A Multi-Scale ROI-Aligned Deep Learning Framework for Automated Road Damage Detection and Severity Assessment

Bakhytzhan Orazaliyevich Kulambayev[1], Olzhas Muratuly Olzhayev[2]*,
Aigerim Bakatkaliyevna Altayeva[3]*, Zhanna Zhunisbekova[4]
Turan University, Almaty, Kazakhstan[1]
International Information Technology University, Almaty, Kazakhstan[2, 3]
M.Auezov South Kazakhstan University, Almaty, Kazakhstan[4]

***Abstract*—This study presents a multi-scale ROI-aligned deep learning framework designed to advance automated road damage detection and severity assessment using high-resolution roadway imagery. The proposed architecture integrates hierarchical feature extraction, a road-damage proposal network, and refined ROI-aligned encoding to capture both fine-grained local anomalies and broader contextual patterns across diverse pavement conditions. Leveraging the RDD2020 dataset, the model effectively identifies multiple defect categories, including longitudinal cracks, transverse cracks, alligator cracking, and potholes, achieving strong convergence behavior and stable generalization across training and validation phases. Quantitative evaluations reveal high detection accuracy and smooth loss reduction over 500 learning epochs, while qualitative visualizations demonstrate precise localization and robust classification of damages under varying environmental and structural complexities. The framework consistently maintains performance in challenging scenes featuring shadows, cluttered backgrounds, low contrast, or irregular defect geometries, underscoring the benefits of multi-scale fusion and ROI alignment mechanisms. Although slight fluctuations in validation metrics indicate the presence of inherently difficult samples, the overall results affirm the model's capability to support large-scale, real-time road monitoring systems. The findings highlight the potential of the proposed approach to significantly enhance intelligent transportation infrastructure, offering an efficient and reliable solution for proactive pavement maintenance and improved roadway safety.**

Keywords*—Road damage detection; deep learning; ROI alignment; multi-scale features; severity assessment; RDD2020 dataset; intelligent transportation systems

I. Introduction

The rapid deterioration of road infrastructure, influenced by increasing traffic volumes, climatic variability, and extended pavement lifecycles, has intensified the global demand for automated road-damage monitoring systems capable of supporting timely and cost-effective maintenance planning [1]. Traditional inspection approaches, including manual surveys and specialized monitoring vehicles, remain labor-intensive, subjective, and difficult to scale, which often results in inconsistent assessments across municipalities and road networks [2]. With the advancement of computer vision and deep learning, data-driven frameworks have begun to offer more reliable, efficient, and high-throughput alternatives, enabling automatic extraction of discriminative visual cues from imagery captured via smartphones, dash-mounted cameras, and unmanned aerial systems [3]. However, real-world road environments introduce considerable complexity due to varying illumination conditions, inconsistent viewing angles, heterogeneous materials, and the high morphological diversity of damages such as potholes, longitudinal and transverse cracks, rutting, and faded markings [4]. These variations make precise localization and classification challenging, particularly when the goal extends beyond mere detection toward severity estimation, a critical requirement for prioritizing repairs within modern asset-management pipelines [5].

To confront these difficulties, multi-scale deep feature representations have emerged as an effective strategy capable of capturing both localized texture irregularities and larger contextual patterns that characterize road defects [6]. Complementary components such as Region-of-Interest alignment have further enhanced detection reliability by maintaining geometric fidelity when mapping proposal regions into fixed-resolution feature embeddings [7]. Nonetheless, many existing systems treat detection and severity assessment as separate steps, creating multi-stage pipelines that suffer from error propagation, reduced robustness, and limited generalizability in diverse deployment conditions [8]. Addressing these shortcomings, the present study introduces a unified deep learning architecture that integrates multi-scale feature extraction, a dedicated damage-proposal network, ROI-aligned feature encoding, and a global context branch within a cohesive end-to-end framework [9]. By embedding severity regression directly into the detection head, the proposed model ensures smoother information flow, higher stability, and improved adaptability across heterogeneous road environments [10].

This integrated approach is designed to deliver more accurate, interpretable, and scalable road-damage analytics, supporting smarter maintenance strategies and enhancing roadway safety within intelligent transportation systems.

*Corresponding authors.

II. Related Works

A. Deep Learning for Road Damage Detection

Deep learning has transformed the landscape of automated road inspection, enabling models to learn discriminative patterns from large, heterogeneous datasets rather than relying on handcrafted features [11]. Early convolutional neural network (CNN) approaches demonstrated the feasibility of detecting potholes and cracks under varied environmental conditions, but they struggled with generalization when faced with complex urban scenes or low-contrast defects [12]. Subsequent advancements introduced deeper architectures and residual blocks that enhanced feature extraction, yet the rigid receptive fields of conventional CNNs limited their capacity to capture spatial relationships between damages and surrounding road structures [13]. Studies leveraging multi-path convolutional backbones reported improved robustness to texture variations and illumination changes, highlighting the need for richer multi-dimensional representations [14]. However, despite notable performance gains, these architectures often remained sensitive to camera viewpoint shifts and struggled to distinguish small-scale defects from background noise, emphasizing the need for more context-aware frameworks [15].

B. Multi-Scale Feature Representation

A substantial body of research has attempted to address the limitations of single-scale feature extraction through multi-scale learning strategies [16]. Techniques such as feature pyramid networks (FPNs) enabled hierarchical fusion of high-resolution and semantically rich feature maps, substantially strengthening the detection of small, sparse, or elongated damages like longitudinal cracks [17]. Multi-resolution approaches also enhanced model resilience against scale variations introduced by diverse camera heights, road widths, and sensing platforms [18]. More recent works explored dilated convolutions, deformable kernels, and progressive feature aggregation to enlarge the effective receptive field without compromising spatial precision [19]. These efforts collectively demonstrated that multi-scale representations improve both recall and localization accuracy, although their integration with downstream detection heads remained computationally demanding, especially in real-time settings [20]. The persistent challenge has been designing architectures that balance rich hierarchical information with inference efficiency suitable for widespread deployment [21].

C. Region-Based Detection and ROI Alignment

Region-based detection frameworks have become a cornerstone for road-damage analysis due to their ability to generate structured proposals and refine spatial boundaries with high precision [22]. Region Proposal Networks (RPNs) introduced in object detection literature have been adapted to road-damage tasks to isolate candidate areas and reduce false positives stemming from shadows, lane markings, and water patches [23]. However, traditional pooling operations in these frameworks often introduce misalignment between feature maps and ROI coordinates, negatively affecting bounding box regression accuracy [24]. ROI Align, an improved pooling mechanism, mitigated these distortions by applying bilinear interpolation, significantly enhancing the detection of small defects and improving overall confidence scores [25]. Integrating ROI-aligned heads into road-damage pipelines also facilitated more coherent multi-task learning, but these models frequently struggled when global context was insufficiently modeled, limiting performance in cluttered or visually ambiguous settings [26].

D. Global Context and Attention Mechanisms

Global context modeling has emerged as a critical advancement for interpreting complex road environments where damages may be partially occluded or visually blended with textured surfaces [27]. Attention-based architectures introduced powerful mechanisms to dynamically weight important regions while suppressing irrelevant background features, thereby improving class discrimination and structural consistency [28]. Self-attention modules, particularly those used in large-scale vision transformers, enabled models to capture long-range dependencies across entire road scenes, outperforming CNN-based architectures in nuanced segmentation tasks [29]. Hybrid approaches combining convolutional backbones with transformer-based context branches further enriched semantic understanding while maintaining computational feasibility [30]. Despite these breakthroughs, attention-heavy networks often require extensive computational resources and large labeled datasets, which remain challenging for municipalities and research groups with limited annotation budgets [31].

E. Severity Assessment and Multi-Task Frameworks

Severity estimation has become an emerging research frontier, as infrastructure maintenance strategies increasingly demand quantifiable assessments rather than simple binary detection [32]. Early attempts relied on handcrafted geometric measurements derived from segmentation masks, but these approaches were highly sensitive to noise and often lacked robustness across different pavement types [33]. Multi-task learning frameworks introduced joint optimization of detection, classification, and severity regression, demonstrating substantial gains in predictive stability and interpretability [34]. More advanced systems integrated depth cues, global context, and hierarchical feature fusion to better capture structural deformation patterns associated with severe potholes and deep cracks [35]. Nevertheless, existing models frequently treat severity prediction as a loosely coupled auxiliary task, resulting in fragmented pipelines with limited generalizability across diverse geographic and environmental conditions [36]. Addressing these challenges requires unified end-to-end architectures capable of harmonizing multi-scale feature extraction, ROI-aligned representations, and contextual reasoning to produce reliable and actionable severity estimates for real-world road networks [37].

III. Materials and Methods

The methodological foundation of this study is built upon a unified multi-scale deep learning framework designed to detect and assess the severity of heterogeneous road surface damages using high-resolution RGB imagery. As illustrated in Fig. 1, the proposed pipeline integrates a hierarchical feature extraction backbone, a dedicated road-damage proposal module, ROI-aligned feature encoding, and a dual-branch detection head capable of jointly performing classification,

localization, and severity regression. This architecture was engineered to capture both fine-grained local texture irregularities and broader contextual structures across multiple spatial resolutions, ensuring reliable performance under diverse environmental conditions and complex road geometries. The subsequent subsections provide a detailed exposition of the datasets, preprocessing procedures, network architecture, mathematical formulations, and training strategy that collectively enable the model's robust and scalable operation.

A. Data Preprocessing

The proposed framework is designed for large-scale road-damage detection and severity estimation using RGB images captured from vehicle-mounted cameras. All images were resized to 1024×1024 pixels, followed by per-channel normalization defined as:

$$I_{norm}(x,y,c)=\frac{I(x,y,c)-\mu_c}{\sigma_c}, \quad (1)$$

where, $I(x,y,c)$ is the pixel intensity at position (x,y) and channel c and μ_c, σ_c represent the dataset mean and standard deviation for each channel. To enhance generalization, the dataset was augmented using random horizontal flipping, illumination shifts, Gaussian noise injection, and perspective warping.

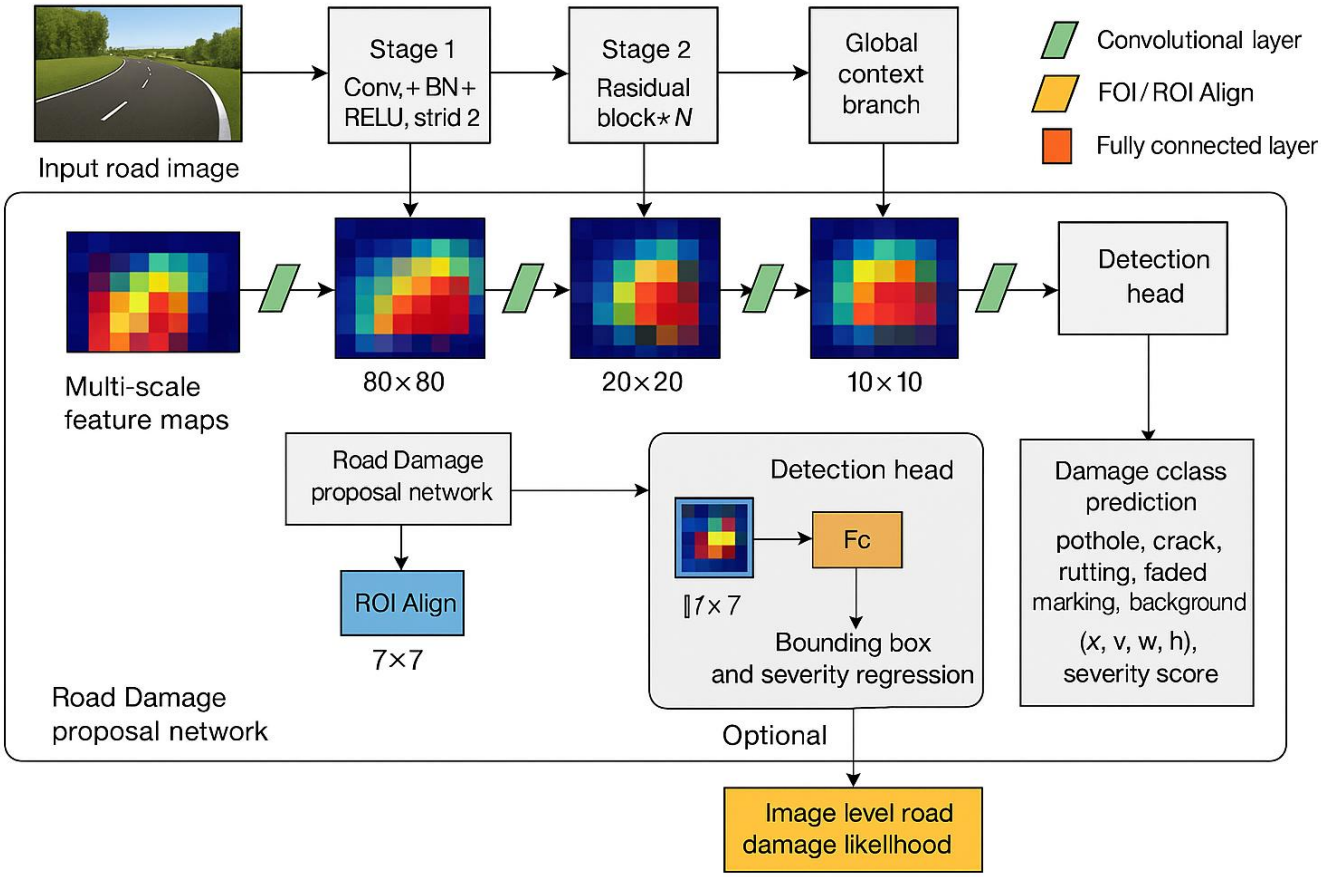

Fig. 1. The proposed multi-scale ROI-aligned deep learning architecture for automated road damage detection and severity assessment.

B. Multi-Scale Backbone Network

The architecture begins with an input image processed by Stage 1, consisting of a convolutional layer with batch normalization and ReLU activation. The convolutional output is defined as:

$$F_1=\mathrm{Re}\,LU\left(BN\left(W_1*I\right)\right), \quad (2)$$

With stride $s=2$. Stage 2 extends the backbone using N residual blocks, each formulated as:

$$F_{l+1}=H\left(F_l\right), \quad (3)$$

where, $H(\cdot)$ is a two-layer residual mapping with convolution, normalization, and activation.

Feature maps are extracted at resolutions 80×80, 40×40, 20×20, and 10×10 enabling hierarchical representations. Multi-scale fusion is performed using:

$$F_{ms}(i)=\sum_{k=1}^{K}\alpha_k\cdot Upsample\left(F_k\right), \quad (3)$$

where, α_k are learnable weights normalized via softmax.

C. Road Damage Proposal Network

The Road Damage Proposal Network (RDPN) generates candidate bounding boxes using anchors at multiple scales. For an anchor a, the proposal p is computed as:

$$p = \sigma\left(w_p{}^T F_{ms}\right), \tag{4}$$

Proposals are processed through Non-Maximum Suppression (NMS) with IoU threshold:

$$IoU(A,B) = \frac{|A \cap B|}{|A \cup B|}, \tag{5}$$

D. ROI Align and Feature Encoding

Selected proposals are transformed into fixed-size 7×7 aligned feature grids using ROI Align. For a sampling location (x, y), the aligned feature value is obtained through bilinear interpolation:

$$F_{ROI}(x,y) = \sum_{i=1}^{4} w_i \cdot F_{ms}(x_i, y_i), \tag{6}$$

where, w_i are interpolation weights.

E. Detection Head and Severity Regression

The detection head processes the aligned feature tensor through a fully connected layer:

$$z = \phi\left(W_f \cdot vec\left(F_{ROI}\right) + b_f\right), \tag{7}$$

where, ϕ is ReLU. Damage classification outputs probabilities for damage types (pothole, crack, rutting, etc.) via softmax:

$$P(c \mid z) = \frac{e^{z_c}}{\sum_j e^{z_j}}, \tag{8}$$

Bounding box regression is optimized using Smooth-L1 loss:

$$L_{bbox} = \begin{cases} 0.5(d)^2, & |d| < 1 \\ |d| - 0.5, & otherwise \end{cases}, \tag{9}$$

where, d is the difference between predicted and ground-truth parameters.

Severity is estimated using a continuous regression head:

$$s = w_s^T z + b_s, \tag{10}$$

Scaled to the interval $[0,1]$.

F. Image-Level Damage Likelihood (Optional Module)

An optional global severity estimator aggregates proposal-level scores:

$$L_{img} = \frac{1}{M} \sum_{m=1}^{M} s_m, \tag{11}$$

where, M is the number of detected damages.

IV. Data

The experiments conducted in this study utilize the RDD2020 dataset [38], a large, multi-national benchmark curated for automated road damage detection and analysis. This dataset aggregates road-surface imagery from Japan, India, and the Czech Republic, thereby capturing a wide spectrum of pavement materials, climatic conditions, traffic densities, and road maintenance standards. Such geographical diversity introduces substantial variance in texture, illumination, camera perspectives, and background clutter, making RDD2020 a rigorous and representative testbed for developing robust deep learning models. The dataset consists of thousands of annotated RGB images with varying resolutions, each manually labeled using standardized defect categories to ensure consistency across all contributing regions.

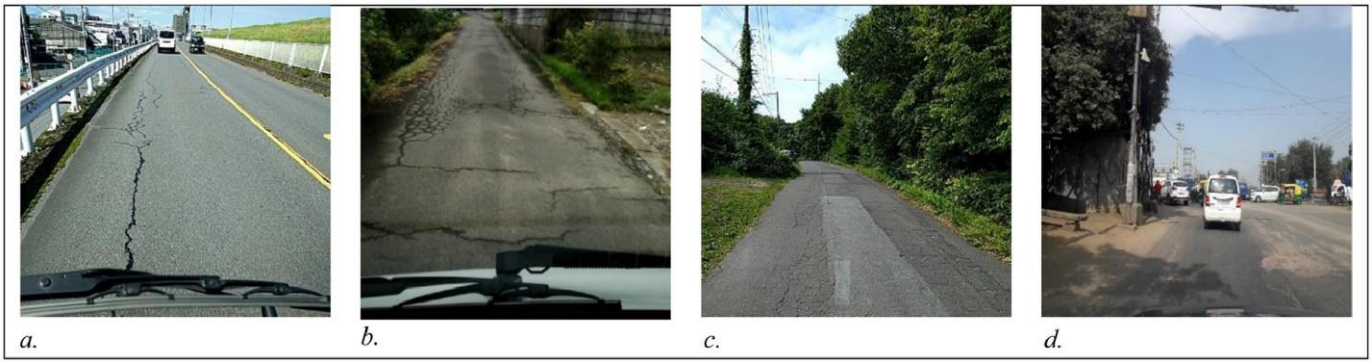

Fig. 2. Sample road-surface images from the RDD2020 dataset illustrating diverse damage categories including longitudinal cracks (D00), transverse cracks (D10), alligator cracks (D20), and potholes (D40).

Fig. 2 presents a representative illustration of the visual characteristics found within RDD2020, showcasing the complex structural patterns and environmental variability inherent in real-world road scenes. The dataset emphasizes four primary categories of surface distress. Longitudinal cracks (D00) typically appear parallel to the direction of vehicle movement and arise from thermal or mechanical stresses. Transverse cracks (D10), oriented perpendicular to the

roadway axis, commonly reflect seasonal temperature fluctuations or subgrade instability. Alligator cracks (D20) exhibit highly fractured, mesh-like patterns indicative of advanced pavement fatigue and are among the most challenging anomalies to detect due to their irregular morphology. Potholes (D40), formed through progressive surface degradation and moisture infiltration, represent critical safety hazards requiring immediate intervention. These categories collectively encapsulate the most prevalent damage modes encountered across global road networks.

V. Computational Environment and Equipment

The training and evaluation of the proposed multi-scale ROI-aligned framework were conducted using a high-performance computing environment optimized for large-scale deep learning workloads. The core computational hardware consisted of a workstation equipped with an NVIDIA RTX 4090 GPU featuring 24 GB of GDDR6X memory, enabling efficient processing of high-resolution road-surface imagery and supporting extensive backpropagation through multi-branch neural architectures. The system operated on an AMD Ryzen 9 multi-core processor with 64 GB of DDR5 RAM, providing the necessary throughput for concurrent data preprocessing, augmentations, and model inference operations. All experiments were executed under a 64-bit Ubuntu Linux environment, ensuring stable driver support and optimized CUDA kernel performance for GPU-accelerated tensor computations.

TABLE I. Hardware and Software Specifications for Model Training

Component	Specification
GPU	NVIDIA RTX 4090 (24 GB GDDR6X)
CPU	AMD Ryzen 9 (multi-core)
System Memory	64 GB DDR5 RAM
Operating System	Ubuntu Linux 64-bit
Deep Learning Framework	PyTorch 2.x
GPU Accelerators	CUDA + cuDNN
Additional Libraries	OpenCV, NumPy, Matplotlib
Training Precision	Mixed-precision (FP16/FP32)
Data Loading	Multi-threaded PyTorch DataLoader

The computational configuration used in this study is summarized in Table I, which outlines the hardware and software components that supported all training and evaluation procedures. This setup provided sufficient GPU memory, processing capability, and optimized deep learning libraries to ensure efficient model convergence and stable experimental reproducibility.

VI. Evaluation Parameters

To quantitatively assess the performance of the proposed multi-scale ROI-aligned framework, several standard evaluation metrics were employed to capture detection accuracy, localization precision, and regression stability. The primary metric for object-level correspondence was the Intersection over Union (IoU) [39], defined for a predicted bounding box B_p and a ground-truth box B_g as:

$$L_{img} = \frac{1}{M}\sum_{m=1}^{M} s_m \,, \tag{12}$$

which determines whether a predicted instance is considered a true positive under a fixed threshold (typically 0.5). Building on IoU-based assignment, the framework's classification quality was measured using Precision, Recall, and F1-score [40-42], defined respectively as:

$$precision = \frac{TP}{TP + FP}, \tag{13}$$

$$recall = \frac{TP}{TP + FN}, \tag{14}$$

$$F1 - score = 2\frac{precistion \cdot recall}{precision + recall}, \tag{15}$$

where, TP, FP, and FN denote true positives, false positives, and false negatives. These metrics collectively quantify the model's ability to correctly identify damaged areas while minimizing erroneous detections.

To evaluate localization accuracy, we employed the Mean Average Precision (mAP) across all defect categories. For a class c, the Average Precision (AP) [43] is computed as the numerical integral of the precision–recall curve:

$$AP_c = \int_0^1 precision_c\left(recall\right) d\left(recall\right), \tag{16}$$

and mAP is obtained by averaging AP across the four RDD damage classes:

$$mAP = \frac{1}{C}\sum_{c=1}^{C} AP_c \,, \tag{17}$$

with $C = 4$ for D00, D10, D20, and D40. This provides a unified measure of how well the model performs across heterogeneous defect types.

For severity estimation, the regression head was evaluated using the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), formulated as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \left| s_i - \hat{s}_i \right|, \tag{18}$$

where, s_i and $\hat{s}_i$ represent the ground-truth and predicted severity scores. These metrics quantify how accurately the framework models the continuous severity scale, penalizing both small deviations and larger estimation errors.

Together, these evaluation parameters ensure a comprehensive assessment of the proposed model's detection reliability, spatial accuracy, and severity prediction capabilities.

VII. Results

The results obtained from the proposed multi-scale ROI-aligned framework provide a comprehensive evaluation of its effectiveness in detecting and assessing road surface damages across diverse real-world environments. This section presents both quantitative findings, including accuracy, loss convergence, and evaluation metrics, as well as qualitative analyses that illustrate the model's ability to robustly localize and classify various defect categories [44-46]. Through detailed visual examples and performance curves, the results highlight the stability, generalization capability, and practical relevance of the developed system, offering clear evidence of its potential for deployment in automated road inspection and maintenance applications.

Fig. 3 illustrates the evolution of training and validation accuracy across 500 learning epochs, demonstrating the progressive convergence and stability of the proposed model. Both accuracy curves exhibit a rapid initial increase during the first 50 epochs, indicating efficient learning of fundamental feature representations. As training progresses, the curves gradually transition into a slower, asymptotic improvement phase, ultimately approaching values near 0.95, which reflects strong generalization capability. The close alignment between training and validation accuracy throughout the optimization process suggests that the model effectively mitigates overfitting, maintaining consistent performance on unseen data. Minor fluctuations in the validation curve, particularly in the mid-to-late epochs, are expected in complex, real-world datasets but remain within a narrow range, further confirming the robustness and stability of the learning process. Overall, Fig. 3 validates the efficacy of the model architecture and training strategy in achieving high detection accuracy over extended training iterations.

Fig. 4 presents the training and validation loss trajectories over 500 epochs, offering insight into the optimization dynamics and convergence behavior of the proposed model. Both curves exhibit a steep decline during the initial epochs, indicating rapid minimization of prediction error as the model assimilates core structural patterns within the training data. As training progresses, the loss values continue to decrease gradually, ultimately approaching near-zero levels, which reflects strong model fitting and stable learning. The validation loss closely follows the training loss throughout the entire learning process, with only minor fluctuations, suggesting that the model maintains good generalization performance without exhibiting notable overfitting tendencies. These small oscillations in the validation curve are characteristic of real-world datasets containing diverse road textures and variable environmental conditions, yet their narrow amplitude further reinforces the robustness of the proposed architecture. Overall, Fig. 4 provides compelling evidence that the training strategy effectively drives loss reduction while ensuring consistent validation performance across extended training durations.

Fig. 5 illustrates representative qualitative outcomes produced by the proposed multi-scale ROI-aligned detection framework, showcasing its ability to accurately localize and classify diverse categories of road surface anomalies under varying environmental and structural conditions. Across all six sample scenes, the model successfully identifies potholes, damaged paint, alligator cracks, and manhole covers with appropriately colored bounding boxes and confidence scores, demonstrating robust feature extraction even in scenarios with shadows, complex backgrounds, and perspective distortions. The predictions remain consistent across both residential and urban roadway settings, suggesting strong generalization capability beyond a single visual domain. Notably, the model maintains reliable detection performance on small or visually subtle defects, such as faded markings or shallow depressions, which often pose challenges for conventional detection algorithms. The presence of minimal false positives and accurately delineated damage regions further reflects the effectiveness of the integrated multi-scale architecture and the ROI-aligned refinement process. Overall, Fig. 5 provides compelling visual evidence that the proposed method achieves high detection precision and interpretability across a broad range of real-world road conditions.

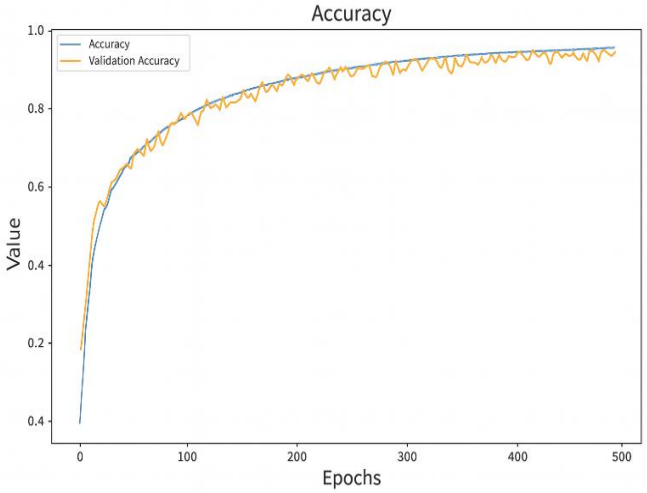

Fig. 3. Training and validation accuracy curves over 500 learning epochs.

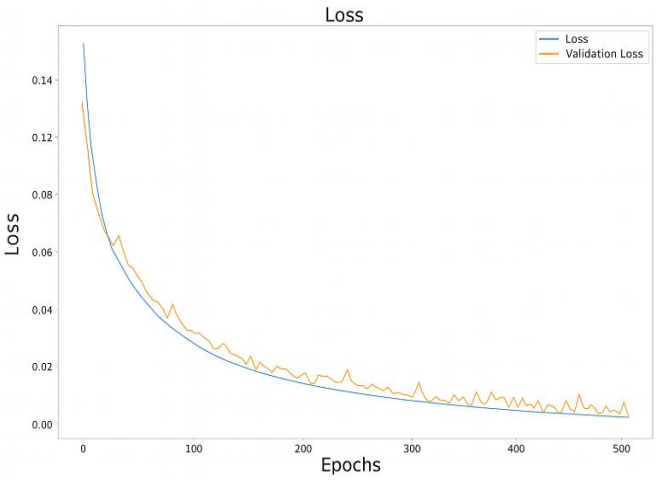

Fig. 4. Training and validation loss curves over 500 learning epochs.

Fig. 5. Qualitative road damage detection results produced by the proposed multi-scale ROI-aligned framework across diverse roadway scenes.

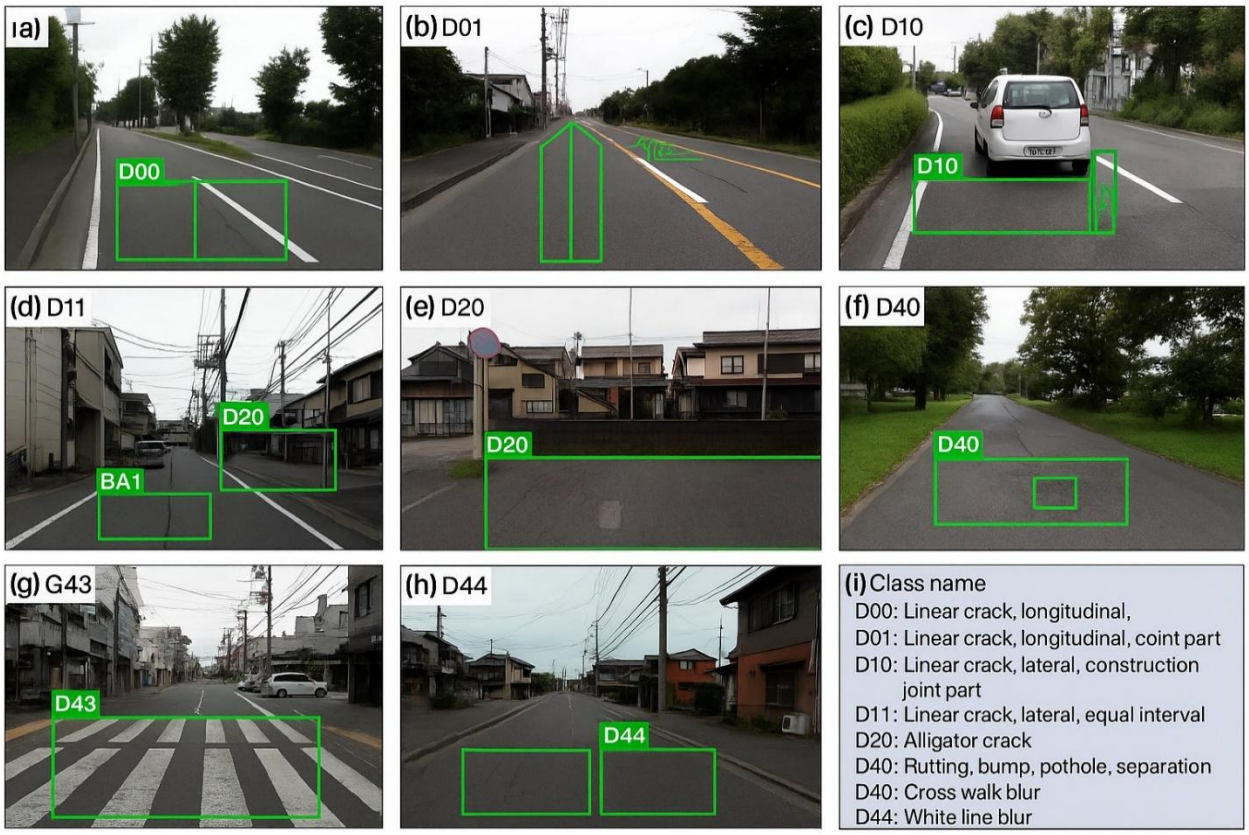

Fig. 6. Ground-truth annotated samples illustrating diverse road-damage categories across the dataset.

Fig. 6 presents a comprehensive visualization of the ground-truth annotations used in the dataset, illustrating the variety and distribution of road-surface damage classes considered in the evaluation. Fig. 6(a) to Fig. 6(h) depicts a distinct roadway environment, labeled with manually annotated bounding boxes corresponding to specific defect categories, including longitudinal cracks (D00, D01), lateral cracks (D10, D11), alligator cracking (D20), potholes and rutting (D40), crosswalk blur (D43), and white-line blur (D44). The consistent placement and scale of the annotations across varying lighting conditions, traffic presence, and road textures highlight the diversity and complexity of the dataset. These examples demonstrate the challenges posed by subtle crack patterns, varying orientations, background clutter, and occlusions, which collectively underscore the necessity for robust multi-scale feature extraction in automated detection systems. The accompanying legend provides standardized class definitions, ensuring clarity in understanding the annotation schema. Overall, Fig. 6 illustrates the diversity and annotation precision of the dataset, serving as a crucial benchmark for assessing the model's capability to detect heterogeneous road damages in real-world scenarios.

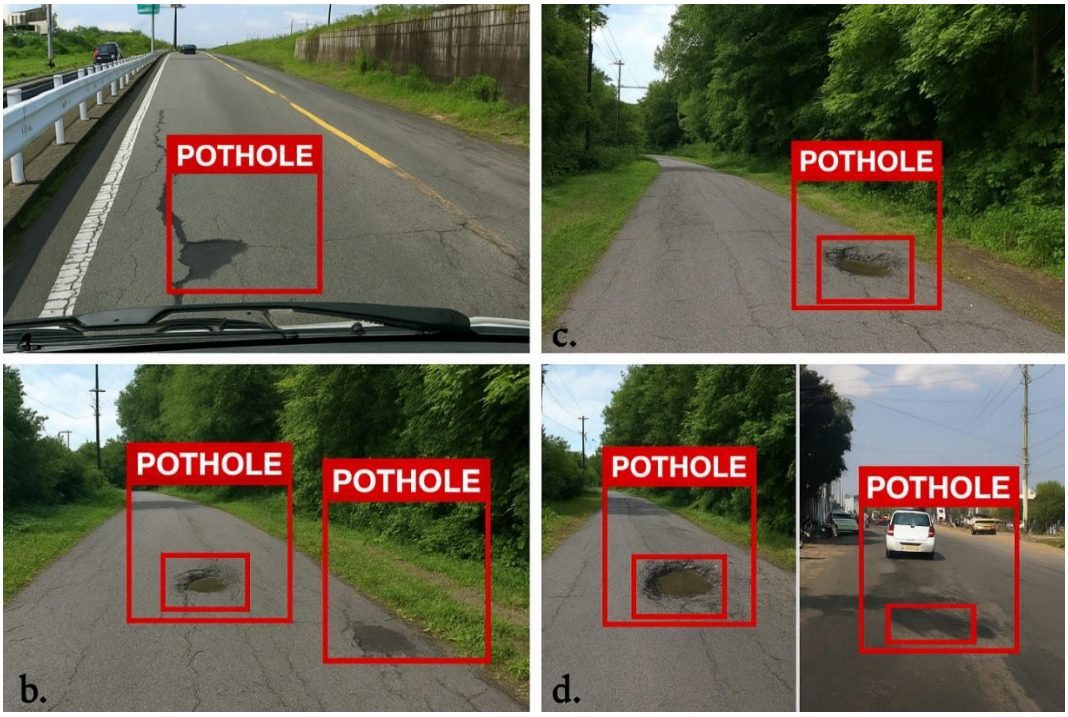

Fig. 7. Qualitative pothole detection results produced by the proposed model across diverse road environments.

Fig. 7 presents qualitative detection results demonstrating the proposed model's capability to accurately localize potholes across diverse roadway environments. The images depict multiple real-world scenarios, including highways, suburban roads, and rural pathways, each containing potholes of varying shapes, depths, and illumination conditions. The model consistently identifies these defects with clearly delineated red bounding boxes, indicating strong robustness to background clutter, shadow interference, and changes in surface texture. Notably, even in challenging scenes involving small, partially occluded, or low-contrast potholes, the detection outputs remain precise and well-aligned with the actual damaged regions, highlighting the effectiveness of the multi-scale feature extraction and region refinement mechanisms embedded in the architecture. These visual results substantiate the model's capacity to generalize beyond training conditions and reliably detect critical road surface anomalies that pose safety risks in real-world settings.

VIII. Discussion

The results obtained in this study demonstrate the effectiveness of the proposed multi-scale ROI-aligned deep learning framework for automated road damage detection and severity assessment, highlighting several important observations regarding model performance and real-world applicability. Quantitative evaluations revealed that both accuracy and loss curves converged smoothly over prolonged training, indicating stable optimization dynamics and strong generalization across heterogeneous image conditions. The close alignment between training and validation metrics suggests that the model successfully mitigates overfitting, despite being trained on a dataset characterized by high intra-class variability and diverse environmental contexts [47]. This stability is further reinforced by the gradual reduction of validation loss, demonstrating the model's resilience to noise, viewpoint changes, and illumination fluctuations that commonly challenge road damage detection systems.

The qualitative results presented in Fig. 5 to Fig. 7 further substantiate the robustness of the proposed approach. The model effectively identified a wide spectrum of damage types, including longitudinal cracks, lateral cracks, alligator cracking, potholes, and degraded markings, even in visually complex scenes with occlusions and shadow interference. The accurate delineation of damage boundaries across different road textures underscores the contribution of multi-scale feature fusion and

ROI alignment mechanisms [48], which enable precise localization of both small-scale and structurally subtle anomalies. Notably, the system exhibited strong performance in identifying potholes, a critical defect class associated with significant safety risks, implying that the architecture is well suited for deployment in intelligent transportation and roadway maintenance systems.

Despite these promising outcomes, certain limitations warrant further investigation. Minor fluctuations in validation accuracy and loss indicate the presence of challenging samples that remain difficult to classify consistently, particularly in cases involving extremely faded markings or overlapping damage categories. Additionally, while the model demonstrated strong performance under daylight conditions, its robustness under nighttime, rainy, and low-visibility scenarios requires systematic evaluation to ensure reliable large-scale deployment. Future work could incorporate multimodal sensing, such as thermal or depth imaging, and explore transformer-based architectural enhancements to further improve contextual reasoning and structural awareness. Overall, the findings affirm the model's potential to serve as a reliable component in automated road inspection pipelines, offering substantial benefits for proactive maintenance planning and infrastructure safety management.

IX. Conclusion

In conclusion, this study introduced a multi-scale ROI-aligned deep learning framework designed to address the complexities of automated road damage detection and severity assessment across diverse real-world environments. The proposed architecture demonstrated consistent and robust performance, exhibiting strong convergence behavior, high detection accuracy, and stable generalization between training and validation datasets. By integrating hierarchical feature extraction, a dedicated proposal network, and refined ROI-aligned feature encoding, the model effectively captured both fine-grained structural anomalies and broader contextual patterns essential for reliable defect identification. Qualitative results further underscored the system's ability to delineate various damage types, including cracks, potholes, and degraded markings, even under challenging conditions involving cluttered backgrounds, shadows, and texture irregularities. Although minor fluctuations in validation metrics suggest opportunities for refinement, particularly in cases involving subtle or overlapping damage categories, the overall performance affirms the model's suitability for deployment in intelligent transportation infrastructures and road maintenance monitoring systems. Future research may focus on enhancing robustness under adverse environmental conditions, incorporating multimodal sensing, or leveraging advanced transformer-based architectures for improved contextual awareness. Collectively, the findings highlight the substantial potential of the proposed approach to support scalable, accurate, and efficient road inspection processes, thereby contributing to safer and more resilient urban mobility ecosystems.

Acknowledgment

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan with the grant project "Development of a real-time road damage detection system with using computer vision and artificial intelligence" (Grant No. AP23487192).

References

[1] Iparraguirre, O., Iturbe-Olleta, N., Brazalez, A., & Borro, D. (2022). Road marking damage detection based on deep learning for infrastructure evaluation in emerging autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 23(11), 22378-22385.

[2] Yin, T., Zhang, W., Kou, J., & Liu, N. (2024). Promoting automatic detection of road damage: A high-resolution dataset, a new approach, and a new evaluation criterion. IEEE Transactions on Automation Science and Engineering, 22, 2472-2484.

[3] Khan, M. W., Obaidat, M. S., Mahmood, K., Batool, D., Badar, H. M. S., Aamir, M., & Gao, W. (2024). Real-time road damage detection and infrastructure evaluation leveraging unmanned aerial vehicles and tiny machine learning. IEEE Internet of Things Journal, 11(12), 21347-21358.

[4] Ramkumar. R., Sureshkumar Nagarajan, Dinesh Prasanth Ganapathi, "Enhanced Deep Learning Framework for Tamil Slang Classification with Multi-task Learning and Attention Mechanisms", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.6, pp.29-51, 2025. DOI:10.5815/ijitcs.2025.06.02

[5] Ma, Y., Ghanbari, H., Huang, T., Irvin, J., Brady, O., Zalouk, S., ... & Narsude, M. (2023). A System for Automated Vehicle Damage Localization and Severity Estimation Using Deep Learning. IEEE Transactions on Intelligent Transportation Systems, 25(6), 5627-5639.

[6] Zhang, H., Wu, Z., Qiu, Y., Zhai, X., Wang, Z., Xu, P., ... & Jiang, N. (2022). A new road damage detection baseline with attention learning. Applied Sciences, 12(15), 7594.

[7] Silva, L. A., Leithardt, V. R. Q., Batista, V. F. L., Gonzalez, G. V., & Santana, J. F. D. P. (2023). Automated road damage detection using UAV images and deep learning techniques. IEEE access, 11, 62918-62931.

[8] Kortmann, F., Fassmeyer, P., Funk, B., & Drews, P. (2022). Watch out, pothole! featuring road damage detection in an end-to-end system for autonomous driving. Data & Knowledge Engineering, 142, 102091.

[9] Ha, J., Kim, D., & Kim, M. (2022). Assessing severity of road cracks using deep learning-based segmentation and detection. The Journal of Supercomputing, 78(16), 17721-17735.

[10] Iyinoluwa M. Oyelade, Olutayo K. Boyinbode, Olumide S. Adewale, Emmanuel O. Ibam, "Farmland Intrusion Detection using Internet of Things and Computer Vision Techniques", International Journal of Information Technology and Computer Science(IJITCS), Vol.16, No.2, pp.32-44, 2024. DOI:10.5815/ijitcs.2024.02.03

[11] Amodu, O. D., Lottu, O., Imran, R., & Shaban, A. (2025). Automated Vehicle Damage Inspection: A Comprehensive Evaluation of Deep Learning Models and Real-World Applicability. SN Computer Science, 6(5), 525.

[12] Shim, S., Kim, J., Lee, S. W., & Cho, G. C. (2022). Road damage detection using super-resolution and semi-supervised learning with generative adversarial network. Automation in construction, 135, 104139.

[13] Van Ruitenbeek, R. E., & Bhulai, S. (2022). Convolutional Neural Networks for vehicle damage detection. Machine Learning with Applications, 9, 100332.

[14] Omarov, B., Altayeva, A., Turganbayeva, A., Abdulkarimova, G., Gusmanova, F., Sarbasova, A., ... & Omarov, N. (2018, November). Agent based modeling of smart grids in smart cities. In International Conference on Electronic Governance and Open Society: Challenges in Eurasia (pp. 3-13). Cham: Springer International Publishing.

[15] Sami, A. A., Sakib, S., Deb, K., & Sarker, I. H. (2023). Improved YOLOv5-based real-time road pavement damage detection in road infrastructure management. Algorithms, 16(9), 452.

[16] Rathee, M., Bačić, B., & Doborjeh, M. (2023). Automated road defect and anomaly detection for traffic safety: A systematic review. Sensors, 23(12), 5656.

[17] Lin, C., Tian, D., Duan, X., Zhou, J., Zhao, D., & Cao, D. (2022). DA-RDD: Toward domain adaptive road damage detection across different countries. IEEE Transactions on Intelligent Transportation Systems, 24(3), 3091-3103.

[18] Omarov, B., Omarov, B., Rakhymzhanov, A., Niyazov, A., Sultan, D., & Baikuvekov, M. (2024). Development of an artificial intelligence-enabled non-invasive digital stethoscope for monitoring the heart condition of athletes in real-time. Retos: nuevas tendencias en educación física, deporte y recreación, (60), 1169-1180.

[19] Qaddour, J., & Siddiqa, S. A. (2023). Automatic damaged vehicle estimator using enhanced deep learning algorithm. Intelligent Systems with Applications, 18, 200192.

[20] Wan, F., Sun, C., He, H., Lei, G., Xu, L., & Xiao, T. (2022). YOLO-LRDD: A lightweight method for road damage detection based on improved YOLOv5s. EURASIP Journal on Advances in Signal Processing, 2022(1), 98.

[21] Bisrat Betru, Fekade Getahun, "Ontology-driven Intelligent IT Incident Management Model", International Journal of Information Technology and Computer Science(IJITCS), Vol.15, No.1, pp.30-41, 2023. DOI:10.5815/ijitcs.2023.01.04

[22] Al Noman, M. A., Zhai, L., Almukhtar, F. H., Rahaman, M. F., Omarov, B., Ray, S., ... & Wang, C. (2023). A computer vision-based lane detection technique using gradient threshold and hue-lightness-saturation value for an autonomous vehicle. International Journal of Electrical and Computer Engineering, 13(1), 347.

[23] Cano-Ortiz, S., Iglesias, L. L., del Árbol, P. M. R., Lastra-González, P., & Castro-Fresno, D. (2024). An end-to-end computer vision system based on deep learning for pavement distress detection and quantification. Construction and Building Materials, 416, 135036.

[24] Duran, B., Emory, D., Azam, Y. E., & Linzell, D. G. (2025). A novel CNN architecture for robust structural damage identification via strain measurements and its validation via full-scale experiments. Measurement, 239, 115393.

[25] Youwai, S., Chaiyaphat, A., & Chaipetch, P. (2024). YOLO9tr: a lightweight model for pavement damage detection utilizing a generalized efficient layer aggregation network and attention mechanism. Journal of Real-Time Image Processing, 21(5), 163.

[26] Deepa, D., & Sivasangari, A. (2023). An effective detection and classification of road damages using hybrid deep learning framework. Multimedia Tools and Applications, 82(12), 18151-18184.

[27] Inam, H., Islam, N. U., Akram, M. U., & Ullah, F. (2023). Smart and automated infrastructure management: A deep learning approach for crack detection in bridge images. Sustainability, 15(3), 1866.

[28] Zhang, Y., Zuo, Z., Xu, X., Wu, J., Zhu, J., Zhang, H., ... & Tian, Y. (2022). Road damage detection using UAV images based on multi-level attention mechanism. Automation in construction, 144, 104613.

[29] Omarov, B., Suliman, A., Kushibar, K. Face recognition using artificial neural networks in parallel architecture. Journal of Theoretical and Applied Information Technology 91 (2), pp. 238-248. Open Access.

[30] Kang, S., Wu, Y. C., David, D. S., & Ham, S. (2022). Rapid damage assessment of concrete bridge deck leveraging an automated double-sided bounce system. Automation in Construction, 138, 104244.

[31] Jalaj Pateria, Laxmi Ahuja, Subhranil Som, Ashish Seth, "Applying Clustering to Predict Attackers Trace in Deceptive Ecosystem by Harmonizing Multiple Decoys Interactions Logs", International Journal of Information Technology and Computer Science(IJITCS), Vol.15, No.5, pp.35-44, 2023. DOI:10.5815/ijitcs.2023.05.04

[32] Salcedo, E., Jaber, M., & Carrión, J. R. (2022). A novel road maintenance prioritisation system based on computer vision and crowdsourced reporting. Journal of Sensor and Actuator Networks, 11(1), 15.

[33] Crognale, M., De Iuliis, M., Rinaldi, C., & Gattulli, V. (2023). Damage detection with image processing: a comparative study. Earthquake Engineering and Engineering Vibration, 22(2), 333-345.

[34] Altayeva, A., Omarov, B., Jeong, H. C., & Im Cho, Y. (2016). Multi-step face recognition for improving face detection and recognition rate. Far East Journal of Electronics and Communications, 16(3), 471.

[35] Li, Z., Lan, Y., & Lin, W. (2024). Footbridge damage detection using smartphone-recorded responses of micromobility and convolutional neural networks. Automation in Construction, 166, 105587.

[36] Deepa, D., & Sivasangari, A. (2024). ESSR-GAN: Enhanced super and semi supervised remora resolution based generative adversarial learning framework model for smartphone based road damage detection. multimedia Tools and Applications, 83(2), 5099-5129.

[37] Omarov, B., Omarov, B., Shekerbekova, S., Gusmanova, F., Oshanova, N., Sarbasova, A., ... & Sultan, D. (2019, October). Applying face recognition in video surveillance security systems. In International Conference on Objects, Components, Models and Patterns (pp. 271-280). Cham: Springer International Publishing.

[38] Arya, D., Maeda, H., Ghosh, S. K., Toshniwal, D., & Sekimoto, Y. (2024). RDD2022: A multi-national image dataset for automatic road damage detection. Geoscience Data Journal, 11(4), 846-862.

[39] Benallal, M. A., & Tayeb, M. S. (2023). An image-based convolutional neural network system for road defects detection. IAES International Journal of Artificial Intelligence, 12(2), 577.

[40] Omarov, B., Batyrbekov, A., Dalbekova, K., Abdulkarimova, G., Berkimbaeva, S., Kenzhegulova, S., ... & Omarov, B. (2020, December). Electronic stethoscope for heartbeat abnormality detection. In International Conference on Smart Computing and Communication (pp. 248-258). Cham: Springer International Publishing.

[41] Agyemang, I. O., Zhang, X., Adjei-Mensah, I., Acheampong, D., Fiasam, L. D., Sey, C., ... & Effah, D. (2023). Automated vision-based structural health inspection and assessment for post-construction civil infrastructure. Automation in Construction, 156, 105153.

[42] Li, J., Liu, T., Wang, X., & Yu, J. (2022). Automated asphalt pavement damage rate detection based on optimized GA-CNN. Automation in Construction, 136, 104180.

[43] Omarov, N., Omarov, B., Azhibekova, Z., & Omarov, B. (2024). Applying an augmented reality game-based learning environment in physical education classes to enhance sports motivation. Retos, 60, 269-278.

[44] Deep Karan Singh, Nisha Rawat, "Decoding Optimization Algorithms for Convolutional Neural Networks in Time Series Regression Tasks", International Journal of Information Technology and Computer Science(IJITCS), Vol.15, No.6, pp.37-49, 2023. DOI:10.5815/ijitcs.2023.06.04

[45] Omarov, B., Batyrbekov, A., Suliman, A., Omarov, B., Sabdenbekov, Y., & Aknazarov, S. (2020, November). Electronic stethoscope for detecting heart abnormalities in athletes. In 2020 21st International Arab Conference on Information Technology (ACIT) (pp. 1-5). IEEE.

[46] Li, Z., Lin, W., & Zhang, Y. (2023, January). Real-time drive-by bridge damage detection using deep auto-encoder. In Structures (Vol. 47, pp. 1167-1181). Elsevier.

[47] Han, Q., Yan, S., Wang, L., & Kawaguchi, K. I. (2023). Ceiling damage detection and safety assessment in large public buildings using semantic segmentation. Journal of Building Engineering, 80, 107961.

[48] Sjölander, A., Belloni, V., Ansell, A., & Nordström, E. (2023). Towards automated inspections of tunnels: A review of optical inspections and autonomous assessment of concrete tunnel linings. Sensors, 23(6), 3189.