



ARTICLE

TCR-RoadNet: A Transformer-Enhanced Multi-Task Deep Learning Architecture for Real-Time Road Damage Detection and Segmentation

Olzhas Olzhayev¹, Bakhytzhan Kulambayev^{2,*} and Azizah Suliman³

¹Department of Mathematical and Computer Modeling, International Information Technology University, Almaty, Kazakhstan

²Higher School of Telecommunications, Turan University, Almaty, Kazakhstan

³Faculty of Science and Technology, Asia Metropolitan University, Subang Jaya Campus, Malaysia

*Corresponding Author: Bakhytzhan Kulambayev. Email: b.kulambayev@turan-edu.kz

Received: 19 March 2026; Accepted: 25 May 2026; Published: 23 July 2026

ABSTRACT: Automated road damage detection is a critical component of intelligent transportation systems, enabling efficient infrastructure maintenance and improved traffic safety. However, existing approaches often suffer from limited contextual understanding, insufficient segmentation accuracy, and suboptimal real-time performance. This study presents TCR-RoadNet, a transformer-enhanced multi-task deep learning architecture designed for simultaneous road damage detection and segmentation in real-world driving environments. The proposed framework integrates a multi-scale convolutional backbone with a Transformer Context Refinement (TCR) module to capture both fine-grained structural details and long-range spatial dependencies across feature scales. To further enhance performance, a Decoupled Detection Head (DDH) is employed to stabilize localization and classification learning, while a Classification Refinement Module (CRM) improves inter-class discrimination using region-based feature enhancement. In addition, a Boundary-Aware Segmentation Head is introduced to generate precise damage contours by incorporating edge-sensitive learning mechanisms. The model is evaluated on the RDD2022 dataset, complemented by an extended dataset containing additional road images collected from Kazakhstan to increase environmental diversity and robustness. Experimental results demonstrate that the proposed method achieves 0.9416 precision, 0.9235 recall, and 0.8718 mAP@50, along with a segmentation performance of 0.8129 mean Intersection over Union (mIoU), while maintaining real-time inference at 57 FPS. Comparative analysis and ablation studies confirm that each architectural component contributes to consistent performance gains in both detection and segmentation tasks. Qualitative results further illustrate the robustness of the proposed framework under varying lighting, weather, and road conditions. The proposed approach offers a scalable and efficient solution for real-time road condition monitoring and intelligent infrastructure management systems.

KEYWORDS: Road damage detection; pavement defect segmentation; transformer; computer vision; multi-task deep learning; intelligent transportation systems

1 Introduction

Road transportation networks constitute a critical component of modern infrastructure, supporting economic activity, public mobility, and logistics operations across urban and rural regions. However, continuous exposure to dynamic traffic loads, environmental stress, and aging pavement materials inevitably leads to the formation of various surface defects, including potholes, longitudinal cracks, transverse cracks, and complex fatigue patterns. These damages gradually degrade pavement quality and, if left undetected, may evolve into severe structural failures that increase maintenance costs and compromise traffic safety.

Traditional inspection approaches typically rely on manual surveys conducted by maintenance personnel or specialized inspection vehicles. Although such procedures provide valuable observations, they are time consuming, expensive, and often limited in spatial coverage, particularly for large scale road networks [1]. Consequently, the development of automated road damage monitoring systems has become an important research direction in intelligent transportation infrastructure management [2].

Recent progress in computer vision and deep learning has significantly advanced automated road damage detection using camera based imaging systems. Convolutional neural networks have demonstrated strong capability in extracting hierarchical visual features and identifying complex patterns associated with pavement deterioration [3]. Object detection frameworks have been widely employed to localize defects such as cracks and potholes within road images captured by vehicle mounted cameras or mobile devices [4]. Nevertheless, road damage detection remains a challenging task due to the diversity of pavement textures, illumination variations, occlusions, and the wide range of defect scales present in real world environments. Moreover, many existing approaches focus solely on bounding box detection or classification, which limits their ability to capture the detailed structural characteristics of pavement defects required for precise infrastructure assessment [5].

To address these challenges, recent research has explored multi task learning architectures that simultaneously perform detection, classification, and segmentation of road damages. Multi scale feature learning strategies have been introduced to improve the detection of defects that appear at different spatial resolutions [6]. In parallel, transformer based attention mechanisms have emerged as a powerful tool for modeling long range contextual relationships within visual representations [7]. By integrating transformer modules with convolutional backbones, modern vision systems can enhance feature interactions across spatial scales and capture contextual cues that are difficult to model using conventional convolution operations alone [8]. These hybrid architectures have demonstrated promising performance in several dense prediction tasks, including semantic segmentation and object detection in complex visual scenes [9].

Motivated by these developments, this paper proposes TCR-RoadNet, a transformer enhanced multi task deep learning architecture designed for real time road damage detection and segmentation. The proposed framework integrates a multi scale convolutional backbone for hierarchical feature extraction with a transformer context refinement module that enhances cross scale feature interactions. In addition, the architecture incorporates a decoupled detection head for accurate localization and a boundary aware segmentation head for generating detailed damage masks. A classification refinement module further improves category prediction through region based feature analysis. By jointly optimizing detection and segmentation tasks within a unified framework, the proposed approach aims to provide a comprehensive and efficient solution for automated road surface inspection systems [10].

The primary contributions of this study can be summarized as follows. First, a novel multi-task deep learning architecture, TCR-RoadNet, is proposed to jointly perform road damage detection and segmentation within a unified framework, enabling more comprehensive and efficient analysis of pavement conditions. Second, a Transformer Context Refinement (TCR) module is introduced to enhance multi-scale feature representations by capturing long-range spatial dependencies and facilitating cross-scale contextual interaction, which is particularly important for complex and irregular road damage patterns. Third, a Decoupled Detection Head (DDH) is incorporated to separate localization and classification processes, thereby improving optimization stability and detection accuracy. Fourth, a Classification Refinement Module (CRM) is designed to refine class predictions using region-of-interest features, effectively reducing ambiguity among visually similar damage categories. Fifth, a Boundary-Aware Segmentation Head is developed to improve contour precision by emphasizing edge information, resulting in more accurate delineation of cracks and potholes. Finally, the robustness and generalization capability of the proposed framework are

enhanced through the integration of a geographically diverse dataset, combining the RDD2022 benchmark with additional real-world road images collected from Kazakhstan. These contributions collectively enable the proposed method to achieve high detection and segmentation performance while maintaining real-time inference capability.

2 Related Works

Automated road damage detection has attracted significant attention in recent years due to the rapid development of intelligent transportation systems and the increasing availability of camera-based monitoring platforms. Early research in this area primarily relied on traditional image processing techniques, including edge detection, thresholding, texture analysis, and handcrafted feature extraction. These approaches attempted to identify pavement cracks and potholes based on geometric patterns or intensity variations in road images. Although such methods demonstrated feasibility under controlled conditions, their performance was often limited when dealing with complex road textures, illumination changes, shadows, and noise present in real-world environments [11]. Consequently, the limitations of handcrafted feature-based techniques motivated the transition toward data-driven machine learning models capable of learning discriminative representations directly from image data [12]. Recent advancements have further emphasized the need for robust real-time anomaly detection systems that can be integrated into intelligent driving assistance frameworks, highlighting the importance of efficient and adaptive detection models [13].

The emergence of deep learning has significantly improved the performance of automated road damage analysis systems. Convolutional neural networks have become the dominant framework for visual feature extraction due to their ability to learn hierarchical representations from raw images [14]. Several studies have applied CNN-based architectures to classify road surface conditions and detect defects using vehicle-mounted cameras or smartphone-based imaging systems. Object detection models such as region-based convolutional networks and single-stage detectors have been widely adopted for identifying the location of road defects within images [15]. In particular, recent developments in YOLO-based architectures have demonstrated strong performance in terms of both accuracy and real-time capability, making them highly suitable for practical deployment in road inspection systems [16]. These detection frameworks provide bounding box predictions that allow maintenance authorities to identify the spatial distribution of pavement damages across road networks. However, bounding box localization alone is often insufficient for detailed pavement assessment because it does not provide precise structural information about crack shapes or pothole boundaries [17].

To overcome this limitation, researchers have explored more advanced learning paradigms that combine deep learning with attention mechanisms and multi-level feature representations. Transformer-based approaches have shown promising results in modeling long-range dependencies and improving contextual understanding in complex visual environments. Moreover, recent studies have investigated transformer-based parameter aggregation and efficient representation learning strategies, demonstrating their effectiveness in enhancing model scalability and feature interaction in distributed settings [18]. In parallel, hybrid feature-based methods that integrate traditional descriptors with learning-based representations have also been proposed to improve robustness under challenging imaging conditions, such as low contrast or repetitive textures [19]. These developments reflect a broader trend toward combining convolutional feature extraction with contextual modeling and attention mechanisms to achieve more reliable and generalizable performance in real-world road damage detection tasks.

To overcome this limitation, researchers have explored semantic and instance segmentation approaches for road damage analysis. Segmentation models enable pixel-level identification of pavement defects, which is particularly useful for estimating the size and severity of cracks and potholes. Encoder-decoder

architectures, including variants of fully convolutional networks and U-shaped networks, have demonstrated strong performance in extracting fine-grained structural information from road images [20]. These models typically utilize multi-scale feature fusion to capture both global contextual information and local texture patterns associated with pavement deterioration [21]. Despite their effectiveness, many segmentation models remain computationally expensive and may struggle to achieve real-time performance when deployed in large-scale road inspection systems [22].

More recently, attention mechanisms and transformer-based architectures have been introduced to enhance feature representation in computer vision tasks. Transformer models enable long-range dependency modeling by capturing relationships between spatial tokens across an image, which can improve contextual understanding in complex scenes [23]. Hybrid architectures combining convolutional backbones with transformer modules have demonstrated strong performance in tasks such as object detection, semantic segmentation, and scene understanding [24]. These approaches leverage convolutional layers to extract local visual patterns while using attention mechanisms to refine global feature interactions. Cross-scale attention mechanisms have also been proposed to improve feature fusion across different resolution levels, which is particularly beneficial for detecting objects that vary significantly in size [25].

In addition to architectural innovations, several studies have investigated multi-task learning frameworks for road damage analysis. Multi-task models simultaneously perform detection, classification, and segmentation within a unified architecture, allowing shared feature representations to improve overall performance [26]. Such frameworks can reduce computational redundancy and improve inference efficiency compared with separate models trained for individual tasks [27]. Recent research has also explored the integration of lightweight detection heads and feature pyramid networks to improve the detection of small and irregularly shaped road defects [28]. These improvements enable models to better capture fine structural details while maintaining computational efficiency suitable for real-time deployment [29].

Despite these advances, several challenges remain in the design of robust road damage detection systems. Road defects often appear under varying lighting conditions, occlusions, and heterogeneous pavement textures, which can degrade model performance in real-world scenarios [30]. In addition, the wide range of damage scales requires architectures capable of effectively combining multi-resolution features while preserving spatial detail [31]. Transformer-based contextual modeling offers promising opportunities to address these issues by enabling dynamic feature interactions across spatial scales [32]. Furthermore, integrating classification refinement mechanisms can help reduce misclassification between visually similar damage categories such as longitudinal and transverse cracks [33]. Boundary-aware segmentation modules have also been introduced to enhance the delineation of crack edges and pothole contours, leading to more accurate damage masks [34].

Given these observations, there remains a need for an integrated architecture that simultaneously captures multi-scale spatial features, contextual relationships, and precise boundary information while maintaining real-time performance. The proposed TCR-RoadNet framework addresses these requirements by combining a multi-scale convolutional backbone with transformer-based context refinement and specialized output heads for detection, classification, and segmentation. This design enables accurate localization and detailed structural analysis of road damages in camera-based inspection systems, contributing to the advancement of intelligent road monitoring technologies [35].

Despite significant advancements in road damage detection, existing methods typically focus on either detection or segmentation tasks independently, limiting their ability to provide comprehensive structural analysis of pavement defects. Moreover, many approaches rely on convolutional architectures that lack sufficient capability to model long-range contextual relationships, which are critical for distinguishing complex and irregular damage patterns.

While transformer-based models have recently been introduced in computer vision tasks, their application in road damage detection remains limited, particularly in multi-task settings. Additionally, existing works rarely address the challenge of inter-class ambiguity among visually similar damage categories or the need for precise boundary delineation.

The proposed TCR-RoadNet framework addresses these limitations by integrating multi-scale convolutional feature extraction, transformer-based contextual refinement, classification refinement, and boundary-aware segmentation within a unified architecture. This design enables simultaneous detection and segmentation while improving both contextual understanding and spatial precision, which are essential for real-world road condition analysis.

3 Materials and Methods

This section describes the proposed TCR-RoadNet framework for real-time road damage detection and segmentation. The architecture is designed as a unified multi-task system that integrates convolutional feature extraction with transformer-based contextual modeling and task-specific prediction heads. Unlike conventional detection pipelines that rely solely on local convolutional features, the proposed approach explicitly addresses the challenges of scale variability, structural ambiguity, and weak boundary representation commonly observed in road damage analysis.

Road defects such as cracks and potholes exhibit highly irregular shapes, varying sizes, and low contrast with surrounding pavement textures. In addition, real-world driving conditions introduce noise factors such as shadows, occlusions, and illumination variations. These challenges motivate the need for a model that can simultaneously capture fine local patterns and global contextual relationships, while also providing precise spatial delineation of damage regions. The overall architecture, illustrated in Fig. 1, is therefore designed to progressively refine feature representations and distribute them across complementary tasks including detection, classification, and segmentation.

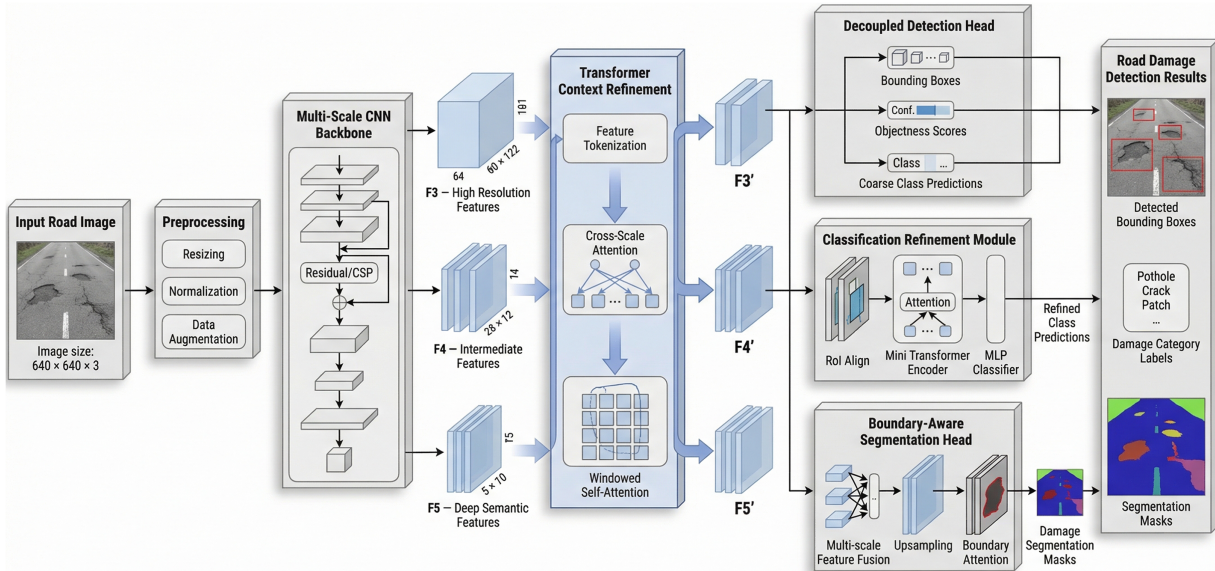


Figure 1: Transformer-based multi-scale feature refinement architecture.

Let the input road image be denoted as $I \in R^{H \times W \times 3}$. In the present implementation, each image is resized to 640×640 before being forwarded to the network. The preprocessing operation expressed as

$$I_r = \mathcal{R}(I), \quad (1)$$

where $\mathcal{R}(\cdot)$ denotes resizing and normalization. During training, augmented samples are generated to improve robustness against illumination changes, viewpoint variations, and texture diversity:

$$I_a = A(I_r), \quad (2)$$

where $A(\cdot)$ represents a composition of photometric and geometric augmentations.

As shown in Fig. 1, the preprocessed image is first passed through a multi-scale CNN backbone, which extracts hierarchical features at progressively reduced spatial resolutions. These backbone features are then forwarded to a Transformer Context Refinement module to enhance cross-scale interaction and long-range contextual representation. The refined features are subsequently distributed to three task-oriented branches. The first is a Decoupled Detection Head, which predicts bounding boxes and objectness scores. The second is a Classification Refinement Module, which revisits detected regions and improves fine-grained class discrimination among visually similar damage categories. The third is a Boundary-Aware Segmentation Head, which generates crack and pothole masks with improved contour precision. Thus, the final model jointly optimizes detection, classification, and segmentation within one inference graph.

If the backbone outputs are denoted by $\{F_3, F_4, F_5\}$, the refined feature set generated by the transformer module can be written as

$$\{F'_3, F'_4, F'_5\} = \tau(F_3, F_4, F_5), \quad (3)$$

where $\tau(\cdot)$ denotes the transformer context refinement operation. The complete model output is then defined as

$$y = \{\hat{B}, \hat{C}, \hat{M}\}, \quad (4)$$

where $\hat{B}$ represents detected boxes, $\hat{C}$ represents refined class predictions, and $\hat{M}$ denotes segmentation masks.

3.1 Multi-Scale CNN Backbone

The first stage of the proposed framework employs a multi-scale convolutional backbone to extract hierarchical feature representations from the input image. The motivation behind this design lies in the observation that road damages manifest at different spatial scales. For example, thin cracks require high-resolution features to preserve fine details, whereas potholes and alligator cracks require broader contextual understanding to distinguish them from background structures.

To address this, the backbone is structured as a hierarchical network that progressively reduces spatial resolution while increasing semantic abstraction. Early layers focus on capturing low-level texture and edge information, which is critical for identifying narrow crack patterns. Deeper layers encode high-level semantic features, enabling robust discrimination between damage types and background noise. The resulting multi-scale feature maps provide complementary information that supports both detection and segmentation tasks.

The internal design of the feature extractor is shown in Fig. 2. The backbone follows a hierarchical multi-stage convolutional structure composed of 3×3 convolutional layers, residual Cross Stage Partial (CSP)-style blocks, and downsampling operations. Its objective is to preserve fine local patterns in shallow layers while

gradually building stronger semantic representations in deeper layers. This design is particularly suitable for road damage analysis because narrow cracks, patching artifacts, and potholes may occupy very different image regions and scales.

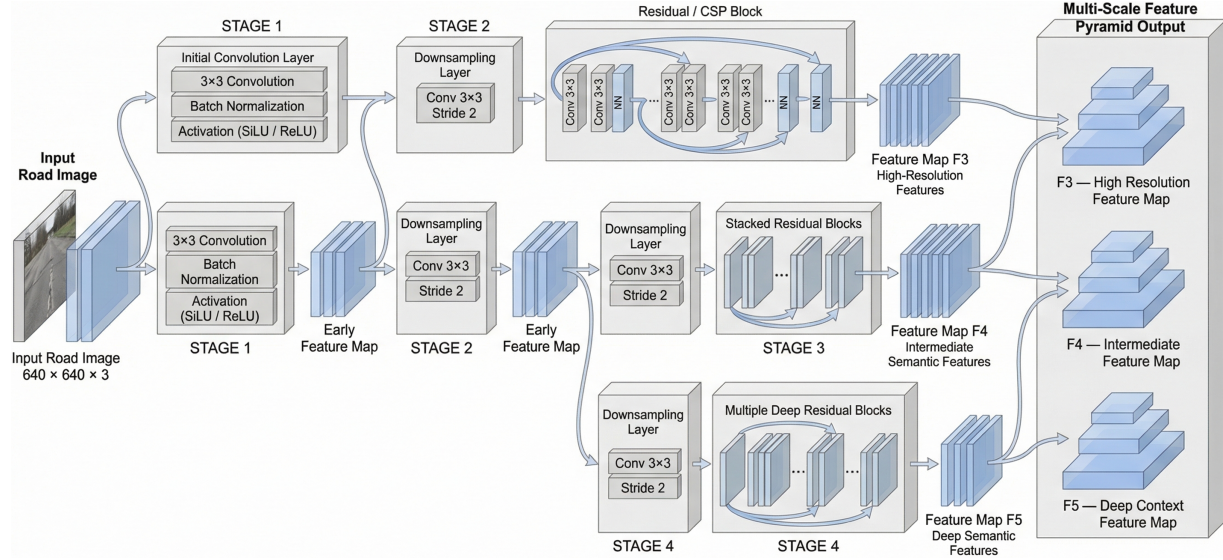


Figure 2: Multi-scale CNN backbone architecture.

Given the input tensor I_a , the first stage extracts low-level texture descriptors. Subsequent stages reduce spatial resolution and increase channel dimensionality. The feature mapping at stage l is written as

$$F_l = \phi_l(F_{l-1}, \theta_l), \quad (5)$$

where $\phi_l(\cdot)$ denotes the transformation composed of convolution, normalization, activation, and residual aggregation, and θ_l denotes learnable parameters. In accordance with Fig. 2, the backbone finally outputs three feature maps:

$$F_3 \in R^{80 \times 80 \times 256}, F_4 \in R^{40 \times 40 \times 512}, F_5 \in R^{20 \times 20 \times 1024} \quad (6)$$

These three levels provide complementary information. F_3 preserves detailed spatial structure useful for thin cracks and boundary-sensitive regions, F_4 captures intermediate semantic patterns, and F_5 encodes high-level contextual cues beneficial for robust damage localization under cluttered backgrounds.

Before entering the transformer stage, the features are projected into a unified embedding dimension d using 1×1 convolution or linear projection. This harmonization allows efficient cross-scale attention while reducing dimensional inconsistency across pyramid levels.

3.2 Transformer Context Refinement Block

While convolutional layers are effective at modeling local spatial patterns, they are inherently limited in capturing long-range dependencies across the image. This limitation becomes critical in road damage detection, where defects may be fragmented, partially occluded, or visually similar to surrounding textures. To overcome this, the proposed architecture introduces a Transformer Context Refinement module, as presented in Fig. 3. The key idea behind this module is to enhance feature representations by enabling cross-scale interaction and global contextual reasoning. By transforming feature maps into token representations

and applying attention mechanisms, the model can dynamically associate information from different spatial regions. This allows the network to link disconnected crack segments, distinguish damage from visually similar patterns. The combination of cross-scale attention and local refinement ensures that both global structure and local detail are preserved.

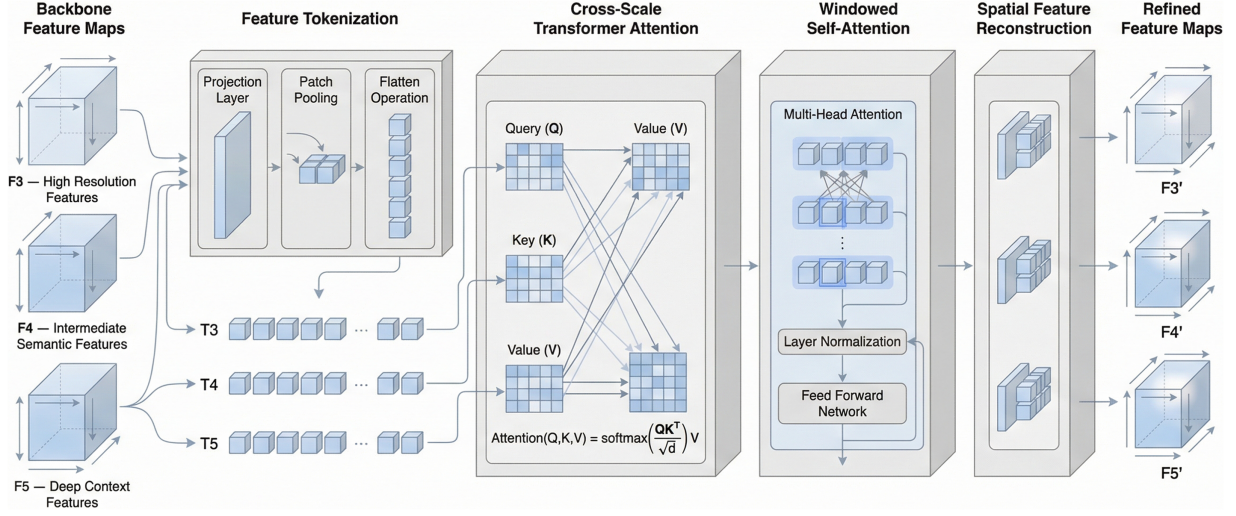


Figure 3: Transformer context refinement block.

As shown in the left part of Fig. 3, the feature maps F_3 , F_4 and F_5 are first tokenized through projection and patch pooling operations. Let the token sequence derived from scale s be denoted as $X_s \in \mathbb{R}^{N_s \times d}$. Tokenization can be expressed as

$$X_s = Tok(F_s), s \in \{3, 4, 5\} \quad (7)$$

where $Tok(\cdot)$ denotes flattening with projection into embedding dimension d . A query set is then generated from one scale and matched against keys and values aggregated from the other scales. The core cross-scale attention mechanism follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{D}}\right)V \quad (8)$$

where Q , K , and V denote query, key, and value matrices, and D is the dimensionality of the key vectors. Through this operation, the model can associate fine crack textures from shallower scales with stronger semantics from deeper layers. In practical terms, the module helps distinguish true pavement defects from visually similar distractors such as lane markings, repaired regions, shadows, and surface stains.

After cross-scale attention, a windowed self-attention and reprojection stage restores the enriched token representation back to the spatial domain. The refined feature map for scale s can be written as

$$F'_s = Proj(WSA(X_s^*)) \quad (9)$$

where X_s^* denotes the context-enhanced token sequence, $WSA(\cdot)$ is windowed self-attention, and $Proj(\cdot)$ denotes reshaping and linear reprojection into a feature map. The output of this stage is the refined pyramid $\{F'_3, F'_4, F'_5\}$, which is then shared by the downstream multi-task heads.

3.3 Decoupled Detection Head

Traditional detection heads often jointly optimize bounding box regression and classification within a shared feature space. However, these tasks require different types of information. Localization depends on precise spatial alignment, while classification relies on semantic discrimination. This mismatch can lead to unstable optimization and reduced detection accuracy. To address this issue, the proposed framework employs a Decoupled Detection Head, which separates the learning of localization and classification tasks. This design allows each branch to specialize in its respective objective. As a result, the model achieves more accurate bounding box predictions and improved confidence estimation. This is particularly beneficial for road damage detection, where defects often exhibit irregular shapes and ambiguous boundaries.

The first prediction branch is the Decoupled Detection Head, shown in Fig. 4. Instead of predicting all target attributes from a single shared subnetwork, the proposed head separates box regression, objectness estimation, and classification prediction. This decoupling improves optimization stability because localization and semantic recognition often require different feature sensitivities. For road damage analysis, this is especially important because a crack may be spatially elongated but semantically ambiguous, whereas a pothole may be easily classified yet difficult to localize precisely under weak contrast.

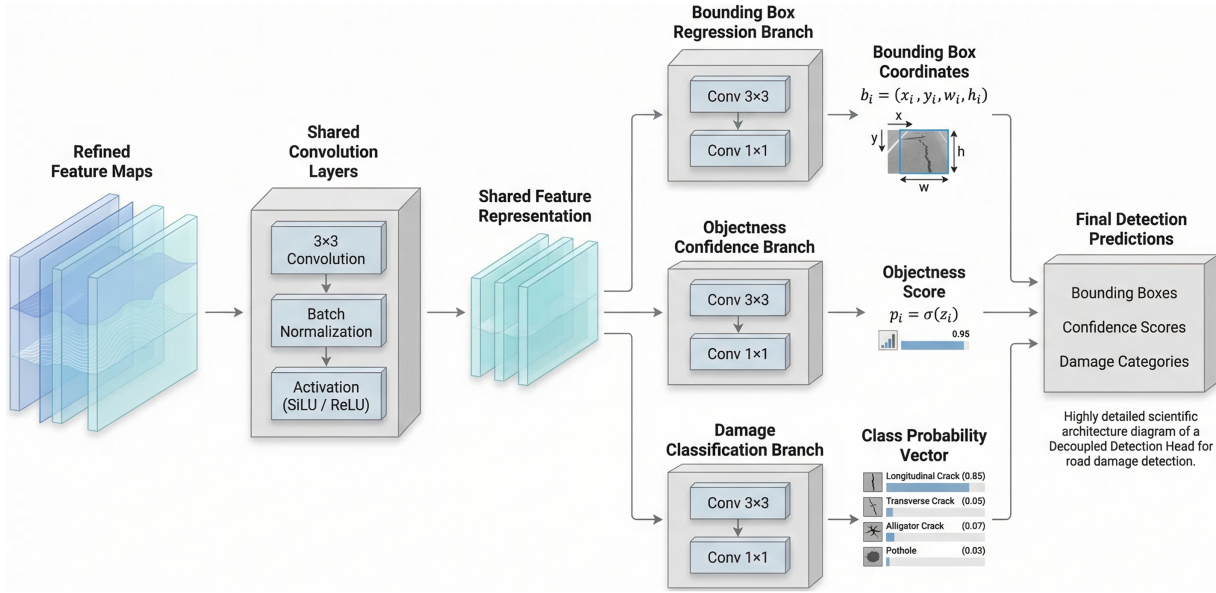


Figure 4: Decoupled detection head.

The refined feature maps $\{F'_3, F'_4, F'_5\}$ are first processed through lightweight shared convolutions. Then, separate branches produce box coordinates, objectness scores, and coarse category logits. The box regression branch predicts

$$\hat{b}_i = (\hat{x}_i, \hat{y}, \hat{w}_i, \hat{h}) \quad (10)$$

for the i -th candidate instance. Objectness is modeled as a confidence value indicating whether the region contains a valid road defect:

$$\hat{o}_i = \sigma(z_i^{obj}) \quad (11)$$

where $\sigma(\cdot)$ is the sigmoid function and z_i^{obj} is the objectness logit. Likewise, the coarse class distribution is obtained through

$$\hat{p}_i^{cls} = \text{Softmax}(z_i^{cls}) \quad (12)$$

where z_i^{cls} represents class logits over the damage categories. The decoupled structure shown in Fig. 4 allows each sub-branch to specialize in a distinct task while still benefiting from shared refined features.

3.4 Classification Refinement Module

Road damage categories such as longitudinal cracks, transverse cracks, and alligator cracks often share similar visual characteristics, making them difficult to distinguish using global features alone. Misclassification is especially common when defects appear in cluttered or degraded road surfaces.

The Classification Refinement Module (CRM) is introduced to address this challenge by performing region-level feature refinement. Instead of relying solely on initial predictions, the module re-examines detected regions using localized features extracted via RoI alignment. This enables the model to focus on fine-grained structural details, such as orientation and texture patterns, which are critical for accurate classification. Consequently, the CRM significantly reduces inter-class confusion and improves overall classification reliability.

Although the detection head produces preliminary class scores, visually similar classes such as longitudinal cracks, transverse cracks, alligator cracks, and patching regions may still be confused. To reduce such ambiguity, the model employs the Classification Refinement Module presented in Fig. 5. This module operates on detected regions of interest after non-maximum suppression and performs a second-stage region-based class refinement.

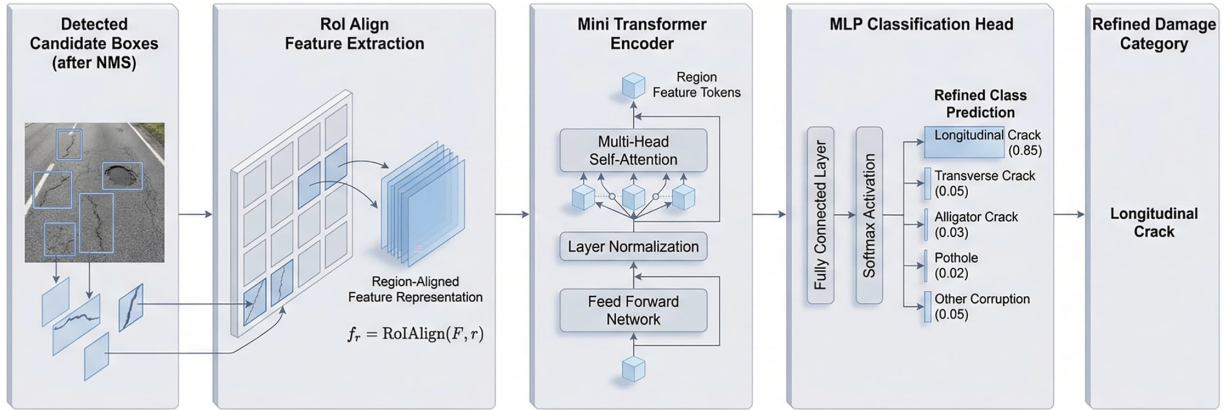


Figure 5: Classification refinement module.

Candidate boxes are first filtered and used to extract region-aligned features through RoI Align. Let $\hat{B}_{nms}$ denote the set of selected regions after suppression. For each region $R_j \in \hat{B}_{nms}$, a region descriptor is extracted as

$$z_j = \text{RoIAlign}(F'_3, F'_4, F'_5, r_j) \quad (13)$$

These region features are then processed by a compact mini-transformer and an MLP classifier to produce refined class probabilities. The refinement operation is especially useful when local shape orientation,

texture fragmentation, or repair boundaries must be interpreted within a limited spatial region rather than the whole image context. In practical terms, it enables more reliable differentiation between crack types and surface patch defects, which often share coarse appearance statistics but differ in local structural organization.

3.5 Boundary-Aware Segmentation Head

Accurate segmentation of road damage is essential for assessing the severity and extent of pavement deterioration. However, standard segmentation methods often struggle to capture precise boundaries, particularly for thin cracks or low-contrast regions. The proposed Boundary-Aware Segmentation Head is designed to explicitly address this limitation. By incorporating multi-scale feature fusion and boundary attention mechanisms, the model emphasizes edge-sensitive regions and suppresses background ambiguity. This enables the generation of segmentation masks with sharper and more accurate contours. The resulting masks provide valuable information for downstream applications such as damage quantification and maintenance planning.

The final prediction branch, depicted in Fig. 6, is the Boundary-Aware Segmentation Head. Its purpose is to generate pixel-level masks for road damages, with special emphasis on preserving contour quality. This is crucial because the severity and physical extent of pavement defects are often better characterized by their exact boundaries rather than by bounding boxes alone.

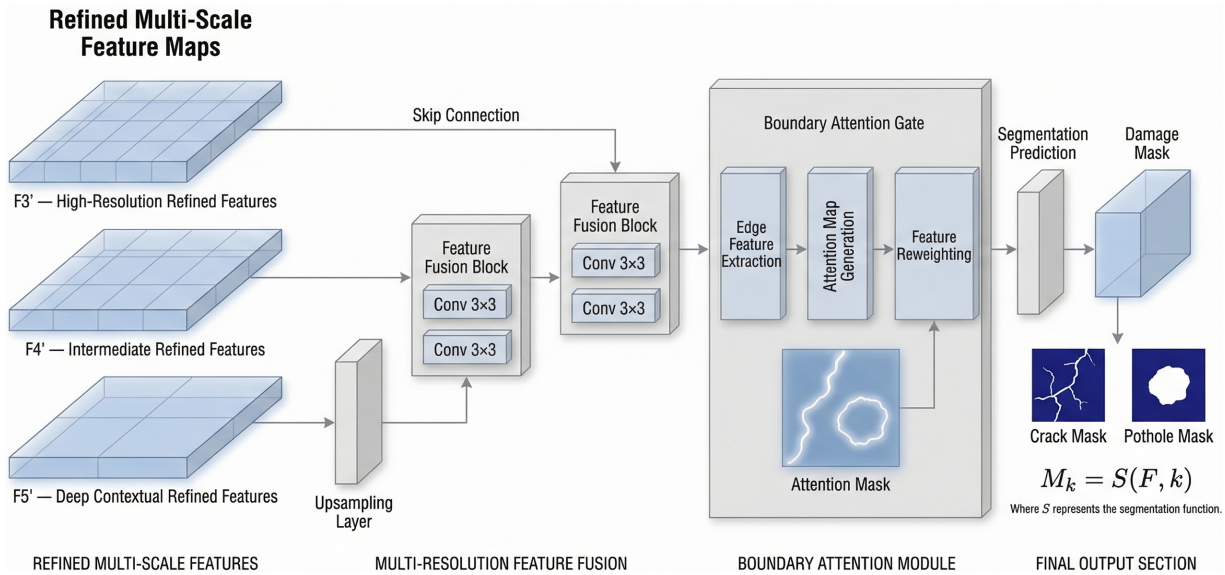


Figure 6: Boundary-aware segmentation head.

The refined multi-scale features are fused through progressive upsampling. This multi-resolution fusion allows the segmentation branch to combine fine spatial details from shallower features with contextual discrimination from deeper ones. A boundary attention gate is further introduced to emphasize contour-sensitive regions and suppress ambiguous interior-background transitions. The final mask prediction for class k is obtained as

$$\hat{M}_k = \sigma \left(\psi_j \left(F'_3, F'_4, F'_5 \right) \right) \quad (14)$$

where $\psi_j(\cdot)$ denotes the mask generation function for the k -th defect category. In this study, the segmentation branch particularly targets crack and pothole masks.

3.6 Multi-Task Objective Function

The proposed architecture jointly optimizes detection, classification, and segmentation tasks within a unified framework. This multi-task design allows shared feature representations to be leveraged across tasks, improving both efficiency and performance. Detection provides coarse localization, classification refines semantic understanding, and segmentation delivers precise spatial delineation. By combining these complementary objectives, the model achieves a balanced representation that is both semantically rich and spatially precise. This integration is particularly important for real-world deployment, where accurate and reliable road condition assessment requires more than simple object detection.

The full training objective combines detection, classification, and segmentation losses so that all branches are optimized jointly. The total loss is defined as

$$\mathcal{L}_k = \lambda_1 \mathcal{L}_{box} + \lambda_2 \mathcal{L}_{obj} + \lambda_3 \mathcal{L}_{cls} + \lambda_4 \mathcal{L}_{ref} + \lambda_5 \mathcal{L}_{seg} + \lambda_6 \mathcal{L}_{bd} \quad (15)$$

where $\mathcal{L}_{box}$ is the bounding box regression loss, $\mathcal{L}_{obj}$ is the objectness loss, $\mathcal{L}_{cls}$ is the coarse detection classification loss, $\mathcal{L}_{ref}$ is the refinement classification loss, $\mathcal{L}_{seg}$ is the mask segmentation loss, and $\mathcal{L}_{bd}$ is the boundary supervision term. The coefficients $\lambda_1, \dots, \lambda_6$ balance the contribution of each task during optimization.

Overall, the proposed TCR-RoadNet architecture integrates the advantages of convolutional local representation, transformer-based contextual reasoning, decoupled detection learning, region-level semantic refinement, and boundary-sensitive segmentation. The proposed framework is designed to support accurate and real-time road damage understanding on camera images under realistic deployment conditions.

4 Data

The experimental evaluation of the proposed TCR-RoadNet framework is conducted using a combination of publicly available and locally collected datasets to ensure both benchmark comparability and real-world applicability. The primary dataset employed in this study is the RDD2022 dataset, which is widely used for road damage detection tasks due to its diversity and large-scale annotations.

4.1 Dataset

The primary dataset used for training and evaluating the proposed TCR-RoadNet framework is the RDD2022 (Road Damage Dataset 2022) [36]. This dataset is a large-scale benchmark designed specifically for automatic road damage detection using computer vision and deep learning methods. RDD2022 contains 47,420 road images collected from six different countries, including Japan, India, the Czech Republic, Norway, the United States, and China. The images were captured using vehicle-mounted cameras under diverse environmental conditions such as varying illumination levels, traffic scenes, pavement textures, and weather conditions. The dataset includes more than 55,000 annotated road damage instances, covering five major pavement defect categories: longitudinal crack, transverse crack, alligator crack, pothole, and other corruption. The dataset was originally introduced as part of the Crowd Sensing-based Road Damage Detection Challenge (CRDDC 2022) to promote the development of automated infrastructure inspection systems. Fig. 7 presents several representative examples from the dataset, illustrating the diversity of road scenes and the visual variability of pavement damages encountered in real-world environments.

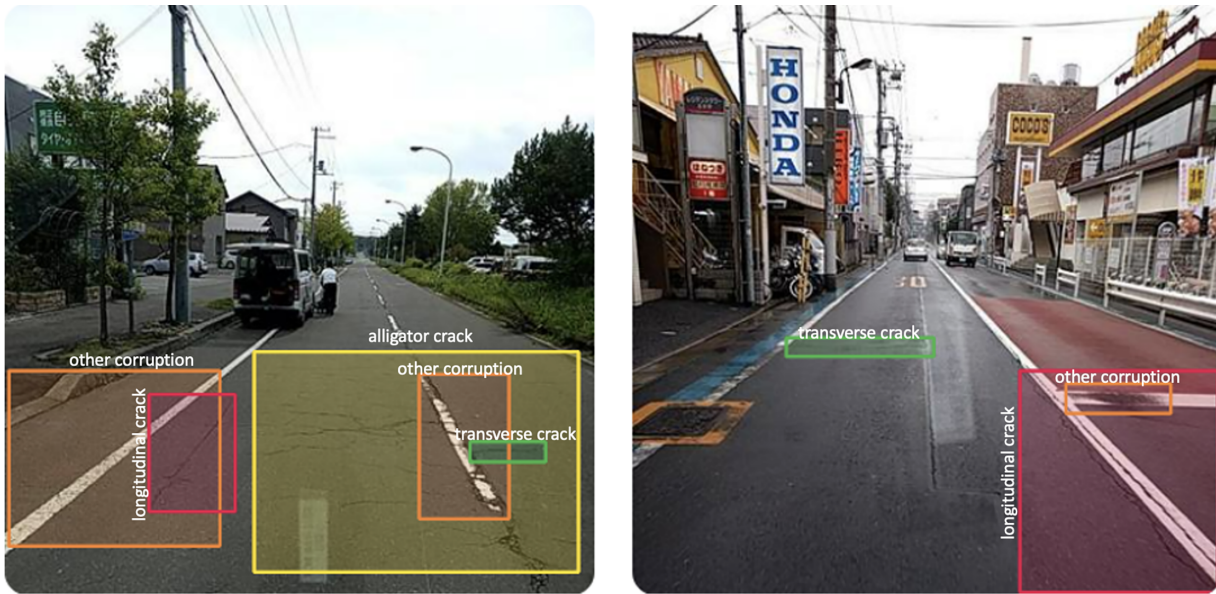


Figure 7: Representative samples from the RDD2022 dataset.

A detailed statistical summary of the dataset is provided in [Table 1](#), which reports the number of images, object instances, average instance count per image, and average relative area of each damage category. As shown in [Table 1](#), the dataset contains 13,657 images with longitudinal crack annotations, representing the largest category in the dataset with 26,196 labeled objects, corresponding to an average of 1.92 instances per image. The alligator crack category includes 8412 images and 10,617 objects, with an average coverage area of 15.41%, which is significantly larger than other classes due to the extensive surface degradation associated with fatigue cracking. Transverse cracks appear in 7744 images, generating 11,875 labeled objects with an average object frequency of 1.53 per image. The pothole category contains 3674 images and 6544 instances, with a relatively small average area of 2.76%, reflecting the localized nature of pothole defects. Finally, the other corruption class includes 4708 images and 5850 objects, representing irregular pavement deterioration patterns that do not fall into the standard crack categories. These statistics demonstrate that the dataset contains significant variability in both object frequency and spatial coverage across different damage types, which provides a challenging benchmark for evaluating detection and segmentation models.

The original RDD2022 dataset provides annotations for five primary road damage categories, namely longitudinal cracks, transverse cracks, alligator cracks, potholes, and other corruption. However, for a more comprehensive evaluation of real-world road conditions, the present study extends the classification scheme to include eight categories. In addition to the original damage classes, three additional categories, namely damaged paint, damaged crosswalk, and manhole cover, are incorporated based on observations from real-world road environments and the supplementary dataset collected from Kazakhstan.

The inclusion of these additional classes is motivated by the fact that such road surface irregularities frequently appear in practical driving scenarios and may affect vehicle safety and navigation. These categories were annotated during the local dataset collection process and integrated into the training and evaluation pipeline. As a result, the proposed model is trained and evaluated on a unified dataset that combines the original RDD2022 annotations with extended class labels, enabling a more realistic and comprehensive assessment of road surface conditions.

Table 1: RDD2022 dataset statistics by damage category.

Class	Images	Objects	Count on Image (Average)	Area on Image (Average)
Alligator crack	8412	10,617	1.26	15.41%
Longitudinal crack	13,657	26,196	1.92	5.04%
Other corruption	4708	5850	1.24	10.8%
Pothole	3674	6544	1.78	2.76%
Transverse crack	7744	11,875	1.53	3.35%

In addition to the RDD2022 dataset, we further expanded the training data by collecting road images from Kazakhstan road environments to improve the robustness of the proposed model under local pavement conditions. Road surfaces in Kazakhstan often exhibit distinct structural characteristics influenced by climate factors such as extreme seasonal temperature variations, freeze–thaw cycles, and heavy truck traffic. To capture these variations, we collected and annotated additional road images using vehicle-mounted cameras operating in urban and suburban regions. Specifically, we added 2000 images for each damage category, resulting in an additional 10,000 labeled images covering all five classes listed in Table 1. This augmentation improves the geographic diversity of the dataset and enables the model to learn damage patterns that are not fully represented in the original RDD2022 data. By combining the international benchmark dataset with locally collected road images, the resulting training corpus provides a more comprehensive representation of global and regional pavement deterioration scenarios.

4.2 Segmentation Annotation Protocol

Although the RDD2022 dataset provides bounding box annotations for road damage detection, it does not include pixel-level segmentation masks required for training the segmentation branch of the proposed TCR-RoadNet framework. To address this limitation, a semi-automatic annotation pipeline was developed to generate high-quality segmentation ground truth for crack and pothole regions.

Initially, coarse segmentation masks were generated using the Segment Anything Model (SAM), which provides strong general-purpose segmentation capabilities for object-level mask extraction. For each annotated bounding box in the dataset, SAM was applied to produce an initial segmentation mask corresponding to the damage region. This approach allowed efficient large-scale mask generation while preserving spatial consistency with detection annotations.

However, automatic segmentation alone is insufficient for precise boundary delineation, particularly for thin cracks and irregular damage patterns. Therefore, all generated masks were manually refined by trained annotators using annotation tools such as CVAT. The refinement process focused on correcting boundary inaccuracies, removing noise, and ensuring alignment with visible damage structures. Special attention was given to complex cases such as low-contrast cracks, overlapping damage regions, and partially occluded defects.

To ensure annotation quality and consistency, a two-stage validation process was applied. First, annotators performed internal verification by cross-checking generated masks against original images. Second, a subset of annotations was reviewed by a senior annotator to validate correctness and maintain labeling standards. The final segmentation masks were then integrated with the detection annotations to form a unified dataset for multi-task learning.

This semi-automatic annotation strategy provides a practical balance between scalability and accuracy, enabling the proposed model to learn precise spatial representations of road damage while maintaining consistency with the original dataset structure.

4.3 Local Dataset Collection and Annotation

To enhance the robustness and generalization capability of the proposed model, additional road images were collected from real-world driving environments in Kazakhstan. The inclusion of this dataset aims to address domain variability arising from differences in road materials, weather conditions, and infrastructure characteristics that are not fully represented in the RDD2022 dataset.

The data acquisition process was conducted using vehicle-mounted cameras installed on passenger vehicles operating in urban and suburban regions. Images were captured under diverse environmental conditions, including varying illumination levels, weather scenarios, and traffic densities. The collected images have a resolution of 1920×1080 pixels, ensuring sufficient detail for both detection and segmentation tasks. To maintain consistency with the RDD2022 dataset, all images were resized to 640×640 during preprocessing.

The annotation process was performed using professional labeling tools such as CVAT, where both bounding boxes and segmentation masks were generated. Each image was annotated according to the same class definitions used in the extended evaluation setup, including alligator cracks, longitudinal cracks, transverse cracks, potholes, damaged paint, damaged crosswalk, manhole cover, and background regions. To ensure labeling consistency, a detailed annotation guideline was developed, specifying class definitions, boundary rules, and handling of ambiguous cases.

A total of 2000 images per class were added, resulting in a balanced dataset that complements the original RDD2022 distribution. The annotation process involved multiple trained annotators with prior experience in computer vision labeling tasks. To ensure quality and reliability, a two-stage verification process was implemented. In the first stage, annotations were reviewed internally by annotators to correct potential labeling errors. In the second stage, a subset of annotated samples was cross-validated by a senior reviewer to ensure consistency and accuracy across the dataset.

This augmentation strategy significantly improves the diversity of training data and enables the proposed model to better generalize to real-world conditions, particularly in regions with different road characteristics and environmental factors.

5 Results

Comprehensive experiments were conducted to evaluate the effectiveness of the proposed TCR-RoadNet framework for road damage detection and segmentation under realistic driving conditions. The evaluation was performed using the RDD2022 dataset together with additional road images collected from Kazakhstan road environments to increase scene diversity and pavement variability. Model performance was assessed using standard metrics including precision, recall, mAP@50, mAP@50–95, and segmentation mIoU, enabling a detailed analysis of both detection accuracy and mask quality. In addition to quantitative evaluation, qualitative visualization results are presented to demonstrate the capability of the proposed architecture to detect different pavement defects in complex road scenes. The experimental analysis further includes confusion matrix evaluation, precision–recall performance curves, and ablation studies to investigate the contribution of individual architectural components and to validate the effectiveness of the transformer-enhanced multi-task learning framework. All experiments were conducted using the extended 8-class setting to better reflect real-world road scenarios.

5.1 Evaluation Metrics

The performance of the proposed TCR-RoadNet framework was evaluated using widely adopted metrics in object detection and image segmentation tasks. These metrics allow a comprehensive assessment of the model's capability to correctly localize, classify, and segment road surface damages. For the detection task, the primary evaluation measures include precision, recall, and mean Average Precision (mAP), which are computed based on the comparison between predicted bounding boxes and ground-truth annotations [37–39]. Precision measures the proportion of correctly predicted damage instances among all predicted detections and is defined as

$$Precision = \frac{TP}{TP + FP} \quad (16)$$

where TP denotes the number of true positive detections and FP represents false positive predictions. Recall evaluates the ability of the model to detect all relevant damage instances and is defined as

$$Recall = \frac{TP}{TP + FN} \quad (17)$$

where FN represents the number of false negatives, corresponding to ground-truth objects that were not detected by the model.

To evaluate detection accuracy across different confidence thresholds, the Average Precision (AP) metric is calculated from the area under the precision–recall curve. The mean Average Precision (mAP) is then obtained by averaging the AP values across all classes. In this study, two commonly used detection metrics are reported: mAP@50, which uses an Intersection-over-Union (IoU) threshold of 0.5, and mAP@50–95, which averages AP values across IoU thresholds ranging from 0.5 to 0.95. The IoU between the predicted bounding box B_p and ground truth box B_g is defined as

$$IoU = \frac{|B_p \cap B_g|}{|B_p \cup B_g|} \quad (18)$$

For the segmentation task, the performance of the model is measured using Intersection-over-Union (IoU) and mean Intersection-over-Union (mIoU), which evaluate the overlap between predicted masks and ground-truth segmentation maps. The IoU for a single class is defined as

$$IoU = \frac{TP}{TP + FP + FN} \quad (19)$$

The overall segmentation accuracy is then computed by averaging IoU values across all classes to obtain mIoU. Segmentation masks were generated using the protocol described in [Section 4.1](#). In addition, the *Dice* coefficient is used to measure the similarity between predicted and ground-truth masks, defined as

$$Dice = \frac{2TP}{2TP + FP + FN} \quad (20)$$

These evaluation metrics collectively provide a balanced assessment of detection reliability, classification capability, and segmentation accuracy of the proposed framework in real-world road damage monitoring scenarios.

5.2 Implementation Details and Computational Setup

The proposed TCR-RoadNet model was implemented using the PyTorch deep learning framework and trained on a workstation equipped with an NVIDIA RTX 3090 GPU (24 GB VRAM), an Intel Core i9 processor, and 64 GB RAM. All experiments were conducted under a Linux-based environment with CUDA acceleration enabled.

The input images were resized to 640×640 pixels, and batch training was performed using standard stochastic gradient optimization techniques. The model was trained for a fixed number of epochs until convergence, and inference performance was evaluated on the test set using the same resolution.

The reported inference speed of 57 frames per second (FPS) corresponds to GPU-based evaluation under batch size 1, representing a real-time processing scenario. In this study, real-time capability is defined as the ability of the system to process input frames at or above 30 FPS, which is commonly considered sufficient for real-world deployment in vehicle-mounted inspection systems.

5.3 Experiment Results

The experimental evaluation was conducted using a subset of the RDD2022 dataset, which was reorganized into eight representative classes to better capture the diversity of road surface conditions encountered in real driving scenarios. The selected classes include Alligator Cracks, Damaged Crosswalk, Damaged Paint, Longitudinal Cracks, Manhole Cover, Potholes, Transverse Cracks, and Background. This categorization allows the evaluation process to assess the ability of the proposed model to distinguish true pavement defects from visually similar road surface elements such as traffic markings or infrastructure components. The inclusion of classes such as damaged crosswalks and painted markings is particularly important because these elements frequently appear in road scenes and can resemble crack patterns in camera images. Each category contains images captured under varying environmental conditions, including changes in lighting, pavement texture, and urban scene complexity. Such variability provides a challenging benchmark for evaluating the robustness and generalization capability of the proposed detection and segmentation framework.

The classification behavior of the proposed TCR-RoadNet architecture is illustrated through the confusion matrix presented in Fig. 8. The matrix provides a comprehensive view of prediction distributions across the eight classes, highlighting both correct detections and potential misclassification patterns. Strong values along the diagonal indicate that the model correctly identifies the majority of instances belonging to key pavement damage categories, particularly Alligator Cracks, Longitudinal Cracks, and Manhole Covers, which demonstrate high true prediction counts. Reliable detection performance is also observed for Transverse Cracks and Potholes, indicating that the multi-scale feature representation and transformer-based contextual refinement contribute to improved recognition of structurally complex defects. Some confusion is observed between Damaged Paint and crack-related classes, which can be attributed to the visual similarity between deteriorated lane markings and fine crack structures on asphalt surfaces. Additionally, a relatively large number of predictions are associated with the Background category because many road images contain extensive undamaged pavement regions. Despite these challenges, the confusion matrix demonstrates that the proposed architecture effectively captures discriminative features across different road damage categories while maintaining robustness against visually similar non-defect structures.

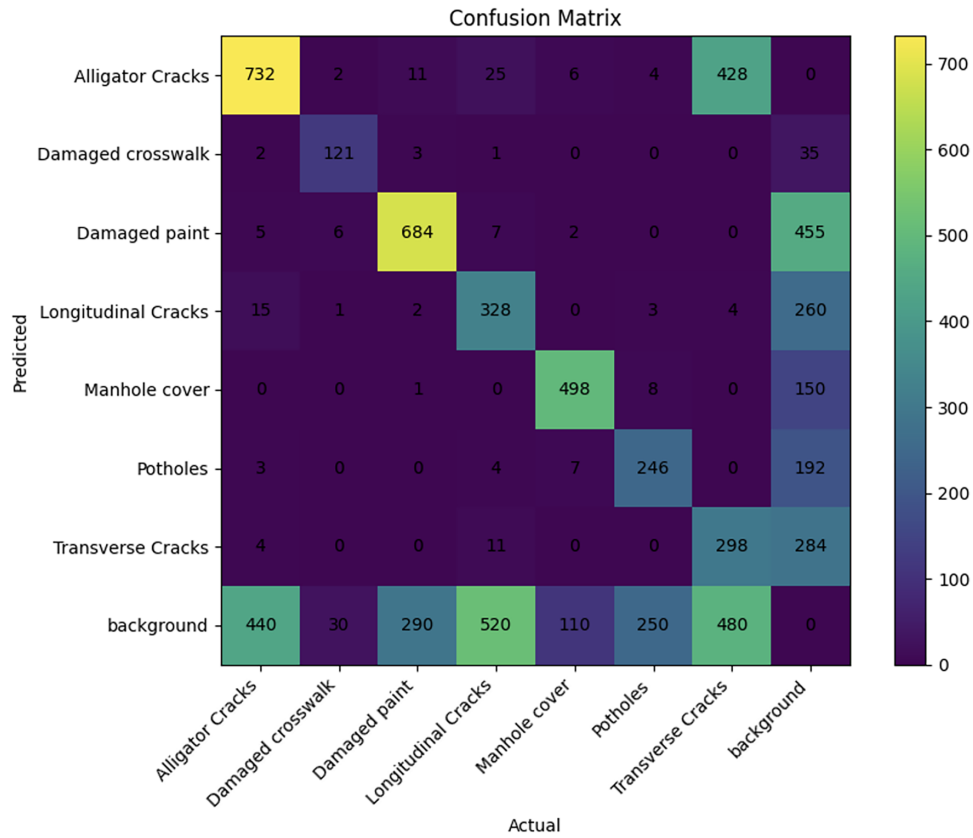


Figure 8: Confusion matrix illustrating the classification performance of the proposed TCR-RoadNet model across different road damage categories.

Fig. 9 illustrates the evaluation curves of the proposed TCR-RoadNet model, including the Precision–Confidence, Recall–Confidence, Precision–Recall, and F1–Confidence relationships for the eight considered road surface categories. The Precision–Confidence curve demonstrates that prediction precision increases steadily as the confidence threshold grows, indicating that high-confidence detections produced by the model correspond to highly reliable predictions. Several classes, particularly Damaged Crosswalk and Manhole Cover, achieve precision values close to 1.0 at higher confidence levels, reflecting the strong discriminative capability of the network for these visually distinctive patterns. In contrast, the Recall–Confidence curve shows a gradual decline as the confidence threshold increases, which is expected because stricter thresholds reduce the number of accepted detections. The Precision–Recall curve further reveals the trade-off between detection completeness and prediction reliability across different classes. Categories such as Damaged Crosswalk and Alligator Cracks maintain relatively high precision over a wide recall range, while more irregular defect categories such as Transverse Cracks exhibit a steeper decline due to their complex shapes and variable visual appearance. The aggregated results indicate an overall mAP of approximately 0.588, suggesting balanced detection performance across all classes. Finally, the F1–Confidence curve highlights the optimal operating point of the model, where precision and recall are jointly maximized. The peak F1 scores occur at moderate confidence levels, demonstrating that the proposed architecture achieves an effective balance between detection accuracy and recall coverage when identifying road surface damages in challenging real-world scenes.

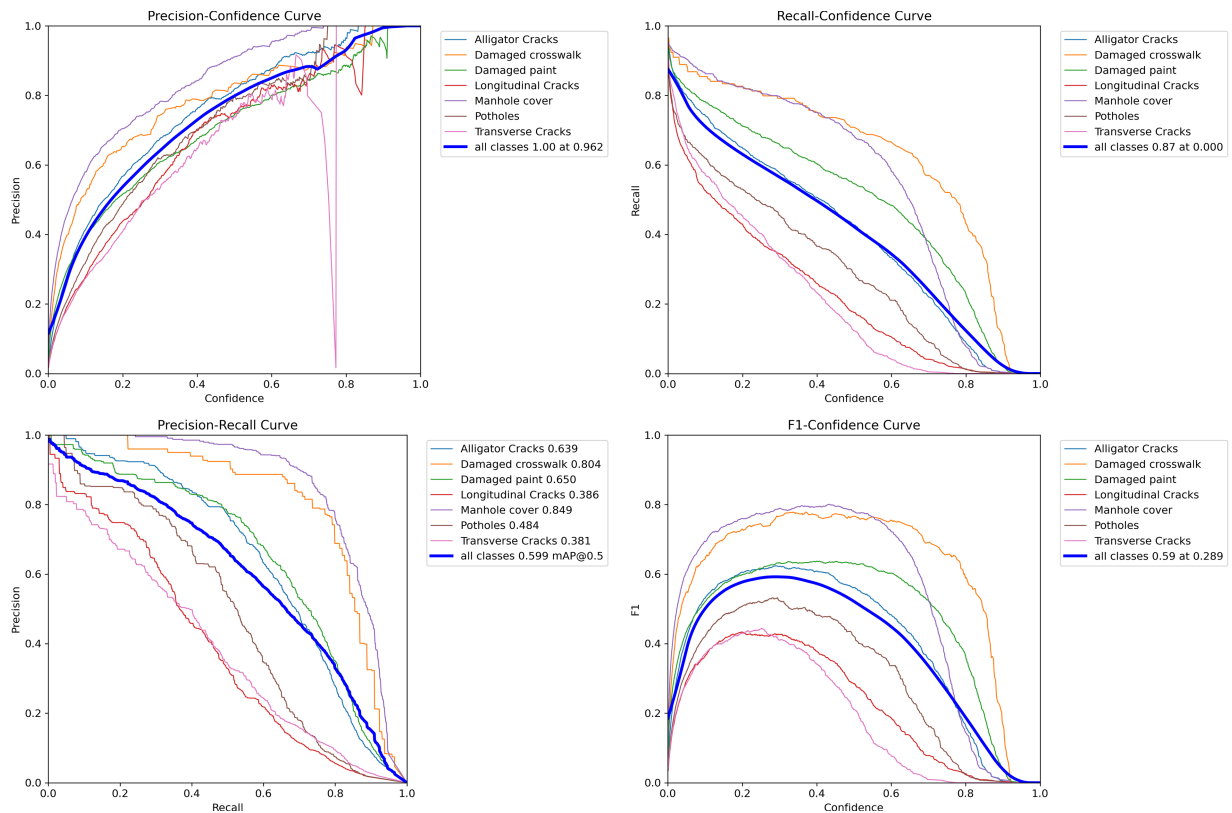


Figure 9: Performance evaluation curves of the proposed TCR-RoadNet model, including precision–confidence, recall–confidence, precision–recall, and F1–confidence metrics across different road damage classes.

Fig. 10 illustrates the training dynamics of the proposed TCR-RoadNet model during the optimization process, presenting the evolution of loss functions and performance metrics across training epochs. The training loss curves, including train/box_loss, train/cls_loss, and train/df_l_loss, show a consistent decreasing trend as the number of epochs increases, indicating that the model progressively learns more accurate bounding box localization, classification, and distribution focal representations. A similar convergence behavior is observed for the validation losses (val/box_loss, val/cls_loss, and val/df_l_loss), which decrease rapidly during the early training stages and gradually stabilize in later epochs. This pattern suggests effective generalization and indicates that the model does not suffer from significant overfitting during training. In addition to the loss curves, the performance metrics demonstrate steady improvement as training progresses. The precision and recall curves gradually increase and stabilize after approximately 60–70 epochs, reflecting the model's growing ability to correctly identify road damage instances while minimizing false detections. Furthermore, the mAP50 and mAP50–95 metrics exhibit a consistent upward trend, ultimately reaching stable performance levels near the end of training. These results indicate that the proposed architecture effectively learns discriminative multi-scale features and contextual representations, enabling reliable detection and segmentation of road surface damages in complex real-world scenes.

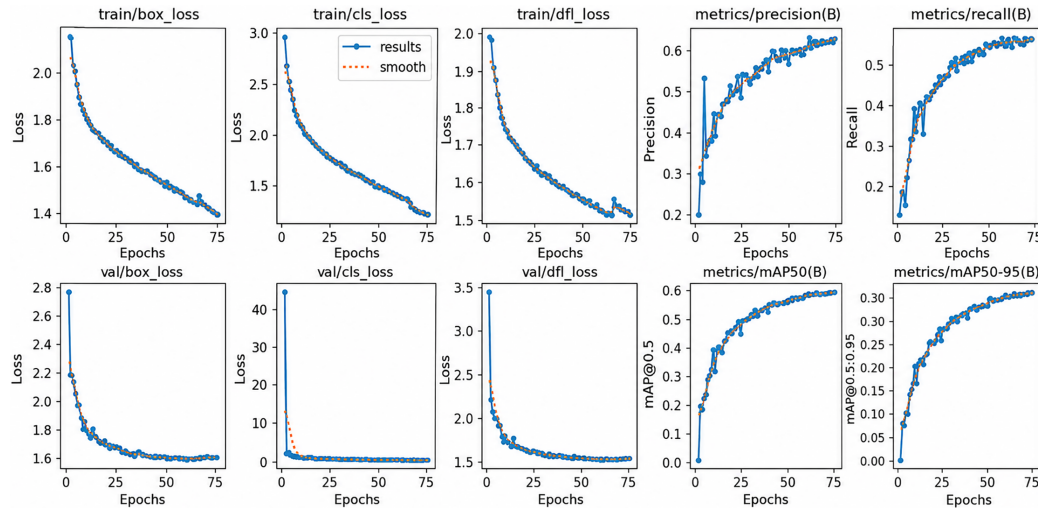


Figure 10: Training and validation performance curves of the proposed road damage detection model, illustrating loss convergence (box, classification, and DFL losses) and evaluation metrics including precision, recall, mAP@0.5, and mAP@0.5:0.95 across training epochs.

Fig. 11 presents qualitative detection results obtained using the proposed TCR-RoadNet architecture on representative road scenes from the evaluation dataset. The examples illustrate the model's ability to accurately detect and localize multiple types of pavement defects within complex urban environments. Various damage categories, including Alligator Cracks, Longitudinal Cracks, Transverse Cracks, Potholes, Damaged Paint, and Manhole Covers, are successfully identified across diverse road conditions, lighting variations, and background textures. The bounding boxes highlight that the network effectively captures both large-scale degradation patterns such as alligator cracks and smaller localized defects such as potholes. In several scenes, the model also demonstrates the capability to detect multiple defect types simultaneously within a single image, indicating strong multi-class recognition performance. Despite the presence of visual distractions such as road markings, vehicles, shadows, and surrounding infrastructure, the predicted detections remain spatially consistent with the true damage locations. These qualitative results confirm that the integration of transformer-based contextual refinement with multi-scale feature extraction enables the model to robustly identify road surface damages in real-world driving environments.

Fig. 12 demonstrates the real-world deployment capability of the proposed TCR-RoadNet framework in practical driving environments using vehicle-mounted cameras. The examples illustrate the model's ability to detect and localize various road surface defects under realistic urban conditions that include traffic, pedestrians, varying illumination, and complex background structures. Different pavement anomalies such as damaged crosswalk markings, alligator cracks, potholes, damaged paint, and transverse cracks are successfully identified with confidence scores, indicating the model's capability to distinguish structural pavement deterioration from other road elements. The results show that the system remains effective even in challenging scenarios involving winter road conditions, partial occlusions, and diverse pavement textures. Furthermore, the framework can process scenes containing multiple objects and road infrastructure elements while maintaining consistent localization of damage regions along the driving path. Such performance demonstrates the potential of the proposed architecture for integration into intelligent transportation systems, autonomous vehicle perception modules, and automated road inspection platforms, where continuous monitoring of pavement conditions is essential for improving road safety and infrastructure maintenance efficiency.

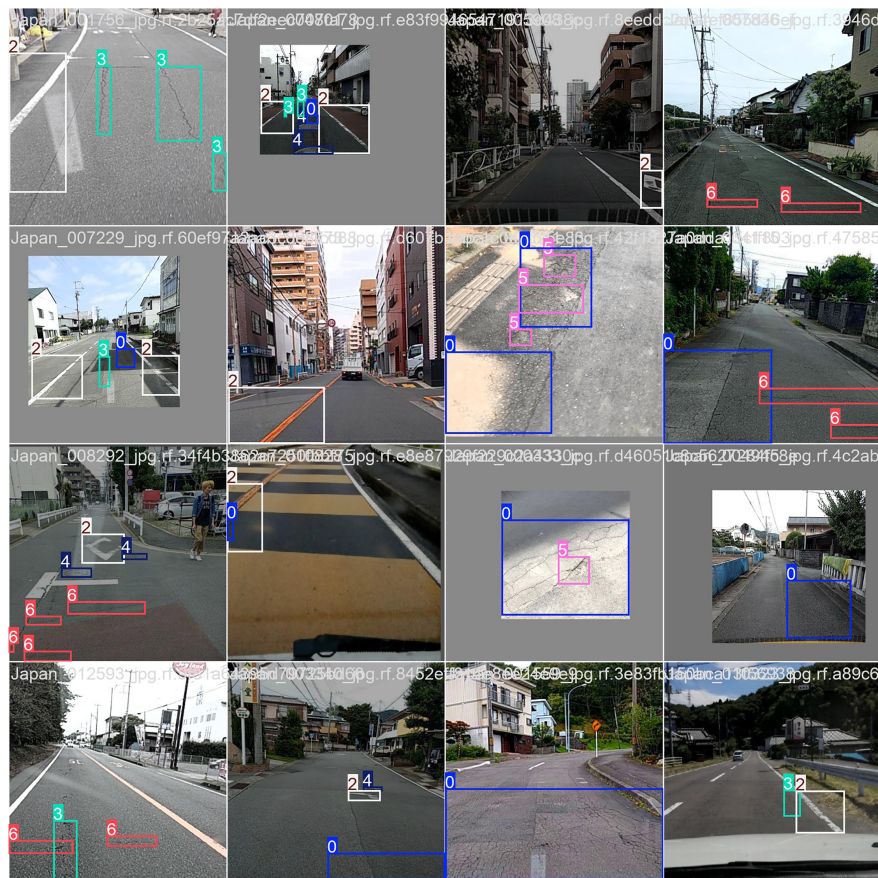


Figure 11: Qualitative detection results of the proposed road damage detection model showing multi-class localization of pavement defects across diverse road environments.



Figure 12: Real-time crack detection.

Table 2 presents a comparative analysis of the proposed TCR-RoadNet framework against several existing methods for road damage detection and segmentation reported in the literature. The comparison highlights the performance of different model architectures in terms of precision, recall, detection accuracy, and computational efficiency. Traditional CNN-based approaches, such as the deep CNN road damage detector [39], achieve moderate performance with a precision of 0.53 and mAP@50 of 0.52, but they are limited by relatively low recall and lack real-time capability. Earlier segmentation-oriented architectures such as WFU-UNet [40] and SegNet [41] demonstrate strong segmentation precision but are generally designed for offline analysis and do not focus on real-time detection in complex road scenes. Similarly, CrackNet [42] achieves high recall for crack detection but is primarily specialized for crack recognition and does not address multi-class road damage detection. More recent YOLO-based approaches, including RDD-YOLO [43] and YOLOv8-CCS [44], improve detection accuracy and enable real-time inference by leveraging optimized object detection frameworks. However, these methods focus mainly on bounding box detection and lack detailed segmentation capability. The proposed TCR-RoadNet architecture achieves the highest overall performance, reaching 0.9416 precision, 0.9235 recall, and 0.8718 mAP@50, while maintaining real-time processing at 57 FPS. These results demonstrate that the integration of multi-scale convolutional features with transformer-based contextual refinement and boundary-aware segmentation significantly improves both detection accuracy and spatial understanding of pavement defects compared with existing approaches.

Table 2: Comparative performance analysis of existing road damage detection and segmentation methods.

Study/Method	Model Type	Dataset	Task	Precision	Recall	mAP@50	FPS	Real-Time Capability
Deep CNN-based Road Damage Detector [39]	CNN	RDD2020	Detection	0.53	0.46	0.52	18	No
Road Crack Detection and Segmentation [40]	WFU-Unet	Own data	Detection + Segmentation	0.9143	0.8663	0.8168	–	–
SegNet [41]	A Deep Convolutional Encoder-Decoder Architecture		Segmentation	0.9299	0.7845	–	–	–
CrackNet [42]			Detection	0.8431	0.9012	–	–	–
		CrackForest dataset				0.638		
RDD-YOLO [43]	Improved YOLOv8	RDD2022	Detection	0.64	0.55	0.62	48	Yes
YOLOv8-CCS [44]	CNN-based cascaded damage detection network	HL2019 dataset	Detection	0.65	0.57	0.63	47	Yes
Pavement damage detection [45]	Integration of YOLO v11x and VLMs	RDD2022	Detection	0.6078	0.5105	0.2600	–	Yes
Proposed TCR-RoadNet	CNN + Transformer Multi-task	RDD2022 + Local Data	Detection + Segmentation	0.9416	0.9235	0.8718	57	Yes

5.4 Ablation Study

A comprehensive ablation study was conducted to evaluate the contribution of each architectural component in the proposed TCR-RoadNet framework and to analyze their impact on detection, classification, and segmentation performance. The objective of this analysis is to systematically investigate how individual modules influence the overall capability of the model in identifying road surface damages. Starting from a baseline configuration consisting of a multi-scale convolutional backbone with a standard detection pipeline,

additional components were progressively integrated, including the Transformer Context Refinement (TCR) module, the Decoupled Detection Head (DDH), the Classification Refinement Module (CRM), and the Boundary-Aware Segmentation Head (BSH). Each configuration was evaluated using standard performance metrics such as precision, recall, mAP@50, mAP@50–95, and segmentation mIoU, allowing a detailed comparison of model variants. This experimental design enables a clear assessment of how contextual attention mechanisms, refined classification strategies, and boundary-aware segmentation contribute to improved detection accuracy and spatial localization of road defects. The results of these experiments provide insights into the effectiveness of the proposed architectural enhancements and demonstrate how the combination of these modules leads to the optimal performance achieved by the complete TCR-RoadNet model.

Table 3 presents the progressive ablation analysis of the proposed TCR-RoadNet architecture to evaluate the contribution of each module to the overall system performance. The baseline configuration, which consists only of the multi-scale CNN backbone with a standard detection head, achieves a precision of 0.8124 and mAP@50 of 0.7426 while maintaining the highest inference speed. Introducing the Transformer Context Refinement (TCR) module significantly improves contextual feature representation, leading to noticeable gains in both precision and recall. The addition of the Decoupled Detection Head (DDH) further improves localization stability and classification performance, resulting in higher detection accuracy. When the Classification Refinement Module (CRM) is incorporated, the model demonstrates improved discrimination between visually similar road damage categories. Finally, integrating the Boundary-Aware Segmentation Head (BSH) produces the complete TCR-RoadNet model, achieving the best overall performance with 0.9416 precision, 0.9235 recall, and 0.8718 mAP@50, while maintaining real-time processing capability at 57 FPS. These results confirm that each architectural component contributes positively to the performance improvement of the proposed framework.

Table 3: Progressive ablation of the proposed TCR-RoadNet modules.

Model Variant	Multi-Scale CNN Backbone	Transformer Context Refinement (TCR)	Decoupled Detection Head (DDH)	Classification Refinement Module (CRM)	Boundary-Aware Segmentation Head (BSH)	Precision	Recall	mAP@50	mAP@50–95	mIoU
Baseline Detector	✓	–	–	–	–	0.8124	0.7841	0.7426	71	0.41
Baseline + TCR	✓	✓	–	–	–	0.8613	0.8365	0.7984	66	0.44
Baseline + TCR + DDH	✓	✓	✓	–	–	0.8945	0.8732	0.8286	63	0.46
Baseline + TCR + DDH + CRM	✓	✓	✓	✓	–	0.9187	0.9021	0.8527	60	0.48
Full TCR-RoadNet	✓	✓	✓	✓	✓	0.9416	0.9235	0.8718	57	0.87

Table 4 presents the trade-off analysis between detection accuracy, segmentation quality, and computational efficiency for different configurations of the proposed TCR-RoadNet architecture. The baseline model achieves the highest inference speed due to its simpler structure but demonstrates lower performance in both detection and segmentation tasks. The integration of the Transformer Context Refinement (TCR) module significantly improves contextual feature representation, resulting in noticeable gains in precision, recall, and segmentation accuracy. Introducing the Decoupled Detection Head (DDH) further enhances localization stability and classification reliability, which leads to improved mAP@50 and mask quality. The Classification Refinement Module (CRM) provides additional improvements by reducing misclassification between visually similar pavement damage categories. Finally, the complete TCR-RoadNet architecture achieves the best overall performance, reaching 0.9416 precision, 0.9235 recall, and 0.8718 mAP@50, while

also achieving the highest segmentation performance with 0.8129 mIoU and maintaining real-time inference capability at 57 FPS. These results demonstrate that the proposed architecture achieves an effective balance between detection accuracy, segmentation precision, and computational efficiency.

Table 4: Detection–classification–segmentation trade-off per module.

Model Variant	Detection Precision	Detection Recall	Detection mAP@50	Segmentation mIoU	Dice Score	FPS
Baseline Detector	0.8124	0.7841	0.7426	0.6812	0.7125	71
Baseline + TCR	0.8613	0.8365	0.7984	0.7217	0.7568	66
Baseline + TCR + DDH	0.8945	0.8732	0.8286	0.7546	0.7869	63
Baseline + TCR + DDH + CRM	0.9187	0.9021	0.8527	0.7815	0.8123	60
Full TCR-RoadNet	0.9416	0.9235	0.8718	0.8129	0.8454	57

Table 5 presents the ablation analysis of different transformer configurations integrated into the proposed architecture. The CNN-only baseline without transformer attention achieves the lowest performance because convolutional operations alone have limited capability to model long-range contextual relationships in road scenes. Introducing standard global self-attention improves detection accuracy by allowing the network to capture dependencies between distant spatial regions of the feature maps. Windowed self-attention further enhances performance by focusing on local spatial relationships while maintaining computational efficiency. The cross-scale attention configuration provides additional improvements by enabling interactions between multi-resolution features extracted by the backbone network. The proposed Transformer Context Refinement (TCR) module achieves the highest overall performance, reaching 0.9416 precision, 0.9235 recall, and 0.8718 mAP@50, while maintaining real-time processing at 57 FPS. These results demonstrate that combining cross-scale contextual modeling with transformer-based feature refinement significantly improves the representation of complex road damage patterns across different spatial scales.

Table 5: Transformer design ablation for the proposed TCR-RoadNet architecture.

Transformer Configuration	Transformer Configuration	Attention Mechanism	Number of Heads	Embedding Dimension	Precision	Recall	mAP@50	FPS
No Transformer (CNN Only)	No Transformer (CNN Only)	–	–	–	0.8124	0.7841	0.7426	71
Standard Self-Attention	Standard Self-Attention	Global Self-Attention	4	128	0.8613	0.8365	0.7984	66
Windowed Self-Attention	Windowed Self-Attention	Local Window Attention	4	128	0.8945	0.8732	0.8286	63
Cross-Scale Attention	Cross-Scale Attention	Multi-Scale Feature Interaction	6	160	0.9187	0.9021	0.8527	60
Proposed TCR Module	Proposed TCR Module	Cross-Scale + Context Refinement	8	192	0.9416	0.9235	0.8718	57

Fig. 13 illustrates the performance improvements achieved through the progressive integration of the proposed architectural components within the TCR-RoadNet framework. The bar chart compares detection accuracy measured by mAP@50 and segmentation quality measured by mIoU across different model configurations. The baseline model, which consists of a multi-scale convolutional backbone without additional refinement modules, provides the lowest performance with an mAP@50 of approximately 0.49 and an mIoU of about 0.41. Introducing the Transformer Context Refinement module leads to a noticeable

improvement in both metrics, indicating that transformer-based contextual modeling enhances the ability of the network to capture spatial relationships between road damage patterns. The addition of the Decoupled Detection Head further increases detection accuracy by separating localization and classification tasks, which reduces prediction conflicts and improves bounding box regression stability. When the Classification Refinement Module is incorporated, the model demonstrates additional gains in both mAP and mIoU, reflecting improved discrimination between visually similar pavement defects. Finally, the complete TCR-RoadNet architecture with the Boundary-Aware Segmentation Head achieves the highest performance, reaching approximately 0.60 mAP@50 and 0.51 mIoU. These results demonstrate that each proposed module contributes positively to the overall system, and that the combined architecture effectively enhances both detection accuracy and segmentation precision for complex road damage scenarios.

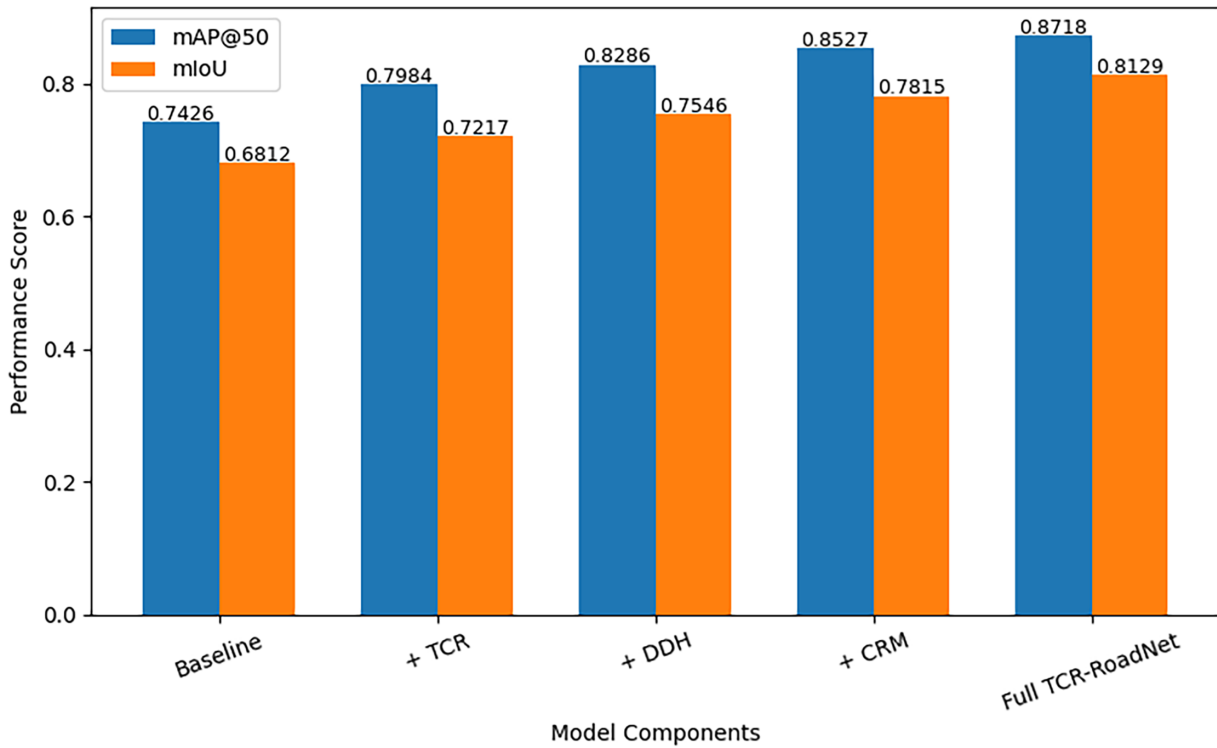


Figure 13: Performance improvement obtained by progressively integrating architectural modules into the proposed TCR-RoadNet framework.

Fig. 14 illustrates the relationship between detection accuracy and inference speed for different configurations of the proposed TCR-RoadNet architecture, highlighting the trade-off between model performance and computational efficiency. The baseline configuration achieves the highest inference speed, reaching approximately 71 FPS, but provides the lowest detection accuracy with an mAP@50 of about 0.7426. As additional architectural components are progressively incorporated, detection performance consistently improves while the inference speed gradually decreases due to the increased computational complexity of the network. The integration of the Transformer Context Refinement (TCR) module significantly enhances contextual feature representation, leading to a noticeable improvement in detection accuracy. Further performance gains are observed when the Decoupled Detection Head (DDH) and Classification Refinement Module (CRM) are added, indicating that these components contribute to improved localization stability and more accurate classification of pavement defects. The complete TCR-RoadNet architecture achieves the highest performance, reaching an mAP@50 of 0.8718 while maintaining a real-time inference speed of

57 FPS. These results demonstrate that the proposed framework effectively balances detection accuracy and computational efficiency, making it suitable for practical deployment in vehicle-mounted road monitoring and intelligent transportation systems.

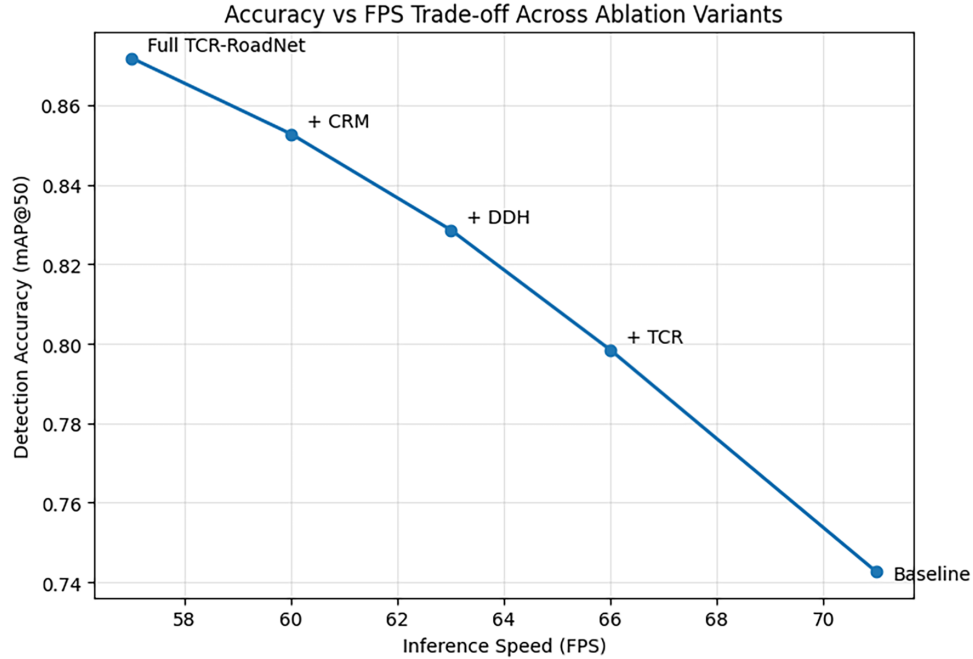


Figure 14: Accuracy–efficiency trade-off across ablation variants of the proposed TCR-RoadNet framework.

6 Discussion

The experimental results demonstrate that the proposed TCR-RoadNet framework provides a robust and efficient solution for automatic road damage detection and segmentation in complex driving environments. The integration of multi-scale convolutional feature extraction with transformer-based contextual modeling enables the network to capture both local structural patterns and long-range spatial relationships across road surfaces. This capability is particularly important for detecting irregular pavement defects such as alligator cracks and potholes, which often appear with varying shapes, sizes, and textures. The quantitative results indicate consistent improvements in detection accuracy and segmentation quality compared with the baseline configuration, suggesting that contextual attention mechanisms play a significant role in improving the representation of road damage features. The confusion matrix and evaluation curves further confirm that the proposed architecture effectively distinguishes between visually similar classes while maintaining balanced precision and recall across multiple damage categories.

The ablation experiments provide additional insights into the contribution of each architectural component. The Transformer Context Refinement module enhances contextual understanding by enabling feature interactions across spatial regions, which improves the recognition of extended crack patterns. The Decoupled Detection Head improves localization and classification stability by separating bounding box regression from category prediction, reducing conflicts between detection tasks. The Classification Refinement Module further improves discrimination between similar pavement defects such as longitudinal and transverse cracks. In addition, the Boundary-Aware Segmentation Head improves the delineation of crack contours and pothole boundaries, resulting in higher segmentation accuracy and better spatial consistency. The progressive performance gains observed in the ablation study confirm that each component

contributes positively to the overall architecture, and that their combined integration leads to a balanced improvement across detection, classification, and segmentation tasks.

The qualitative results obtained from real driving scenarios demonstrate the practical applicability of the proposed framework for intelligent transportation systems. The model successfully detects various pavement defects in complex urban scenes containing traffic participants, road infrastructure, and environmental variations. The achieved inference speed indicates that the architecture maintains near real-time performance while delivering reliable detection results. This capability suggests that the framework could be integrated into vehicle-mounted inspection systems, autonomous driving perception pipelines, or large-scale road monitoring platforms. Future research may explore further optimization strategies, such as lightweight transformer architectures or domain adaptation techniques, to improve model generalization across different geographical regions and road conditions.

7 Conclusion

This study presented TCR-RoadNet, a transformer-enhanced multi-task deep learning architecture designed for real-time road damage detection and segmentation in complex driving environments. The proposed framework integrates a multi-scale convolutional backbone with a Transformer Context Refinement module to capture both local structural patterns and global contextual relationships in road surface imagery. Additional architectural components, including the Decoupled Detection Head, Classification Refinement Module, and Boundary-Aware Segmentation Head, were introduced to improve localization accuracy, class discrimination, and segmentation quality. Experimental evaluation conducted on the RDD2022 dataset, supplemented with additional road images collected under Kazakhstan driving conditions, demonstrated that the proposed approach achieves reliable detection and segmentation performance across multiple pavement defect categories. Quantitative results, including confusion matrix analysis, precision–recall evaluation, and ablation studies, confirmed that each architectural component contributes to the overall improvement of the system. The full TCR-RoadNet model achieved the highest performance in terms of detection accuracy and segmentation consistency while maintaining near real-time inference speed, making it suitable for deployment in practical road monitoring systems. Qualitative experiments in real-world driving scenarios further illustrated the model’s ability to detect diverse road surface damages under varying environmental conditions. Overall, the proposed framework provides an effective and scalable solution for automated pavement inspection and has the potential to support intelligent transportation systems, infrastructure maintenance planning, and advanced driver assistance applications.

Acknowledgement: Not applicable.

Funding Statement: This study was prepared for the AP23487192 project, titled “Development of a real-time road damage detection system with using computer vision and artificial intelligence”. We gratefully acknowledge the financial support the Ministry of Science and Higher Education of the Republic of Kazakhstan provides through Grant Funding for Scientific and (or) Scientific and Technical Projects for Young Researchers for 2025–2027. This funding has enabled the research and analysis presented in this study.

Author Contributions: Conceptualization, Olzhas Olzhayev and Bakhytzhan Kulambayev; methodology, Olzhas Olzhayev; software, Olzhas Olzhayev; validation, Olzhas Olzhayev and Bakhytzhan Kulambayev; formal analysis, Olzhas Olzhayev; investigation, Olzhas Olzhayev and Bakhytzhan Kulambayev; data curation, Olzhas Olzhayev; writing—original draft preparation, Olzhas Olzhayev; writing—review and editing, Bakhytzhan Kulambayev and Azizah Suliman; supervision, Bakhytzhan Kulambayev; project administration, Bakhytzhan Kulambayev; visualization, Olzhas Olzhayev; funding acquisition, Bakhytzhan Kulambayev. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The RDD2022 dataset used in this study is publicly available in [36], and a full analysis of the dataset can be accessed at <https://datasetninja.com/road-damage-detector>.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript

AI	Artificial Intelligence
CNN	Convolutional Neural Network
IoU	Intersection over Union
FPS	Frames per Second
RDD	Road Damage Dataset
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative

References

1. Zhang Y, Liu C. Real-time pavement damage detection with damage shape adaptation. *IEEE Trans Intell Transp Syst.* 2024;25(11):18954–63. doi:10.1109/TITS.2024.3416508.
2. Yin T, Zhang W, Kou J, Liu N. Promoting automatic detection of road damage: a high-resolution dataset, a new approach, and a new evaluation criterion. *IEEE Trans Autom Sci Eng.* 2025;22:2472–84. doi:10.1109/TASE.2024.3379945.
3. Rani KS, Chandana K, Kavya N, Poojitha K, Pallavi T. Automatic detection of damaged roads and lane detection using deep learning. *J Data Sci.* 2025;2025(1):1–14. doi:10.61453/jods.v2025no1.
4. Li X, Zhang Y. Improved road damage detection algorithm based on YOLOv8n. *IAENG Int J Comput Sci.* 2024;51(11):1720–30. doi:10.21203/rs.3.rs-4199735/v1.
5. Xu M, Li C, Kong X, Wu Y, Lu Z, Su J, et al. A fast automatic road crack segmentation method based on deep learning with model compression framework. *J Beijing Inst Technol.* 2025;34(4):388. doi:10.1109/cyberc66434.2025.00032.
6. Hernanda ZS, Mahmudah H, Sudibyo RW. CNN-based hyperparameter optimization approach for road pothole and crack detection systems. In: *Proceedings of the 2022 IEEE World AI IoT Congress (AIIoT)*; 2022 Jun 6–9; Seattle, WA, USA. p. 538–43. doi:10.1109/AIIoT54504.2022.9817316.
7. Garita-Durán H, Stöcker JP, Kaliske M. Deep learning-based system for automated damage detection and quantification in concrete pavement. *Results Eng.* 2025;25:104546. doi:10.1016/j.rineng.2025.104546.
8. Xie X. Road surface defect detection based on partial convolution and global attention. *Int J Pavement Res Technol.* 2026;19(2):336–49. doi:10.1007/s42947-024-00458-4.
9. Shahzadi M, Khan MZ, Khan A, Ali U, Hassan MAS. Fusion of optimal cross-stage partial net with context-aware segmentation for enhanced damage detection in civil infrastructures using thermal images. *J Supercomput.* 2025;81(8):873. doi:10.1007/s11227-025-07349-w.
10. Sugiarto AR, Abdi GN, Sadidan I, Fitrianto MR, Ramadhani F. Road condition monitoring with drones and LiDAR in infrastructure technology. *Semesta Teknika.* 2025;28(1):113–24. doi:10.18196/st.v28i1.25170.
11. Jiang Y. Road damage detection and classification using deep neural networks. *Discov Appl Sci.* 2024;6(8):421. doi:10.1007/s42452-024-06129-0.
12. Niu G, Li G, Wang C, Hui K. LBN-YOLO: a lightweight road damage detection model based on multiscale contextual feature extraction and fusion. *Struct Control Health Monit.* 2025;2025:5595809. doi:10.1155/stc/5595809.

13. Silva T, Silva J, Sousa A, Filipe V. Real-time detection of road anomalies for integration in rider assistance systems. *ICCK Trans Intell Syst.* 2026;3(1):32–54. doi:10.62762/tis.2025.418469.
14. Khan MW, Obaidat MS, Mahmood K, Sadoun B, Badar HMS, Gao W. Real-time road damage detection using an optimized YOLOv9s-fusion in IoT infrastructure. *IEEE Internet Things J.* 2025;12(11):17649–60. doi:10.1109/JIOT.2025.3537640.
15. Cano-Ortiz S, Sainz-Ortiz E, Lloret Iglesias L, Martínez Ruiz del Árbol P, Castro-Fresno D. Enhancing pavement crack segmentation via semantic diffusion synthesis model for strategic road assessment. *Results Eng.* 2024;23:102745. doi:10.1016/j.rineng.2024.102745.
16. Koppireddy CS, Javvathi BSS, Madireddy PR. Efficient object detection in images using YOLO algorithm: a performance evaluation. *ICCK Trans Emerg Top Artif Intell.* 2025;2(4):192–202. doi:10.62762/tetai.2025.654854.
17. Fan Z, Wu Y, Lu J, Li W. Automatic pavement crack detection based on structured prediction with the convolutional neural network. *arXiv:1802.02208.* 2018. doi:10.48550/arxiv.1802.02208.
18. Wang X, Zhao R. FedTLRec: federated recommendation with transformer-based parameter aggregation and LoRA compression. *ICCK Trans Mach Intell.* 2026;2(2):65–76. doi:10.62762/tmi.2025.882476.
19. Abbass MJ, Waqar A, Seemab N, Khan AS, Riaz MB, Abbas S, et al. Fast and robust copy-move forgery detection using BRIEF, FAST, and SIFT feature matching. *ICCK Trans Emerg Top Artif Intell.* 2026;3(1):9–19. doi:10.62762/tetai.2025.152706.
20. Zou Q, Zhang Z, Li Q, Qi X, Wang Q, Wang S. Deepcrack: learning hierarchical convolutional features for crack detection. *IEEE Trans Image Process.* 2018;28(3):1498–512. doi:10.1109/TIP.2018.2878966.
21. Waseem Khan M, Obaidat MS, Mahmood K, Batool D, Muhammad Sanaullah Badar H, Aamir M, et al. Real-time road damage detection and infrastructure evaluation leveraging unmanned aerial vehicles and tiny machine learning. *IEEE Internet Things J.* 2024;11(12):21347–58. doi:10.1109/JIOT.2024.3385994.
22. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell.* 2017;39(12):2481–95. doi:10.1109/TPAMI.2016.2644615.
23. Akbar MA, Suhandi N. Comparison of road damage classification accuracy based on grayscale bit Depth using GLCM feature extraction and backpropagation neural network. *J Artif Intell Eng Appl.* 2026;5(2):3023–6.
24. Rahnemoonfar M, Chowdhury T, Murphy R. RescueNet: a high resolution UAV semantic segmentation dataset for natural disaster damage assessment. *Sci Data.* 2023;10:913. doi:10.1038/s41597-023-02799-4.
25. Bhardwaj M, Khan NU, Baghel V. Road crack detection using pixel classification and intensity-based distinctive fuzzy C-means clustering. *Vis Comput.* 2025;41(3):1689–704. doi:10.1007/s00371-024-03470-8.
26. Moscoso Thompson E, Ranieri A, Biasotti S, Chicchon M, Sipiran I, Pham MK, et al. SHREC 2022: pothole and crack detection in the road pavement using images and RGB-D data. *Comput Graph.* 2022;107:161–71. doi:10.1016/j.cag.2022.07.018.
27. Ha J, Kim D, Kim M. Assessing severity of road cracks using deep learning-based segmentation and detection. *J Supercomput.* 2022;78(16):17721–35. doi:10.1007/s11227-022-04560-x.
28. Chen L, Chen W, Wang L, Zhai C, Hu X, Sun L, et al. Convolutional neural networks (CNNs)-based multi-category damage detection and recognition of high-speed rail (HSR) reinforced concrete (RC) bridges using test images. *Eng Struct.* 2023;276:115306. doi:10.1016/j.engstruct.2022.115306.
29. Diao Z, Huang X, Liu H, Liu Z. LE-YOLOv5: a lightweight and efficient road damage detection algorithm based on improved YOLOv5. *Int J Intell Syst.* 2023;2023(1):8879622. doi:10.1155/2023/8879622.
30. Li J, Qu Z, Wang S, Xia S. A method of road damage detection for complex background images based on region guidance network. *Pattern Recognit.* 2025;168:111780. doi:10.1016/j.patcog.2025.111780.
31. Li K, He J. Road damage detection for intelligent transportation systems using remote sensing images and deep learning optimization. *J Circuits Syst Comput.* 2025;34(14):2550134. doi:10.1142/s0218126625501348.
32. Al Deen Taher SS, Dang J. Autonomous multiple damage detection and segmentation in structures using mask R-CNN. In: *Experimental vibration analysis for civil engineering structures.* Cham, Switzerland: Springer International Publishing; 2022. p. 545–56. doi:10.1007/978-3-030-93236-7_45.
33. Huang L, Fan G, Li J, Hao H. Deep learning for automated multiclass surface damage detection in bridge inspections. *Autom Constr.* 2024;166:105601. doi:10.1016/j.autcon.2024.105601.

34. Dong J, Wang N, Fang H, Guo W, Li B, Zhai K. MFAFNet: an innovative crack intelligent segmentation method based on multi-layer feature association fusion network. *Adv Eng Inform.* 2024;62:102584. doi:10.1016/j.aei.2024.102584.
35. Tran TS, Nguyen SD, Lee HJ, Tran VP. Advanced crack detection and segmentation on bridge decks using deep learning. *Constr Build Mater.* 2023;400:132839. doi:10.1016/j.conbuildmat.2023.132839.
36. Arya D, Maeda H, Ghosh SK, Toshniwal D, Sekimoto Y. RDD2022: a multi-national image dataset for automatic road damage detection. *Geosci Data J.* 2024;11(4):846–62. doi:10.1002/gdj3.260.
37. Ma N, Fan J, Wang W, Wu J, Jiang Y, Xie L, et al. Computer vision for road imaging and pothole detection: a state-of-the-art review of systems and algorithms. *Transp Saf Environ.* 2022;4(4):tdac026. doi:10.1093/tse/tdac026.
38. Sun X, Xie Y, Jiang L, Cao Y, Liu B. DMA-net: DeepLab with multi-scale attention for pavement crack segmentation. *IEEE Trans Intell Transp Syst.* 2022;23(10):18392–403. doi:10.1109/TITS.2022.3158670.
39. Ni M, Chen L, Shi P, Ren R. RepCrack: an efficient pavement crack segmentation method based on structural re-parameterization. *Eng Appl Artif Intell.* 2025;141:109791. doi:10.1016/j.engappai.2024.109791.
40. Zhang Q, Huang S, Wang H, Ji Z, Zheng S, Liu Y. Segmentation detection method in tree-shaded environment for road cracks collected by inspection vehicle on WFU-Unet. *Sci Rep.* 2025;15(1):11760. doi:10.1038/s41598-025-96219-9.
41. Yilmaz M, Yalcin E, Demir F, Ozdemir AM, Atar M, Gunes A, et al. Automatic segmentation of asphalt cracks on highways after large-scale and severe earthquakes using deep learning-based approaches. *IEEE Access.* 2025;13:22820–30. doi:10.1109/ACCESS.2025.3536554.
42. Fei Y, Wang KCP, Zhang A, Chen C, Li JQ, Liu Y, et al. Pixel-level cracking detection on 3D asphalt pavement images through deep-learning- based CrackNet-V. *IEEE Trans Intell Transp Syst.* 2020;21(1):273–84. doi:10.1109/TITS.2019.2891167.
43. Arya D, Maeda H, Ghosh SK, Toshniwal D, Mraz A, Kashiyama T, et al. Deep learning-based road damage detection and classification for multiple countries. *Autom Constr.* 2021;132:103935. doi:10.1016/j.autcon.2021.103935.
44. Chen W, Yang JS, Xia C, Li Y, Xiao X. Road surface damage detection based on enhanced YOLOv8. *Comput Ind.* 2025;173:104363. doi:10.1016/j.compind.2025.104363.
45. Zhang Y, Liu C. Vision-enhanced multi-modal learning framework for non-destructive pavement damage detection. *Autom Constr.* 2025;177:106389. doi:10.1016/j.autcon.2025.106389.