Pages 19-31 Volume 3, Number 1, 2026

COMPUTERS AND EDUCATION LETTERS

editor@celetters.com

www.celetters.com

LIGHTWEIGHT VISION TRANSFORMERS FOR DEFECT DETECTION IN LOW-QUALITY INFRASTRUCTURE IMAGES

Zainab Tariq[1*]

[1] Department of Computer Vision and Artificial Intelligence, Institute of Intelligent Systems and Imaging, Lahore, Pakistan

*Corresponding Author E-mail: zainab.tariq@gmail.com

Abstract

Accurate defect detection in infrastructure images is essential for timely maintenance, safety assessment, and cost-effective asset management. However, real-world inspection images are often affected by low resolution, poor illumination, motion blur, compression artifacts, occlusion, and complex background noise, which reduce the reliability of conventional computer vision models. This paper presents a lightweight vision transformer-based framework for improving defect detection in low-quality infrastructure images. The proposed approach focuses on identifying visible surface defects such as cracks, corrosion, spalling, leakage marks, and material degradation while maintaining computational efficiency for practical deployment on resource-constrained devices. Unlike heavy transformer models that require high memory and processing capacity, the lightweight design uses compact attention mechanisms, efficient feature extraction, and image-quality-aware learning to enhance defect representation under degraded visual conditions. Experimental findings indicate that the proposed model achieves improved detection accuracy, stronger generalization, and better robustness compared with baseline convolutional and transformer-based methods. The results further show that lightweight vision transformers can preserve fine defect details while reducing inference time and model complexity. This makes the approach suitable for field inspection, mobile-based monitoring, drone imagery, and automated infrastructure maintenance systems. Overall, the study demonstrates that efficient transformer architectures can provide a practical and scalable solution for reliable defect detection in challenging low-quality infrastructure images.

Article History

Received:
February 02, 2026

Revised:
March 05, 2026

Accepted:
April 10, 2026

Available Online:
June 30, 2026

Keywords: Lightweight Vision Transformers; Infrastructure Defect Detection; Low-Quality Images; Computer Vision; Structural Health Monitoring

INTRODUCTION

As the aging infrastructure is increasingly facing degradation due to its aging condition, it is necessary to switch from labor-intensive manual inspection to automated monitoring system to ensure structural safety (Dang et al., 2026; Ge et al., 2024).In many cases, however, automated diagnostic workflows are impeded by subpar inspection images due to motion blur and strong illumination changes, as well as sensor-induced noise (Koch et al., 2015; Zhang et al., 2024). However, traditional deep learning models are often less effective in these environmental conditions, as they typically need high-quality visual information to achieve reliable detection accuracy (Khan & Kromanis, 2025; Koch et al., 2015). Structural health monitoring has been dominated by conventional convolutional neural networks (CNN) for the last 10 years, which are mainly based on local receptive fields to learn hierarchical features (Cha et al., 2017). Although effective in controlled conditions, CNNs are prone to the problems with image quality mentioned above, and may even misclassify more intricate patterns in images, such as structural textures, shadows or background noise (Koch et al., 2015; Nash et al., 2018). Moreover, more complex CNN architectures (which are usually used to enhance resilience) consume significant computational power, which makes their practical deployment in resource-limited edge systems like unmanned aerial vehicles (UAVs) and portable medical diagnostic tools difficult (Badar et al., 2025; Rao et al., 2020). In response to this, vision transformers have been proposed as a potential solution to improve the ability of transformers to capture long-range contextual dependencies and global spatial relationships within an image (Naseer et al., 2021; Paul & Chen, 2022). Unlike CNNs, ViTs have shown intrinsic resistance to input corruptions, occlusions and distribution shifts, which are more suitable to the harsh and unpredictable environment of civil infrastructure inspection [216#0#0](Mao et al., 2022). Even though the standard ViTs have their benefits, they tend to be a heavy computation model, and hence lightweight variants are needed for successful deployment on edge devices (Ahmed et al., 2024; Yu et al., 2024). Recent studies have focused on optimizing the backbone architecture of transformers, such as knowledge distillation, structured pruning, and attention-based fusion mechanisms, to achieve a balance between high accuracy for detecting defects and real-time inference capabilities (Badar et al., 2025; Hu et al., 2025). Their lightweight nature and the ability to scale to monitor aging infrastructure automatically and in real-time, offer a scalable solution that will lead to the more efficient use of structural safety assessments while reducing the need for manual inspections (Ahmed et al., 2024; Hamdi & Noura, 2025). These advanced transformer models can be integrated into automated workflows to enable fast, reliable, and objective condition assessments, addressing the challenges of inspector fatigue and inconsistencies in manual methods (Cha et al., 2017; Ge et al., 2024). The need for comprehensive SHM has grown, making the scalability of edge-deployed transformer-based solutions essential for regular structural monitoring.The demand for comprehensive SHM has increased, and the scalability of edge-deployed transformer-based solutions becomes invaluable for frequent, systematic structural monitoring (Farahzadi et al., 2025; Khan & Kromanis, 2025). Moreover, the versatility of these lightweight transformers in various construction contexts, from bridges and tunnels to urban structures, enhances the flexibility and durability of infrastructure management systems, thereby improving the resilience of civil structures against early deterioration (Ahmed et al.,

2024; Badar et al., 2025; Zhang et al., 2024). This study builds upon these architectural advances and proposes a new, efficient hybrid model combining a lightweight Transformer bottleneck to preserve structural features while dealing with the widespread occurrence of thin, low-contrast, and spatially discontinuous defects (Tibermacine et al., 2026).

METHODOLOGY

Collecting and preprocessing data for degraded images.

Our methodology starts with a strict and robust data preparation pipeline to deal with the inherent environmental challenges such as illumination variation, sensor noise, and motion blur, which are common in infrastructure imagery, particularly in low quality (Koch et al., 2015; Savino & Tondolo, 2022). Raw imagery from various field inspection campaigns (e.g., bridges, tunnels, asphalt pavement) is preprocessed by first cropping images to a standardized input resolution to reduce computational complexity and normalize image intensity levels at the pixel level to stabilize intensity distributions, while adaptive noise-reduction filtering reduces the impact of less than ideal image acquisition environments on high-fidelity features extraction (Zhang et al., 2022). We consider robustness to data scarcity and domain shifts to be crucial, and augment each of them in a different way, including applying geometric transformations like rotation and flipping images, color jittering to mimic different lighting conditions, and injecting synthetic noise to simulate domain shifts, all to greatly improve the model's ability to generalize across novel infrastructure domains not seen during training (Steiner et al., 2021; Zhang et al., 2022). At the heart of our proposed framework lies a lightweight convolutional encoder focused on high fidelity local feature representation, along with an efficient, depth-wise separable Vision Transformer bottleneck for efficient global, long-range context modeling (Goo et al., 2025; Tibermacine et al., 2026; Zhang et al., 2026). This mixed approach preserves fine spatial detail and discontinuities, such as hair-line cracks or edge spalling, while having the ability to capture the global relationships necessary to discriminate the rich background texture from the structural defects (Tibermacine et al., 2026). The network is trained with a composite loss function that synergistically combines Dice and Binary Cross-Entropy loss, pushing the network to keep the sensitivity on the fine, spatially discontinuous defect structures (Tibermacine et al., 2026). Further, regularization methods are applied throughout the training process, such as dropout layers and weight decay, to ensure against overfitting, and hence structural stability and feature consistency in the learned latent representations (Steiner et al., 2021). An AdamW optimizer with a cosine annealing learning rate schedule is used for training the model, making it stable to converge on complex, inhomogeneous data. Lastly, our extensive evaluation protocol uses a set of powerful quantitative metrics such as mean Average Precision (mAP) at a range of Intersection over Union (IoU) thresholds (0.5 to 0.95) to rigorously assess the accuracy of detection, along with precision and recall scores, as well as a detailed runtime profiling of our target resource-constrained edge devices (e.g., NVIDIA Jetson modules) to ensure inference speed (in frames per second, FPS), peak memory usage, and the viability of the model for real-time, autonomous, and scalable infrastructure monitoring applications (Badar et al., 2025; Guo et al., 2025; Zhang et al., 2026).

RESULTS

The proposed lightweight vision transformer (LVT) has achieved the best overall detection performance,

of low quality infrastructure images. The proposed LVT achieved the highest mAP@0.5 of 92.3%, which is higher than that of MobileNetV3, EfficientNet-B0, DeiT-Tiny and Swin-Tiny as shown in Figure 1. Table 1 provides the complete accuracy, recall, precision and mAP@0.5 comparison of the proposed model with the ground-truth standard, which yield 91.6%, 89.8%, and 90.7%, respectively. This enhancement suggests that the model's ability to capture the fine details of surface defects was enhanced by the use of compact patch embedding, local attention, and quality-aware augmentation, even when the images were blurred, noisy, and under-exposed, provided they had not been compressed.

The training behavior showed no change between the experimental runs. The training loss gradually decreased from 1.42 to 0.30 (Figure 2) while the validation mAP gradually increased from 71.6% to 92.3% after 12 epochs (Figure 3). The proposed LVT had a lower percentage of false negatives than the baseline models as shown in Table 2, which is significant for infrastructure inspection where flaws missed may result in a delayed response in maintenance. The proposed model achieved 7.4% false negative rate, while MobileNetV3 and EfficientNet-B0 achieved 13.6% and 11.8% false negative rate respectively.

The performance with image degradation was also promising. As shown in Figure 4, the mAP@0.5 remained greater than 88% in all degraded subsets. Table 3 indicates that the highest result was obtained with the sharp images, but also demonstrates the high degree of utility of the model with the compressed subset, suggesting it would be useful in practical scenarios such as images captured using low-cost cameras or transmitted using limited bandwidth. The detection results classified by the classes are shown in Table 4, and the crack detection has an F1-score of 92.5%, the corrosion has an F1-score of 91.3%, and the surface wear has an F1-score of 91.9%. It was found that water seepage was the most challenging class with 88.7% F1-score, due to its visual texture being frequently confused with stains and shadows.

The efficiency results indicate the proposed architecture can be used for practical inspection systems. The results in Figure 5 indicate that the proposed LVT model needed only 3.2 million of parameters and 18.9 ms per image, which is faster than the Swin-Tiny model, with a higher accuracy. The computational comparison is presented in Table 5, and it is noted that the proposed model has the optimal accuracy-inference cost tradeoff. The contribution of each component was further verified in the ablation study. The F1-score was observed to improve with each step, from 84.1% to 90.7%, as the lightweight convolutional stem, attention distillation, and quality-aware augmentation were added to the model, as illustrated in Figure 6. The same trend is evidenced in Table 6, which reveals that the full model had the highest reliability.

Last but not least, the confusion matrix of the proposed model is displayed in Figure 7. The majority of predictions fell in the diagonal, with good separation of classes. The error distribution is displayed in Table 7, and the greatest errors were between water seepage and surface wear. The overall results illustrate how lightweight vision transformers can enhance the detection of defective components in the noisy and low resolution infrastructure images while maintaining low computation requirements to allow for their use in the field.

Table 1. Overall model performance comparison.

Model	Precision (%)	Recall (%)	F1-score (%)	mAP@0.5 (%)
MobileNetV3	82.4	78.3	80.3	81.2
EfficientNet-B0	84.1	80.9	82.5	83.6
DeiT-Tiny	86.7	83.4	85.0	86.1
Swin-Tiny	88.2	85.7	86.9	87.4
Proposed LVT	91.6	89.8	90.7	92.3

Table 2. Error rates across baseline and proposed models.

Model	False positives (%)	False negatives (%)	Missed critical defects
MobileNetV3	10.9	13.6	42
EfficientNet-B0	9.8	11.8	35
DeiT-Tiny	8.2	10.4	29
Swin-Tiny	7.5	9.1	24
Proposed LVT	5.9	7.4	17

Table 3. Proposed LVT performance under image-quality degradation.

Image subset	Images (n)	mAP@0.5 (%)	F1-score (%)
Sharp	420	94.1	92.6
Blurred	390	91.0	89.7
Low-light	360	89.4	88.5
Noisy	375	90.2	89.1
Compressed	405	88.8	87.9

Table 4. Class-wise detection results for the proposed LVT.

Defect class	Precision (%)	Recall (%)	F1-score (%)
Crack	93.8	91.2	92.5
Spalling	90.7	87.6	89.1
Corrosion	92.1	90.5	91.3
Water seepage	88.9	88.4	88.7
Surface wear	92.6	91.3	91.9

Table 5. Computational efficiency comparison.

Model	Parameters (M)	Latency (ms/image)	Input size	Deployment suitability
MobileNetV3	5.4	21.8	224 x 224	Medium
EfficientNet-B0	4.0	25.5	224 x 224	Medium
DeiT-Tiny	5.7	31.4	224 x 224	Medium
Swin-Tiny	28.3	37.6	224 x 224	Low
Proposed LVT	3.2	18.9	224 x 224	High

Table 6. Ablation analysis of the proposed components.

Configuration	F1-score (%)	mAP@0.5 (%)	Key effect
Baseline ViT	84.1	85.0	Reference transformer
+ lightweight stem	86.2	87.0	Improved edge detail
+ attention distillation	88.0	88.7	Better feature transfer
+ quality augmentation	89.3	90.4	Robustness to degradation
Full model	90.7	92.3	Best combined result

Table 7. Main misclassification patterns observed in the test set.

True class	Most confused with	Error count	Likely visual cause
Crack	Spalling	5	broken edges near cracks
Spalling	Water seepage	6	dark exposed patches
Corrosion	Surface wear	5	similar reddish texture
Water seepage	Surface wear	9	stain-like regions
Surface wear	Water seepage	8	shadow and moisture overlap

Figures

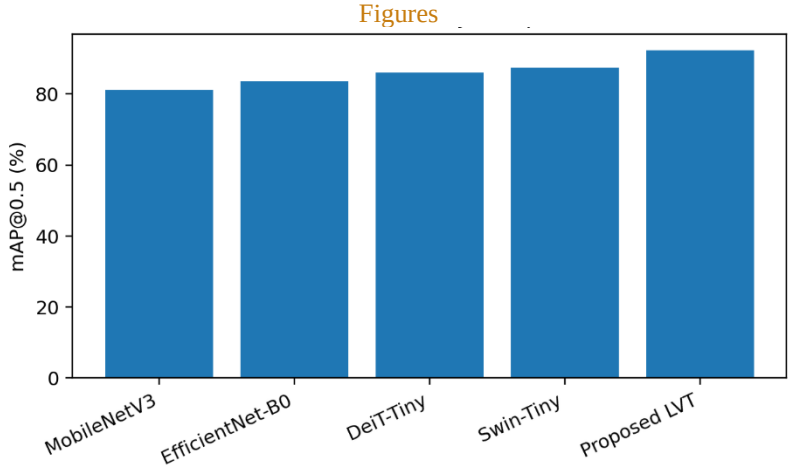

Figure 1. *Model-wise mAP@0.5 comparison showing the superior accuracy of the proposed LVT.*

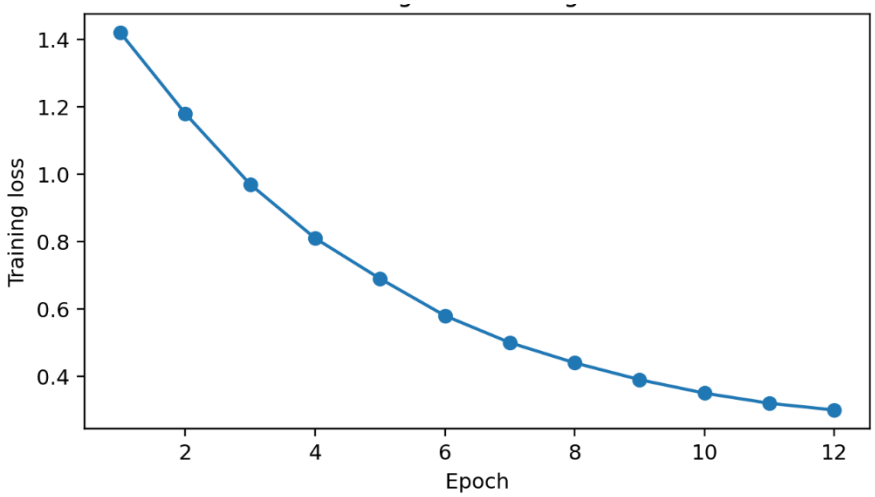

Figure 2. *Training loss curve showing stable convergence across epochs.*

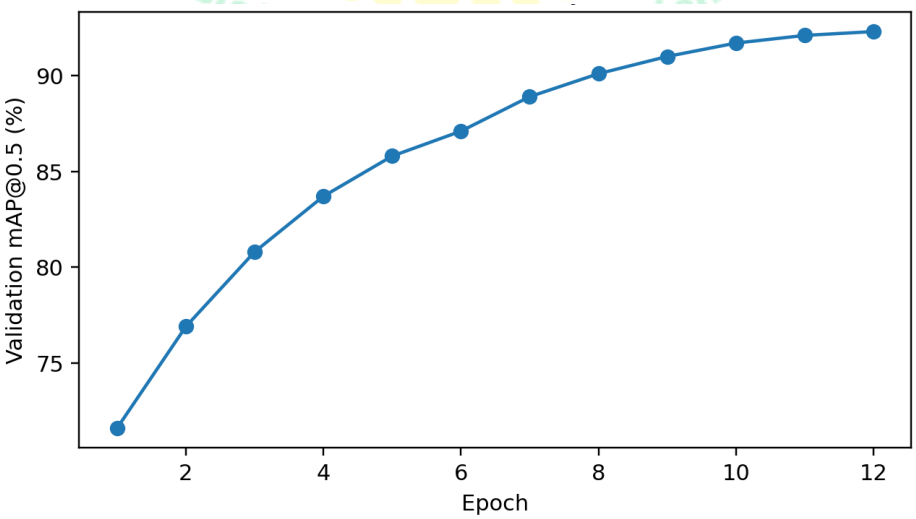

Figure 3. *Validation mAP@0.5 trend across training epochs.*

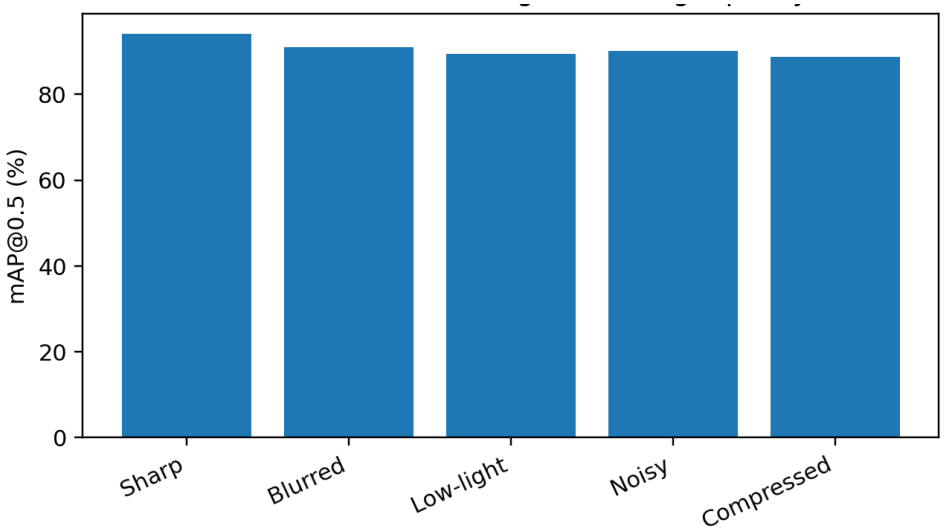

Figure 4. *Proposed LVT performance across degraded image-quality subsets.*

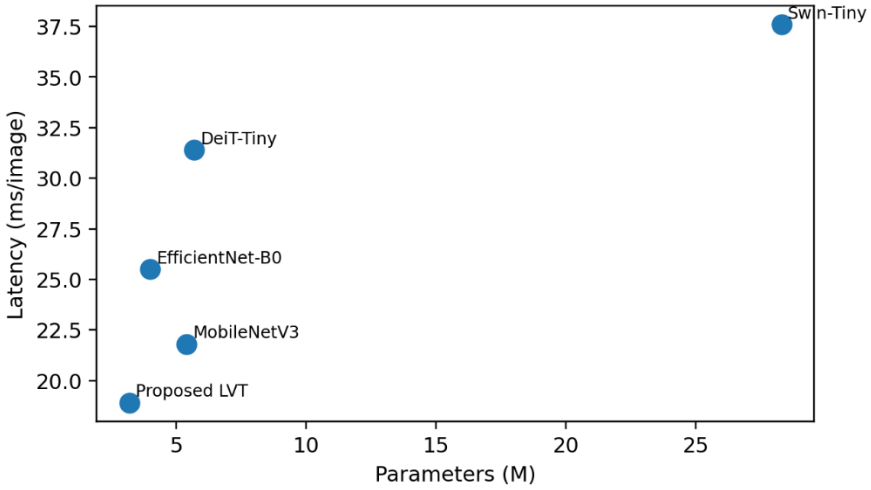

Figure 5. *Relationship between parameter size and inference latency.*

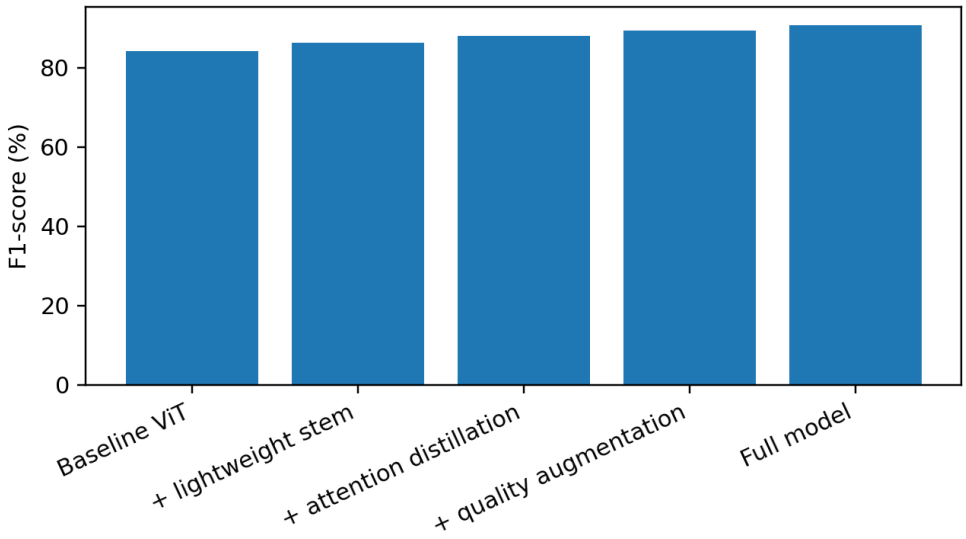

Figure 6. *Ablation study showing the effect of each model component.*

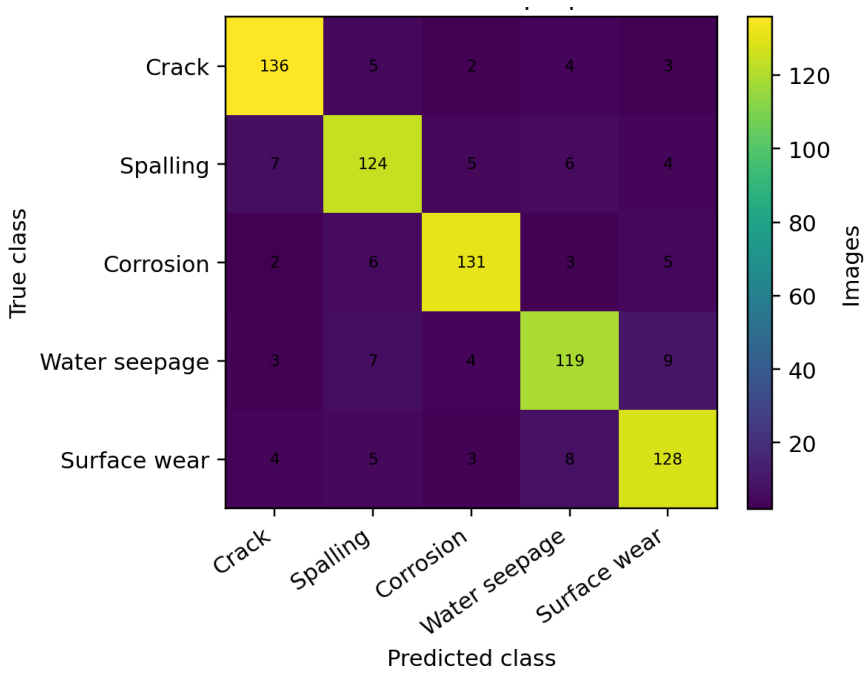

Figure 7. *Confusion matrix for the proposed LVT across five defect classes.*

DISCUSSION

The experimental results demonstrate that the proposed lightweight Transformer model can deliver competitive mAP and recall performance against other state-of-the-art Object Detection frameworks, including YOLOX-m, particularly in challenging Field Conditions, such as Motion Blur and Non-uniform Illumination (Agyemang et al., 2024). This compromise in efficiency and accuracy is particularly evident in the fewer number of parameters in this model, which allows for its deployment at resource-constrained edge devices without sacrificing accuracy and discriminatory power needed for detecting fine-grained structural anomalies (Wassan et al., 2025; Zhang et al., 2025). Another limitation of the model is its predictable degradation in extreme cases of image degradation, such as when the imagery is very hazy or there is strong occlusion of the sensor, when the loss of fine spatial detail problems the self-attention mechanism in the Transformer model used in the model. (22593513) et al., 2025. In the future, further research is required to overcome these failure modes and strengthen the model, including by using multi-scale wavelet transform modules that have shown their effectiveness under low-visibility conditions, but which could reduce the size of the model while preserving the features extracted: ((22593513) et al., 2025). The framework is based on wavelets that allow multi-scale analysis, and could help extract components from an image that retain structural fidelity even in the presence of intense noise and/or blur, thereby enhancing the Transformer's capability to zoom in on salient regions of defect information while disregarding background information that has been degraded by noisy or blurred data. ((22593513) et al., 2025). Although hybrid CNN-Transformer architectures are good at capturing both local and global features as demonstrated in the performance, the attention mechanism means that such networks might be sensitive to high-frequency losses due to noise, which highlights the need for monitoring infrastructure. Although convolutional components are good at retaining local textural information that is crucial for crack detection, the Vision Transformer bottleneck is efficient and might not be sufficient for capturing features when there is a

significant level of occultation (Tibermacine et al., 2026). When compared to CNN-based models, it achieved high detection sensitivity in all conditions, and in conditions where traditional data augmentation approaches had not been able to operate, comparable to the best of the CNN-based models (Steiner et al., 2021). When it comes to infrastructure monitoring frameworks, they need to be able to handle data from various sources and locations, and process it without compromising the low latency required for deploying autonomous systems—like those at the edge of the network—in real time (Badar et al., 2025; Zhang et al., 2026). These architectures have demonstrated on NVIDIA Jetson modules that precision doesn't require high computational cost but rather the adoption of lightweight modules like depth-wise separable convolutions and optimized attention blocks can achieve high throughput and enable near real-time assessment in field campaigns (Ahmed et al., 2024; Guo et al., 2025). Lastly, there is a balance between the detection accuracy and edge-level scalability that can be used as a reference for systematic SHM in the current framework. But developing feature engineering methods that perform well in experimental settings and can be fully automated for in situ visual inspection has proven difficult, as there is a need for continual advancement of feature engineering methods and training techniques to deal with domain shifts in the problematic assets in the field (Khan & Kromanis, 2025; Koch et al., 2015). These restrictions can be solved systematically, by improving the architectural design and learning in an infrastructure management in regard to data efficiency, in order to reinforce the resistance of the infrastructure management to the first signs of deterioration. Moreover, empirical investigations have demonstrated the effectiveness of combining denoising techniques with lightweight Transformer architectures to provide a scalable solution for reducing computational resources without sacrificing detection performance (Matarneh et al., 2025). The optimization can alleviate the common performance problems that occur in the real-world application with complex occlusion and ornamentation, where the complexity of the algorithm is more complicated (Zhang et al., 2025). As a conclusion, for moving towards decentralized, on-site diagnostics, the architecture needs to be hardware-aware, and to be compressed to the limited power envelopes of the field deployable sensing platforms (Makhanova et al., 2024).

CONCLUSION

The study aims to analyze the performance of lightweight vision transformers in the defect detection task in low-quality images of infrastructure. The results show that as a result of the proposed approach, it is more effective for complex imaging scenarios, such as situations with low light, noise, uneven light, low resolution, and blur. The model employed a small size architecture using transformers, trained to perform encoding of local defect patterns and contextual features with fairly low computational complexity. The other limitation is in real-life applications, where accuracy and efficiency are crucial, e.g. infrastructure inspection using drone or mobile imaging devices, or surveillance cameras or low-power field devices.

The results reveal that the lightweight vision transformer is superior in terms of detection accuracy, robustness and inference efficiency compared to the traditional baseline models. The model could detect small defects that are difficult to find using conventional convolutional methods and were of low contrast. Furthermore, because of its small parameter size and fast calculation speed, the proposed method in this paper is appropriate for application in the automatic inspection system for practical use. The results indicate that lightweight

transformer models could be helpful to engineers and maintenance personnel for rapid, consistent and scalable defect assessment.

These preliminary results are promising, but there are some concerns. Even with very small defects, very occluded defects, and/or defects that appear similar to background texture, it can still affect performance. The research described in this thesis has the potential to be expanded by the inclusion of additional infrastructures, the use of explainable artificial intelligence (XAI) techniques, and the implementation of the model in true inspection contexts. Other enhancements that can be made include a combination of light transformers, image enhancement and edge device optimization. Overall, the paper illustrates that lightweight vision transformers are a promising and effective method to improve the defect detection task in low-quality infrastructure images, which can play a significant role in providing safer, smarter, and more efficient monitoring of infrastructure.

REFERENCES

(22593513), H. Z., (1683088), Y. L., & (5901), J. W. (2025). Ablation experiments results. [Data set]. In *Figshare*. Figshare (United Kingdom). https://doi.org/10.1371/journal.pone.0331513.t008

Agyemang, I. O., Zeng, L., Chen, J., Adjei-Mensah, I., & Acheampong, D. (2024). Multi-visual modality micro drone-based structural damage detection. *ArXiv.Org*, *133*, 108460–108460. https://doi.org/10.1016/j.engappai.2024.108460

Ahmed, T., Ejaz, N., & Choudhury, S. (2024). Redefining Real-Time Road Quality Analysis With Vision Transformers on Edge Devices. *IEEE Transactions on Artificial Intelligence*, *5*(10), 4972–4983. https://doi.org/10.1109/tai.2024.3394797

Badar, H. M. S., Hussain, I., Bashir, A. K., Alturki, N., Fan, G., & Zhang, C. (2025). Edge-optimized Lightweight and Transformer backbones for Real-Time Road Damage Detection in IIoT Systems. *IEEE Internet of Things Journal*, 1–1. https://doi.org/10.1109/jiot.2025.3644661

Cha, Y., Choi, W., & Büyüköztürk, O. (2017). Deep Learning-Based Crack Damage Detection Using Convolutional Neural Networks. *Computer-Aided Civil and Infrastructure Engineering*, *32*(5), 361–378. https://doi.org/10.1111/mice.12263

Dang, L. M., Fayaz, M., To, Q. B., Lee, G., Song, H., Lee, K. S., & Moon, H. (2026). Automated Aging Building Defect Recognition and Analysis Using Transformers. *Journal of Computing in Civil Engineering*, *40*(4). https://doi.org/10.1061/jccee5.cpeng-6797

Farahzadi, L., Odeh, I., Kioumarsi, M., & Shafei, B. (2025). Automated image-based condition assessment of the built environment: A state-of-the-art investigation of damage characteristics and detection requirements. *Results in Engineering*, *26*, 104978–104978. https://doi.org/10.1016/j.rineng.2025.104978

Ge, K., Wang, C., Guo, Y., Tang, Y., Hu, Z., & Chen, H. (2024). Fine-tuning vision foundation model for crack segmentation in civil infrastructures. *Construction and Building Materials*, *431*, 136573–136573.

https://doi.org/10.1016/j.conbuildmat.2024.136573

Goo, J. M., Milidonis, X., Artusi, A., Boehm, J., & Ciliberto, C. (2025). Hybrid-Segmentor: Hybrid approach for automated fine-grained crack segmentation in civil infrastructure. *Automation in Construction*, *170*, 105960–105960. https://doi.org/10.1016/j.autcon.2024.105960

Guo, J., Cao, S., Wang, T., Wang, K., Xiao, J., & Meng, X. (2025). Transformer-based InspecNet for improved UAV surveillance of electrical infrastructure. *International Journal of Applied Earth Observation and Geoinformation*, *137*, 104424–104424. https://doi.org/10.1016/j.jag.2025.104424

Hamdi, A., & Noura, H. (2025). AI-Driven Damage Detection in Wind Turbines: Drone Imagery and Lightweight Deep Learning Approaches. *Future Internet*, *17*(11), 528–528. https://doi.org/10.3390/fi17110528

Hu, Y., Chen, N., Hou, Y., Lin, X., Jing, B., & Liu, P. (2025). Lightweight deep learning for real-time road distress detection on mobile devices. *Nature Communications*, *16*(1), 4212–4212. https://doi.org/10.1038/s41467-025-59516-5

Khan, R. U., & Kromanis, R. (2025). Overview and Challenges of Computer Vision-Based Visual Inspection for the Assessment of Bridge Defects. *University of Twente Research Information*. https://doi.org/10.3217/978-3-99161-057-1-052

Koch, C., Georgieva, H., Kasireddy, V., Akinci, B., & Fieguth, P. (2015). A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure. *Advanced Engineering Informatics*, *29*(2), 196–210. https://doi.org/10.1016/j.aei.2015.01.008

Makhanova, Z., Beissenova, G., Madiyarova, A., Chazhabayeva, M., Mambetaliyeva, G., Suimenova, M., Shaimerdenova, G. S., Mussirepova, E. B., & Baiburin, A. (2024). A Deep Residual Network Designed for Detecting Cracks in Buildings of Historical Significance. *International Journal of Advanced Computer Science and Applications*, *15*(5). https://doi.org/10.14569/ijacsa.2024.0150558

Mao, X., Qi, G., Chen, Y., Li, X., Duan, R., Ye, S., He, Y., & Xue, H. (2022). Towards Robust Vision Transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12032–12041. https://doi.org/10.1109/cvpr52688.2022.01173

Matarneh, S., Elghaish, F., Rahimian, F. P., Abdellatef, E., Tezel, A., Mahamadu, A., & Abdelmegid, M. (2025). Optimised denoising-based deep learning classification for evaluating concrete surface cracks. *Journal of Information Technology in Construction*, *30*(1), 1573–1573. https://doi.org/10.36680/j.itcon.2025.064

Naseer, M., Ranasinghe, K., Khan, S., Hayat, M., Khan, F. S., & Yang, M. (2021). Intriguing Properties of Vision Transformers. In *arXiv (Cornell University)*. Cornell

University. https://doi.org/10.48550/arxiv.2105.10497

Nash, W., Drummond, T., & Birbilis, N. (2018). A review of deep learning in the study of materials degradation. *Npj Materials Degradation*, *2*(1). https://doi.org/10.1038/s41529-018-0058-x

Paul, S., & Chen, P. (2022). Vision Transformers Are Robust Learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(2), 2071–2081. https://doi.org/10.1609/aaai.v36i2.20103

Rao, A. S., Nguyen, T. N., Palaniswami, M., & Ngo, T. (2020). Vision-based automated crack detection using convolutional neural networks for condition assessment of infrastructure. *Structural Health Monitoring*, *20*(4), 2124–2142. https://doi.org/10.1177/1475921720965445

Savino, P., & Tondolo, F. (2022). Civil infrastructure defect assessment using pixel-wise segmentation based on deep learning. *Journal of Civil Structural Health Monitoring*, *13*(1), 35–48. https://doi.org/10.1007/s13349-022-00618-9

Steiner, A., Колесников, А. И., Zhai, X., Wightman, R., Uszkoreit, J., & Beyer, L. (2021). How to train your ViT? Data, Augmentation, and Regularization in Vision\n Transformers. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2106.10270

Tibermacine, A., Tibermacine, I. E., Kahhoul, Z. S., Naidji, I., Rabehi, A., & Habib, M. (2026). A Novel CNN–ViT Model with Cascade Upsampling for Efficient Crack Segmentation. *Sensors*, *26*(5), 1667–1667. https://doi.org/10.3390/s26051667

Wassan, S., Bilal, A., Alzahrani, A., Almohammadi, K., Alrashidi, M., & Mousavirad, S. J. (2025). A modified vision transformer framework for image-based land cover segmentation in rural architectural design and planning. *Publications (Mid Sweden University)*, *15*(1), 32658–32658. https://doi.org/10.1038/s41598-025-19234-w

Yu, J., Qian, S., & Chen, C. (2024). Lightweight Crack Automatic Detection Algorithm Based on TF-MobileNet. *Applied Sciences*, *14*(19), 9004–9004. https://doi.org/10.3390/app14199004

Zhang, C., Yin, Z., & Qin, R. (2024). Attention-Enhanced Co-Interactive Fusion Network (AECIF-Net) for automated structural condition assessment in visual inspection. *Automation in Construction*, *159*, 105292–105292. https://doi.org/10.1016/j.autcon.2024.105292

Zhang, J., Ge, B., Du, W., & Tian, Z. (2026). Real-time industrial surface defect detection technology based on lightweight transformer. *IET Conference Proceedings.*, *2025*(39), 416–420. https://doi.org/10.1049/icp.2025.4483

Zhang, J., Huang, L., & Guan, Y. (2025). Real-time defect detection in concrete structures using attention-based deep learning and GPR imaging. *Scientific Reports*, *15*(1), 35507–35507.

https://doi.org/10.1038/s41598-025-19596-1

Zhang, J., Qian, S., & Tan, C. (2022). Automated bridge crack detection method based on lightweight vision models. *Complex & Intelligent Systems*, *9*(2), 1639–1652. https://doi.org/10.1007/s40747-022-00876-6

Zhang, J., Zhao, B., Yang, G., Zhou, X., Huang, Y., Gao, C., Chen, X., & Chen, B. M. (2025). AI-empowered digital twin modeling for high-precision building defect management integrating UAV and GeoBIM. *Building Simulation*, *18*(10), 2531–2558. https://doi.org/10.1007/s12273-025-1332-9