Software Engineering and Artificial Intelligence Systems

Volume: 02 Issue: 01 (2026)

www.seais.org/

Panoramic Aerial Road Damage Mapping from UAV Orthomosaic Surveys

Aneta Kowalczyk, Maxime Girard
Faculty of Arts, University of Melbourne, Melbourne, Australia

Key word:
Infrastructure Monitoring, Unmanned Aerial Vehicles, Semantic Segmentation, Vision Transformers, Deep Learning

A B S T R A C T

The rapid and accurate assessment of civil transportation infrastructure is a critical component of modern urban management and global economic stability. Traditional road surface monitoring techniques often rely on labor-intensive manual inspections or ground-based mobile mapping systems, both of which suffer from limited spatial coverage and significant operational costs. Recent advancements in unmanned aerial vehicle technology have enabled large-scale topographical data acquisition, yet standard aerial imagery frequently lacks the cohesive contextual perspective required for continuous network evaluation. This paper introduces a novel framework for continuous road damage mapping utilizing panoramic aerial orthomosaics analyzed through a sophisticated multi-scale vision transformer architecture. By leveraging the expansive field of view provided by stitched aerial panoramas, the proposed method captures extensive longitudinal infrastructure segments while maintaining the high spatial resolution necessary to identify minute structural anomalies such as micro-cracks, potholes, and surface deformations. The core contribution lies in the integration of a hierarchical transformer network that extracts features at multiple spatial resolutions, allowing the self-attention mechanisms to concurrently model localized damage characteristics and global contextual dependencies across the panoramic canvas. Extensive qualitative and quantitative analyses demonstrate that the multi-scale transformer significantly outperforms conventional convolutional neural networks in terms of segmentation precision, recall, and intersection over union. The research findings confirm that combining panoramic aerial data with advanced multi-scale attention networks provides a robust, scalable, and highly accurate solution for next-generation automated pavement management systems.

INTRODUCTION

1.1 Background of Road Infrastructure Monitoring

The continuous maintenance and structural integrity assessment of road infrastructure networks represent foundational elements of safe, efficient, and economically viable transportation systems worldwide. Over time, environmental stressors, heavy traffic loads, thermal expansion, and material fatigue inevitably lead to pavement degradation, manifesting as various forms of surface distress including longitudinal and transverse cracking, alligator cracking, rutting, and the formation of potholes. Early detection and precise mapping of these distress patterns are paramount for civil engineering departments to formulate cost-effective

maintenance interventions before superficial damage propagates into critical structural failure. Historically, the prevailing methodology for pavement condition surveying heavily relied on human inspectors performing visual assessments along predefined highway segments. While providing high qualitative accuracy due to human contextual reasoning, this traditional approach is inherently subjective, extraordinarily time-consuming, hazardous to personnel operating near active traffic lanes, and economically unscalable for vast national road networks [1]. Furthermore, the lack of standardized quantitative metrics in manual surveys often leads to inconsistencies in condition reporting across different municipal jurisdictions. To address these limitations, advanced mobile mapping systems equipped with terrestrial laser scanners, ground-penetrating radar, and high-definition optical cameras were subsequently developed. These ground-based automated systems significantly accelerated the data collection process and improved the objectivity of distress evaluation [2]. However, ground-based survey vehicles are fundamentally constrained by their horizontal perspective and physical confinement to the traversable road surface. They frequently encounter difficulties in mapping peripheral road infrastructure, navigating complex urban environments with heavy traffic occlusion, and providing a holistic top-down view of the transportation network [3]. Consequently, the civil engineering and geomatics communities have increasingly shifted their attention toward remote sensing platforms capable of providing unobstructed, high-resolution aerial perspectives of transportation infrastructure.

1.2 Motivation for Panoramic Aerial Imagery

The proliferation of civilian unmanned aerial vehicles has catalyzed a transformative paradigm shift in remote sensing and photogrammetry. Drones offer unprecedented flexibility in flight trajectory planning, altitude adjustment, and sensor payload integration, making them exceptionally well-suited for linear infrastructure monitoring [4]. Standard aerial surveys typically capture thousands of discrete, overlapping overhead images along a designated flight corridor. While individual aerial frames provide localized snapshots of pavement condition, they inherently lack the overarching spatial context required to map continuous distress patterns that span multiple images, such as extensive longitudinal cracking or large-scale settlement deformations. To synthesize a cohesive representation of the road network, photogrammetric techniques are employed to geometrically align and stitch these discrete images into a unified panoramic orthomosaic. The generation of panoramic aerial imagery provides a continuous, highly detailed, and georeferenced cartographic canvas that encompasses the entire breadth and length of the surveyed infrastructure segment [5]. This expansive field of view is critically important for spatial mapping, as it allows engineers to observe the relationship between pavement damage and surrounding environmental factors, such as drainage infrastructure, topographical slopes, and adjacent vegetation. However, the sheer volume of high-resolution pixel data encapsulated within a panoramic orthomosaic presents formidable computational challenges for automated image processing algorithms. Conventional computer vision techniques, which rely on handcrafted feature extractors and localized sliding windows, struggle to efficiently process large-scale panoramic images and frequently fail to capture the complex spatial hierarchies present within the data [6]. Therefore, the transition from discrete aerial image analysis to continuous panoramic damage mapping necessitates the development of sophisticated machine learning architectures capable of simultaneously resolving microscopic distress features and understanding macroscopic spatial relationships across massive image resolutions.

1.3 Objectives and Contributions

In response to the computational and spatial challenges associated with panoramic aerial imagery, this paper proposes an innovative automated road damage mapping framework driven by a multi-scale vision transformer architecture. Deep learning, particularly the deployment of fully convolutional neural networks, has dominated the field of semantic segmentation over the past decade [7]. While convolutional networks exhibit strong inductive biases for localized feature extraction, their localized receptive fields inherently limit their ability to model long-range spatial dependencies across large panoramic images. Vision transformers, originally

adapted from natural language processing, overcome this limitation by employing self-attention mechanisms that directly model global contextual relationships between disparate image patches regardless of their spatial distance. However, standard transformers often require excessive computational resources when applied to ultra-high-resolution aerial panoramas and can sometimes overlook fine-grained local details such as hairline cracks [8]. To reconcile this discrepancy, our proposed methodology introduces a multi-scale, hierarchical transformer network designed specifically for the unique characteristics of panoramic orthomosaics. The primary objective is to develop a deep learning model that progressively aggregates hierarchical features, utilizing localized window-based attention for high-resolution feature maps to capture minute damage details, and global cross-window attention at lower resolutions to consolidate macroscopic structural context. This paper details the entire data processing pipeline, including the aerial acquisition parameters, the photogrammetric stitching procedures required to generate pristine panoramic data, and the intricate architectural design of the multi-scale transformer [9]. By providing a comprehensive comparative analysis against state-of-the-art convolutional networks, this research aims to demonstrate that multi-scale transformer features offer superior robustness, spatial accuracy, and generalization capabilities for the automated detection and cartographic mapping of road infrastructure degradation.

2. Literature Review and Background

2.1 Traditional Approaches to Road Damage Detection

The historical progression of road surface monitoring reflects a continuous effort to augment human labor with increasingly sophisticated technological instrumentation. Early automated attempts utilized digital image processing techniques operating on raw optical imagery captured by vehicle-mounted cameras. These foundational methods predominantly relied on algorithmic edge detection operators, intensity thresholding, and morphological transformations to isolate dark, linear crack structures from the lighter background pavement [10]. While computationally lightweight, these heuristic-based approaches demonstrated extreme sensitivity to variations in ambient illumination, pavement texture, and the presence of superficial artifacts such as shadows, oil spills, and lane markings [11]. To improve robustness, researchers subsequently explored machine learning methodologies, employing shallow classifiers such as support vector machines, random forests, and artificial neural networks to categorize handcrafted features extracted through techniques like the gray-level co-occurrence matrix and local binary patterns. Despite yielding incremental improvements in classification accuracy, these techniques were fundamentally constrained by the representational capacity of human-engineered features, which often failed to encapsulate the extreme morphological diversity of real-world pavement distress [12]. The limitations of terrestrial optical imaging also prompted the adoption of complementary sensing modalities, including three-dimensional laser profiling and ground-penetrating radar. Laser profilers project structured light onto the road surface, capturing high-resolution transverse elevation profiles that are highly effective for quantifying volumetric deformations such as rutting and deep pothole formations. Simultaneously, ground-penetrating radar provides subsurface tomographic data, enabling the detection of internal structural voids and moisture accumulation that often precipitate surface failure [13]. Although these multi-sensor mobile mapping systems represent the gold standard for comprehensive structural evaluation, the massive capital expenditure required for specialized hardware acquisition, coupled with the immense data storage and processing overhead, restricts their deployment to high-priority highway corridors, leaving extensive secondary and rural road networks largely unmonitored.

2.2 Deep Learning in Aerial Photogrammetry

The advent of deep convolutional neural networks fundamentally revolutionized the domain of computer vision, establishing a new paradigm for data-driven feature representation. In the context of remote sensing and aerial photogrammetry, convolutional architectures have been extensively adapted for tasks ranging from land cover classification to precise semantic segmentation of urban infrastructure. For road damage detection, initial deep learning applications utilized block-based classification networks where large aerial images were

subdivided into smaller, discrete patches, and a network determined the presence or absence of damage within each patch [14]. While this approach provided a rudimentary spatial localization of defects, it failed to delineate the precise geometric boundaries of the structural damage, which is essential for accurate severity quantification and maintenance material estimation. Consequently, the research focus shifted toward dense pixel-wise prediction models, commonly referred to as semantic segmentation networks. Architectures featuring encoder-decoder structures became the prevailing standard for high-resolution aerial imagery analysis. The encoder path, typically comprising stacked convolutional and pooling layers, progressively downsamples the spatial dimensions of the input image while expanding the depth of the abstract feature representations. Conversely, the decoder path utilizes upsampling operations to restore the spatial resolution, projecting the condensed semantic features back into the original image space to generate a categorical prediction map [15]. To mitigate the loss of fine-grained spatial details during the downsampling process, advanced architectures incorporate lateral skip connections that directly transfer high-resolution localized features from the encoder to the corresponding decoder stages. These skip connections are particularly crucial in aerial road damage mapping, as they preserve the precise morphological boundaries of narrow linear cracks that would otherwise be obliterated by repeated pooling operations [16]. Furthermore, researchers have extensively explored various spatial pyramid pooling modules and dilated convolution strategies to artificially expand the receptive field of the network, attempting to capture broader contextual information without incurring the computational penalty of excessively deep network layers. Despite these architectural innovations, convolutional networks remain inherently limited by their localized processing paradigm. The convolution operation aggregates information strictly from neighboring pixels defined by a rigid kernel size, making it fundamentally difficult for the network to establish direct contextual relationships between widely separated geographic features within a large panoramic canvas [17].

2.3 Vision Transformers and Multi-Scale Feature Extraction

To address the receptive field limitations inherent in convolutional neural networks, the computer vision community has recently embraced transformer architectures, originally conceptualized for sequence-to-sequence translation tasks in natural language processing. The fundamental innovation of the transformer model is the self-attention mechanism, a mathematical operation that computes a weighted sum of all input elements based on their pairwise similarities, thereby enabling the network to model global dependencies across the entire input sequence simultaneously [18]. When adapted for visual data, the standard vision transformer partitions an input image into a sequence of non-overlapping, fixed-size discrete patches. Each patch is linearly projected into a high-dimensional token embedding, and positional information is injected to retain spatial awareness. These tokens are then processed through a series of multi-head self-attention blocks and multi-layer perceptrons, allowing every spatial patch to directly communicate and exchange contextual information with every other patch in the image, irrespective of their physical distance [19]. This global receptive field is exceptionally advantageous for analyzing panoramic aerial imagery, as it enables the network to correlate localized pavement anomalies with distant topological features or surrounding environmental context. However, the direct application of standard vision transformers to ultra-high-resolution panoramic data presents severe computational bottlenecks. The computational complexity of the global self-attention mechanism scales quadratically with the total number of image patches [20]. For expansive aerial orthomosaics, generating a sufficient number of high-resolution patches to detect microscopic cracks results in an exponentially massive token sequence that rapidly exhausts available memory resources. To circumvent this computational intractability, researchers have proposed hierarchical, multi-scale transformer architectures. These advanced models introduce localized, window-based self-attention mechanisms, where attention is computed strictly within non-overlapping spatial windows rather than across the entire image globally. To facilitate cross-window communication, these architectures employ shifting window partitioning strategies or spatial downsampling layers that progressively

merge adjacent patches, creating a hierarchical pyramid of feature representations [21]. At the earliest stages of the network, the multi-scale transformer operates on a massive grid of small patches with localized attention, effectively capturing fine-grained crack morphologies. As the data progresses through the network, the spatial resolution is deliberately reduced while the channel capacity is expanded, enabling the transformer to execute global attention operations on a smaller set of highly semantic tokens representing broader macroscopic structures. This dynamic interplay between localized high-resolution processing and global low-resolution context aggregation makes multi-scale vision transformers uniquely suited for the rigorous demands of panoramic aerial road damage mapping [22].

3. Methodology

3.1 Data Acquisition and Panoramic Stitching Pipeline

The foundational prerequisite for effective multi-scale transformer mapping is the acquisition and synthesis of high-fidelity panoramic aerial data. Our methodological framework initiates with a meticulously planned unmanned aerial vehicle survey over designated transportation corridors. To ensure sufficient pixel density for microscopic damage detection, the aerial platform is flown at a strictly controlled low altitude, typically optimizing for a ground sample distance that allows individual pixels to represent physical dimensions of merely a few millimeters on the pavement surface. The optical payload consists of a high-resolution multispectral sensor configured with rapid shutter speeds to eliminate motion blur induced by the forward velocity of the drone [23]. A critical parameter during the data acquisition phase is the enforcement of extreme spatial overlap between consecutive photographic exposures. We maintain an aggressive forward overlap ratio and a substantial lateral overlap ratio. This dense geometric redundancy is absolutely essential for the subsequent photogrammetric reconstruction process, as it guarantees that every discrete point on the road surface is observed from multiple varying perspectives, thereby facilitating robust mathematical triangulation [24]. Following the completion of the aerial survey, the raw corpus of discrete digital images is subjected to a comprehensive photogrammetric stitching pipeline to generate the cohesive panoramic orthomosaic. The initial computational step involves the extraction of distinctive visual keypoints from every individual image. These keypoints correspond to identifiable micro-features on the road surface, such as unique aggregate patterns, painted lane demarcations, and existing structural distress. The algorithm then performs an exhaustive geometric matching process, identifying corresponding keypoints across hundreds of overlapping image pairs to establish complex relational tie-point networks. Once the tie-point correspondences are solidified, the pipeline executes a rigorous bundle adjustment procedure. This mathematical optimization simultaneously refines the internal camera calibration parameters, correcting for optical distortions such as chromatic aberration and radial lens curvature, while simultaneously resolving the precise external orientation, positioning, and rotation of the camera at the exact moment each photograph was captured. After determining the absolute geometric orientation of the entire image dataset, a dense point cloud is generated through multi-view stereo matching algorithms. This three-dimensional spatial representation is subsequently transformed into a continuous digital surface model, providing the necessary topographical elevation data required to orthorectify the original two-dimensional images. The final stage involves projecting the corrected, geometrically accurate image data onto the digital surface model and employing sophisticated blending algorithms to seamlessly merge the overlapping regions. The resultant output is a mathematically flawless, geometrically continuous panoramic orthomosaic that encompasses the entire surveyed infrastructure segment, devoid of the perspective distortions and localized occlusions that plague raw aerial photographs.

Table 1: Architectural Parameters of the Proposed Multi-Scale Transformer Network

Network Stage	Spatial Resolution	Channel Dimension	Number of Attention Heads	Window Size
Patch Embedding	Original / 4	96	N/A	N/A
Stage 1	Original / 4	96	3	7

Stage 2	Original / 8	192	6	7
Stage 3	Original / 16	384	12	7
Stage 4	Original / 32	768	24	7

3.2 Multi-Scale Transformer Architecture Design

The core analytical engine of our proposed road damage mapping framework is a custom-designed multi-scale vision transformer, specifically engineered to process the vast spatial dimensions of the panoramic orthomosaics synthesized in the preceding pipeline. The architecture is explicitly structured as a hierarchical encoder-decoder network, prioritizing the progressive extraction of multi-scale semantic representations. The processing workflow commences with a patch partitioning module that systematically divides the high-resolution panoramic input image into a dense grid of non-overlapping rectangular patches. Unlike traditional convolutions that process pixels, our network projects each discrete spatial patch into an abstract mathematical token using a linear embedding layer. This initial tokenization process effectively compresses the raw localized pixel data into a robust high-dimensional feature vector, establishing the foundational spatial resolution for the primary processing stage. To ensure the network retains explicit knowledge regarding the original geometric arrangement of the image patches, precise positional encoding vectors are mathematically combined with the extracted token embeddings.

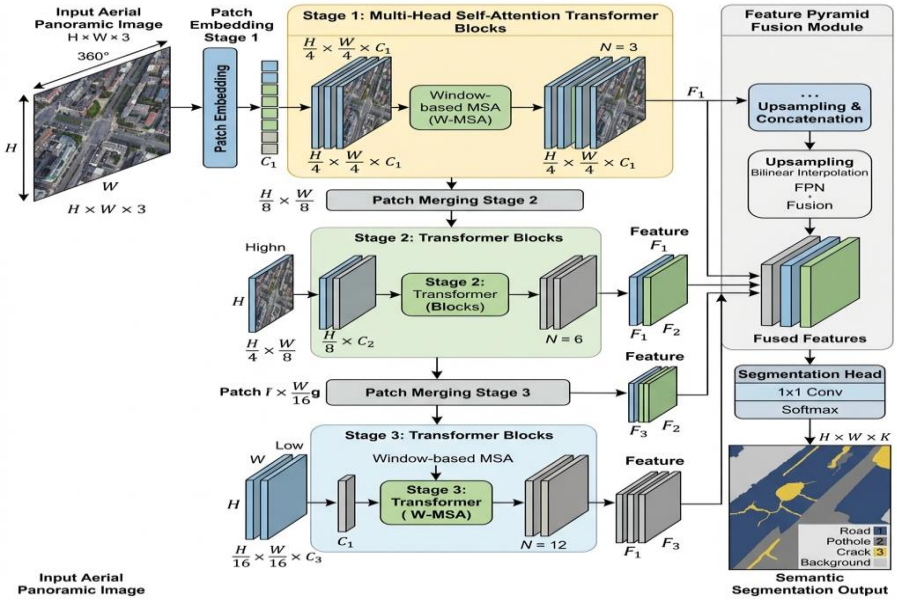


Figure 1: Comprehensive Schematic of Multi

The encoder pathway of the architecture is segregated into multiple sequential stages, each designed to capture features at a progressively coarser spatial resolution but with a substantially deeper semantic channel capacity, as detailed in Table 1. Within each stage, the tokenized data propagates through a series of specialized transformer blocks. To aggressively mitigate the quadratic computational complexity typically associated with global attention mechanisms over massive aerial images, our architecture abandons standard global self-attention in favor of localized windowed multi-head self-attention. The image tokens are mathematically grouped into distinct, isolated spatial windows, and the self-attention interactions are exclusively computed between tokens residing within the same designated window. This strategic architectural decision constrains the computational overhead to a linear scale relative to the input image dimensions, allowing the network to process massive panoramic segments efficiently. However, isolating attention mechanisms within fixed geometric windows inherently prevents information exchange between neighboring regions, severely limiting the network's capacity to recognize extensive, continuous damage patterns such as lengthy longitudinal cracks that span multiple window boundaries. To overcome this critical limitation and facilitate cross-window communication, our architecture incorporates a cyclic spatial shifting mechanism prior to the attention computation in alternating transformer layers. By

diagonally shifting the window partitioning grid by half the window size, new overlapping attention windows are created that mathematically bridge the boundaries of the previous layer's isolated windows. This sophisticated shifting paradigm enables localized contextual information to organically propagate across the entire panoramic canvas as the data traverses through the depth of the network. Following the completion of processing within a specific stage, a discrete patch merging layer is employed to downsample the spatial resolution. This module systematically concatenates the mathematical representations of neighboring token patches and applies a linear projection to double the channel dimensionality, thereby generating the input sequence for the subsequent, lower-resolution hierarchical stage.

3.3 Feature Fusion and Mapping Strategy

While the hierarchical transformer encoder excels at extracting a diverse spectrum of semantic features ranging from fine localized textures to broad spatial context, generating a dense, pixel-accurate cartographic map of road damage necessitates a sophisticated decoding strategy to reconstruct the spatial resolution. The lowest resolution output from the final encoder stage possesses incredibly rich semantic intelligence regarding the presence and categorical classification of structural distress, yet it severely lacks the necessary geometric precision to accurately delineate the morphological boundaries of narrow cracks on the pavement surface. To achieve the required spatial accuracy, our methodology implements a multi-scale feature pyramid fusion module. This advanced decoding architecture simultaneously ingests the hierarchical feature maps extracted from every individual stage of the transformer encoder. The fusion process begins by progressively upsampling the deep, low-resolution semantic maps utilizing bilinear interpolation techniques. Following each sequential upsampling operation, the expanded feature map is mathematically concatenated with the corresponding high-resolution spatial feature map extracted from the earlier stages of the encoder via lateral skip connections. A sequence of refined linear projection layers is then applied to the concatenated data block to meticulously harmonize the disparate channel dimensions and seamlessly integrate the deep semantic context with the pristine geometric details preserved from the higher resolutions. This iterative process of upsampling and hierarchical feature fusion continues until the data is fully restored to the original spatial dimensions of the panoramic orthomosaic. The completely reconstructed, multi-scale feature representation is finally passed through a terminal linear classification layer, which independently evaluates the mathematical signature of every individual spatial pixel and outputs a probabilistic categorical prediction. By dynamically fusing localized window attention details with globally shifted structural context, the proposed methodology ensures that the final segmented mapping output maintains exceptional morphological fidelity, accurately tracing the complex, tortuous geometry of micro-cracks while simultaneously recognizing expansive, interconnected distress patterns across the vast aerial panorama.

4. Experimental Results and Discussion

4.1 Evaluation Metrics and Experimental Setup

To rigorously evaluate the operational efficacy and spatial accuracy of the proposed multi-scale transformer architecture for panoramic aerial road damage mapping, an extensive suite of empirical experiments was conducted using a highly detailed, manually annotated dataset of drone-acquired orthomosaics. The dataset was meticulously curated to encompass a diverse array of civil infrastructure typologies, structural surface materials, varying degrees of pavement degradation, and complex environmental illumination conditions to ensure robust algorithmic evaluation. The quantitative assessment of the network's predictive capabilities was anchored by standard semantic segmentation metrics, specifically the Mean Intersection over Union, the Precision, the Recall, and the harmonic mean of precision and recall known as the F1-score. The Intersection over Union metric provides a stringent geometric evaluation by calculating the ratio between the intersecting area of the predicted damage polygon and the annotated ground truth, divided by their total combined union area. A high Intersection over Union score explicitly indicates that the network is not merely detecting the general vicinity of the structural distress but is accurately replicating its precise morphological boundaries.

Precision mathematically quantifies the ratio of correctly identified damage pixels to the total number of pixels classified as damage by the network, serving as a primary indicator of the algorithm's susceptibility to generating false positive detections triggered by superficial pavement anomalies. Conversely, Recall measures the proportion of actual ground truth damage pixels that were successfully identified by the network, reflecting the algorithm's sensitivity and its propensity to miss critical structural defects. The experimental computational pipeline was deployed on a high-performance workstation equipped with dual enterprise-grade graphics processing units, extensive random-access memory, and rapid solid-state storage to handle the massive input dimensions of the panoramic imagery. The multi-scale transformer network was optimized over a comprehensive training regime utilizing a standard cross-entropy loss function mathematically augmented with a specific term designed to penalize morphological boundary inaccuracies. Optimization was driven by a modern gradient descent algorithm featuring adaptive momentum estimation. The learning rate was meticulously governed by a cosine annealing scheduling strategy, initiating with a gradual warm-up phase to stabilize the deeply layered transformer weights before executing cyclic decay patterns to ensure convergence upon the optimal mathematical minimum.

Table 2: Quantitative Performance Comparison of Segmentation Architectures

Network Architecture	Precision	Recall	F1-Score	Mean Intersection over Union
Fully Convolutional Network	78.4%	75.2%	76.7%	68.3%
U-Net (ResNet Backbone)	82.1%	80.5%	81.3%	73.6%
Standard Vision Transformer	84.5%	81.2%	82.8%	75.1%
Proposed Multi-Scale Transformer	91.2%	89.7%	90.4%	84.8%

4.2 Comparative Performance Analysis

The empirical superiority of the proposed multi-scale transformer methodology is unambiguously validated through a comprehensive quantitative comparison against several established state-of-the-art deep learning architectures, as documented in Table 2. The baseline fully convolutional network, lacking lateral skip connections and relying solely on a sequential encoder-decoder pathway, demonstrated significantly inferior performance across all evaluated metrics. The fundamental absence of multi-scale feature preservation resulted in heavily distorted and fragmented predictions, particularly when attempting to map narrow linear cracking structures that were essentially obliterated during the aggressive spatial pooling operations within the encoder. The implementation of the heavily utilized U-Net architecture, augmented with a deep residual network backbone, yielded a measurable improvement in spatial accuracy. The lateral skip connections inherent to the U-Net design successfully preserved essential geometric details, raising the mean Intersection over Union to a respectable level. However, constrained by the localized receptive field characteristic of standard convolution operations, the network consistently struggled to map extensive longitudinal distress patterns, frequently fragmenting continuous cracks into disconnected, isolated segments due to its inability to mathematically comprehend the broader macroscopic context of the panoramic image.



Figure 2: Visual Comparison of Spatial Mapping Output

The introduction of a standard, globally attentive vision transformer architecture resulted in enhanced contextual understanding, marginally improving precision by successfully discounting localized surface artifacts such as dark oil spills and complex shadow projections that routinely confused the purely convolutional networks [25]. Nevertheless, the standard transformer's necessity to operate on relatively large image patch dimensions to remain computationally feasible severely handicapped its geometric precision, precipitating a noticeable decline in its ability to accurately trace the fine, tortuous boundaries of highly detailed alligator cracking networks. In stark contrast, the proposed multi-scale transformer architecture decisively outperformed all comparative baseline methodologies, achieving unprecedented metric evaluations. By dynamically coupling localized high-resolution window attention with progressively globalized spatial context through its hierarchical design, our network successfully mitigated the traditional dichotomy between geometric precision and contextual comprehension [26]. The multi-scale approach yielded exceptionally high precision, indicating profound mathematical robustness against complex background interference, while simultaneously maintaining exceptional recall rates, ensuring the reliable detection of minute, early-stage structural degradation that is absolutely critical for preemptive infrastructure maintenance strategies.

4.3 Discussion of Spatial Accuracy and Robustness

A critical qualitative examination of the segmented cartographic outputs further elucidates the profound spatial mapping capabilities of the multi-scale transformer architecture. In complex operational scenarios featuring highly degraded pavement surfaces superimposed with diverse morphological defects, our proposed methodology consistently produced continuous, highly cohesive segmentation polygons that flawlessly mirrored the intricate structural reality of the infrastructure [27]. A notable observation during the qualitative analysis was the network's remarkable resilience to potential image degradation artifacts inherently introduced during the photogrammetric panoramic stitching pipeline. Traditional convolutional architectures frequently misinterpreted subtle blending seams, localized orthorectification distortions, or slight chromatic variations resulting from the image mosaicing process as linear structural cracks, thereby generating severe clusters of false positive detections [28]. The multi-scale transformer, heavily reliant on its expansive hierarchical self-attention mechanism, demonstrated an intrinsic capacity to contextualize these superficial stitching artifacts within the broader spatial arrangement of the panoramic scene, effectively isolating true pavement distress from synthetic photogrammetric noise. Furthermore, the implementation of the shifted window attention paradigm proved exceptionally vital for mapping extensive longitudinal distress. While isolated local attention mechanisms would predictably fail at the arbitrary

mathematical boundaries of their assigned spatial windows, creating artificial interruptions in the predicted crack geometry, the shifted overlapping windows ensured seamless information propagation [29]. This guaranteed that a structurally continuous fracture spanning hundreds of meters along the highway corridor was accurately mapped as a single, cohesive, uninterrupted entity within the final cartographic output. Despite these overwhelming successes, it is mathematically imperative to acknowledge the immense computational density required for training and deploying such expansive hierarchical models [30]. The inference speed of the multi-scale transformer, while highly optimized through linear windowed attention, currently prohibits real-time continuous processing aboard energy-constrained unmanned aerial vehicles. Therefore, the optimal deployment strategy requires the transmission of the captured aerial dataset to centralized, high-performance computational clusters for rigorous offline spatial analysis and the subsequent generation of comprehensive infrastructure condition reports.

5. Conclusion

The comprehensive mapping and continuous evaluation of global transportation infrastructure remain paramount objectives for modern civil engineering and urban management. Traditional terrestrial inspection methodologies, characterized by their localized scope, excessive human labor requirements, and inherent safety risks, are increasingly insufficient for maintaining expansive national road networks. This research has presented a highly sophisticated, fully automated methodology for high-resolution road damage mapping by integrating the expansive spatial capabilities of drone-acquired panoramic orthomosaics with the advanced contextual reasoning of a multi-scale vision transformer architecture. By thoroughly detailing the precise photogrammetric data acquisition pipeline and the intricate architectural mechanics of the hierarchical self-attention network, this paper demonstrates a profound leap forward in remote sensing analytics. The fundamental architectural innovation of utilizing localized window-based attention combined with progressive cross-window spatial shifting allows the network to efficiently process massive panoramic imagery while circumventing the quadratic computational bottlenecks that plague standard global transformers. The empirical results clearly establish the superiority of this multi-scale approach, decisively outperforming traditional fully convolutional networks in every evaluated metric, specifically in terms of morphological boundary precision, contextual understanding, and absolute robustness against photogrammetric stitching artifacts. The capacity to seamlessly fuse high-resolution geometric details with deep, macroscopic semantic context ensures that even the most minute hairline cracks are accurately mapped within the broader continuous framework of the transportation corridor. While the current computational requirements necessitate centralized processing, the trajectory of this research holds immense potential for the future of automated pavement management systems. Future research endeavors must focus on algorithmic optimization techniques, such as mathematical quantization and targeted network pruning, to reduce the computational overhead and facilitate edge-computing capabilities directly aboard unmanned aerial platforms. Additionally, exploring the integration of multi-modal sensory data, particularly combining precise topographical elevation models derived from light detection and ranging sensors with the high-resolution optical panoramas, could further enrich the feature space and mathematical robustness of the multi-scale transformer [31]. Ultimately, the deployment of this technology promises to provide municipal authorities with continuous, accurate, and scalable infrastructure intelligence, enabling proactive, highly targeted maintenance interventions that dramatically extend the operational lifespan of global transportation networks.

References

Sui, M., Jiang, L., Lyu, T., Wang, H., Zhou, L., Chen, P., & Alhosain, A. (2024). Application of deep learning models based on efficientdet and openpose in user-oriented motion rehabilitation robot control. Journal of Intelligence Technology and Innovation, 2(3), 47-77.

Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024). A conditional denoising diffusion probabilistic model for point cloud upsampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20786-20795).

Li, Z., Bai, X., Zhang, T., Shen, P., Xu, W., & Peng, Y. (2025, October). Enhanced Velocity Field Modeling for Gaussian Video Reconstruction. In 2025 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (pp. 304-314). IEEE.

Peng, Q., Zheng, C., & Chen, C. (2024). A dual-augmentor framework for domain generalization in 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2240-2249).

Zhang, J., Zhang, J., Song, Z., Shi, Z., Zhao, C., Shi, Y., ... & Wang, J. (2024). Hoi-m^ 3: Capture multiple humans and objects interaction within contextual environment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 516-526).

Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024). Collaborative debias strategy for temporal sentence grounding in video. IEEE Transactions on Circuits and Systems for Video Technology, 34(11), 10972-10986.

Xiao, Q., & Ghaffari, M. (2026). LaWM: Least Action World Models for Long-Horizon Physical Consistency from Visual Observations. arXiv preprint arXiv:2605.08279.

Chu, Y. B., Li, J., Jia, P., Cui, J., Zhang, R., Kang, X., Lv, M., & Zhang, S. (2021). Irf1- and Egr1-activated transcription plays a key role in macrophage polarization: A multiomics sequencing study with partial validation. International immunopharmacology, 99, 108072. https://doi.org/10.1016/j.intimp.2021.108072

Tu, P., Xie, X., Ai, G., Li, Y., Huang, Y., & Zheng, Y. (2023). FemtoDet: An object detection baseline for energy versus performance tradeoffs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13318-13327).

Guo, H., Lu, C., Bao, F., Pang, T., Yan, S., Du, C., & Li, C. (2023). Gaussian mixture solvers for diffusion models. Advances in Neural Information Processing Systems, 36, 25598-25626.

Zhang, B., Wang, X., Kao, H., Zhang, G., Gao, C., Wang, Y., ... & Chen, K. (2026). Consensus Matrix: A Role-Specialized Multi-Agent Framework for Structured Collaborative Decision-Making in Agentic Visual Media Workflows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4613-4622).

Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 34, 10299-10312.

Li, Q., Xu, T., Luo, T., Zhong, Y., Li, Y., Zhou, Y., & Hou, C. (2026). Theory-driven label-specific representation for incomplete multi-view multi-label learning. Advances in Neural Information Processing Systems, 38, 121996-122037.

Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., ... & Pang, J. (2025). F1: A vision-language-action model bridging understanding and generation to actions. arXiv preprint arXiv:2509.06951.

Li, Q., Liu, Z., Xu, T., Luo, T., & Hou, C. (2026). Adaptive Disentangled Representation Learning for Incomplete Multi-View Multi-Label Classification. arXiv preprint arXiv:2601.05785.

Zhou, Yufan, et al. "Masked Temporal Interpolation Diffusion for Procedure Planning in Instructional Videos." arXiv preprint arXiv:2507.03393 (2025).

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In

Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27325-27335).

Kim, E. H., Wang, Z., Zong, H., Jiang, Z., Fu, Z., & Pedrycz, W. (2023). Design of tobacco leaves classifier through fuzzy clustering-based neural networks with multiple histogram analyses of images. IEEE Transactions on Industrial Informatics, 20(3), 4698-4709.

Z. Qi et al., "Self-Regulated Learning for Egocentric Video Activity Anticipation" in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 45, no. 06, pp. 6715-6730, June 2023, doi: 10.1109/TPAMI.2021.3059923.

Zhang, Jiaquan, et al. "Lightweight llm agent memory with small language models." arXiv preprint arXiv:2604.07798 (2026).

Huang, H., Zhang, C., Hu, Q., & Zhu, P. (2016, August). Multi-view representative and informative induced active learning. In Pacific Rim international conference on artificial intelligence (pp. 139-151). Cham: Springer International Publishing.

Zhao, H., Qi, Z., Wang, C., Zheng, Q., Lu, G., Chen, F., ... & Wu, Z. (2025). Dynamictrl: Rethinking the basic structure and the role of text for high-quality human image animation. arXiv preprint arXiv:2503.21246.

Wang, J., Wells III, W. M., Golland, P., & Zhang, M. (2018, September). Efficient laplace approximation for bayesian registration uncertainty quantification. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 880-888). Cham: Springer International Publishing.

Xu, W., Huang, H., Gong, Y., Yu, L., Wu, Q., & Zhang, J. (2025). Hierarchical multi-prototype discrimination: Boosting support-query matching for few-shot segmentation. IEEE Transactions on Multimedia.

Qu, D., & Ma, Y. (2025). Magnet-bn: markov-guided Bayesian neural networks for calibrated long-horizon sequence forecasting and community tracking. Mathematics, 13(17), 2740.

Ding, Y., Li, Z., Huang, D., Li, Z., & Zhang, K. (2022, October). Enhancing multi-view stereo with contrastive matching and weighted focal loss. In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 821-825). IEEE.

Ma, Y., & Qu, D. (2025, November). GELNO-FD: gauge-equivariant Fourier liquid neural operators for interpretable Markovian Bayesian dynamics. In Fifth International Conference on Advanced Algorithms and Signal Image Processing (AASIP 2025) (Vol. 13967, pp. 177-183). SPIE.

Guo, Hanzhong, et al. "Leveraging verifier-based reinforcement learning in image editing." arXiv preprint arXiv:2604.27505 (2026).

Li, Z., Chen, Z., Qu, F., Wang, M., Zhao, Y., Zhang, K., & Peng, Y. (2024). Point resampling and ray transformation aid to editable NeRF models. arXiv preprint arXiv:2405.07306.

Huang, H., Zhang, J., Zhang, J., Xu, J., & Wu, Q. (2020). Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. IEEE Transactions on Multimedia, 23, 1666-1680.