mathbf{V}_i), \quad (11)$$

where $\odot$ is the element-wise production and the division of $\mathbf{w}_L$ and $\mathbf{w}_i$ is also element-wise. The output of the S6 block can be expressed as:

$$\mathbf{y}_b = \mathbf{Q}_L(\mathbf{w}_L \odot \mathbf{h}_a) + \mathbf{Q}_L \sum_{i=1}^L \frac{\mathbf{w}_L}{\mathbf{w}_i} \odot (\mathbf{K}_i^T \mathbf{V}_i). \quad (12)$$

Thus, the j^{th} slice along the dimension d_v of output $\mathbf{Y}^{(j)}$ is:

$$\mathbf{Y}^{(j)} = (\mathbf{Q}_L \odot \mathbf{w}^{(j)}) \mathbf{h}_a^{(j)} + \left[(\mathbf{Q}_L \odot \mathbf{w}^{(j)}) \left(\frac{\mathbf{K}}{\mathbf{w}^{(j)}} \right)^T \odot \mathbf{M} \right] \mathbf{V}^{(j)}, \quad (13)$$

where $\mathbf{M}$ presents the temporal mask matrix with the lower triangular parts set to 1 and elsewhere 0. As shown in Eq. (13), the multiplication of Q, K, and V matrices without SoftMax function is similar to the self-attention module of Vision Transformers in Eq. (18) and is regarded as the generalised form of Gated Linear Attention (Liu, et al., 2024; Yang, et al., 2023). This linear attention in VMamba contributes to its computational efficiency and considerable performance enhancement in multiple vision assignment types.

Due to the discrepancy among CNNs, Transformers, and Mamba, the computational complexity of image processing by each of these architectures is different. As presented in Figure 4.3, the input and output images are assumed to have the same size and are equally flattened into several patches, with each patch embedding size L . To keep the same size of input and output, the convolutional operation in CNNs can be regarded as L times dot-productions of two flattened vectors with size k , where k is the convolutional kernel size, thus resulting in its $O(kL)$ computational complexity (Mohammadi Farsani and Pazouki, 2020). In the self-attention module of Transformers, after concatenating Queries (Q), Keys (K), and Values (V) matrices, each embedded image patch needs to dot-product all patch embeddings to get the

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

attention score, and the operation of $\text{SoftMax}(QK^T)$ makes the computational complexity become $O(L^2)$ (Liu, et al., 2024). Unlike Transformers, the S6 block in Mamba - see Eq. (13) - could remove the SoftMax operation and first conduct the operation of K^T and V instead of Q and K^T , thus resulting in a linear computational complexity of $O(L)$ (Liu, et al., 2024). The lower-order computational complexity of the core part contributes to the significant efficiency of Mamba with much lower floating-point operations (FLOPs) than the other two architectures, which will be discussed later.

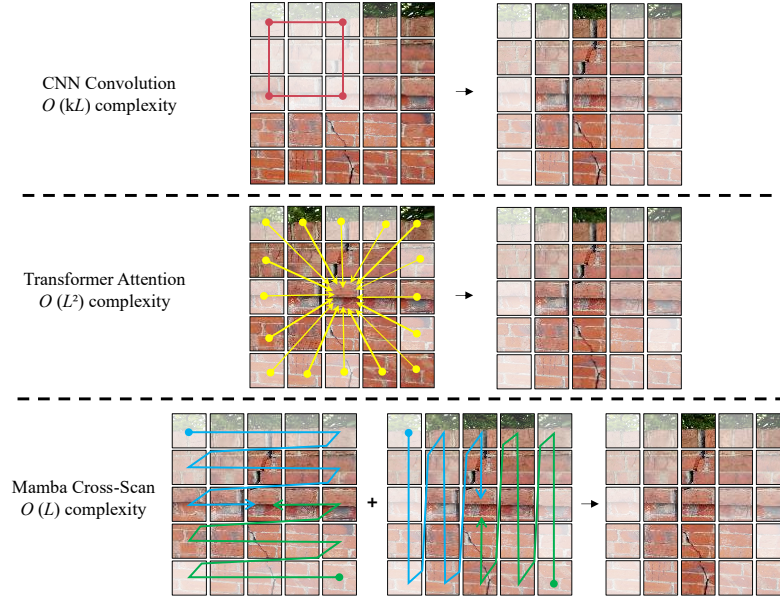


Figure 4.3 Comparison of computational complexity.

4.2.2 Vision Mamba based encoder-decoder network

In this section, the VMamba-based baseline encoder-decoder network (VM-UNet (Ruan and Xiang, 2024)) is introduced for the crack segmentation task. As illustrated in Figure 4.4, the patch embedding layer of VM-UNet first partitions the input image into non-overlapping patches with size 4×4 , then transforms the image dimensions into C . The embedded image is then fed into the encoder, which consists of four stages of feature extraction, with a patch merging down-sampling module (Liu, et al., 2022) employed for the initial three stages to decrease the size of the input feature maps while augmenting the channels. Likewise, the decoder part is structured into four stages, with a patch expanding up-sampling module (Cao, et al., 2022) applied to reduce the channel number while enlarging the feature map size.

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

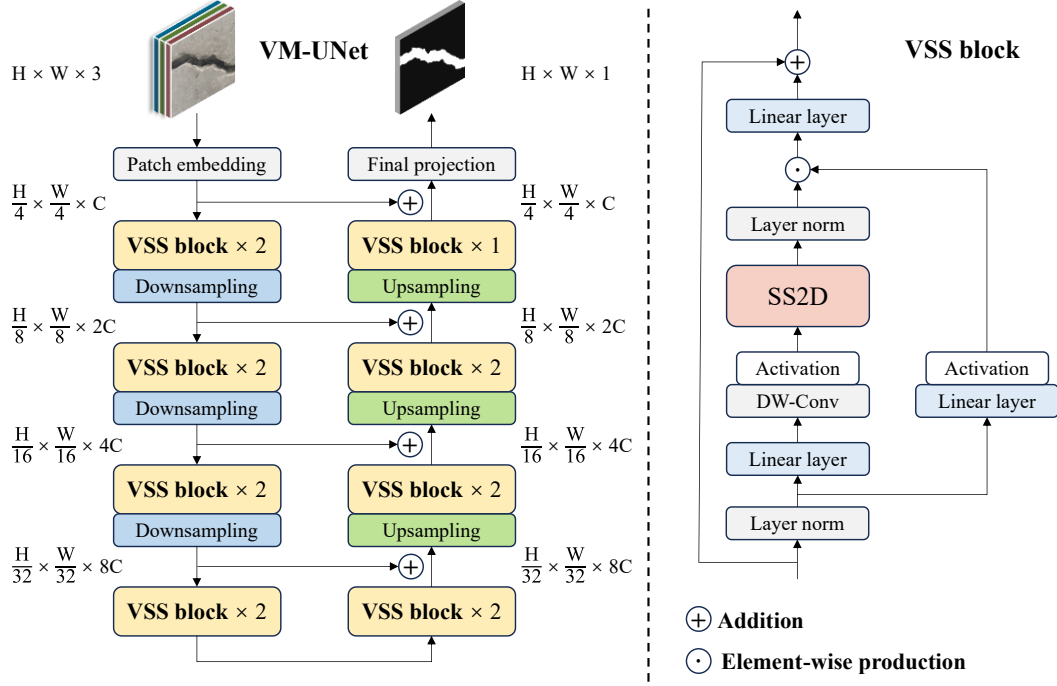


Figure 4.4 Overview of the VM-UNet (Ruan and Xiang, 2024) architecture.

The Visual State Space (VSS) block (Liu, et al., 2024) in the encoder and decoder serves as the core component of VM-UNet. After layer normalisation, the input feature map undergoes division into two pathways. In the initial path, the input feature map proceeds through a linear layer and then the Sigmoid Linear Unit (SiLU) (Elfwing, et al., 2018) activation function. Meanwhile, in the second path, the input feature map undergoes processing via a linear layer, depth-wise separable convolution, and the SiLU activation function, followed by the SS2D (See Figure 4.2) module and layer normalisation for additional feature extraction. Afterwards, the element-wise production is operated to merge the outputs of these two pathways, followed by a linear layer and the combination with the residual connection to shape the final output.

4.2.3 Evaluation metrics

For the crack segmentation task, which aims to classify the crack and non-crack pixels, mean Dice score (mDS) and mean intersection over union (mIoU) score are commonly utilised due to their balanced evaluation for segmentation accuracy (Shamsabadi, et al., 2022). Thus, these two metrics are employed to assess the model performance in this study, which can be formulated as:

$$DS = \frac{2|P \cap T|}{|P| + |T|}, \quad (14)$$

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

$$IoU = \frac{|P \cap T|}{|P \cup T|}, \quad (15)$$

herein, P is the pixel map of the model output, and T is the ground truth pixel map.

IoU measures the overlap between mask pixels and segmentation map sets by dividing their intersection area by union, whereas DS measures overlap by dividing twice the intersection area by the total area of both sets. Both metrics range from 0 to 1, where 1 indicates perfect overlap, and 0 indicates no overlap. IoU penalises poor detections more than DS, especially when overlaps are small. However, as overlaps increase, the difference between IoU and DS decreases. Therefore, mIoU is more sensitive to poor predictions, while mDS shows average performance more accurately. Since these two metrics do not depend on the designated threshold for the classification of crack pixels and assess the overlap of the entire region, including its boundary, they offer a more impartial and reasonable evaluation of segmentation performance (Chen, et al., 2024).

4.2.4 Datasets

In this study, three public crack datasets, with concrete, concrete and asphalt mixed, and masonry backgrounds, are utilised to assess the performance of VMamba-based encoder-decoder network (VM-UNet). Detailed information on the selected datasets is introduced below.

- Crack500 (Zhang, et al., 2016): this dataset depicts cracks on asphalt and concrete pavement surfaces, includes 1896 training images and 1124 testing images. The original images were collected by cell phone cameras on the main campus of Temple University and split by the dataset developers, who retained only the split images (640×360) containing crack pixels. To satisfy the input of models analysed in this study, each image and its corresponding annotation were resized to 448×448 pixels.
- Ozgenel (Özgenel, 2019): this dataset consists of 458 high-resolution concrete crack images (4032×3024 pixels) with corresponding annotation masks collected from various Middle East Technical University buildings. In this study, the images were split into patches with a resolution of 448×448 pixels to prepare the dataset for evaluation. Finally, the training and test set consist of 1800 and 454 images with their masks, respectively.
- MC448 (Shamsabadi, et al., 2022): this dataset is sourced from 45 copyright-free images, 153 manually captured full-resolution photos, and 15 artificially generated images. After splitting, three thousand images and the corresponding masks were allocated for the model training, and the remaining 351 images and their masks were assigned for testing. This

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

dataset is challenging for crack segmentation because it includes non-crack images and has textures on the masonry surface that resemble cracks, making it harder to differentiate between actual cracks and background textures.

Overall, the specifications of these datasets are presented in Table 4.1 in terms of the input image size and the image number of training and test sets.

Table 4.1 Details of three datasets

Dataset	Input image size	Total number	Train set	Test set
Crack500	448×448	3020	1896	1124
Ozgenel	448×448	2254	1800	454
MC448	448×448	3351	3000	351

4.2.5 Training configuration

The models in this study were constructed by Python employing the PyTorch (Paszke, et al., 2019) open-source framework. The training and evaluation of models were carried out on the Ubuntu 22.04 LTS system, utilising a workstation equipped with an AMD Ryzen 9 5900X 12-Core CPU, an NVIDIA GeForce RTX 3090 Ti GPU, and 32 gigabytes of RAM. All the models in this study underwent training, validation, and testing within the same framework to ensure a fair comparison. The training settings included a learning rate of 5×10^{-5} and a batch size of 2 images, running for 80 epochs with the AdamW optimiser and Dice loss function. The training set was initially shuffled randomly followed by random image augmentation techniques, such as colour-jittering and horizontal flipping, to prevent over-fitting during the training process. Additionally, to facilitate faster training and enhance the final model performance, the model backbones were pre-trained on the ImageNet (Deng, et al., 2009) dataset and the trained weights were utilised for model initialisation.

4.3 Results and discussions

This section first implements the accuracy comparison for representative CNN, Transformer, and Mamba-based models, thus validating the comparable performance of Mamba-based encoder-decoder network (VM-UNet). An efficiency comparison is then conducted to highlight the advantage of the low-order complexity of Mamba.

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

4.3.1 Accuracy comparison

Considering the literature advancements achieved in crack segmentation with CNN and Transformer-based models, a comprehensive study is conducted with the following representative architectures as the established benchmarks in the field:

- CNN-based: UNet (Ronneberger, et al., 2015) and LinkNet (Chaurasia and Culurciello, 2017) with the EfficientNet (Tan and Le, 2019) backbone (UNet-EB7 & LinkNet-EB7);
- Transformer-based: SwinUNet (Cao, et al., 2022), and SegFormer-B5 (Xie, et al., 2021);
- CNN-Transformer hybrid-designed: TransUNet (Chen, et al., 2021) and DTrC-Net (Xiang, et al., 2023);
- Efficient self-attention designed: PoolingCrack (Chen, et al., 2024) with advanced performances in various crack datasets.

The results of these models are studied and compared - see Table 4.2 - to a Mamba-based baseline network (VM-UNet) with a simple design to reveal the capabilities of the Mamba architecture. As presented in Table 4.2, Mamba-based VM-UNet exhibits comparable mDS and mIoU, with up to 0.5% lower mDS and 0.3% - 0.4% lower mIoU than the best ones in the three datasets while having 15.6% - 74.5% fewer training parameters than all other networks. Except for SegFormer-B5 and PoolingCrack, VM-UNet achieves 0.1% - 6.2% higher mDS and up to 6.8% higher mIoU than the other architectures. Although SegFormer-B5 outperforms VM-UNet in the Crack500 dataset, it exhibits up to 0.8 % lower mDS and mIoU in the Ozgenel and MC448 datasets, with around three times more training parameters. Although PoolingCrack shows overall higher accuracy than VM-UNet, it has 15.6% more parameters, requires more floating-point operations, and has a longer inference time, which is discussed in the next section in detail. In summary, the Mamba-based network demonstrates significant potential for effective crack segmentation, particularly due to its highly efficient computational performance.

Figure 4.5 illustrates the qualitative analysis results of the different network performances over three datasets. As shown in Figure 4.5-a, VM-UNet outputs clear and continuous crack maps in the Crack500 dataset without creating noises in most cases. Although SegFormer-B5 exhibits the best average accuracy (quantitative results) in this dataset, it sometimes cannot portray the crack profile consistently, e.g., the crack prediction maps of SegFormer-B5 in the third and fifth rows. As for the produced crack maps of other networks, they miss some crack pixels in the second-row case or mistakenly classify some background noises as cracks in the third-row

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

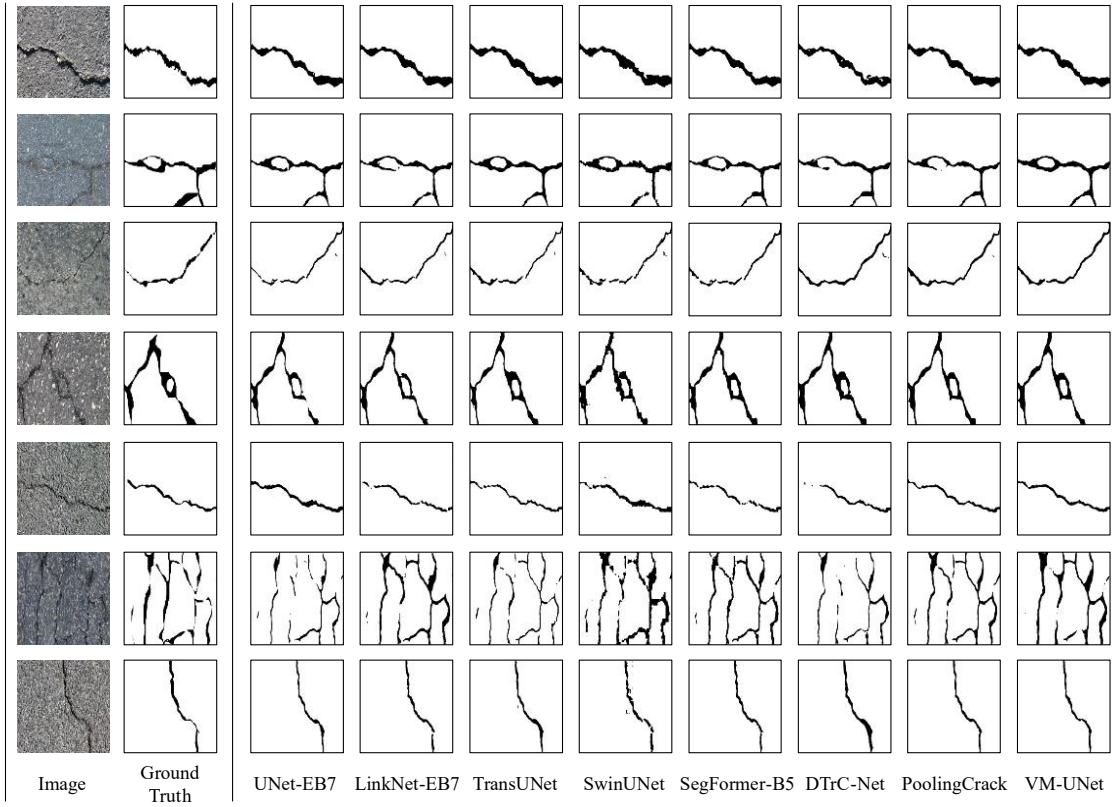
case, thus having an unstable performance for the crack pixel classification. VM-UNet, which could achieve the highest mDS in the Ozgenel dataset - Table 4.2, shows enhanced segmentation results compared to other networks - Figure 4.5-b. It presents robustness without being disturbed by surface complexities (red stripe and scratch background texture) in the third and last rows while accurately producing the complete crack map in the fourth-row case with challenging crack pixels.

Table 4.2 Accuracy comparison of different networks in three datasets (bold values are the best)

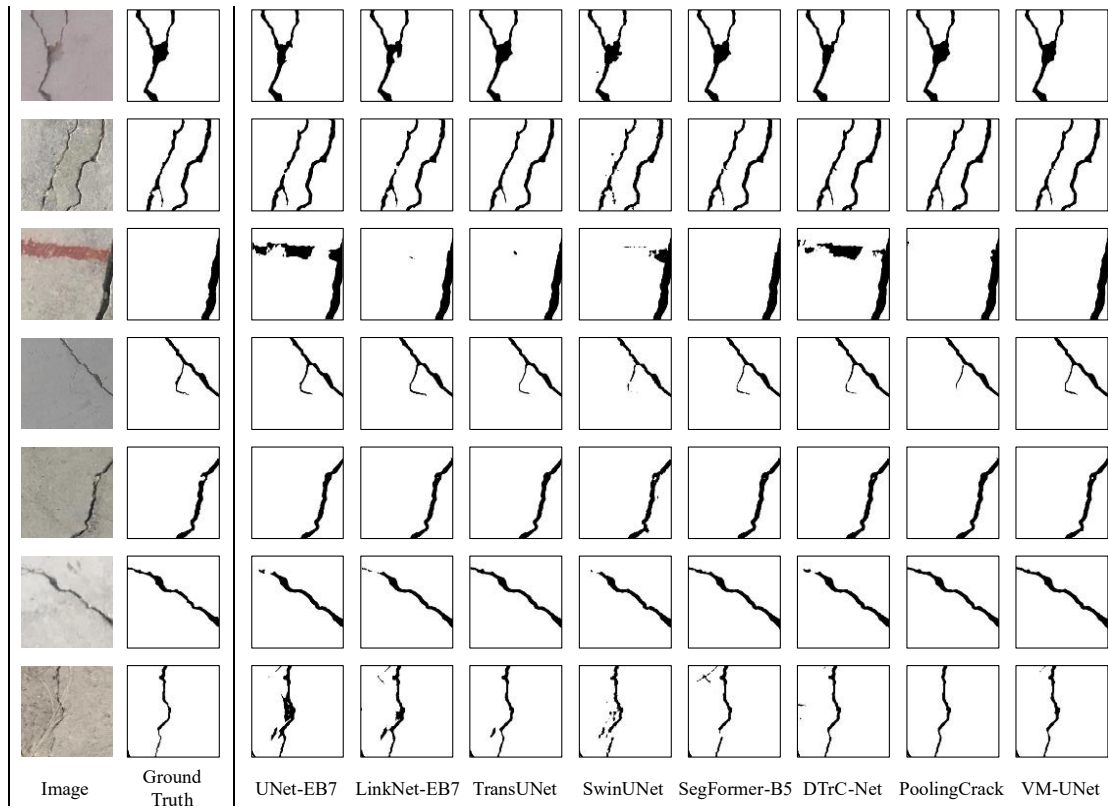
Model	Parameters(M)	Crack500		Ozgenel		MC448	
		mDS (%)	mIoU(%)	mDS (%)	mIoU(%)	mDS (%)	mIoU(%)
UNet-EB7	67	69.9	55.7	84.1	77.3	72.9	62.1
LinkNet-EB7	64	69.9	55.6	84.6	78.1	72.0	60.9
TransUNet	106	70.2	56.0	85.3	79.2	75.3	64.7
SwinUNet	42	68.1	53.3	83.3	76.1	70.2	58.8
SegFormer-B5	85	70.7	56.5	84.9	78.6	75.6	64.7
DTrC-Net	42	67.5	53.3	84.7	78.3	69.5	58.3
PoolingCrack	32	70.6	56.4	85.7	79.7	76.2	65.5
VM-UNet	27	70.3	56.0	85.7	79.4	75.7	65.1

In addition, according to the output crack masks in the MC448 dataset in Figure 4.5-c, although the crack map produced by the VM-UNet for the last case has slight noises, it creates relatively complete crack patterns compared with other network outputs in the other cases. Despite the challenges associated with the MC448 dataset, a masonry crack dataset with many crack-like non-crack patterns, the Mamba-based VM-UNet achieved high accuracy.

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

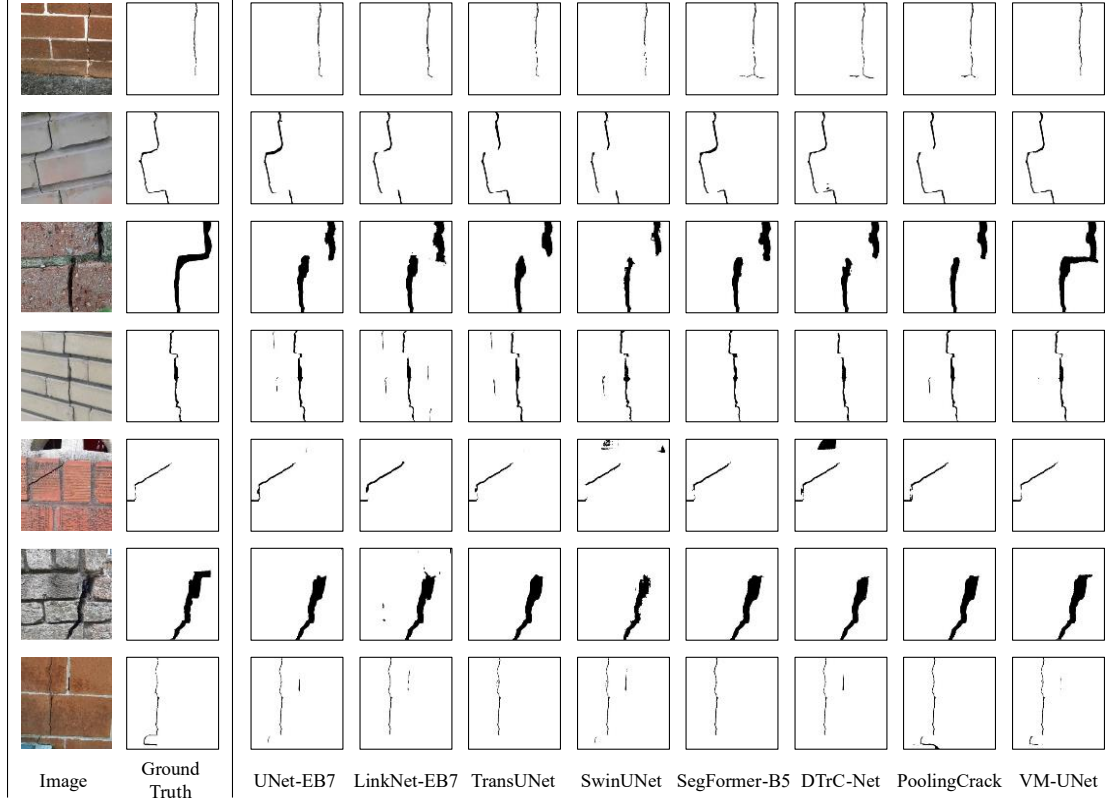


(a) Crack500 (concrete & asphalt)



(b) Ozgenel (concrete)

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection



(c) MC448 (masonry)

Figure 4.5 Qualitative analysis of the segmentation results for different networks in three datasets: (a) Crack 500, (b) Ozgenel, and (c) MC448.

Because of the selective mechanism of Mamba (Gu and Dao, 2023), the Mamba-based model will automatically select or ignore divided image patches depending on whether these patches include cracks, which is similar to the behaviour of attention scores in Transformers. As a result, the Mamba-based model could have better accuracy than CNNs and similar precision to Transformers. Yet, the low computational complexity of the selective mechanism in Mamba significantly contributes to an enhanced efficiency relative to the Transformer-based models. The corresponding comparison in terms of efficiency is discussed in detail in the next section.

4.3.2 Efficiency comparison

Three factors, including model parameters (which quantify the complexity of the model), floating-point operations (which measure the computational workload), and inference time (which indicates the time taken for the model to make predictions), are selected as evaluation metrics for the efficiency comparison. The results are presented in Figure 4.6.

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

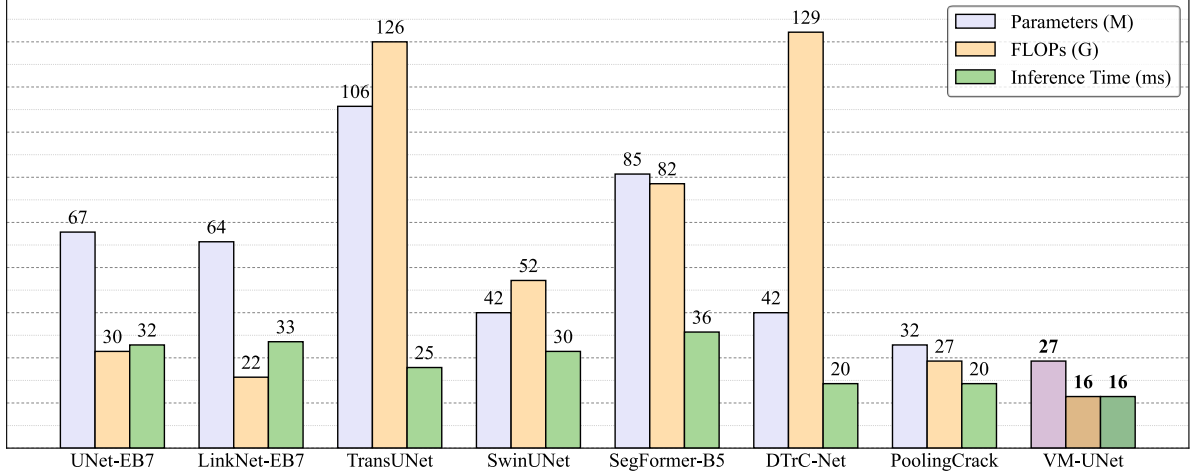


Figure 4.6 Efficiency comparison for processing one image with 448×448 resolution.

As illustrated in the Figure 4.7, the Mamba-based model comprises 27 million training parameters and operates with 16 G FLOPs for each image at a resolution of 448×448 , achieving an inference time of 16 ms. It has 15.6% - 74.5% fewer training parameters, 27.3% - 87.6% fewer FLOPs, and 20% - 55.6% less inference time than other representative models. Due to the low-order complexity of the Mamba-based architecture, VM-UNet achieves the best efficiency among the similar or even larger scale representative models while showing the same accuracy performance as discussed in the previous section. Another computational cost comparison with different image size inputs, shown in Figure 4.7, can better reveal the efficiency of the Mamba-based model; the figure shows the number of FLOPs versus the size of the input image. TransUNet and DTrC-Net (designed with a hybrid of CNN and Transformer) have the top two highest FLOPs as the input image resolution increases. The Transformer-based SwinUNet and SegFormer-B5 present the second-tier highest FLOPs, and the CNN-based UNet-EB7 and LinkNet-EB7 are close to the efficiently designed Transformer-based PoolingCrack in the third tier of highest FLOPs. Mamba-based VM-UNet shows the lowest FLOPs increase among all the models, while it could achieve similar or higher accuracy than the other models - see section 4.3.1.

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

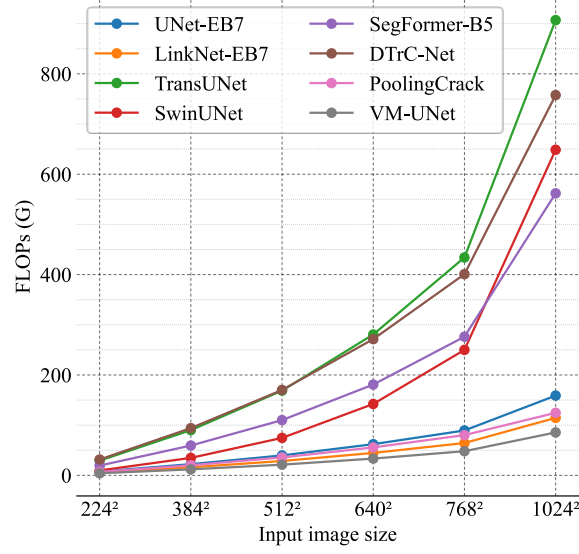


Figure 4.7 Computational costs with different image resolution inputs; different line slopes show the discrepancy among the intrinsic complexity of these architectures.

As mentioned in Section 4.2.1, the attention mechanism in Transformer, the convolutional operation in CNN, and the cross-scan in Mamba have the computational complexity of $O(L^2)$, $O(kL)$, and $O(L)$, respectively. Due to the two branches (CNN and Transformer) of image processing, CNN-Transformer hybrid-designed networks exhibit the highest computational costs. Because of the different computational complexity of different cores, the pure Transformer, CNN, and Mamba-based models show in-order decreased computational costs.

Image resolution can be very important when running deep learning models on edge devices and for real-time monitoring, as the capacity of the equipment could be very limited. Representative CNN or Transformer-based models may have few parameters and low FLOPs under the low-resolution image input; nonetheless, the computational cost under high-resolution image input differentiates Mamba from them. For example, DTrC-Net with 42 M parameters has similar FLOPs as VM-UNet with 27 M parameters under 224×224 image input but with up to 9 times the FLOPs of VM-UNet as the input image resolution increases. Similarly, some lightweight Transformer-based models may have fewer parameters and lower FLOPs than Mamba-based models under the low-resolution image input, yet their computational costs would overtake Mamba-based models as the image resolution is enhanced. The intrinsic shortcoming of computational complexity for these traditional CNN or Transformer-based model designs would hinder their further development for future high-resolution crack detection. Based on the above analysis, the Mamba-based model demonstrates potential for efficient and effective crack segmentation.

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

4.4 Conclusion

Focusing on local information weakens the generalisation of CNN-based models, and the quadratic complexity of the global self-attention hinders the ability of Transformer-based models to process long-sequence inputs. Stemming from these issues, this study proposes a VMamba-based framework for crack segmentation, with low-order complexity of processing crack images and varying time steps to capture crack or ignore non-crack information and improve the generalisation. A VMamba-based encoder-decoder network was used to develop the framework, and its performance was compared with representative CNN-based (UNet-EB7 & LinkNet-EB7), Transformer-based (SwinUNet, SegFormer-B5, and PoolingCrack), and hybrid models (TransUNet & DTrC-Net).

The encoder-decoder network equipped with VMamba (VM-UNet) could achieve up to 4.2% higher, a slightly lower (0.5%), and 0.1% to 6.8% higher mDS and mIoU than representative CNN, Transformer, and CNN-Transformer hybrid-based models while maintaining 15.6% - 74.5% fewer parameters and 20% - 55.6% less inference time. Furthermore, a computational cost comparison was conducted with the increasing network input image sizes. The results showed that the CNN-Transformer hybrid-designed models have the highest FLOPs due to integrating two-branch operations (namely convolution and self-attention), and the Transformer, CNN, and VMamba-based models have the in-order decreasing FLOPs due to the different computational complexity of their cores. As the input image size grows, the VMamba-based model could achieve up to 90.6% lower FLOPs.

The study implemented in this paper could highlight the potential of VMamba for crack detection assignments. The strength of VMamba in processing long-sequence inputs effectively and efficiently could contribute to the development of high-resolution crack detection expected in the future. Given that VMamba is currently in the early development stage, subsequent studies could explore the advanced network design to further improve its overall performance for defect segmentation tasks.

Appendix 4.A

4.A.1 Convolutional Neural Networks (CNNs)

As the classic deep learning network in crack detection (Ali, et al., 2021; Fan, et al., 2020; Li, et al., 2023), CNN architecture typically consists of convolutional, pooling, and fully connected layers - see Figure 4.8. The convolutional operation extracts the local representations

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

of input images through the trainable filters, transforming them into essential feature maps – see Eq. (16):

$$O(i, j) = \sum_{k=1}^m \sum_{l=1}^n I(i+k-1, j+l-1) K(k, l), \quad (16)$$

where I and K stand for the input image with size $M \times N$ and the filter with kernel size $m \times n$, respectively; and O is the feature map with size $(M - m + 1) \times (N - n + 1)$ produced by the convolutional operation. In most CNN-based architectures, convolutional and pooling layers are used to systematically capture multi-level feature representations from the data (Geirhos, et al., 2018). The fully connected layers traditionally followed these operations to decide on each input element.

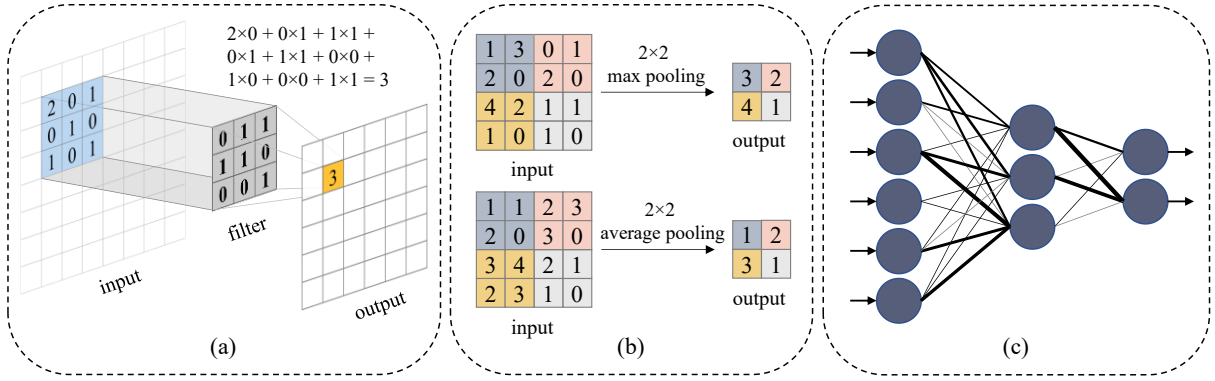


Figure 4.8 Diagram of CNN architecture: (a) convolutional operation; (b) max and average pooling; (c) fully connected layers.

CNNs have achieved good results across various vision-used tasks, yet due to the fixed size of the filter kernel and limited receptive field in the convolutional layer, CNNs have the intrinsic attribute of inductive bias, which focuses on the local information and may limit generalisation (Shamsabadi, et al., 2022) in settings outside of controlled lab environments. Additionally, the local dependency of CNNs restricts the ability to capture long-range feature information and global context in the input data, thus resulting in sensitivity to noise or irrelevant details (Zhou, et al., 2022).

4.A.2 Vision Transformers

Transformers have shown significant performance in crack detection (Chen, et al., 2024), which can be primarily attributed to their global attention mechanism. This mechanism facilitates the extraction of multi-range information from inputs and mitigates the biases in CNNs. Initially introduced in natural language processing, transformers have since been

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

adapted for computer vision tasks. A pioneering example of this adaptation is the Vision Transformer (Dosovitskiy, et al., 2020), which converts input images into sequences by partitioning them into smaller patches before feeding a multi-head self-attention module to extract global contextual information, which is followed by the multilayer perceptron (MLP) (Gardner and Dorling, 1998) to output the prediction results. The schematic diagram of the Vision Transformer is illustrated in Figure 4.9.

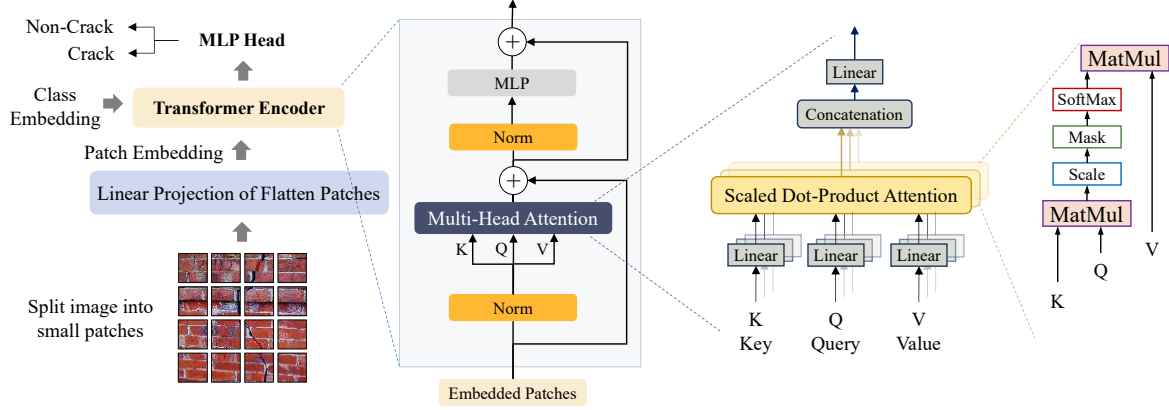


Figure 4.9 The schematic diagram of a Vision Transformer.

Assuming the input image is split into N patches, the flattened patches x_p^i ($i = 1, \dots, N$) with size $p \times p$ are first linearly projected into lower-dimension embeddings as:

$$z_0 = [x_{\text{class}}; x_p^1 E; x_p^2 E; \dots; x_p^N E] + E_{\text{pos}}, \quad (17)$$

where x_{class} represents the learnable class embedding, E is the trainable linear projection, and E_{pos} denotes the position embedding. Embedded patch vectors are then multiplied by learnable matrices W_q , W_k , and W_v to obtain the Key (K), Query (Q), and Value (V) matrices for the self-attention calculation given by:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{Q K^T}{\sqrt{d_k}}\right) V, \quad (18)$$

where K^T represents the transposed matrix K , d_k denotes the dimension of matrix K . In the Transformer encoder, the multi-head self-attention (MSA) with the linear normalisation (LN) (Dosovitskiy, et al., 2020) and multilayer perceptron layers are used to extract the feature representations as:

$$z'_l = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}, \quad (19)$$

$$z_l = \text{MLP}(\text{LN}(z'_l)) + z'_l, \quad (20)$$

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

where z_l is the encoded feature representation. Although Transformers with self-attention design have shown better generalisation and anti-noise robustness than CNNs (Chen, et al., 2024), the demanding scaled dot-product operations for the calculation of the attention score increase the computational cost, thus impairing model efficiency and restricting the practical deployment on edge devices (Chen, et al., 2024).

References

- Ai D., G. Jiang, S.-K. Lam, P. He, C. Li, Computer vision framework for crack detection of civil infrastructure—A review, *Engineering Applications of Artificial Intelligence* 117 (2023), pp. 105478.
- Ali R., J.H. Chuah, M.S.A. Talip, N. Mokhtar, M.A. Shoaib, Automatic pixel-level crack segmentation in images using fully convolutional neural network based on residual blocks and pixel local weights, *Engineering Applications of Artificial Intelligence* 104 (2021), pp. 104391.
- Cao H., Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, *European Conference on Computer Vision*, Springer, (2022), pp. 205-218.
- Cha Y.-J., W. Choi, O. Büyüköztürk, Deep learning-based crack damage detection using convolutional neural networks, *Computer-Aided Civil and Infrastructure Engineering* 32 (5) (2017), pp. 361-378.
- Chaurasia A., E. Culurciello, Linknet: Exploiting encoder representations for efficient semantic segmentation, *2017 IEEE visual communications and image processing (VCIP)*, IEEE, (2017), pp. 1-4.
- Chen J., Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: transformers make strong encoders for medical image segmentation, *arXiv preprint arXiv:2102.04306* (2021).
- Chen Z., E. Asadi Shamsabadi, S. Jiang, L. Shen, D. Dias-da-Costa, An average pooling designed Transformer for robust crack segmentation, *Automation in Construction* 162 (2024), pp. 105367.
- Chen Z., E.A. Shamsabadi, S. Jiang, L. Shen, D. Dias-da-Costa, Robust feature knowledge distillation for enhanced performance of lightweight crack segmentation models, *arXiv preprint arXiv:2404.06258* (2024).

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

Chu H., P.j. Chun, Fine-grained crack segmentation for high-resolution images via a multiscale cascaded network, *Computer-Aided Civil and Infrastructure Engineering* (2023).

Deng J., W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, 2009 IEEE Conference on Computer Vision and Pattern Recognition, (2009), pp. 248-255.

Deng J., A. Singh, Y. Zhou, Y. Lu, V.C.-S. Lee, Review on computer vision-based crack detection and quantification methodologies for civil structures, *Construction and Building Materials* 356 (2022), pp. 129238.

Dosovitskiy A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint* (2020).

Elfwing S., E. Uchibe, K. Doya, Sigmoid-weighted linear units for neural network function approximation in reinforcement learning, *Neural networks* 107 (2018), pp. 3-11.

Fan Z., C. Li, Y. Chen, J. Wei, G. Loprencipe, X. Chen, P. Di Mascio, Automatic crack detection on road pavements using encoder-decoder architecture, *Materials* 13 (13) (2020), pp. 2960.

Fu D.Y., T. Dao, K.K. Saab, A.W. Thomas, A. Rudra, C. Ré, Hungry hungry hippos: Towards language modeling with state space models, *arXiv preprint arXiv:2212.14052* (2022).

Gardner M.W., S. Dorling, Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences, *Atmospheric Environment* 32 (14-15) (1998), pp. 2627-2636.

Geirhos R., P. Rubisch, C. Michaelis, M. Bethge, F.A. Wichmann, W. Brendel, ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, *arXiv preprint arXiv:1811.12231* (2018).

Gu A., T. Dao, S. Ermon, A. Rudra, C. Ré, Hippo: Recurrent memory with optimal polynomial projections, *Advances in Neural Information Processing Systems* 33 (2020), pp. 1474-1487.

Gu A., K. Goel, C. Ré, Efficiently modeling long sequences with structured state spaces, *arXiv preprint arXiv:2111.00396* (2021).

Gu A., T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, *arXiv preprint arXiv:2312.00752* (2023).

Han X., Z. Zhao, L. Chen, X. Hu, Y. Tian, C. Zhai, L. Wang, X. Huang, Structural damage-causing concrete cracking detection based on a deep-learning method, *Construction and Building Materials* 337 (2022), pp. 127562.

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

- Katharopoulos A., A. Vyas, N. Pappas, F. Fleuret, Transformers are rnns: Fast autoregressive transformers with linear attention, *International Conference on Machine Learning*, (2020), pp. 5156-5165.
- Li C., Z. Fan, Y. Chen, H. Lin, L. Moretti, G. Loprencipe, W. Sheng, K.C. Wang, CrackCLF: Automatic Pavement Crack Detection Based on Closed-Loop Feedback, *IEEE Transactions on Intelligent Transportation Systems* (2023).
- Liu Y., Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, Y. Liu, Vmamba: Visual state space model, *arXiv preprint arXiv:2401.10166* (2024).
- Liu Z., H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, Swin transformer v2: Scaling up capacity and resolution, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 12009-12019.
- Mohammadi Farsani R., E. Pazouki, A transformer self-attention model for time series forecasting, *Journal of Electrical and Computer Engineering Innovations* 9 (1) (2020), pp. 1-10.
- Nguyen S.D., T.S. Tran, V.P. Tran, H.J. Lee, M.J. Piran, V.P. Le, Deep Learning-Based Crack Detection: A Survey, *International Journal of Pavement Research and Technology* 16 (4) (2023), pp. 943-967.
- Özgenel Ç.F., Concrete crack segmentation dataset, *Mendeley Data* 1 (2019), pp. 2019.
- Paszke A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimselshein, L. Antiga, Pytorch: an imperative style, high-performance deep learning library, *Advances in Neural Information Processing Systems* 32 (2019).
- Ronneberger O., P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, *Medical Image Computing and Computer-Assisted Intervention–18th International Conference, Munich, Germany Springer*, (2015), pp. 234-241.
- Ruan J., S. Xiang, Vm-unet: Vision mamba unet for medical image segmentation, *arXiv preprint arXiv:2402.02491* (2024).
- Sarkar K., A. Shiuly, K.G. Dhal, Revolutionizing concrete analysis: An in-depth survey of AI-powered insights with image-centric approaches on comprehensive quality control, advanced crack detection and concrete property exploration, *Construction and Building Materials* 411 (2024), pp. 134212.
- Schuster M., K.K. Paliwal, Bidirectional recurrent neural networks, *IEEE Transactions on Signal Processing* 45 (11) (1997), pp. 2673-2681.
- Shamsabadi E.A., C. Xu, D. Dias-da-Costa, Robust crack detection in masonry structures with Transformers, *Measurement* 200 (2022), pp. 111590.

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

Shamsabadi E.A., C. Xu, A.S. Rao, T. Nguyen, T. Ngo, D. Dias-da-Costa, Vision transformer-based autonomous crack detection on asphalt and concrete surfaces, *Automation in Construction* 140 (2022), pp. 104316.

Shi Z., N. Jin, D. Chen, D. Ai, A comparison study of semantic segmentation networks for crack detection in construction materials, *Construction and Building Materials* 414 (2024), pp. 134950.

Shiuly A., D. Dutta, A. Mondal, Assessing compressive strengths of mortar and concrete from digital images by machine learning techniques, *Frontiers of Structural and Civil Engineering* 16 (3) (2022), pp. 347-358.

Sun Y., L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, J. Wang, F. Wei, Retentive network: A successor to transformer for large language models, *arXiv preprint arXiv:2307.08621* (2023).

Tan M., Q. Le, Efficientnet: rethinking model scaling for convolutional neural networks, *Proceedings of Machine Learning Research*, (2019), pp. 6105-6114.

Tran T.S., S.D. Nguyen, H.J. Lee, V.P. Tran, Advanced crack detection and segmentation on bridge decks using deep learning, *Construction and Building Materials* 400 (2023), pp. 132839.

Xiang C., J. Guo, R. Cao, L. Deng, A crack-segmentation algorithm fusing transformers and convolutional neural networks for complex detection scenarios, *Automation in Construction* 152 (2023), pp. 104894.

Xie E., W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, *Advances in Neural Information Processing Systems* 34 (2021), pp. 12077-12090.

Xiong C., T. Zayed, E.M. Abdelkader, A novel YOLOv8-GAM-Wise-IoU model for automated detection of bridge surface cracks, *Construction and Building Materials* 414 (2024), pp. 135025.

Yang G., K. Liu, J. Zhang, B. Zhao, Z. Zhao, X. Chen, B.M. Chen, Datasets and processing methods for boosting visual inspection of civil infrastructure: A comprehensive review and algorithm comparison for crack classification, segmentation, and detection, *Construction and Building Materials* 356 (2022), pp. 129226.

Yang S., B. Wang, Y. Shen, R. Panda, Y. Kim, Gated linear attention transformers with hardware-efficient training, *arXiv preprint arXiv:2312.06635* (2023).

Zhang J., Y.-Y. Cai, D. Yang, Y. Yuan, W.-Y. He, Y.-J. Wang, MobileNetV3-BLS: A broad learning approach for automatic concrete surface crack detection, *Construction and Building Materials* 392 (2023), pp. 131941.

Zhang L., F. Yang, Y.D. Zhang, Y.J. Zhu, Road crack detection using deep convolutional neural network, *2016 IEEE International Conference on Image Processing* (2016), pp. 3708-3712.

4. Vision Mamba-based architectures for high-resolution local-scale defect inspection

Zhou D., Z. Yu, E. Xie, C. Xiao, A. Anandkumar, J. Feng, J.M. Alvarez, Understanding the robustness in vision transformers, International Conference on Machine Learning, (2022), pp. 27378-27394.

5. AI agent framework for global-scale defect perception

This chapter presents the paper “Zhaohui Chen, Elyas Asadi Shamsabadi, Sheng Jiang, Luming Shen, Daniel Dias-da-Costa. (2026). Integration of large vision language models for efficient post-disaster damage assessment and reporting. *Nature Communications*”, which introduces the DisasTeller framework for global-scale disaster perception and assessment.

The chapter develops a structured multi-agent AI architecture that integrates large vision-language models with role-based task decomposition to support post-disaster interpretation, coordination, and decision support. The framework organises specialised agents to perform distributed reasoning across perception, damage interpretation, alert generation, emergency planning, and resource allocation, enabling the transformation of local visual observations into coherent global situational awareness.

By coupling local-scale visual perception with structured multi-agent reasoning, this chapter establishes a system-level architecture for global disaster assessment that enables coherent information integration, structured interpretation, and coordinated decision-making across post-disaster response processes.

5.1 Introduction

Natural disasters consist of various catastrophic events, including earthquakes, hurricanes, floods, and wildfires. These events can cause extensive damage, loss of life, and have long-term impacts on societies and economies. For instance, earthquakes can cause buildings to collapse, trapping people under rubble, while hurricanes and floods can result in drowning and physical damage. These natural disasters also lead to extended social challenges, such as the disruption of community cohesion and the psychological effects of trauma and loss. According to the report from the United Nations office, the overall global economic losses from natural disasters in 2023 were estimated to be US\$250 billion, which roughly equals the entire gross domestic product of New Zealand (2023). Beyond immediate losses, disasters would have long-lasting economic effects, including reduced economic growth, increased poverty, and disruption to markets and industries. The frequency and intensity of natural disasters have been rising due to factors such as climate change, urbanisation, and environmental degradation, making the need for effective post-disaster management even more critical.

5. AI agent framework for global-scale defect perception

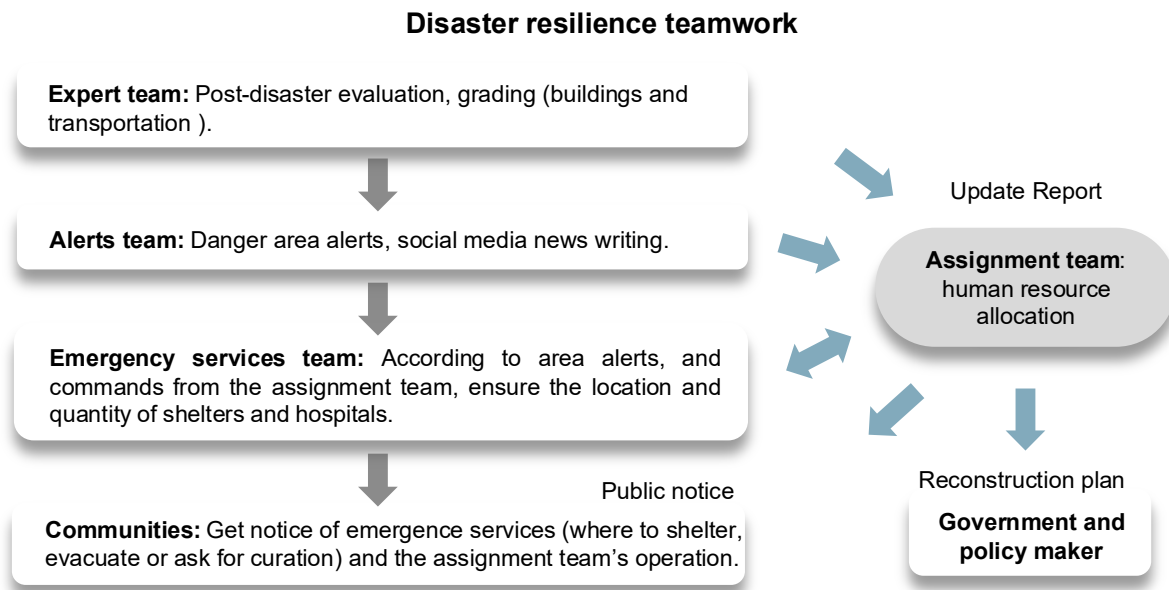


Figure 5.1 The schematic of disaster response human teamwork.

Post-disaster management involves activities including disaster response and post-disaster recovery aimed at mitigating the impacts of natural disasters. The Sendai Framework of the United Nations for Disaster Risk Reduction (2015–2030) highlights the crucial role of enhanced disaster risk governance in ensuring an effective response and promotes the concept of ‘Building Back Better’ in sustainable recovery, reconstruction, and rehabilitation efforts (2015). Traditionally, the entire process of disaster response and post-disaster recovery requires broad collaboration and teamwork (2021; Edwin and Naomi, 2016). As shown in Figure 5.1, the first step in the disaster response process involves the post-disaster on-site assessment through an expert team, which would focus on evaluating the structural integrity of buildings, transportation networks and other critical infrastructures. Based on the assessment of the expert team, the alerts team is activated. The role of the alerts team is to issue emergency alerts to the public, the emergency service team, and government agencies. These alerts are crucial for ensuring that all parties are aware of the situation and can act accordingly. Simultaneously, the emergency services team is mobilised to provide immediate relief to affected populations and determines the number and locations of relevant facilities based on the information provided by the alerts team and the expert evaluations. The assignment team works closely with other teams and is tasked with distributing human resources for emergency service construction, determining where and how resources should be assigned depending on the other three teams’ reports. In addition, an essential component of the disaster response is keeping the affected communities informed. This includes advising them about the ongoing operations, where they

5. AI agent framework for global-scale defect perception

can access shelter and medical care, and any other relevant updates. Effective communication helps to manage public expectations, reduce panic, and foster cooperation among affected communities. After the response phase, the government and policymakers could rely on updates from the assignment team to make decisions about rebuilding strategies and long-term resilience measures.

Generally, the disaster response work would last several weeks due to the limitations of human execution and mobility (Ao and Bahmani, 2023; Feng and Cui, 2021), and the response process can be even slower in underdeveloped areas due to insufficient support and infrastructure investment (Park, et al., 2024), inevitably leading to vast losses. In all periods, enhancing the speed of disaster response efforts and ensuring marginalised groups are not left behind are widely recognised as crucial for reducing disaster-related losses and saving human lives (Forrester, 2023; Pollack, et al., 2024). Current disaster management research focuses on the analysis of disaster-related factors under specific scenarios (Ceferino, et al., 2024; Ding, et al., 2023; Opabola and Galasso, 2024). Math equations or simple machine learning models cannot handle the complex scenarios in post-disaster management, let alone simulate human collaborations to reduce labour work, accelerate the response, or even mitigate the disaster response disparities between developed and underdeveloped areas.

As artificial intelligence (AI) technology develops, it will provide the possibility to better emulate human intelligence, thus rendering it possible to assist humans in improving operational efficiency under the oversight of humans. One of the most significant breakthroughs in AI has been the development of large vision language models (LVLMs), which extend the capabilities of large language models (LLMs) to handle multimodal tasks involving both textual and visual data. These models, which include OpenAI's GPT-4 series (Achiam, et al., 2023), are built on neural networks that mimic the structure of the human brain and trained on vast amounts of image and text data, enabling them to generate human-like text, process information from image and context, and even perform complex tasks like summarisation and conversation. While these models do not truly 'understand' language in a human sense, their ability to model linguistic and visual patterns has made them useful tools in domains requiring large-scale information processing and analysis (M. Bran, et al., 2024). The recent progress in LLMs has laid a strong foundation for the development of LVLMs. Bharathi Mohan et al. (Bharathi Mohan, et al., 2024) offer a comprehensive review of LLM architectures and their expanding applications, illustrating the technological evolution toward more

5. AI agent framework for global-scale defect perception

sophisticated multimodal models. Foundational works by Brown et al. (Brown, 2020) and Radford et al. (Radford, et al., 2019) have demonstrated how few-shot learning and multitask generalisation in LLMs have shaped the design of current-generation LVLMs. Liu et al. (Liu, et al., 2023) further show how GPT-based evaluation systems enhance the alignment and reliability of generative models, while Yang et al. (Yang, et al., 2024) highlight the potential risks of misinformation, particularly in sensitive applications such as healthcare, underscoring the need for caution when deploying these models in critical domains like disaster response.

In the disaster management domain, integrating vision-language understanding is particularly valuable. Nguyen et al. (Nguyen, et al., 2017) show that social media imagery can be effectively leveraged for real-time damage assessment during disasters, pointing to the potential of LVLMs, which can jointly process visual and textual inputs to automate and enhance such tasks. Moreover, standardised frameworks such as the European Macroseismic Scale (EMS-98) (Grünthal, 1998) provide structured criteria for assessing seismic intensity, which can be used as the rubric to evaluate the ability of LVLMs to interpret post-disaster images and reports. These developments support the increasing relevance of LVLMs in disaster response workflows, as the disaster response field advances toward more automated and accountable systems. Building on these insights, the use of multi-LVLM systems to support disaster response operations offers promising benefits, like accelerating the complete procedure, minimising disaster-related losses and narrowing the disaster response gap between developed and underdeveloped areas. Unlike humans, LVLMs can operate continuously without fatigue and facilitate tasks across regions with varying levels of development. They are capable of supporting team-based workflows through rapid interactions, efficiently referencing large volumes of historical data, and offering standardised outputs. Although they may lack true understanding or judgment like humans, their standardised and algorithmic output could facilitate a faster, more uniform response during the initial disaster assessment phase, potentially reducing inconsistencies and biases inherent in human judgment (Bharathi Mohan, et al., 2024).

In this work, we propose a multi-LVLM-powered disaster response framework, DisasTeller, designed to streamline the collaboration of different teams in post-disaster management. As shown in Figure 5.2, DisasTeller operates by prompting four independent LVLM agents (GPT-4o in our experiment) with specific task instructions and implementing three assistant tools to complete the process. The LVLM agents are provided with the descriptions of their roles, the utility of defined tools, and the details of expected input and output. The expert team agent first

5. AI agent framework for global-scale defect perception

analyses on-site local post-disaster and global region images, then shares their information with the other three teams. The alerts team and emergency team agents then generate alerting news and emergency service reports, respectively, while the assignment team agent consolidates all prior outputs to produce the human resource allocation command, public announcement and future reconstruction plan. Designed as a modular engine, DisasTeller can be readily extended with additional agents and tools for emerging applications. It will also support non-experts through user-facing interfaces (e.g., mobile apps or dashboards), enabling real-time image submission, localised alerts, actionable guidance and transparent monitoring of resource distribution. Aggregated multi-source inputs from non-expert users further enhance system responsiveness. We assess the reasonability of all six output reports from the agents of the four teams in Supplementary Information. Overall, DisasTeller consistently produces coherent and actionable outputs across key disaster management tasks, with GPT-4o exhibiting higher grading accuracy and lower annotation deviation than Gemma3, demonstrating the potential of multi-LVLM agent systems to provide standardised and rapid support for post-disaster response.

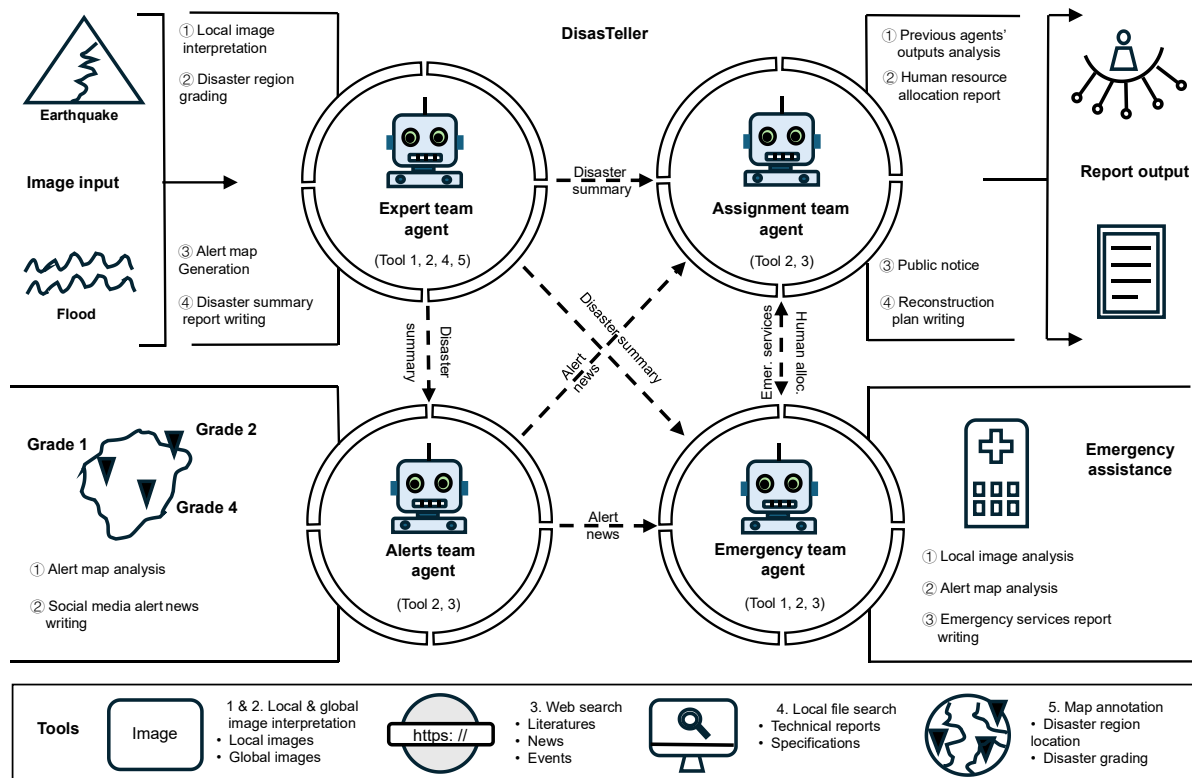


Figure 5.2 Overall framework of DisasTeller: The expert team agent interprets the input images, grades the affected regions, and generates a disaster summary, which is sent to the alerts, assignment, and emergency team agents. The alerts team agent analyses alert maps and social-media information and generates alert news, which is sent to the assignment and emergency team agents. The assignment and emergency team agents receive both the disaster summary and alert news, and in turn produce public notices, human-resource allocation plans, and emergency-service reports.

5.2 Results and discussions

5.2.1 Autonomous disaster response

To validate the autonomous process of DisasTeller, a case study of earthquake disaster (another two validations with flood and bushfire are introduced in Supplementary Information) was conducted in Figure 5.3, utilising the data from previous literature (Nguyen, et al., 2017). We define LVLM agents by prompting task descriptions for several independent GPT-4o models with assigned tools, as shown in Figure 5.3a before the execution of DisasTeller. Like the initial action of the experts' on-site inspections for the post-disaster response, the input of the framework starts with local on-site post-disaster and global map images. As shown in Figure 5.3b, we manually defined the affected area by earthquake as the Wajima City in Japan, and the locations of the local-view images as Wajima Drama Memorial Hall, Hama Street, Concrete Bridge, Central Nishikigawa Street, North Asaichi Street and South Central Asaichi Street. DisasTeller first processes the local-view images for the location and visible damage descriptions. Utilising the file search tool, DisasTeller reads the local technical guideline file European Macroseismic Scale (EMS98) (Grünthal, 1998) for the damage evaluation and grading (G1 – G5, from slight damage to very heavy damage). The output of the post-disaster damage grading in Figure 5.3b matches the local-view images well, confirming the successful collaborations of LVLM agents to analyse the severity of the disaster. After the damage evaluation of local areas, DisasTeller deploys the map annotation tool to search and annotate the damage grades in the relevant locations of the global map image, thus generating the alert map to visualise the overall situation in this disaster area. This alert map is then further reviewed by other LVLM agents in DisasTeller to generate different reports (partial output of reports shown in Figure 5.3c).

5. AI agent framework for global-scale defect perception

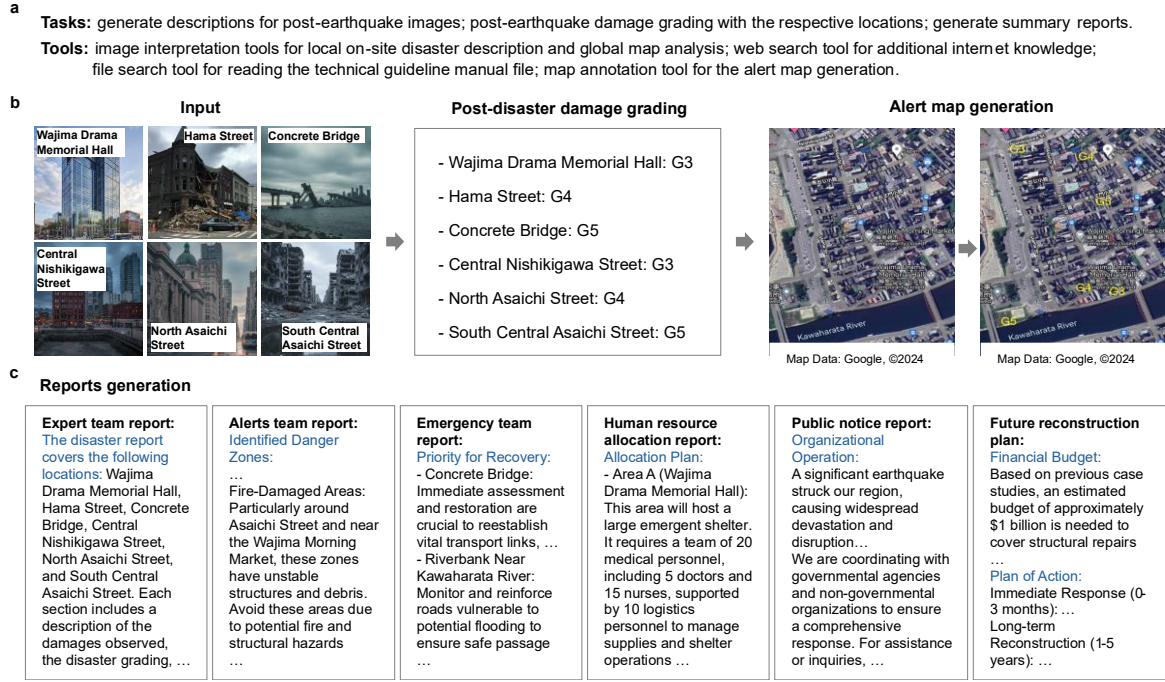


Figure 5.3 The experimental process of our framework: a. the descriptions for tasks of agents and tools assigned; b. the running process of DisasTeller and intermediate tasks with disaster grades G1 – G5; c. the final output reports of agents in DisasTeller.

The first report is generated by the expert team agent in DisasTeller to briefly summarise the local-view disaster and its damage grades. After inspecting the expert team report and the alert map, the alerts team agent in DisasTeller starts reasoning and generating alert news, providing information and safety alerts of existing dangerous areas (e.g., ‘Fire-Damaged Areas: Particularly around Asaichi Street and near the Wajima Morning Market, these zones have unstable structures and debris. Avoid these areas due to potential fire and structural hazards’). Analysing the reports of previous agent teams and taking reference to the historical disaster data with the real website link (e.g., ‘An Overview of the Design of Disaster Relief Shelters, <https://www.sciencedirect.com/science/article/pii/S2212567114010193>, ...’) from the internet through the web search tool, the emergency team agent ensures the areas requiring immediate recovery and emergency services (e.g., ‘Concrete Bridge: Immediate assessment and restoration are crucial to reestablish vital transport links’). Afterwards, the assignment team agent integrates the information from the previous three reports and disaster archives obtained online (e.g., ‘Historic Disasters - FEMA.gov, <https://www.fema.gov/disaster/historic>, ...’) via ‘web search tool’ to generate the human resource allocation report (e.g., ‘Wajima Drama Memorial Hall: This area will host a large emergency shelter. It requires a team of 20 medical personnel, including five doctors and 15 nurses, supported by 10 logistics personnel’), public notice report (e.g., ‘We are coordinating with governmental agencies and non-governmental

5. AI agent framework for global-scale defect perception

organisations to ensure a comprehensive response’), and future reconstruction plan (e.g., ‘Based on previous case studies, an estimated budget of approximately \$1 billion is needed to cover structural repairs’). DisasTeller saves the intermediate outputs of LVLM agent teams in memory and allows LVLM agents to reuse these intermediate outputs and the defined tools provided they have access, thus automating the whole disaster response process by leveraging the reasoning capabilities of LVLMs.

Reliable outputs of LVLM agents are critical for effectively executing DisasTeller. Yet, the quality of LVLMs’ outputs highly relies on the prompt inputs for the LVLM agents, and even the same instruction prompt can lead to different outputs (Brown, 2020; Radford, et al., 2019). Designing prompts to achieve optimal results is an iterative process. Since maintaining identical and best outputs of LVLM agents is complex, LVLM agents in DisasTeller are restricted to output contents with designated format templates when defining task prompts, thus providing relatively controlled results. This case study demonstrates DisasTeller’s ability to autonomously carry out post-disaster response tasks through LVLM agents’ successful simulation of the coordination of disaster management teams.

5.2.2 Agent output results evaluation

In recent years, integrating LVLMs into machine learning frameworks has accelerated the discoveries in different scientific areas (M. Bran, et al., 2024; Yang, et al., 2024). Yet, LVLMs can generate incorrect or deceptive content, and there is a lack of standardised evaluation metrics, posing a great challenge to assess the output quality of LVLM-based frameworks. In addition, since there is the raising concern about the time-dependent performance of close-source LVLMs like OpenAI’s GPT models (Chen, et al., 2024), we additionally test our framework with Google’s lightweight open-source LVLM model Gemma3-27B (Team, et al., 2025) which has 27 billion parameters (compared with the around 200 billion parameters of GPT-4o (Achiam, et al., 2023)), thus facilitating the reproducibility and better building-upon for future research. Considering the previous studies (M. Bran, et al., 2024; Yang, et al., 2024), we conducted five independent runs respectively for both GPT-4o and Gemma3-driven DisasTeller and implemented the comprehensive assessment with a combination of the LVLM (GPT-4o) and human inspection. The intermediate process tasks of GPT-4o and Gemma3-driven DisasTeller are evaluated manually due to their simplicity, and the six output reports of GPT-4o and Gemma3-driven DisasTeller are assessed by both LVLM (GPT-4o) and human inspection, considering their complexity. GPT-4o is prompted to act as a post-disaster

5. AI agent framework for global-scale defect perception

management expert (which is called EvaluatorGPT (Liu, et al., 2023)), provided with post-disaster on-site images, documents retrieved from the internet, and the evaluation rubric to enhance the accuracy of its assessments. The prompt for EvaluatorGPT is modified to grade the output reports of DisasTeller based on the rubric, which focuses on coherence and consistency of sentences and the accuracy of claims. Each evaluation of EvaluatorGPT requires a detailed explanation of the weakness of the report to support the corresponding grade. In addition, we select people who have civil engineering backgrounds with construction experience to manually inspect DisasTeller's outputs and utilise the same rules to provide grades as a comparison of EvaluatorGPT for a more thorough evaluation. The full evaluation procedure for on-site disaster grading, global map annotations of disaster districts, six output reports and an additional error propagation study are included in Supplementary Information.

The initial response phase of real post-disaster scenarios typically occurs within the first 12 to 24 hours, focusing on immediate response actions such as deploying survey and assessment expert teams to generate the initial alert reports and public announcement (Edwin and Naomi, 2016; Ninivaggi, 2024). The subsequent actions by emergency service teams to provide necessary assistance generally start within the first 72 hours, followed by the structured long-term recovery plan in 1 to 2 weeks (2021; Edwin and Naomi, 2016). Post-disaster response and recovery in real scenarios is a challenging and slow process due to the complexity of multi-stakeholder coordination and comprehensive damage assessment (Feng and Cui, 2021). However, unlike human teams, taking several hours to communicate and intervene, from image inspection to disaster-relevant report generation, LVLM-driven DisasTeller would only utilise few minutes (GPT-4o with 3/35 input/output tokens per second, Gemma3-27B with 2/10 input/output tokens per second, on a workstation with the NVIDIA A100 80G GPU and Intel Xeon Platinum 8352V CPU) to coordinate all the agents and complete the given damage assessment tasks, thus simplifying the disaster response process and exhibit the potential to accelerate this process through the autonomous procedure.

5. AI agent framework for global-scale defect perception

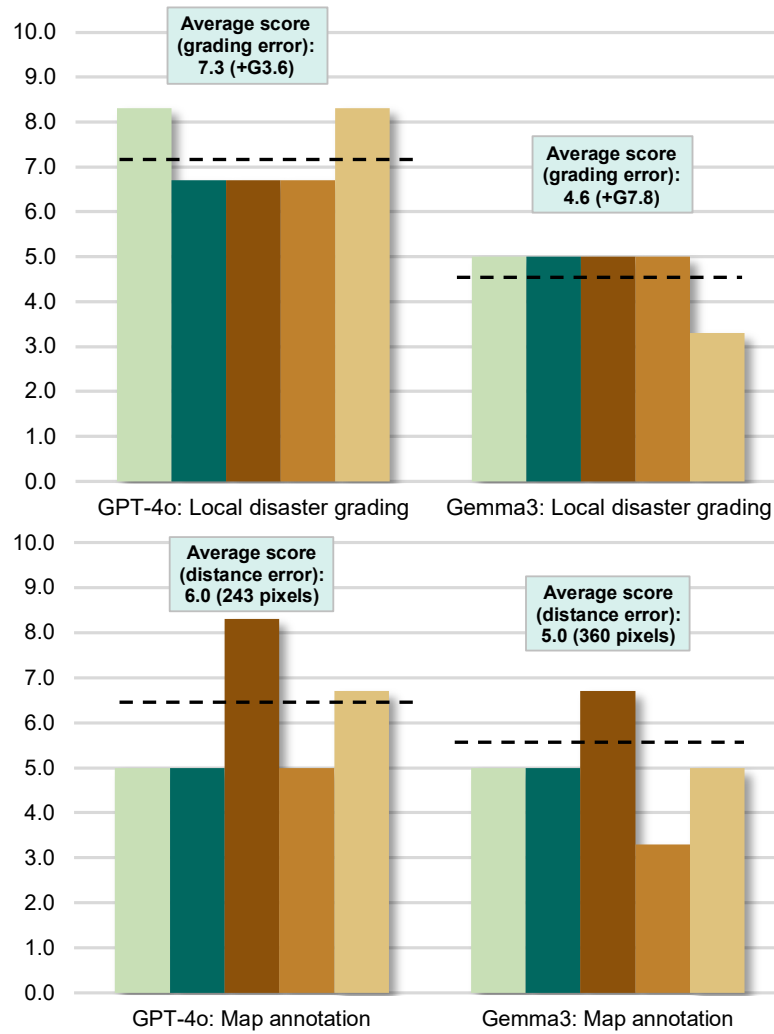


Figure 5.4 The evaluation of intermediate process tasks in GPT-4o- and Gemma3-driven DisasTeller. Each bar plot shows the accuracy scores (0–10 scale) for local disaster grading and map annotation. Five coloured bars represent five independent runs of DisasTeller. The dashed line indicates the mean score across the five runs. The text box above each subplot reports the average accuracy score together with the corresponding deviation error: grading deviation (+G) for local disaster grading and distance deviation (in pixels) for map annotation.

The assessment of intermediate process tasks is crucial for each run of GPT-4o and Gemma3-driven DisasTeller since the incorrect output of an individual agent would spread wrong information within the whole workflow and result in misleading final reports for all the stakeholders in the disaster event. This section evaluates two main intermediate tasks, including local disaster grading and alert map generation. Figure 5.4 summarises the evaluation results across five independent rounds for both local disaster grading and alert map annotation tasks. The disaster grading error is defined through accumulating the deviation of DisasTeller’s damage grading output compared to our manual assessment, and the map annotation distance error is defined through comparing the spatial locations of DisasTeller’s autonomous annotations with our manual annotations.

5. AI agent framework for global-scale defect perception

The average scores achieved by the GPT-4o-driven framework for local disaster grading (7.3/10) and map annotations (6.0/10) are higher than those (4.6/10 for disaster grading and 5.0/10 for map annotations) achieved by the Gemma3-driven framework, which only has 63% and 83% accuracy of the GPT-4o-driven framework in these two tasks. This drop in accuracy would be owing to the parameter discrepancy between Gemma3 (27 billion) and GPT-4o (around 200 billion), which would cause the drop in generalisation ability of Gemma3 compared to GPT-4o. Each bar in Figure 5.4 represents the score for one of the five independent rounds, demonstrating fluctuations in performances of both GPT-4o and Gemma3-driven DisasTeller during the intermediate process tasks. The discrepancies between local disaster grading and map annotation in both results of GPT-4o and Gemma3-driven DisasTeller exhibit the agent's weakness in integrating disaster-image-based and map-based data. Improving the LVLM agent's ability to process information from different sources could help address this gap. The details of the evaluation are introduced in the Supplementary Information.

The output report quality of DisasTeller is crucial for post-disaster management because it would be a foundation for rapid and accurate decision-making. As presented in Figure 5.5, the human evaluator scores exhibit more variability than those provided by the automatic evaluation of EvaluatorGPT, reflecting the varying subject assessments for different persons. The overall scores of GPT-4o-driven results are higher than those of Gemma3-driven results, which would be attributed to the different model scales (Gemma3-27B has 27 billion parameters while GPT-4o has around 200 billion parameters). Reports like the future reconstruction plan, human allocation report and expert team report, which involve more technical, structured planning (budgeting, recovery, and restoration tasks), were rated higher by EvaluatorGPT, likely because the GPT model favours well-defined, structured tasks. On the other hand, reports that require more subjective judgment or social considerations, such as the public notice and alert team report, may have been better appreciated by human evaluators. Although DisasTeller's output reports achieve the matching score from the LVLM and human evaluators, further performance improvement is necessary for its integration into real-scenario disaster management workflow. The details for the human and GPT evaluation of six reports are introduced in the Supplementary Information.

5. AI agent framework for global-scale defect perception

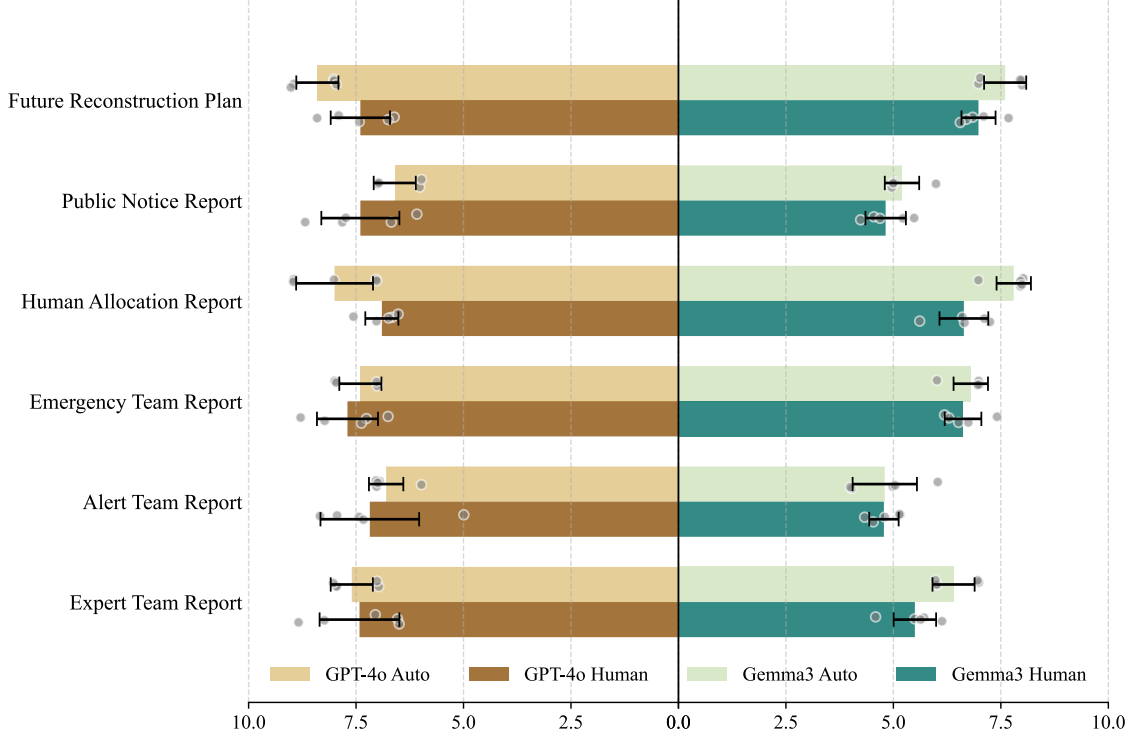


Figure 5.5 The evaluation of output reports generated by GPT-4o- and Gemma3-driven DisasTeller. For each task, five independent runs were evaluated by both EvaluatorGPT and human assessors. Grey points represent individual evaluation scores (0–10 scale). Light and dark yellow bars correspond to the mean automated (EvaluatorGPT) and human scores for GPT-4o-based DisasTeller, respectively; light and dark green bars represent the corresponding mean scores for Gemma3-based DisasTeller. Error bars indicate standard deviations across the five runs.

5.2.3 Limitations and deployment considerations

Although DisasTeller demonstrates the possibility of autonomous post-disaster response tasks completion, integrating multi-LVLM agent-based engines like DisasTeller for post-disaster response still exists several inherent limitations and risks that need to be carefully mitigated to ensure effective and safe deployment.

One prominent concern is the potential for spreading inaccurate or misleading information (LVLM hallucination) in high-stakes situations where flawed decisions may exacerbate disaster impacts (Augenstein, et al., 2024; Yang, et al., 2024). Despite the use of domain-specific prompting, technical guidelines (e.g., EMS-98), and calibrated workflows, our experiments show that errors in early-stage agents (e.g., disaster grading) can propagate to downstream outputs like human allocations and reconstruction planning. However, the system also exhibits corrective behaviour in some cases, revalidating local summaries and overriding initial errors based on visual or contextual cues. To mitigate hallucination risks, our framework is designed to output step-by-step intermediate results (e.g., annotated maps, retrieved references, and

5. AI agent framework for global-scale defect perception

reasoning steps) at each stage, enabling human experts to audit and intervene during execution rather than waiting for final outcomes.

In post-disaster settings, establishing an optimal ground truth is inherently difficult due to the uniqueness of each event, and the absence of repeatable baseline data evaluating variables such as the ‘optimum’ number of doctors or shelters remains a challenge. In our case studies, we addressed this by comparing outputs with publicly available historical records retrieved via web tools. In future work, agents could interface with private databases to access more reliable baselines for self-validation.

Some errors identified during evaluation, such as shelter suggestions near collapsed structures, highlight deficiencies of LVLM-based spatial reasoning and the reliability of the map annotation tool used for geospatial understanding. While the agents can interpret image captions and geolocation inputs, their spatial logic and environmental safety awareness remain constrained. The present annotation module, built primarily for visual-semantic reasoning, lacks integration with established geographic information system (GIS) infrastructures. In future work, coupling DisasTeller’s agents with GIS platforms such as QGIS (Rosas-Chavoya, et al., 2022) could enable access to structured spatial layers (e.g., hazard zones, elevation, infrastructure networks) that can guide safer and more context-aware reasoning. Enhancing spatial awareness through GIS-linked heuristics and formal geospatial constraints (Chrisman, 2005) will therefore be a key direction for improving both reliability and interpretability.

Effective deployment of DisasTeller requires timely access to local-view images or other situational data. In real-world conditions, acquiring such data may involve using drones, satellites, or crowdsourced images. However, logistical barriers in damaged or inaccessible areas, as well as limited access to private or proprietary datasets, remain major challenges. Moreover, any data such as multi-angle, high-resolution images (e.g., aerial and satellite data for wildfire contexts) that can improve standard human assessment would also improve the contribution of DisasTeller. Future collaboration with relevant stakeholders and government agencies for shared image datasets could further strengthen the system’s credibility and better reflect real-world deployment conditions. Our framework assumes data collected by human teams for general assessment can be reused to trigger automated reasoning, thus providing an initial assessment that can become the basis for decision making, reducing delay and guiding more targeted on-site inspections.

5. AI agent framework for global-scale defect perception

As with all neural network-based AI, LVLMS suffer from limited interpretability (Gao, et al., 2023). This poses challenges for acceptance in mission-critical applications like disaster response. While the inherent complexity of neural networks cannot be eliminated, the integration of open-source LVLMS (like Gemma3 in our study) in DisasTeller offers certain advantages in mitigating interpretability concerns. Due to the transparency of model architecture and parameters, open-source LVLMS allow full access to the architecture, weights, and training procedures, enabling researchers and practitioners to inspect how the model operates internally. Open-source LVLMS will also be under collective peer review and benchmarking within the research community, which often leads to faster identification of failure cases and better understanding of model behaviours. Benefiting from the attributes of open-source LVLMS, agents' behaviours in DisasTeller will be more transparent and traceable. In addition, Our framework is designed to output not only final decisions but also the step-by-step reasoning results, and human can inspect the intermediate results (e.g., the visual evidence of annotated damage regions and the historical information retrieved from the internet as a reference for human allocation) without waiting the running finished, thus each intermediate reasoning progress and task output can be checked and allow human experts to verify how each conclusion was reached. Since fully resolving the interpretability challenge remains an open research problem at current stage, to enhance the trustworthiness of DisasTeller for practical deployment, it is important to keep a human check in the running process. This human-in-the-loop format ensures that AI outputs serve as assistive recommendations rather than unquestioned directives, thereby reducing the risk of over-reliance on black-box results.

In real-world disasters, available resources are often insufficient, and planning must balance urgency, impact, and logistical constraints. The scarcity of rescue forces and relief supplies makes mathematically optimal allocation nearly impossible within short response times. DisasTeller's human allocation module does not perform strict mathematical optimisation in the sense of operations research. Instead, it emulates rule-based prioritisation and dynamic reallocation, reflecting how human decision-makers act under uncertainty. Specifically, it integrates spatial, temporal, and historical cues to prioritise areas with higher estimated damage severity, thereby improving allocation relevance rather than guaranteeing optimality. DisasTeller is designed as a decision-support framework that assists human planners in rapidly generating and revising preliminary allocation plans. Future works could integrate real-time resource inventories or supply chain data and introduce more specific priority rules to improve scheduling under constrained scenarios.

5. AI agent framework for global-scale defect perception

The current retrieval-augmented generation component relies highly on web search to obtain reference information. While this ensures adaptability, the reliability of open-web data can fluctuate. Future iterations should integrate more human-curated or institutional knowledge bases to enhance factual grounding, especially for region-specific response standards and resource benchmarks.

As a decision-support tool, DisasTeller is not intended to replace human authority. All final decisions remain the responsibility of human operators and implementing organisations. To maintain ethical deployment, robust risk management, legal compliance, and stakeholder oversight are necessary. In the case of incorrect AI-generated decisions, traceability and transparency of agent logic will be essential for accountability.

Given the rapid evolution of AI, maintaining system relevance over time is a challenge. To address this, DisasTeller is built with a modular architecture that allows seamless substitution of new LVLMs, tools, and reasoning modules. Periodic retraining, version control, and embedded performance monitoring can ensure future upgrades remain aligned with the system’s safety and usability requirements.

5.2.4 Summary

This study develops a multi-LVLM-agent-based framework DisasTeller for autonomous disaster response, aiming to accelerate the information sharing and team response in the disaster event, reduce disaster-relevant losses, and alleviate regional inequality in disaster response by its availability across developed and underdeveloped areas. Integrating the advanced reasoning ability of LVLMs with expert knowledge from external tools, DisasTeller can coordinate multiple LVLM agents and autonomously simulate various activities of human teams in the disaster event, including local on-site damage evaluation, alert map and disaster reports generation.

Assessment results showed that DisasTeller can reduce the time needed for coordination and task completion compared to the current human-based disaster response procedure. However, the efficiency gains must be interpreted alongside non-negligible limitations in accuracy. Our evaluation reveals that errors in early-stage agents, such as inaccurate disaster grading, can propagate to downstream decision outputs. While report-level evaluations show alignment with human judgment, intermediate results, particularly in image-grounded damage assessment and map annotation, expose weaknesses in the current LVLM spatial and contextual reasoning

5. AI agent framework for global-scale defect perception

capacity. These findings reinforce that DisasTeller is intended as a complementary support tool, not a replacement for human disaster experts.

Further development is required to improve task-specific LVLM accuracy, integrate real-time multimodal sensor data, and strengthen interactions with human decision-makers in the loop. Augmenting the system with fine-tuned LVLMs and specialised tools beyond the current framework could also enhance reliability under real-world constraints.

Additionally, challenges in evaluation stem from the lack of universal ground-truth benchmarks and the non-deterministic nature of LVLM outputs. Due to random sampling during inference (Brown, 2020), repeated executions may yield slightly different results. Although the temperature parameter can regulate generation variability [17], the deterministic output would cause bias and reduce LVLM's problem-solving capability, particularly in disaster scenarios requiring diverse reasoning approaches (Brown, 2020). Balancing output consistency with reasoning richness remains an open research direction.

In conclusion, despite current limitations, our work demonstrates the potential of multi-agent LVLM frameworks like DisasTeller to improve the speed, structure, and accessibility of post-disaster response. Continued collaboration between AI developers, domain experts, and field responders will be essential to realise safe, effective, and trustworthy AI-assisted disaster management systems.

5.3 Methods

5.3.1 Large Vision Language Models (LVLMs) with augmented generation

The rapid rise of LVLMs in recent years has created unprecedented opportunities across various domains. Based on pretrained knowledge from the large amount of data, one of the key factors behind the growing capabilities of LVLMs is their ability to perform zero-shot reasoning (solving tasks based on pretrained knowledge without specific re-training on those tasks) (Yao, et al., 2022). This ability is significantly amplified when LVLMs are used within advanced frameworks exhibiting the power of chain-of-thought reasoning (Wei, et al., 2022). LVLMs generate step-by-step logical explanations that can interact with external systems and databases through tool usage, rendering LVLMs' ability to be intelligent agents (Schick, et al., 2024). However, the zero-shot reasoning of LVLMs has restrictions and low accuracy in domain-specific tasks. To mitigate this drawback, we supplement with domain-specific prompts and external technical guidelines (e.g., EMS98) to enhance the performance of our framework in

5. AI agent framework for global-scale defect perception

this disaster domain. This is identified as in-text learning (Dong, et al., 2022) and retrieved augmented generation (Lewis, et al., 2020), where external knowledge is injected into prompts to enhance model reliability without retraining. For practical deployment, further results could be achieved with specialised training. Additionally, to make LVLMs accessible to developers and businesses, Application Programming Interfaces (APIs) are commonly employed. In this study, we streamlined our framework by utilising the API of the most representative LVLM model GPT (version ‘gpt-4o-2024-08-06’), and we also conducted the additional validation through the local deployment and integration of Google’s open-source LVLM model Gemma3-27B. The modularity of our framework allows for future integration of fine-tuned LVLMs for improved specialisation and robustness.

5.3.2 Custom multi-LVLM agent framework design

While foundational tools like LangChain (Chase, 2022) and CrewAI (Moura, 2024) provide essential infrastructure for chaining prompts and coordinating multi-agent systems, our work contributes a domain-specific, multi-role LVLM agent framework specifically tailored to automate tasks in disaster response. We designed a system where each LVLM-based agent is assigned a clearly defined functional role, such as damage grading, emergency alert generation, and human resource planning, based on the operational needs of real-world disaster workflows. Building upon CrewAI, we customised the role-specific memory management to ensure each agent retains its context across interactions, developed the inter-agent communication protocols to enable agents to request and refine each other’s output, and designed the dynamic task allocation allowing the system to reassign subtasks based on runtime performance or conflict resolution.

Unlike prior generic implementations of CrewAI or LangChain, our framework demonstrates a vertical integration of multi-agent reasoning, domain grounding, and multi-tool interaction designed explicitly for structured coordination in disaster scenarios.

5.3.3 Tools development

To maximise the capabilities of each LVLM agent, we designed a set of tools, mapped to agents’ responsibilities and modularly integrated into the pipeline of our framework. It is necessary to highlight that additional tools can be easily integrated into this toolset according to requirements due to the modularity of our framework. A key innovation in our design is the use of adaptive tool invocation, where each agent determines, based on task complexity and confidence thresholds, whether external tools are needed to enhance output quality. This

5. AI agent framework for global-scale defect perception

conditional reasoning enables greater efficiency, avoids over-reliance on tools, and mimics how human teams selectively consult external resources.

ImageInterpretationTool. To simplify the multi-agent framework of DisasTeller, we integrate two LVLMs into the local and global image interpretation tools instead of defining extra agents as the image interpreters. When these two image interpretation tools are invoked, they will start to read the content of local and global images and give responses for detailed descriptions of the image contents. This tool set helps different team agents check the local and global image information freely.

WebSearchTool. This tool is designed to empower the LVLM with the capability to access information from online sources, utilising LangChain (Chase, 2022), DuckDuckGo Search (2024) and BeautifulSoup (Richardson, 2007) APIs. For a given invocation, this tool retrieves the top two search results using DuckDuckGo search engine and returns them as a list. After accessing the webpage links in the list, BeautifulSoup library is employed to parse the website content. The text from the webpage is extracted and truncated to avoid excessive token usage. This helps limit the amount of content processed, ensuring the relevant parts of the webpage are provided without overwhelming the system. Incorporating this tool allows DisasTeller to effectively broaden its knowledge repository through the up-to-date information online.

FileReadTool. The file read tool focuses on extracting relevant data from various file formats, such as PDFs, text files, and DOCX documents, to assist LVLM agents in analysing and processing information. When this tool is invoked, agents will leverage the file read tool API in CrewAI (Moura, 2024). The documents are first pre-processed into vector embeddings, which are mathematical representations of the text. Once relevant documents or information have been retrieved, the tool generates a summary relevant to the query. This summary is then integrated into the prompt to activate the LVLM’s potential for a more informed response. Instead of relying solely on the LVLM’s internal knowledge, this step augments the LVLM’s ability with specific information retrieved from the external documents, ensuring that DisasTeller’s response is grounded in actual data, thus mitigating the hallucination to a certain degree.

MapAnnotationTool. This tool aims to reflect the local on-site disaster grades in the global map, thus creating an alert map to illustrate the overall conditions in this disaster area. We customised this tool by integrating an LVLM to analyse the global map image and the input instruction from other agents. When this tool is invoked, it first retrieves the prompt from other agents

5. AI agent framework for global-scale defect perception

with the local area names and their corresponding disaster grades and then identifies the location name in the global map image. Employing the reasoning ability of the LVLM, this tool outputs the coordinates of all relevant locations and calls Python image processing libraries (Lundh and Ellis, 2012) to generate the news alert map.

5.3.4 Agents definition for post-disaster management

In our framework, each LVLM-based agent was defined through a combination of three core components: a role specification, a model assignment, and a set of task-relevant tools. This modular structure enables each agent to behave intelligently and autonomously across different stages of post-disaster response (Schick, et al., 2024). To simulate the collaborative workflow of real-world disaster response teams, we defined four distinct agents including the expert team, alerts team, emergency team, and assignment team, each representing a key functional unit in disaster operations. These agents were instantiated with independent LVLM instances and customised prompts to reflect their unique responsibilities. Table 5.1 presents the LVLM agents we defined with a certain prompt input in our experiments. Agents interact through the CrewAI coordination layer, which facilitates inter-agent communication, memory exchange, and dynamic task sequencing. Through this agent-oriented architecture, we modelled a multi-LVLM-powered disaster response framework that closely mirrors the structure and decision flow of actual post-disaster management teams. The complete prompt templates and system instructions are available at <https://github.com/zche3016/DisasTeller/tree/main>.

Table 5.1 Details of the agent definition in DisasTeller framework

Agents	Toolset	Task prompt input
Expert team	Image interpretation tools, local file search tool, map annotation tool	‘1. Generate descriptions for post-disaster images using the image interpretation tools; 2. Get post-disaster grading with their location names using the local file search tool; 3. Send the concise summary of disaster grades to the map annotation tool for the alert map generation; 4. Generate a report with the given format: ...’
Alerts team	Global image interpretation tool, web search tool	‘Analyse the output from the Expert team, generate a comprehensive social media news that describes the damaged events for regional areas, analyse potentially dangerous area according to the global map using the global image interpretation tool.’
Emergency team	Image interpretation tools, web search tool	‘Analyse the outputs from the Expert team and Alerts team. Then, ensure the number and locations of emergent shelters and hospitals using image interpretation tools and web search tool to take a reference from other areas' historical disaster data. Also, confirm

5. AI agent framework for global-scale defect perception

		which place needs immediate transportation recovery. Finally, generate a report with given format: ...’
Assignment team	Global image interpretation tool, web search tool	‘Analyse the outputs from the Expert team, Alerts team and Emergency team. 1. Generate a report for the emergency service team human resource allocation, search the internet to take a reference from other area historical disaster data. The report format: ... 2. Generate a report to inform the public communities, with the given format: ... 3. Generate a reconstruction plan with financial budgets, aiming to submit it to the government and policy maker, with the given format: ...’

Appendix 5.A

A significant concern when integrating LVLMS into scientific workflows is reproducibility, since the same instruction prompt for LVLMS can produce varying outputs. To evaluate the reproducibility of DisasTeller in implementing given tasks and report generating for post-disaster response, five independent runs of DisasTeller embedded with GPT-4o (“gpt-4o-2024-08-06”) (Achiam, et al., 2023) and Google’s open-source Gemma3-27B (Team, et al., 2025) kernels, respectively, were conducted on a high-performance workstation with the A100 80G GPU and Intel Xeon Platinum 8352V CPU to address it. In the five independent runs, intermediate process tasks of DisasTeller were evaluated manually and the output reports were assessed through both EvaluatorGPT and humans having civil engineering backgrounds with construction experience. The detailed process is introduced in the following sections.

5.A.1 Intermediate process tasks

In the execution of DisasTeller, the first task is to grade the local post-disaster on-site images based on local technical guideline file EMS98 (Grünthal, 1998). In Figure 5.6-a, there are six damaged images with different locations in Wajima city. We manually assess the grading of these damaged images according to EMS98: Wajima Drama Memorial Hall – G1 - G2; Hama Street – G4 - G5; Concrete Bridge – G5; Central Nishikigawa Street – G1 - G2; North Asaichi Street – G3 - G4; South Central Asaichi Street – G4 - G5. We also define the total disaster grading error through accumulating the deviation of DisasTeller’s damage grading output compared to ground truth (our manual assessment for the damage grading). For example, when DisasTeller’s output only has the wrong grading G3 for Central Nishikigawa Street (ground

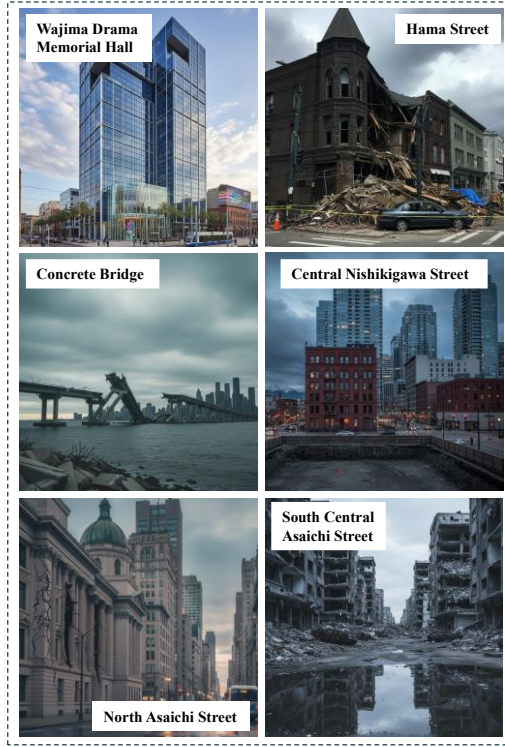
5. AI agent framework for global-scale defect perception

truth ranges from G1 - G2), we calculate the difference between G3 and the lower limit of ground truth (G1) to get the final total grading error +G2. According to the manual evaluation results, Figure 5.6-b presents the five independent-runs grading outputs of GPT-4o and Gemma3 driven DisasTeller where texts in orange show the wrong generation and texts in blue exhibit the final scores with the total grading error, we provide for DisasTeller. The errors in Figure 5.6-b highlight several misclassifications (e.g., Wajima Drama Memorial Hall – G3 is consistently incorrect across several rounds). The Gemma3 driven DisasTeller shows inferior accuracy than the GPT-4o driven one, which would be owing to the parameter discrepancy between Gemma3 (27 billion) and GPT-4o (around 200 billion). The evaluation result implies that LVLM agents make repetitive mistakes with specific locations. Refinement of agents' models for such tasks could improve overall consistency.

After the first local disaster grading task, the subsequent execution of DisasTeller is to annotate the damage grades in the relevant locations of the global map, namely generate the alert map to visualise the overall situation in this disaster area. We inspect the generated annotations in each independent run and evaluate the accuracy of the annotations by determining whether the annotations are close to the ground truth (our manual annotations) on the global map. The map annotation distance error is defined through comparing the spatial location of DisasTeller's autonomous annotations with ground truth. In the global map image with 1080×1522 pixels, we define the distance error grades: 0 – 100 pixels (G1), 100 – 200 pixels (G2), 200 – 250 pixels (G3), 250 – 350 pixels (G4), larger than 350 pixels (G5). The distance error grades larger than G3 (250 pixels) is considered as wrong spatial annotation and marked as red cross. Figure 5.7 and Figure 5.8 show the final evaluation of GPT-4o and Gemma3 driven DisasTeller's annotations in the global map with the wrong annotation marked as red cross and the average distance error of six locations in each round provided. According to the results of Figure 5.7 and Figure 5.8, both GPT-4o and Gemma3 driven LVLM agents are prone to errors in spatial understanding or accurately mapping disaster grades to geographic locations. Compared with the GPT-4o driven results, Gemma3 driven results sometimes cannot provide complete annotations for all local places with their respective damage grades. This inferior image understanding ability would be attributed to the fewer parameters of Gemma3, which would cause the generalisation ability to drop. Considering the limitation of spatial understanding, task-specific fine-tuning for LVLMs could reduce the error rate and improve the agent's ability to analyse geospatial information.

5. AI agent framework for global-scale defect perception

a Local post-disaster on-site images



b GPT-4o and Gemma3 driven DisasTeller output evaluation

	Location	GPT-4o	Gemma3
Round 1	Wajima Drama Memorial Hall	G1	G3-G4 (Wrong)
	Hama Street	G4	G4-G5
	Concrete Bridge	G5	G3-G4 (Wrong)
	Central Nishikigawa Street	G3 (Wrong)	G3-G4 (Wrong)
	North Asaichi Street	G4	G4-G5
	South Central Asaichi Street	G5	G4-G5
	Score (total grading error)	5/6 (+G2)	3/6 (+G7)
Round 2	Wajima Drama Memorial Hall	G3 (Wrong)	G4 (Wrong)
	Hama Street	G4	G3 (Wrong)
	Concrete Bridge	G4-G5	G5
	Central Nishikigawa Street	G5 (Wrong)	G5 (Wrong)
	North Asaichi Street	G4	G4
	South Central Asaichi Street	G5	G4
	Score (total grading error)	4/6 (+G6)	3/6 (+G8)
Round 3	Wajima Drama Memorial Hall	G3 (Wrong)	G3 (Wrong)
	Hama Street	G4	G5
	Concrete Bridge	G5	G5
	Central Nishikigawa Street	G3 (Wrong)	G4 (Wrong)
	North Asaichi Street	G4	G5 (Wrong)
	South Central Asaichi Street	G5	G4
	Score (total grading error)	4/6 (+G4)	3/6 (+G7)
Round 4	Wajima Drama Memorial Hall	G3 (Wrong)	G3 (Wrong)
	Hama Street	G5	G5
	Concrete Bridge	G5	G5
	Central Nishikigawa Street	G3 (Wrong)	G5 (Wrong)
	North Asaichi Street	G4	G5 (Wrong)
	South Central Asaichi Street	G5	G5
	Score (total grading error)	4/6 (+G4)	3/6 (+G8)
Round 5	Wajima Drama Memorial Hall	D1	G3 (Wrong)
	Hama Street	D4	G3 (Wrong)
	Concrete Bridge	D4-D5	G5
	Central Nishikigawa Street	D3 (Wrong)	G5 (Wrong)
	North Asaichi Street	D4	G5 (Wrong)
	South Central Asaichi Street	D5	G5
	Score (total grading error)	5/6 (+G2)	2/6 (+G9)

Figure 5.6 Evaluation of local post-disaster on-site images grading: texts in orange represent the wrong outputs of DisasTeller; texts in blue are the scores of accuracy percentages and total grading error.



Figure 5.7 Evaluation of global map annotations: GPT-4o driven DisasTeller's autonomous annotations are in yellow, and annotations with distance error larger than 250 pixels are identified as wrong and marked as red cross.

5. AI agent framework for global-scale defect perception



Figure 5.8 Evaluation of global map annotations: Gemma3 driven DisasTeller’s autonomous annotations are in yellow, and annotations with distance error larger than 250 pixels are identified as wrong and marked as red cross.

5.A.2 Output reports evaluation

The quality of output reports is essential for DisasTeller’s integration in the post-disaster real-scenario workflows. In the output report, evaluating the decision variables (e.g., the number of shelters, hospitals, human allocations, and the reconstruction budget) is complex. Even in a real post-disaster intervention, human teams would typically not be able to optimise a solution since decisions have to be made within a short span and there is not enough information from previous identical disasters to make an informed evaluation and assessment. Natural disasters are rare and tend not to repeat in the same conditions. That said, we addressed this by comparing the agent recommendations with historical data obtained online (the literatures retrieved by the web search tool, see Figure 5.9) in the evaluation, since there is lack of private disaster data, to confirm that the resource allocation outputs obtained are within realistic boundaries. In future work, stakeholders can fine-tune LVLM with private data or let agents autonomously retrieve the information from their private database to achieve a more reliable performance in specific assignments.

5. AI agent framework for global-scale defect perception

```
[2025-04-28 21:47:54][DEBUG]: == Working Agent: Emergency team
[2025-04-28 21:47:54][INFO]: == Starting Task... use web_search_tool searching:

[{'title': 'Hospitals as disaster victims: Lessons not learned? - PMC', 'href': 'https://pmc.ncbi.nlm.nih.gov/articles/PMC8749465/',
'body': 'Temporary shelter/triage area set up and transfer of patients to next door hospital with disaster hospital staff caring for
them, runners used to communicate ... In a review of hospital evacuations after an earthquake, ... Kanatani Y. Disaster
Vulnerability of hospitals: a nationwide surveillance in Japan. Disaster Med Public Health Prep. 2015 ...', {'title': 'An Overview of
the Design of Disaster Relief Shelters', 'href': 'https://www.sciencedirect.com/science/article/pii/S2212567114010193', 'body':
'Scoping Study to inform the Revision of "Shelter after Disaster: Guidelines for Assistance . Wu, J. Y. & Lindell, M. K. 2004.
Housing reconstruction after two major earthquakes: The 1994 Northridge earthquake in the United States and the 1999 Chi-Chi
earthquake in Taiwan. Disasters, 28 , 63-81.'}]...

[2025-04-28 21:49:54][DEBUG]: == Working Agent: Assignment team
[2025-04-28 21:49:54][INFO]: == Starting Task... use web_search_tool searching:

[{'title': 'Historic Disasters - FEMA.gov', 'href': 'https://www.fema.gov/disaster/historic', 'body': 'The following disasters are
considered historical because of how they impacted the way we handle similar disasters in the future. ... which caused a significant
reevaluation of the execution of federal disaster response efforts and resource allocation. ... The Northridge Earthquake of 1994
tested the new reforms that followed Hurricane Andrew ...', {'title': 'Optimized resource allocation for emergency response after
earthquake ...', 'href': 'https://www.sciencedirect.com/science/article/pii/S0925753500000217', 'body': 'The main influencing factor
for optimizing resource allocation in disaster response is time. The reason for this is the time limit of the SAR period, conditional
on the goal to minimize the total number of fatalities. ... Coburn, A., Spence, R., 1992. Factors determining human casualty levels
in earthquakes: mortality prediction in building ...'}]...

[2025-04-28 21:51:25][DEBUG]: == Working Agent: Assignment team
[2025-04-28 21:51:25][INFO]: == Starting Task... use web_search_tool searching:

[{'title': 'PDF Guidance Note Financing Post-Disaster Recovery and Reconstruction ...', 'href':
'https://openknowledge.worldbank.org/bitstream/handle/10986/22418/Financing0post0rces000guidance0note.pdf?sequence=1',
'body': '3 Clarify post-disaster financial responsibilities of the central and local governments 4 Mobilize resources ex-ante through
a DRM fund and financial protection strategy 5 Allocate resources based on damage and loss assessment methodology 6 Execute
resources and coordinate efforts across the post-disaster recovery and reconstruction process'}, {'title': 'Unearthing the impact of
earthquakes: A review of economic and social ...', 'href': 'https://onlinelibrary.wiley.com/doi/full/10.1002/pam.22642', 'body': 'For
instance, the March 2011 Great East Japan Earthquake led to substantial costs for Japan's government, amounting to 4% of Japan's
2010 GDP. Funding covered disaster relief, recovery, and reconstruction, and utilized general contingency budgets and
supplementary budgets financed through bonds, expenditure cuts, and budget surpluses.'}]...
```

Figure 5.9 A sample for the content retrieved from the internet by the web search tool

Considering the extensive text content in the six generated reports from DisasTeller, we assess them using both the LVLMM (GPT-4o)-based automatic evaluation and human inspection. GPT-4o is prompted to act as an expert for post disaster management (referred to as EvaluatorGPT), specialising in disaster impact assessment, emergency response planning, and recovery evaluation, and is provided with post-disaster on-site images, documents retrieved from the internet, and the evaluation rubric (see Table 5.2) to enhance the accuracy of its assessments. The evaluation rubric for EvaluatorGPT and HumanEvaluator are based on the coherence and consistency of sentences and the accuracy of claims.

Figure 5.10 and Figure 5.13 present the samples of the EvaluatorGPT and human assessment reports for the identical reports in the same round, with the detailed explanations for the DisasTeller reports' weakness. In addition, the assessment summaries of EvaluatorGPT for GPT-4o and Gemma3 respectively driven DisasTeller output reports are shown in Figure 5.11 and Figure 5.12, and HumanEvaluator (averaging the scores of two persons) are shown in Figure 5.14 and Figure 5.15. While EvaluatorGPT may be more stable and less prone to large

5. AI agent framework for global-scale defect perception

score variations, HumanEvaluator introduces more variance, which could stem from changing perceptions of report quality across rounds or more nuanced judgment.

Table 5.2 The evaluation rubric for DisasTeller output reports

Dimension	Sub-Dimension	Evaluation Focus	Score Range	Detailed Notes
1. Sentence consistency	1.1 Logical Organization	Are sections clearly structured (Intro, Damage, Needs, Recommendations, etc.)? Does the report avoid random jumping between ideas?	0–2	2: Fully structured and logical; 1~2: Mostly structured but minor jumps; 0~1: Poorly structured.
	1.2 Consistent Terminology	Are terms like "Grade D4-D5", "Emergent Shelter", "Critical Area" used consistently across sections?	0–2	2: Full consistency; 1~2: Minor inconsistencies; 0~1: Frequent switching of terms/confusion.
	1.3 Consistent Argumentation	Are the cause-effect relationships explained consistently? (e.g., Damage → Need → Action Plan.)	0–1	1: Clear and consistent; 0~1: Missing or contradictory flow.
Subtotal			0–5	
2. Wrong Claims or Overclaims	2.1 Factual Accuracy	Are descriptions faithful to the available evidence (e.g., damage images, situation)?	0–3	3: Highly accurate; 1~3: Medium exaggerations; 0~1: Major inaccuracies.
	2.2 Avoiding Speculative Statements	Does the report avoid overly speculative or unsupported claims (e.g., "hundreds of dead" without evidence)?	0–2	2: No speculative claims; 0~1: Speculative or unverified claims present.
Subtotal			0–5	
Total			0–10	

5. AI agent framework for global-scale defect perception

1. Expert Team Report: The Expert Team Report has acceptable structure with clear sections, but some transitions (like damage descriptions leading to general conclusions) feel abrupt, earning **1/2 for Logical Organization**. Terminology such as "Grade 3" and "Damage Grading" is consistent (**2/2**), and the cause-effect reasoning between damage and assigned grades is coherent (**1/1**). In terms of factual accuracy, while most descriptions match the images, risk factors like secondary collapses are underemphasized, so **2/3** is awarded. Additionally, optimistic statements about resilience without direct evidence cause a minor deduction for speculation (**1/2**). Therefore, the final strict score is **7.0/10**.

2. Alert Team Report: The Alerts Team Report is clearly structured with consistent use of labels like "Grade 4" and "Safety Advisory," scoring full points for **Logical Organization (2/2)** and **Terminology (2/2)**, and the argument from damage to risk advisory is logical (**1/1**). However, a critical factual error occurs by labeling Wajima Drama Memorial Hall as only minimally damaged despite visible structural issues, thus **1/3 for Factual Accuracy**, and the advisories are broad and not carefully differentiated by location severity, slightly speculative, scoring **1/2**. Thus, the overall rubric-aligned score is **7.0/10**.

3. Emergency Team Report: The Emergency Team Report is logically organized, detailing each location's situation, needs, and shelter plans (**2/2**), with consistent technical terminology (**2/2**) and a clear link between observed damage and recommended emergency measures (**1/1**). While factual descriptions mostly match observed conditions, the number of proposed shelters is slightly aggressive without full population data support, resulting in **2/3 for Factual Accuracy**, and the conclusion about system-level strategy improving efficiency is somewhat speculative (**1/2**), leading to a strict final score of **8.0/10**.

4. Human Resource Allocation Report : The Human Resource Allocation Report is very well structured, moving systematically through affected areas and required personnel (**2/2**), maintains perfect consistency in using terms like "Medical Personnel" and "Rescue Teams" (**2/2**), and has clear, logical cause-effect chains between severity and resource allocation (**1/1**). Factual deployment numbers closely match actual damages observed, justifying **3/3 for Factual Accuracy**, and while minor future risk projections are mentioned without deep evidence (**1/2 for Speculation**), the report remains largely precise, resulting in a final score of **9.0/10**.

5. Public Notice Report: The Public Community Report lacks strong logical focus, mixing local earthquake impacts with irrelevant global disaster statistics like FEMA and NOAA data (**1/2 for Organization**), and its terminology shifts between global and local events inconsistently (**1/2**), with a weak link between needs and actions (**0/1 Argumentation**). Factual descriptions are vague compared to real disaster images (**2/3 for Factual Accuracy**), but no speculative or exaggerated claims are made (**2/2**), resulting in a strict rubric-based score of **6.0/10**.

6. Future Reconstruction Plan: The Future Reconstruction Plan is excellently structured into Immediate, Short-term, and Long-term phases (**2/2 Logical Organization**), uses consistent terms like "Phased Recovery" and "Risk Reduction" (**2/2 Terminology**), and builds a logical flow from observed damage to recovery steps (**1/1 Argumentation**). Factual references are generally adapted well to local needs even though some costs are derived from global comparisons (**3/3 for Factual Accuracy**), and minor overgeneralizations about financial capacity result in a small deduction for speculation (**1/2**), producing a final score of **9.0/10**.

Figure 5.10 A sample of the EvaluatorGPT assessment report

EvaluatorGPT	Round 1	Round 2	Round 3	Round 4	Round 5	Average
Expert Team Report	7.0/10	8.0/10	7.0/10	8.0/10	8.0/10	7.6/10
Alert Team Report	7.0/10	7.0/10	6.0/10	7.0/10	7.0/10	6.8/10
Emergency Team Report	8.0/10	8.0/10	7.0/10	7.0/10	7.0/10	7.4/10
Human Allocation Report	9.0/10	9.0/10	7.0/10	8.0/10	7.0/10	8.0/10
Public Notice Report	6.0/10	7.0/10	6.0/10	7.0/10	7.0/10	6.6/10
Future Reconstruction Plan	9.0/10	9.0/10	8.0/10	8.0/10	8.0/10	8.4/10

Figure 5.11 EvaluatorGPT assessment summary for five GPT-4o-driven independent-run output reports.

5. AI agent framework for global-scale defect perception

EvaluatorGPT	Round 1	Round 2	Round 3	Round 4	Round 5	Average
Expert Team Report	7.0/10	6.0/10	7.0/10	6.0/10	6.0/10	6.4/10
Alert Team Report	6.0/10	5.0/10	5.0/10	4.0/10	4.0/10	4.8/10
Emergency Team Report	7.0/10	6.0/10	7.0/10	7.0/10	7.0/10	6.8/10
Human Allocation Report	8.0/10	7.0/10	8.0/10	8.0/10	8.0/10	7.8/10
Public Notice Report	5.0/10	5.0/10	5.0/10	5.0/10	6.0/10	5.2/10
Future Reconstruction Plan	8.0/10	8.0/10	8.0/10	7.0/10	7.0/10	7.6/10

Figure 5.12 EvaluatorGPT assessment summary for five Gemma3-driven independent-run output reports.

1. Alerts Team Report for Social Media

- **Score: 8.1/10**
- Sentence Coherence and Consistency: The most parts of this report are well organized, but the reuse of damage descriptions seems to be kind of redundant (**1.1 1.6/2 points; 1.2 2/2 points; 1.3 1/1 points**).
- Wrong Claims/Overclaims: The description of Central Nishikigawa Street as "Moderate damage" (Grade 3) is an overclaim, but this mistake is not too far from the ground truth (**2.1 2.5/3 points**), and the safety advisories are also affected by this wrong claim (**2.2 1/2 points**).

2. Emergency Service Human Resource Allocation Report

- **Score: 7.7/10**
- Sentence Coherence and Consistency: The report is structured but kind of suffers from redundancy and need to provide more concise summary for human allocation (**1.1 1.7/2 points; 1.2 2/2 points; 1.3 1/1 points**).
- Wrong Claims/Overclaims: The proposed number of rescue workers and medical personnel is kind of large (**2.1 2/3 points**). The report still needs to fully consider the possible scarcity of humans. This may be a potential overclaims about human resource availability and the allocation (**2.2 1/2 points**).

3. Emergency Team Report

- **Score: 7.1/10**
- Sentence Coherence and Consistency: The most parts are good in this report, the sentence structure and consistency in the first part can be further improved (**1.1 1.6/2 points; 1.2 2/2 points; 1.3 0.5/1 points**).
- Wrong Claims/Overclaims: The suggestion of setting up the hospitals for every location is under consideration. Some places may not need the hospital and it could be an overclaim (**2.1 2/3 points; 2.2 1/2 points**).

4. Expert Team Report

- **Score: 9.3/10**
- Sentence Coherence and Consistency: The report's sentences are well organised (**1.1 2/2 points; 1.2 2/2 points; 1.3 1/1 points**).
- Wrong Claims/Overclaims: The description of Central Nishikigawa Street as "Moderate damage" (Grade 3) seems an overclaim, but this mistake is not too far from the ground truth (**2.1 2.5/3 points**). Most descriptions in the report match the images (**2.2 1.8/2 points**).

5. Public Community Report

- **Score: 7.5/10**
- Sentence Coherence and Consistency: The report can reduce the redundant description for the Introduction part and should keep consistent terminology for different events (**1.1 1.8/2 points; 1.2 1.5/2 points; 1.3 1/1 points**).
- Wrong Claims/Overclaims: This report has irrelevant claims about the current disasters, mixes the data from global disaster statistics from FEMA and NOAA, and it is not sure which area the claim points to: "The affected areas faced catastrophic events, including severe storms, flooding, and hurricanes, causing significant structural damage and necessitating immediate emergency response" (**2.1 2/3 points; 2.2 1.2/2 points**).

6. Reconstruction Plan Report

- **Score: 7.8/10**
- Sentence Coherence and Consistency: The report is mostly structured with clear description for reconstruction plan (**1.1 1.8/2 points; 1.2 2/2 points; 1.3 1/1 points**).
- Wrong Claims/Overclaims: In the first part of this report, the irrelevant claims with the global disasters are introduced, and the speculative statements for the grade 5 damaged areas are succeeded from previous outputs (**2.1 2/3 points; 2.2 1/2 points**).

Figure 5.13 A sample of the human assessment for the same reports evaluated by EvaluatorGPT.

5. AI agent framework for global-scale defect perception

HumanEvaluator	Round 1	Round 2	Round 3	Round 4	Round 5	Average
Expert Team Report	8.8/10	6.5/10	7.1/10	6.5/10	8.2/10	7.4/10
Alert Team Report	7.4/10	5.0/10	7.9/10	7.3/10	8.3/10	7.2/10
Emergency Team Report	6.8/10	7.3/10	8.8/10	8.2/10	7.4/10	7.7/10
Human Allocation Report	7.6/10	6.7/10	6.5/10	6.7/10	7.0/10	6.9/10
Public Notice Report	6.7/10	6.1/10	8.7/10	7.8/10	7.7/10	7.4/10
Future Reconstruction Plan	7.4/10	6.6/10	7.9/10	6.7/10	8.4/10	7.4/10

Figure 5.14 Human assessment summary for five GPT-4o-driven independent-run output reports.

HumanEvaluator	Round 1	Round 2	Round 3	Round 4	Round 5	Average
Expert Team Report	6.1/10	5.5/10	5.7/10	5.6/10	4.6/10	5.5/10
Alert Team Report	5.2/10	4.3/10	4.8/10	4.5/10	5.1/10	4.8/10
Emergency Team Report	7.4/10	6.7/10	6.5/10	6.2/10	6.3/10	6.4/10
Human Allocation Report	7.2/10	5.6/10	6.6/10	7.1/10	6.7/10	6.6/10
Public Notice Report	5.5/10	4.5/10	4.2/10	5.2/10	4.7/10	4.8/10
Future Reconstruction Plan	7.7/10	6.8/10	6.7/10	7.1/10	6.6/10	7.0/10

Figure 5.15 Human assessment summary for five Gemma3-driven independent-run output reports.

5.A.3 Error propagation study

Since the sequential output of our framework are disaster grading, map annotation and output reports, our previous report evaluation for the output reports assessed the obvious error propagation throughout the framework (e.g., when the initial expert team agent generates wrong disaster grading “Central Nishikigawa Street: G3”, then the alerts team, assignment team and emergency team agents will reuse this wrong information and keep generating “Central Nishikigawa Street: G3” with the relevant speculative statements, which are previously evaluated by EvaluatorGPT and human evaluators).

For the additional error propagation, we studied how the initial disaster grading error by the expert team agent will affect the later key downstream decision variables generated by the other three agents (the num of shelters, hospitals, human allocations, and the reconstruction budget). The map annotation spatial variance errors do not change the disaster grading information and have no influence for the later downstream decision variables, since the later three agents

5. AI agent framework for global-scale defect perception

generate these downstream decision variables based on local disaster descriptions with their respective grades which are provided by the initial expert team agent.

We conducted two experiments including the positive error (e.g., the ground truth disaster grade of Wajima Drama Memorial Hall is G1, but we introduce an error higher grade which is G5. Thus, the disaster grade error will be +G4) and negative error propagation (we manually make the grade lower than the ground truth). We use the ground truth disaster grades as “Wajima Drama Memorial Hall – G1; Hama Street – G4; Concrete Bridge – G5; Central Nishikigawa Street – G1; North Asaichi Street – G3; South Central Asaichi Street – G4”. To get the ground truth downstream decision variables generated by the other three agents (the num of shelters, hospitals, human allocations, and the reconstruction budget), we fixed the initial disaster grades with the ground truth and then average the output downstream decision variables of three independent running as the ground truth (150 rescue workers, 100 medical staff, 80 logistics support personnel, 3 emergency shelters, 1 hospital, and \$0.05 billion reconstruction budgets). Due to the incomplete disaster data in this proof-of-concept case study, the human evaluation would be subjective, and the estimated ground truth downstream decision variables are in an interval and challenging to use for quantitative analysis. We chose the downstream decision variables output by DisasTeller in this interval as the ground truth, thus able to evaluate the exclusive error propagation of DisasTeller without introducing human subjective bias. Because our framework decides the downstream variables based on the reference of historical data, we manually feed the ground truth for the framework as the historical data and then we can evaluate how the deviation of disaster grading from the historical data (ground truth) will affect the deviation of the downstream variables.

For the positive error propagation part, we manually introduce the wrong disaster grades (Wajima Drama Memorial Hall – G5; Hama Street – G4; Concrete Bridge – G5; Central Nishikigawa Street – G5; North Asaichi Street – G3; South Central Asaichi Street – G4) with the total disaster grade error +G8. Then, the output downstream variables of three independent running are listed in Table 5.3, and Figure 5.16 presents the reasoning results of our framework for the calculation of downstream variables. In Table 5.3, when the initial disaster grades are manually introduced the total error of +G8 (+44% deviation from the total grades of ground truth), the propagated errors for downstream variables vary: rescue workers (+33% ~ 50%), medical staff (+30% ~ 50%), logistics support personnel (+25% ~ 50%), emergency shelters (+0% ~ 67%), hospitals (+0% ~ 100%), reconstruction budgets (+40% ~ 80%).

5. AI agent framework for global-scale defect perception

Table 5.3 The downstream variables of three independent-run output in the positive error propagation test

	round1	round2	round3	Ground Truth
Rescue workers	200	225	200	150
Medical staff	130	140	150	100
Logistics support personnel	100	120	120	80
Emergency shelters	3	4	5	3
Hospitals	1	1	2	1
Reconstruction budget (\$ billion)	0.07	0.07	0.09	0.05

Downstream reports for the key variables:

Human Resource Requirements:

...

Based on the current disaster grading and historical data, the human resource requirements for the affected areas are as follows:

- **Rescue Workers**: Given the increased severity in the current situation, especially in G5 areas, a total of 200 rescue workers are required. This is an increase from the historical requirement of 150 due to the higher number of G5 areas.

- **Medical Staff**: The need for medical staff is critical, especially with multiple G5 areas. A total of 130 medical staff are required, up from the historical 100, to ensure adequate medical response and care.

- **Logistic Support Personnel**: To manage the increased logistical challenges, 100 logistic support personnel are needed, compared to the historical 80, to ensure efficient distribution of resources and support...

Emergent Shelters and Hospitals:

...

In comparison to historical disaster data, where "Wajima Drama Memorial Hall – G1; Hama Street – G4; Concrete Bridge – G5; Central Nishikigawa Street – G1; North Asaichi Street – G3; South Central Asaichi Street – G4" required a total of 3 shelters and 1 hospital, the current situation demands the same number of shelters and hospitals due to the increased severity of the earthquake impact in certain areas.

...

Financial Budget Estimation:

...

Based on the current disaster grading and historical data, the estimated budget required for reconstruction is approximately 0.07 billion. This budget accounts for the increased severity and extent of damage in the high-risk areas compared to historical events.

...

Figure 5.16 The reasoning results of DisasTeller for the calculation of downstream variables in the positive error propagation test

For the negative error propagation part, we manually introduce the wrong disaster grades (Wajima Drama Memorial Hall – G1; Hama Street – G1; Concrete Bridge – G5; Central Nishikigawa Street – G1; North Asaichi Street – G1; South Central Asaichi Street – G1) with the total disaster grade error -G8. Then, the output downstream variables of three independent running are listed in Table 5.4. In Table 5.4, when the initial disaster grades are manually introduced the total error of -G8 (-44% deviation from the total grades of ground truth), the most propagated errors for downstream variables are none and only the reconstruction budget has +0% ~ 40% error. Since our framework will reuse the summary of each local disaster image description for the downstream agents (see Figure 5.2 in main text), it will autonomously check

5. AI agent framework for global-scale defect perception

the local disaster summary when calculating these downstream variables. For example, as the underlined text in Figure 5.17, although the initial wrong disaster grade (G1) was introduced to ‘Hama Street’ (the ground truth is G4), the later running of our framework still finds the significant structural damage from the local disaster summary, thus never cutting down the number of shelters, hospitals and reconstruction budget. This attribute of our framework would provide a safe bottom line for the real disaster scenarios.

Table 5.4 The downstream variables of three independent-run output in the negative error propagation test.

	round1	round2	round3	Ground Truth
Rescue workers	150	150	150	150
Medical staff	100	100	100	100
Logistics support personnel	80	80	80	80
Emergency shelters	3	3	3	3
Hospitals	1	1	1	1
Reconstruction budget (\$ billion)	0.07	0.057	0.05	0.05

Downstream reports for the key variables:

Human Resource Requirements:

...

Historical data indicates a similar situation, a total of 150 rescue workers, 100 medical staff, and 80 logistic support personnel are required. Based on the current disaster analysis and historical data, the following human resource requirements have been determined for effective post-disaster management:

- **Rescue Workers**: 150
- **Medical Staff**: 100
- **Logistic Support Personnel**: 80

...

Emergent Shelters and Hospitals:

...

Based on the analysis of the earthquake impact on the specified locations and comparison with historical disaster data, the following recommendations are made for the establishment of emergent shelters and hospitals:

- **Shelters**: 3
- **Location Recommendations**:
 - Near **Wajima Drama Memorial Hall**: This area remains stable with minimal damage, making it a suitable location for a shelter.
 - **Hama Street**: Despite significant structural damage, establishing a shelter here will provide accessibility to affected individuals in the vicinity.
 - **North Asaichi Street**: Given the hazardous environment due to a collapsed building, a shelter here will aid in immediate relief efforts.

- **Hospitals**: 1

- **Location Recommendation**:

- Centrally located to serve all affected areas efficiently, possibly near **Central Nishikigawa Street**. This area, despite active construction, has withstood the earthquake well and can support medical facilities.

...

Financial Budget Estimation:

...

Considering the current disaster damage and historical data, the total financial budget required for reconstruction is estimated at 0.05 billion. This budget allocation prioritizes Hama Street and Concrete Bridge due to their higher damage.

...

Figure 5.17 The reasoning results of DisasTeller for the calculation of downstream variables in the negative error propagation test.

5. AI agent framework for global-scale defect perception

The agents carry out different assignments and output reports sequentially in our framework. Thus, our framework also allows humans to check the intermediate outputs without waiting for the whole framework's completion, thereby minimising the effects of error propagation.

In addition to single-source uncertainty, real-world disaster scenarios may also involve conflicting information from multiple agents. For example, discrepancies may arise between visual observations, historical records, and external reports. In the current framework, such inconsistencies are not resolved through explicit consensus or arbitration mechanisms. Instead, the system relies on structured multi-agent reasoning and contextual aggregation to reinterpret and qualify conflicting evidence, rather than enforcing a single deterministic conclusion. It should be noted that resolving simultaneous inter-agent conflicts in a principled manner requires more advanced mechanisms, such as uncertainty modelling, confidence estimation, and consensus strategies, which are beyond the scope of this study. These aspects are considered important directions for future work in improving the robustness of multi-agent disaster assessment systems.

5.A.4 Allocation under scarce resources

In real-world disaster scenarios, the scarcity of rescue forces and relief supplies is the most fundamental challenge. In the emergent disaster scenarios, ensuring the optimal allocation under resource scarcity is always difficult and allocation plan is usually decided by rule-based prioritisation (e.g., priority according to the damage severity of different areas) with dynamic reallocation after real-time on-site inspection (Wang, 2021). DisasTeller is also designed to mitigate the impact of scarcity, following the decision-making logic of the real situation where human experts need to inspect the on-site scenarios and then make decisions based on rule-based prioritisation and dynamic reallocation.

We conducted an extra experiment for the earthquake case to simulate the real-time dynamic allocation by rule-based prioritisation (allocation decision according to the damage severity of the local areas) with scarce resources in disasters. In this experiment, we assumed the available rescue resources are insufficient with only 75 rescue workers, 50 medical staff, and 40 logistic support personnel existing. Extra information with current scarce resource (“only 75 rescue workers, 50 medical staff, and 40 logistic support personnel existing, allocated current scarce resource according to the damage severity of the local areas”) was injected into DisasTeller framework to mimic the real-time dynamic information input, thus evaluating the ability of

5. AI agent framework for global-scale defect perception

DisasTeller to allocate scarce resources under specific rule-based prioritisation. The damage severity grading of local areas was manually assessed according to the technical guideline EMS98 (Grünthal, 1998): Wajima Drama Memorial Hall – G1; Hama Street – G4; Concrete Bridge – G5; Central Nishikigawa Street – G1; North Asaichi Street – G3; South Central Asaichi Street – G4. According to the disaster grading percentage of different local areas (the sum of grading for six local areas is G18), we manually allocated the scarce resources with 1/18, 4/18, 5/18, 1/18, 3/18, 4/18 for these six locations, respectively.

Table 5.5 Comparison of the scarce resources allocation results

	Local Areas	Rescue Workers	Medical Staff	Logistic Support Personnel
Manual Allocation	Wajima Drama Memorial Hall	4.2	2.8	2.2
	Hama Street	16.7	11.1	8.9
	Concrete Bridge	20.8	13.9	11.1
	Central Nishikigawa Street	4.2	2.8	2.2
	North Asaichi Street	12.5	8.3	6.7
	South Central Asaichi Street	16.7	11.1	8.9
DisasTeller	Wajima Drama Memorial Hall	5	6	4
	Hama Street	15	10	8
	Concrete Bridge	20	10	10
	Central Nishikigawa Street	5	6	4
	North Asaichi Street	10	8	6
	South Central Asaichi Street	15	10	8
	Mean Square Root		3.4	
Evenly Allocated	Wajima Drama Memorial Hall	12.5	8.3	6.7
	Hama Street	12.5	8.3	6.7
	Concrete Bridge	12.5	8.3	6.7
	Central Nishikigawa Street	12.5	8.3	6.7
	North Asaichi Street	12.5	8.3	6.7
	South Central Asaichi Street	12.5	8.3	6.7
	Mean Square Root		23.3	

As shown in Table 5.5, compared with the evenly allocated strategy, DisasTeller reduces 85.4% mean square root deviation from the manual allocation, showcasing its ability to perform rule-based dynamic reallocation. DisasTeller is designed to provide preliminary suggestions and aims to accelerate the human decision-making process, not to achieve the optimal scarce resource allocation that is even difficult to realise in real scenarios.

5.A.5 Flood scenario preliminary study

The broader validation for DisasTeller in different disaster scenarios is necessary. We conducted a preliminary case study of flood scenario utilising data from the Australian government open report (2024). Due to licensing restrictions, the input images shown here are synthetic illustrations rather than the original data. The Tweed River affected by the flood locate in the northern region of New South Wales, Australia. Six affected local places with their local-view images, including Tweed Regional Aquatic Centre, Tweed Regional Aquatic Centre Carpark, the corner of Wharf St and Commercial Rd, the corner of Wharf St and Queen St, York Street, and Tumbulgum Rd near Racecourse Rd, were selected in this study. According to the damage scale for flood (Kelman, 2003), we manually assessed the disaster grades of six local places as the ground truth: Tweed Regional Aquatic Centre – DS3; Tweed Regional Aquatic Centre Carpark – DS3; the corner of Wharf St and Commercial Rd – DS3; the corner of Wharf St and Queen St – DS2; York Street – DS3; Tumbulgum Rd near Racecourse Rd – DS4. According to the output content of DisasTeller in Figure 5.18-b, the score for disaster grading task is 3/6 (wrong grading for Tweed Regional Aquatic Centre – DS4, the corner of Wharf St and Queen St – DS3, and York Street – DS4) with the total grading error +DS3. Figure 5.19 presents the global map annotation results with the accuracy score 4/6, and Table 5.6 shows the results for the evaluation of six output reports. In the alert map generation task, the LVLM agents need to identify the six affected local places from about one hundred location names in the global map. Finding some affected areas requires understanding the junction relationship of two places, like the corner of Wharf St and Queen St, which makes it a difficult task even for humans. Considering the current ability of LVLM agents to understand spatial locations is still limited, the task-specific fine-tuning could help enhance this ability.

5. AI agent framework for global-scale defect perception

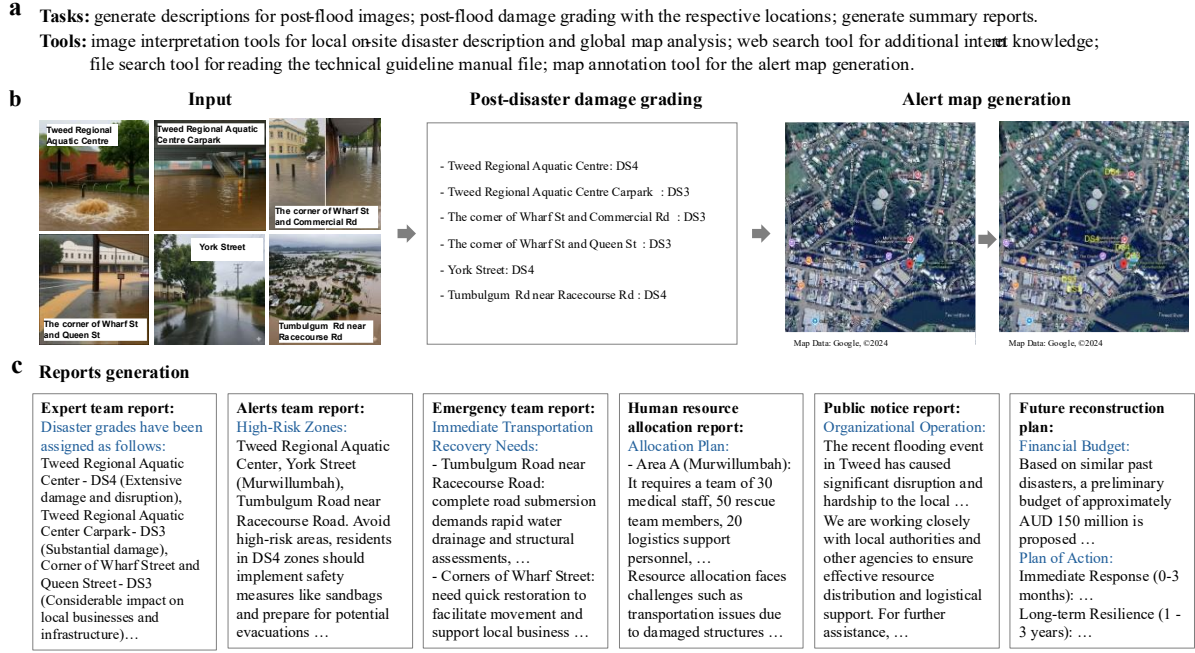


Figure 5.18 The experimental process for post-flood case study: a. the descriptions for tasks of agents and tools assigned; b. the running process of DisasTeller and intermediate tasks; c. the final output reports of agents in DisasTeller.



Figure 5.19 Evaluation of global map annotations for post-flood case study: DisasTeller's autonomous annotations are in yellow, and annotations with distance error larger than 250 pixels are identified as wrong and marked as red cross.

Table 5.6 EvaluatorGPT and human Evaluation for the output reports in post-flood case study.

	Expert Team Report	Alert Team Report	Emergency Team Report	Human Allocation Report	Public Notice Report	Future Reconstruction Plan
EvaluatorGPT	6.0/10	7.0/10	6.0/10	8.0/10	6.0/10	6.0/10
HumanEvaluator	6.2/10	6.5/10	7.1/10	7.3/10	6.8/10	7.1/10

5. AI agent framework for global-scale defect perception

5.A.6 Bushfire scenario preliminary study

Another post-bushfire disaster scenario preliminary case study was conducted utilising images from Australian government historical data (2019). Due to licensing restrictions, the input images shown here are synthetic illustrations rather than the original data. We defined six affected local places in Queensland, Australia, with their local-view images, including Tamborine Mountain, Gold Coast, Robina, Helensvale, Tweed Heads, and Kingscliff. After manual assessment according to the Australian fire danger rating system (2022), the ground truth disaster grades for these six affected local places are: Tamborine Mountain – G4 (catastrophic); Gold Coast – G2 - G3 (from high to extreme); Robina – G2 (high); Helensvale – G1 (moderate); Tweed Heads – G3 - G4 (from extreme to catastrophic); Kingscliff – G4 (catastrophic). Based on the output of DisasTeller in Figure 5.20, the score for disaster grading task is 4/6 (wrong grading for Robina – G3 and Helensvale – G2) with the total grading error +G2. Figure 5.21 and Table 5.7 show the evaluation of global map annotations and the evaluation of six output reports for the post-bushfire case study.

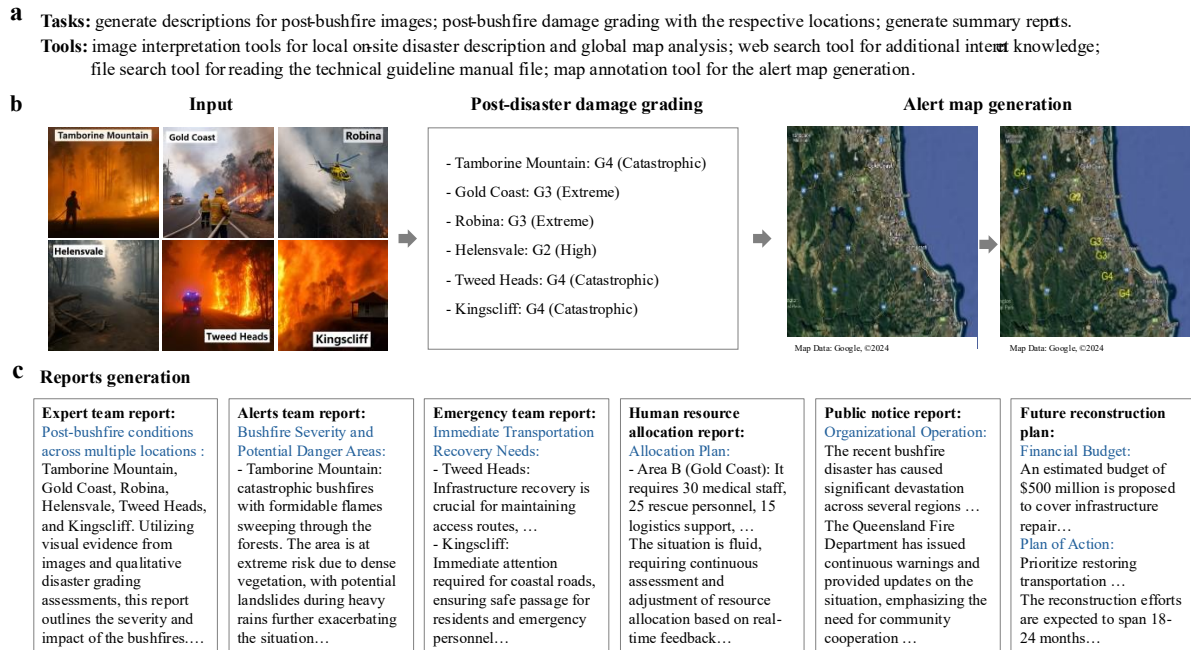


Figure 5.20 The experimental process for post-bushfire case study: a. the descriptions for tasks of agents and tools assigned; b. the running process of DisasTeller and intermediate tasks; c. the final output reports of agents in DisasTeller.

5. AI agent framework for global-scale defect perception

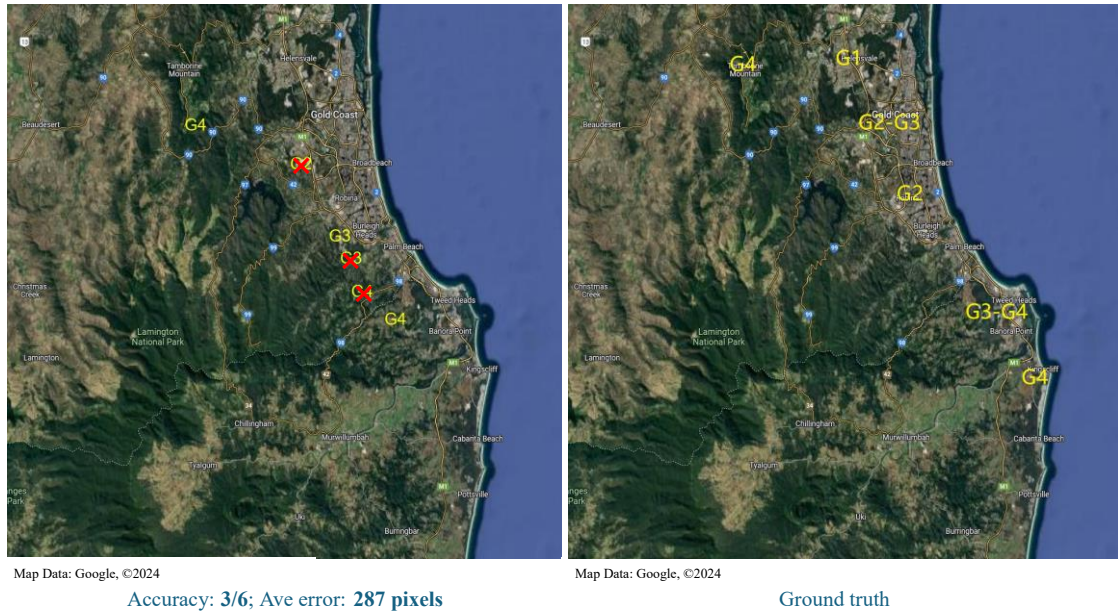


Figure 5.21 Evaluation of global map annotations for post-bushfire case study: DisasTeller's autonomous annotations are in yellow, and annotations with distance error larger than 250 pixels are identified as wrong and marked as red cross.

Table 5.7 EvaluatorGPT and human Evaluation for the output reports in post-bushfire case study

	Expert Team Report	Alert Team Report	Emergency Team Report	Human Allocation Report	Public Notice Report	Future Reconstruction Plan
EvaluatorGPT	7.0/10	6.0/10	9.0/10	8.0/10	7.0/10	8.0/10
HumanEvaluator	7.7/10	6.3/10	8.4/10	7.5/10	7.1/10	7.6/10

The implementation of this case study is based on general local-view images, which can be replaced with high-resolution aerial and satellite imagery to enable a more rigorous validation of wildfire damage assessments. However, since disaster scenario datasets are often restricted or privately held, our current access to such data is limited. In future work, we plan to collaborate with relevant stakeholders and government agencies to extend the framework toward multi-resolution remote-sensing data integration (including drone and satellite imagery). This enhancement will strengthen the credibility of the case studies and better reflect real-world deployment conditions.

References

Sendai Framework for Disaster Risk Reduction 2015-2030, United Nations Office for Disaster Risk Reduction, 2015.

Resilient and ready for the people of queensland, Queensland Fire Department, Australia, 2019.

5. AI agent framework for global-scale defect perception

Natural Disaster and Emergency Learnings and Recommendations Report, New South Wales Rural Doctor Network, Australia, 2021.

Australian Fire Danger Rating System, Australian Government, National Emergency Management Agency, Australia, 2022.

Uncounted costs: data gaps hide the true human impacts of disasters in 2023, United Nations Office for Disaster Risk Reduction, 2023.

Post 2022 Event Flood Behaviour Analysis - Tweed River, New South Wales Government, Department of Planning and Environment Australia, 2024.

Duckduckgo-search, 2024.

Achiam J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altschmidt, S. Altman, S. Anadkat, Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

Ao Y., H. Bahmani, Current Research Status of Disasters and Human Response, Navigating Complexity: Understanding Human Responses to Multifaceted Disasters (2023), pp. 1-18.

Augenstein I., T. Baldwin, M. Cha, T. Chakraborty, G.L. Ciampaglia, D. Corney, R. DiResta, E. Ferrara, S. Hale, A. Halevy, Factuality challenges in the era of large language models and opportunities for fact-checking, Nature Machine Intelligence (2024), pp. 1-12.

Bharathi Mohan G., R. Prasanna Kumar, P. Vishal Krishh, A. Keerthinathan, G. Lavanya, M.K.U. Meghana, S. Sulthana, S. Doss, An analysis of large language models: their impact and potential applications, Knowledge and Information Systems (2024), pp. 1-24.

Brown T.B., Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).
Ceferino L., Y. Merino, S. Pizarro, L. Moya, B. Ozturk, Placing engineering in the earthquake response and the survival chain, Nature Communications 15 (1) (2024), pp. 4298.

Chase H., LangChain, 2022.

Chen L., M. Zaharia, J. Zou, How is ChatGPT's behavior changing over time?, Harvard Data Science Review 6 (2) (2024).

Chrisman N., Full circle: More than just social implications of GIS, Cartographica: The International Journal for Geographic Information and Geovisualization 40 (4) (2005), pp. 23-35.

Ding M., P. Lin, S. Gao, J. Wang, Z. Zeng, K. Zheng, X. Zhou, D. Yamazaki, Y. Gao, Y. Liu, Reversal of the levee effect towards sustainable floodplain management, Nature Sustainability 6 (12) (2023), pp. 1578-1586.

Dong Q., L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu, A survey on in-context learning, arXiv preprint arXiv:2301.00234 (2022).

5. AI agent framework for global-scale defect perception

- Edwin M.L., K. Naomi, City and County of San Francisco Post Disaster Safety Assessment Guide, 2016.
- Feng Y., S. Cui, A review of emergency response in disasters: present and future perspectives, *Natural Hazards* 105 (1) (2021), pp. 1109-1138.
- Forrester N., Meet the scientists planning for disasters, *Nature*, Vol. 619, 2023, p. S1.
- Gao L., J. Schulman, J. Hilton, Scaling laws for reward model overoptimization, *International Conference on Machine Learning*, (2023), pp. 10835-10866.
- Grünthal G., European macroseismic scale 1998 (EMS-98), (1998).
- Kelman I., Physical flood vulnerability of residential properties in coastal, eastern England, (2003).
- Lewis P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, Retrieval-augmented generation for knowledge-intensive nlp tasks, *Advances in neural information processing systems* 33 (2020), pp. 9459-9474.
- Liu Y., D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-eval: Nlg evaluation using gpt-4 with better human alignment, *arXiv preprint arXiv:2303.16634* (2023).
- Lundh F., M. Ellis, Python imaging library (PIL), 2012.
- M. Bran A., S. Cox, O. Schilter, C. Baldassari, A.D. White, P. Schwaller, Augmenting large language models with chemistry tools, *Nature Machine Intelligence* (2024), pp. 1-11.
- Moura J., Crew AI, 2024.
- Nguyen D.T., F. Ofli, M. Imran, P. Mitra, Damage assessment from social media imagery data during disasters, *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, (2017), pp. 569-576.
- Ninivaggi G., Ishikawa earthquake response sees growing opposition scrutiny, *The Japan Times*, 2024.
- Opabola E.A., C. Galasso, Informing disaster-risk management policies for education infrastructure using scenario-based recovery analyses, *Nature Communications* 15 (1) (2024), pp. 325.
- Park S., T. Yao, S.V. Ukkusuri, Spatiotemporal heterogeneity reveals urban-rural differences in post-disaster recovery, *npj Urban Sustainability* 4 (1) (2024), pp. 2.
- Pollack A.B., C. Helgeson, C. Kousky, K. Keller, Developing more useful equity measurements for flood-risk management, *Nature Sustainability* (2024), pp. 1-10.
- Radford A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, *OpenAI blog* 1 (8) (2019), pp. 9.
- Richardson L., Beautiful soup documentation, April, 2007.

5. AI agent framework for global-scale defect perception

- Rosas-Chavoya M., J.L. Gallardo-Salazar, P.M. López-Serrano, P.C. Alcántara-Concepción, A.K. León-Miranda, QGIS a constantly growing free and open-source geospatial software contributing to scientific development, *Cuadernos de Investigación Geográfica* 48 (1) (2022), pp. 197-213.
- Schick T., J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language models can teach themselves to use tools, *Advances in Neural Information Processing Systems* 36 (2024).
- Team G., A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, Gemma 3 technical report, arXiv preprint arXiv:2503.19786 (2025).
- Wang Y., Multiperiod optimal allocation of emergency resources in support of cross-regional disaster sustainable rescue, *International Journal of Disaster Risk Science* 12 (3) (2021), pp. 394-409.
- Wei J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q.V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, *Advances in neural information processing systems* 35 (2022), pp. 24824-24837.
- Yang J., H. Xu, S. Mirzoyan, T. Chen, Z. Liu, Z. Liu, W. Ju, L. Liu, Z. Xiao, M. Zhang, Poisoning medical knowledge using large language models, *Nature Machine Intelligence* (2024), pp. 1-13.
- Yao S., J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, arXiv preprint arXiv:2210.03629 (2022).

6. Conclusions and Future Research

6.1 Concluding remarks

This thesis has investigated AI-driven methodologies for infrastructure defect detection and disaster response, with the objective of bridging local-scale visual perception and global-scale situational understanding within a coherent system architecture. Driven by the growing frequency of natural disasters, the increasing vulnerability of ageing civil infrastructure, and the limitations of traditional expert inspection workflows, the research developed a structured AI framework that integrates efficient and robust computer vision models with agent-based reasoning for large-scale disaster assessment.

At the local scale, the thesis demonstrates that robust and deployable infrastructure perception cannot be achieved through accuracy-driven model design alone. Real-world inspection environments are characterised by noise, visual degradation, domain shift, resource constraints, and operational uncertainty, all of which fundamentally challenge traditional learning paradigms. Through a sequence of architectural and methodological contributions, this thesis establishes a principled foundation for robust local-scale perception:

- First, the PoolingCrack model introduced a Transformer-based segmentation architecture that integrates average pooling-based feature aggregation with convolutional projection to improve both robustness and accuracy. By enhancing feature stability and representation consistency, the model demonstrates strong performance under heterogeneous surface textures and degraded visual conditions, while maintaining computational efficiency. This work establishes that architectural design choices play a central role in balancing robustness, efficiency, and segmentation fidelity in real-world defect detection.
- Second, the RFKD framework addressed the fundamental tension between robustness and lightweight deployment. By transferring not only predictive outputs but also intermediate feature representations from robust teacher models to lightweight student models under mixed clean-noisy training conditions, RFKD enables compact models to learn noise-invariant and stable representations. The co-design of the RFKD framework with the PoolingCrackTiny model demonstrates that robustness, efficiency,

6. Conclusions and future research

and deployment are not mutually exclusive objectives, but can be jointly optimised through structured distillation strategies and representation-level supervision.

- Third, the VMamba-based architectures extend local-scale perception toward high-resolution and large-scale visual modelling. By leveraging sequence-model-based representations, these architectures address long-range dependency modelling, large image processing, and fine-grained defect representation in high-resolution inspection scenarios. This work demonstrates that scalable visual modelling is essential for practical deployment in realistic infrastructure environments, where inspection images often exceed the scale and complexity assumed in traditional benchmark datasets.

Together, these three local-scale contributions establish a coherent perceptual foundation that is robust, efficient, and deployable. Rather than treating defect detection as a single-model pattern-recognition task, the thesis frames local-scale perception as a structured information extraction process that produces stable, interpretable, and reliable representations of infrastructure condition. This perceptual layer is not treated as an end, but as a foundational component for higher-level reasoning and decision support.

At the global scale, the thesis advances beyond single-model perception paradigms by developing an AI agent-based framework for disaster assessment and response. The DisasTeller system demonstrates how local perceptual outputs can be integrated within a structured multi-agent reasoning architecture to support interpretation, coordination, and decision-making. By decomposing complex disaster response workflows into role-based agent functions, the framework mirrors the organisational structure of human disaster management, where specialised teams perform interdependent tasks such as assessment, reporting, alerting, and resource coordination. This agent-based paradigm transforms raw perceptual outputs into coherent global situational awareness through structured information flow, collaborative reasoning, and interpretable decision processes. The framework is not designed to replace human expertise, but to function as structured cognitive support. AI agents operate as assistive components that enhance information processing capacity, reduce cognitive load, improve consistency, and support coordinated decision-making under time pressure and uncertainty.

By integrating local-scale perception with global-scale reasoning, this thesis demonstrates a unified system architecture in which perception, interpretation, and decision-making are structurally connected. This integration represents a shift from fragmented automation toward integrated intelligence. The thesis therefore contributes not only new models and methods, but

6. Conclusions and future research

a system-level conceptual framework for AI-driven disaster response and infrastructure management.

From a broader perspective, this research redefines the role of AI in civil infrastructure systems. Rather than positioning AI as a replacement for engineering judgement, it regards AI as a foundation that empowers human decision-making. Robust perception models provide reliable information extraction and structured agent-based reasoning architectures provide interpretability and coordination for complex workflows. Together, these components support a form of human-AI collaboration that is aligned with the realities of safety-critical, uncertainty-driven, and ethically sensitive disaster contexts.

The thesis therefore contributes across multiple levels:

- Methodological: by developing robust and accurate visual models, the structured distillation framework, and the scalable visual architecture.
- System-level: by integrating perception and reasoning into a coherent multi-agent architecture.
- Conceptual: by positioning AI as a system-level decision-support framework rather than isolated automation tools
- Practical: by targeting deployment, robustness, interpretability, and real-world operational constraints.

These contributions advance the state of the art in AI-driven infrastructure inspection and disaster response, providing a foundation toward scalable, reliable, and human-centred intelligent systems for civil infrastructure management.

6.2 Limitations and future research

Despite the contributions of this thesis, several limitations remain, which also define important directions for future research. First, although robustness has been extensively addressed through architectural design, feature learning strategies, and knowledge distillation frameworks, real-world disaster environments exhibit forms of uncertainty and complexity that cannot be fully captured through synthetic noise models or curated datasets. Extreme structural deformation, rare damage patterns, complex debris interactions, evolving hazard dynamics, and multi-hazard coupling introduce uncertainties that challenge current perception models. Future research should therefore move beyond conventional robustness modelling toward uncertainty-aware perception, integrating probabilistic reasoning, confidence estimation, and uncertainty

6. Conclusions and future research

quantification to support risk-sensitive and trustworthy decision-making in safety-critical environments.

Second, while the proposed agent-based global reasoning framework provides a structured architecture for integrating perception, interpretation, and coordination, it remains constrained by the current capabilities of large vision-language models and agent orchestration technologies. Challenges such as hallucination, reasoning inconsistency, long dependency modelling, uncertainty propagation, and controllability remain open problems in large-model-based systems. Future research should focus on developing more reliable reasoning architectures, incorporating verification mechanisms, consistency constraints, and structured reasoning protocols that improve stability, traceability, and reliability in complex, multi-stage decision workflows.

Third, disaster environments are inherently dynamic, and both physical conditions and data distributions evolve over time. The long-term effectiveness of AI-driven systems therefore requires adaptive learning mechanisms that go beyond static training paradigms. Future research should explore continual learning, online adaptation, and incremental model updating strategies that enable perception models and reasoning agents to evolve with changing environments while avoiding catastrophic forgetting and performance degradation. Such adaptive capabilities will be essential for maintaining long-term system relevance in operational deployments.

In addition to technical challenges, large-scale deployment introduces critical socio-technical constraints that extend beyond algorithmic performance. Issues of governance, accountability, ethical responsibility, data privacy, institutional trust, and regulatory integration cannot be resolved through technical design alone. AI-driven disaster response systems must operate within complex institutional ecosystems involving public authorities, emergency agencies, infrastructure operators, and communities. Future research should therefore adopt interdisciplinary perspectives that integrate engineering, policy, ethics, and social sciences to ensure that AI systems are not only technically effective but also socially legitimate, trustworthy, and institutionally viable.

Finally, practical deployment remains constrained by computational power, sensing availability, and organisational integration. While the frameworks proposed in this thesis are designed for efficiency and scalability, real-world disaster response systems require sustained investment in sensing networks, data pipelines, computational resources, and cross-organisational

6. Conclusions and future research

coordination mechanisms. Future research should explore system-level deployment strategies that integrate AI architectures with digital twins, simulation platforms, and operational infrastructures, enabling predictive assessment, scenario-based reasoning, and reactive decision support.

Overall, these limitations highlight that AI-driven disaster response is not a closed technical problem, but a long-term socio-technical transformation process. Future progress will depend not only on advances in perception models and reasoning architectures, but also on the co-evolution of institutional structures, governance frameworks, and human-AI collaboration paradigms. By addressing these challenges, future research can move toward truly resilient, adaptive, and trustworthy AI systems that support civil infrastructure management and disaster response in complex real-world environments.