Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 6, No. 2, 2026

An Integrated Deep Learning Framework for Road Distress Detection, Segmentation, and Quantitative Evaluation

Meng Xu [1, 2], Wei Gao [1, *]

[1] Department of Computer Science, North China Electric Power University (Baoding), Baoding, China
[2] Industrial and Commercial Bank of China, Hebei Branch, Shijiazhuang, China

Abstract: Road distress inspection plays a critical role in pavement-condition assessment and maintenance planning. However, existing studies often address detection, segmentation, or measurement separately, which limits their practical applicability. This paper proposes an integrated deep learning–based framework for road distress detection, pixel-level segmentation, and quantitative evaluation. First, an improved YOLOv7 detector is developed by introducing SE attention, CARAFE-based content-aware upsampling, and a Dynamic Head to enhance multi-scale feature representation and robustness under complex road backgrounds. Second, a multi-scale encoder–decoder network termed MIResU-Net is designed to accurately extract crack and pothole regions with improved structural continuity and boundary precision. Finally, a calibration-based measurement strategy is employed to convert segmentation results into physically meaningful geometric parameters, such as crack length and pothole area. A real-world road-distress dataset collected by a vehicle-mounted system is constructed for comprehensive evaluation. Experimental results demonstrate that the proposed framework achieves superior detection and segmentation performance compared with mainstream methods and provides metrically reliable quantitative indicators for pavement-condition assessment. The proposed approach offers an effective and practical solution for intelligent road inspection.

Keywords: Road distress detection, Deep learning, YOLOv7, Image segmentation, Quantitative evaluation.

1. Introduction

Road distresses such as cracks and potholes are common defects that gradually develop during long-term road service [1]. If they are not identified and repaired in time, minor surface damage may evolve into serious structural problems, leading to higher maintenance costs and potential safety risks. Therefore, efficient and reliable road distress inspection is an essential task in modern road maintenance.

Conventional road inspection mainly relies on manual surveys or specialized inspection vehicles. Manual inspection is time-consuming and strongly dependent on human experience, which makes it difficult to ensure consistency in large-scale applications. Vehicle-mounted inspection systems can provide accurate measurements, but their high cost and limited flexibility restrict widespread deployment. As an alternative, image-based inspection using ordinary cameras has attracted interest due to its low cost and ease of deployment. However, early image processing methods based on handcrafted features are highly sensitive to illumination changes, shadows, and surface noise, and their performance degrades significantly in real road environments [2].

Recent advances in deep learning have improved the performance of vision-based road distress detection [3]. Convolutional neural networks (CNNs) have been applied to both detection and segmentation tasks, enabling automatic identification of different distress types. Detection models based on the YOLO series [4] offer high efficiency, while segmentation networks such as U-Net [5] can extract distress regions at the pixel level. Despite these improvements, practical challenges remain. Road images collected in real scenes often contain complex backgrounds and multi-scale distresses, which makes small or irregular defects difficult to detect reliably. In addition, many existing studies focus on either detection or segmentation alone, and only limited attention has been paid to the quantitative evaluation of road distresses, which is important for maintenance planning [6].

In this work, an integrated framework for road distress detection, segmentation, and quantitative evaluation is presented. An improved YOLOv7-based model is employed for multi-class road distress detection, with attention and feature aggregation mechanisms introduced to improve robustness under complex background conditions. To obtain fine-grained distress regions, a MIResU-Net–based segmentation network is further designed for cracks and potholes. Based on the segmentation results, geometric parameters such as crack length and pothole area are calculated to support objective assessment of road damage. The main contributions of this study are as follows:

(1) An integrated framework that jointly performs road-distress detection, segmentation, and quantitative evaluation is proposed for intelligent pavement inspection.

(2) An improved YOLOv7 and a multi-scale MIResU-Net are developed to enhance multi-scale feature representation and structurally consistent distress extraction.

(3) A calibration-based measurement method is introduced to provide physically meaningful pavement-condition indicators from visual data.

2. Related work

2.1. Deep Learning-Based Road Distress Detection

With the rapid development of deep learning, data-driven approaches have significantly advanced road-distress detection compared with traditional image-processing methods, owing to their superior feature-learning capability and robustness to complex backgrounds.

Early deep learning–based studies mainly relied on two-stage detectors. The R-CNN framework proposed by Girshick et al. [7] employed selective search for region proposal generation and used a support vector machine (SVM) for classification, together with non-maximum suppression and bounding-box regression. Although it substantially improved detection accuracy, its computational cost was high due to the large number of candidate regions. To improve performance for pavement inspection, Wang et al. [8] developed a Faster R-CNN–based distress detection method and investigated different backbone networks, where ResNet-152 achieved an accuracy of 62.55% after data augmentation and parameter optimization. Chen et al. [9] further introduced a Mask R-CNN–based model with DenseNet as the backbone to simultaneously perform detection and pixel-level segmentation. However, the insufficient feature representation for distress patterns and the imbalanced data distribution still limited its performance.

To meet real-time inspection requirements, one-stage regression-based detectors have been widely adopted. Yang et al. [10] proposed an improved YOLOv3 for bridge crack detection, in which adaptive illumination correction, k-means++ anchor clustering, and the generalized IoU loss were introduced to improve robustness and localization accuracy, achieving an average precision of 94.88%. Cao et al. [11] conducted a comprehensive comparison of eight deep learning models and showed that SSD provided a favorable balance between accuracy and speed. Luo et al. [12] designed an improved YOLOv4 using depthwise separable convolutions, Focal loss, and transfer learning to address multi-scale distress detection with limited samples. Feng et al. [13] combined SSD and U-Net to jointly perform crack classification and segmentation and further derived geometric parameters from segmentation outputs. Rehana et al. [14] adopted YOLOv5 for fast detection of seven distress categories, demonstrating the effectiveness of lightweight one-stage detectors for pavement inspection.

To enhance feature representation under complex road conditions, recent studies have incorporated attention mechanisms and advanced feature-fusion strategies. Zhang et al. [15] introduced a multi-level attention block into YOLOv3 to strengthen feature extraction and significantly improve the detection accuracy of alligator cracks and diagonal cracks. Wang et al. [16] integrated SE and coordinate attention modules into YOLOv5 and employed ensemble learning with test-time augmentation to improve generalization. Wang et al. [17] adopted Shuffle Attention and CARAFE upsampling to enhance crack localization under varying illumination. Ma [18] replaced the conventional detection head with a dynamic head to improve scale, spatial, and task-aware feature modeling. Yang et al. [19] incorporated pyramid squeeze attention and large-field contextual feature integration to capture multi-scale semantic information.

Despite these advances, most existing methods focus primarily on detection and are limited in handling multiple distress types with large scale variations and complex morphology. Moreover, the quantitative assessment of pavement damage is rarely considered in an integrated framework.

2.2. Road Distress Segmentation and Quantitative Analysis

Pixel-level segmentation provides more precise structural information than bounding-box detection and is therefore essential for accurate distress measurement. U-Net and its variants have been widely used in crack and pothole segmentation due to their encoder–decoder architecture and skip connections, which enable the fusion of low-level spatial details and high-level semantic information. However, the original U-Net was designed for medical images with relatively simple backgrounds and often suffers from blurred boundaries and missed thin structures when directly applied to road scenes with complex textures and noise.

To improve segmentation performance, several studies have explored multi-scale feature extraction and enhanced feature-fusion strategies. The fusion of detection and segmentation networks has also been investigated. Feng et al. [13] combined SSD and U-Net to achieve crack classification and precise segmentation and further computed crack length, width, and area based on the segmentation results. This demonstrates the potential of integrating visual recognition with geometric measurement for pavement-condition assessment.

For quantitative analysis, most existing approaches rely on converting pixel-level segmentation results into physical measurements using calibration-based scale transformation. Crack width is typically estimated using geometric analysis of the segmented contour, while crack length and pothole area are computed from pixel statistics. However, such quantitative evaluation is often treated as an independent post-processing step rather than being incorporated into a unified learning framework.

In summary, although deep learning-based detection and segmentation methods have achieved promising results, an integrated system that simultaneously performs robust multi-scale detection, accurate pixel-level segmentation, and physically meaningful distress quantification is still lacking. This limitation motivates the development of the unified framework proposed in this study.

3. Dataset and Problem Formulation

3.1. Dataset Description

To train and evaluate the proposed framework under real road conditions, a dedicated road-damage dataset was constructed using a vehicle-mounted image acquisition system. Existing public datasets usually contain limited distress categories and relatively simple backgrounds, which cannot adequately reflect the complex disturbances in practical road environments. Therefore, large-scale road images were collected on urban roads in Baoding, China, and a multi-class dataset with diverse noise interference was established.

3.1.1. Image Acquisition System

The road images were acquired using a vehicle-mounted inspection platform consisting of an industrial area-scan camera and a detection vehicle. The camera was mounted on a stabilized gimbal to ensure image quality during motion, while the vehicle maintained an approximately constant speed throughout data collection.

A Hikvision industrial camera (MV-CH120-10UC) equipped with a Sony IMX304 sensor was used for image acquisition, providing a resolution of 4096 × 3000 at 23.1 fps through a USB3 Vision interface. A 16-mm fixed-focus lens (MVL-KF1628M-12MPE) was adopted. During acquisition, the effective image resolution was set to 4000 × 2000.

The collected images contain complex real-world disturbances, including uneven illumination, shadows, water

stains, and motion-induced variations. All image data were stored and processed on an NVIDIA Jetson AGX Xavier edge-computing platform. The vehicle-mounted image acquisition system is shown in Fig. 1.

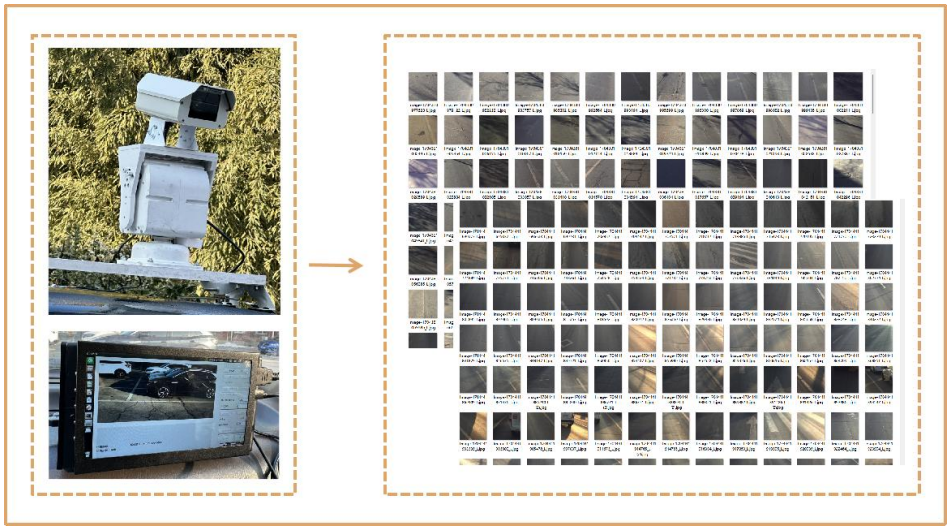

Figure 1. Vehicle-Mounted Image Acquisition System

3.1.2. Data Augmentation

To improve the robustness and generalization ability of the detection and segmentation models, data augmentation was applied to the collected images, as shown in Fig. 2. Specifically, random brightness adjustment and horizontal flipping were performed to simulate illumination variation and viewpoint changes. These operations effectively increased the diversity of the dataset and alleviated the problem of insufficient samples for certain distress categories.

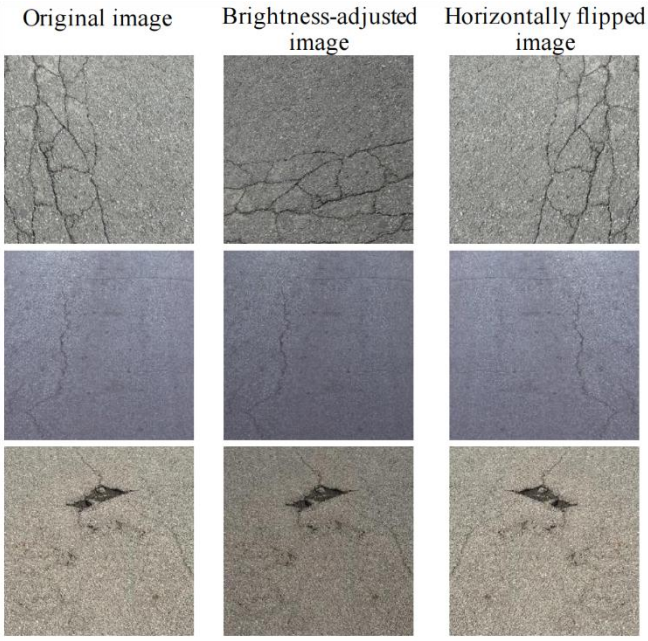

Figure 2. Data Augmentation

3.1.3. Image Preprocessing and Annotation

Since the proposed method is based on supervised learning, accurate manual annotation is required. The original images with a resolution of 4000 × 2000 were first cropped and resized to 640 × 640 to meet the network input requirement and reduce computational cost.

For the detection task, road-damage instances were annotated using LabelImg [20] with rectangular bounding boxes and corresponding category labels. The annotations were stored in Pascal VOC XML format, including image filename, image size, object category, and bounding-box coordinates.

For the segmentation task, pixel-level annotations were further provided for cracks and potholes to enable fine-grained distress extraction and subsequent quantitative analysis.

3.1.4. Damage Categories and Dataset Composition

The constructed dataset contains five types of road damage and one repair category. The distress types include longitudinal crack (D00), transverse crack (D10), alligator crack (D20), pothole (D40), and block crack (D50). Due to the limited number of pothole and block-crack samples in the self-collected data, additional images from the RDD2022 China subset were incorporated [21]. The final dataset consists of 4907 images, including 3219 images collected by the vehicle-mounted system and 1688 images from RDD2022. The distribution of each distress category is shown in Fig. 3. The dataset was divided into a training set, validation set, and test set with a ratio of 7: 2: 1, corresponding to 3435, 981, and 491 images, respectively.

This dataset provides multi-class annotations for detection and pixel-level labels for segmentation, enabling unified evaluation of classification, localization, and quantitative measurement of road distresses in complex real-world scenarios.

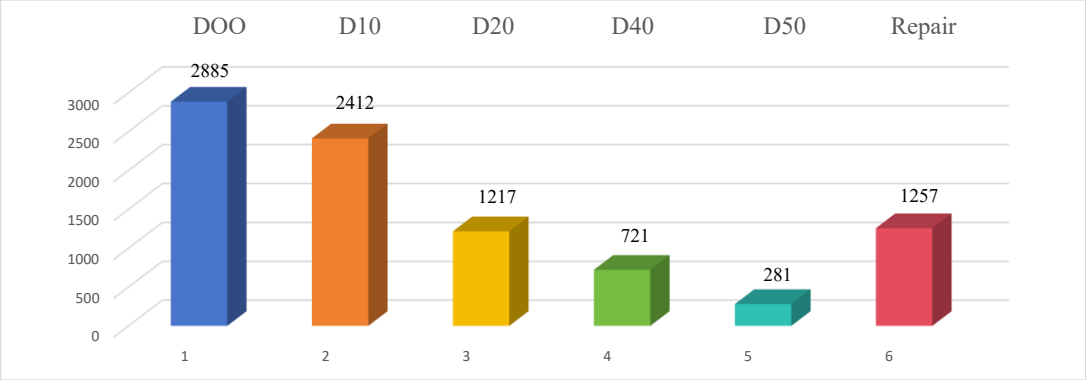

Figure 3. Distribution of distress categories

3.2. Problem Formulation

Based on the constructed dataset, the proposed framework jointly addresses road-distress detection, pixel-level segmentation, and quantitative evaluation. The detection task is formulated as a multi-class object detection problem that localizes each distress instance and predicts its category using bounding boxes and class labels. The segmentation task further refines distress regions by generating binary masks for cracks and potholes to obtain accurate geometric morphology. On this basis, quantitative assessment is performed by computing geometric metrics, such as crack length and pothole area, to measure distress severity. This unified formulation establishes a direct mapping from visual recognition to engineering-oriented condition evaluation for road-maintenance applications.

4. Proposed Framework

4.1. Improved YOLOv7 for Road Distress Detection

To address the challenges of scale variation, background interference, and irregular morphology in road-distress detection, an enhanced YOLOv7-based framework is developed.

4.1.1. Baseline YOLOv7 Architecture

YOLOv7 (2022) [22] achieves a favorable balance between detection accuracy and real-time performance, reaching 56.8% AP on the MS COCO dataset at 30 FPS on an NVIDIA V100 GPU. Owing to its high efficiency and strong feature representation capability, it is adopted in this study as the baseline detector for road-distress detection and further optimized for the target task.

The YOLOv7 architecture follows a three-stage design consisting of a backbone, a neck, and a detection head, as illustrated in Fig. 4.

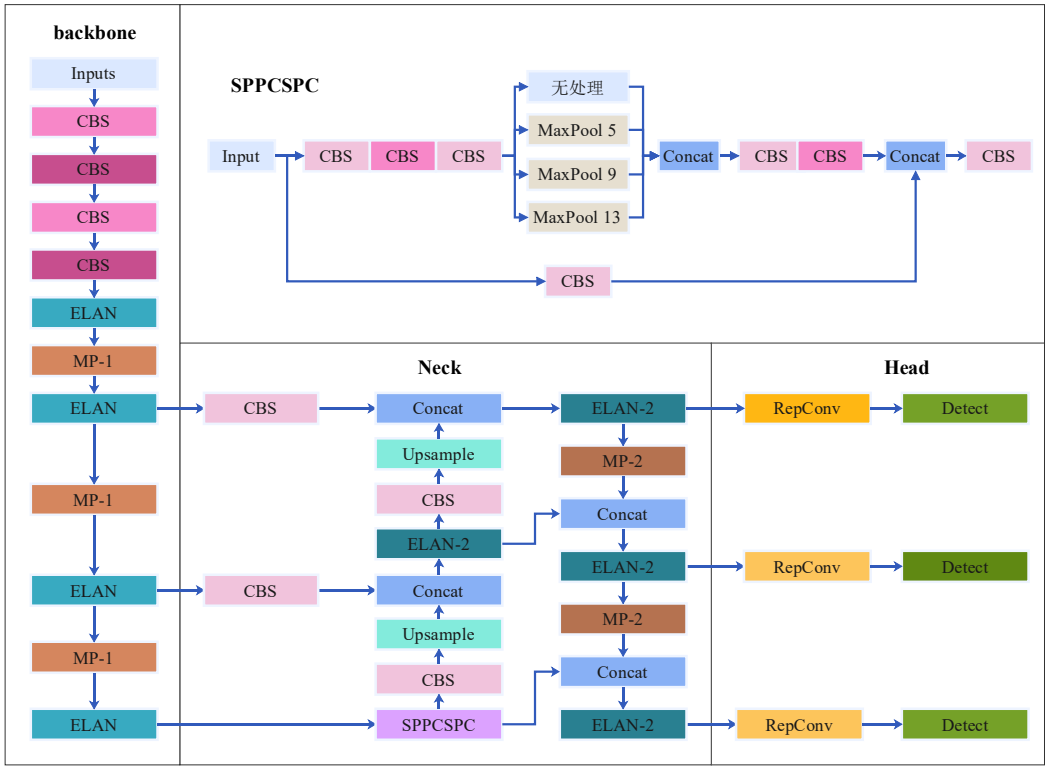

Figure 4. Architecture of YOLOv7

The backbone is responsible for hierarchical feature extraction and is mainly composed of CBS (Conv–BatchNorm–SiLU) layers, ELAN (Efficient Layer Aggregation Network) modules, and MPConv (Multi-Path Convolution) downsampling blocks. The ELAN module adopts a multi-branch aggregation strategy to enhance

gradient flow and improve learning efficiency. By expanding network depth while maintaining computational stability, it enables the extraction of multi-scale and high-level semantic features through feature transformation and channel-wise concatenation. The MPConv module performs efficient spatial downsampling via parallel paths that combine max-pooling and convolutional operations, which reduces feature-map resolution while preserving representative information and enlarging the receptive field.

The neck integrates a Path Aggregation Feature Pyramid Network (PAFPN) with the SPPCSPC module to achieve multi-scale feature fusion. The SPPCSPC block introduces multiple parallel max-pooling operations to capture richer contextual information and enlarge the receptive field while suppressing feature redundancy. The PAFPN structure enhances bidirectional information flow through repeated upsampling and downsampling, generating feature maps at three different resolutions for detecting objects of various scales. Compared with the backbone ELAN, the ELAN-2 structure in the neck employs denser branch connections to further strengthen feature aggregation.

The YOLO head performs classification and bounding-box regression on the fused multi-scale feature maps. It adopts an anchor-based prediction mechanism and outputs objectness score, category probability, and bounding-box offsets. Following the decoupled-head design, the prediction is implemented through 1×1 convolutions, enabling efficient joint optimization of localization and classification tasks.

Through the above architecture, YOLOv7 provides a strong baseline for accurate and real-time road-distress detection. This baseline architecture serves as the foundation for the subsequent improvements proposed in this study.

4.1.2. Feature Enhancement with SE Attention

Road images usually contain large amounts of background interference, such as lane markings, shadows, water stains, and surrounding objects, which may degrade the effectiveness of multi-scale feature aggregation. To address this issue, a squeeze-and-excitation (SE) [23] attention mechanism is introduced to enhance the discriminative capability of fused features.

In the proposed architecture, SE modules are inserted at three feature-fusion nodes in the neck, i.e., two in the top–down pathway before the Concat operations and one on the SPPCSPC branch prior to feature aggregation. By performing channel-wise recalibration immediately before multi-scale feature fusion, the network adaptively emphasizes distress-related responses while suppressing redundant background information from different levels.

Given an input feature map $U \in \mathbb{R}^{H\times W\times C}$, global spatial information is first aggregated via global average pooling to obtain a channel descriptor. Two fully connected layers are then employed to model inter-channel dependencies and generate channel attention weights, which are applied to the original feature map through channel-wise multiplication to produce a refined representation. Through this process, channels corresponding to distress regions are strengthened, whereas responses caused by pavement textures and environmental noise are effectively reduced.

Placing SE at the fusion nodes brings two task-oriented advantages for road-distress detection. First, it improves the consistency of multi-scale feature aggregation by filtering out irrelevant channel responses before feature concatenation. Second, it enhances the representation of fine and low-contrast distresses, such as thin cracks, whose features are easily overwhelmed by complex road backgrounds.

Since the SE module operates only along the channel dimension, it introduces negligible computational overhead. Therefore, the proposed fusion-aware channel attention improves feature discrimination without affecting the real-time performance of YOLOv7, making it suitable for vehicle-mounted inspection scenarios.

This design is complementary to the CARAFE-based content-aware upsampling and the Dynamic Head for adaptive prediction, forming a progressive feature-refinement pipeline for robust road-distress detection.

4.1.3. Multi-Scale Feature Aggregation with CARAFE

Road distresses exhibit significant scale variation and strong dependence on fine-grained texture and structural continuity. In conventional feature pyramid networks, nearest-neighbor or bilinear interpolation is commonly used for upsampling in the top–down pathway. However, these fixed and content-agnostic operations only exploit local sub-pixel information and are unable to adapt to the spatially heterogeneous characteristics of distress regions, leading to the loss of crack continuity and blurred boundary responses. Although transposed convolution is learnable, its location-invariant kernels limit its ability to model content-aware feature reconstruction and may introduce checkerboard artifacts. These limitations degrade the effectiveness of multi-scale feature fusion for irregular and small-scale road distresses.

To enable content-adaptive feature aggregation, the CARAFE operator [24] is introduced into the neck to replace the conventional upsampling module in the feature pyramid. Different from interpolation-based methods, CARAFE dynamically generates reassembly kernels according to the local semantic context of the input feature map, allowing large receptive-field information to be propagated to high-resolution features with negligible computational overhead.

Given an input feature map $\chi \in \mathbb{R}^{W\times H\times C}$ and an upsampling factor σ, CARAFE first predicts position-specific kernels and then performs content-aware feature reassembly to obtain $\chi' \in \mathbb{R}^{\sigma W\times \sigma H\times C}$. For each target location $l' = (i', j')$, the reassembly kernel is generated from the corresponding neighborhood centered at $l = (i'/\sigma, j'/\sigma)$:

$$W_{\ell'} = \psi\left(N(X_\ell, k_{\text{encoder}})\right) \tag{1}$$

Where ψ denotes the kernel prediction function and $N(X_\ell, k_{\text{encoder}})$ represents the local region used to encode contextual information. The predicted kernels are normalized by softmax and used to perform weighted aggregation:

$$\mathcal{X}'_{\ell'} = \phi\left(N(\mathcal{X}_\ell, k_{up}), \mathcal{W}_{\ell'}\right) = \sum_{n=-r}^{r}\sum_{m=-r}^{r} cW_{\ell'}(n,m) \times \chi_{(i+n,j+m)} \tag{2}$$

Where $r = \lfloor k_{up}/2 \rfloor$.

By embedding CARAFE into the top–down pathway of the feature pyramid, the proposed network achieves content-aware multi-scale feature fusion. This design brings three task-oriented benefits for road-distress detection:

(1) Enhanced structural continuity perception: Context-aware reassembly preserves the connectivity of crack pattems and reduces feature fragmentation.

(2) Improved small-distress sensitivity: High-level semantic information is adaptively propagated to shallow high-resolution features, strengthening the representation of fine-scale defects.

(3) Shape-adaptive feature aggregation: Spatially variant

kernels enable better modeling of irregular distress regions with diverse morphologies.

Consequently, the fused multi-scale features contain richer texture and boundary information, which significantly improves detection performance for road distresses with large scale variation and complex geometric structures.

4.1.4. Dynamic Head for Robust Multi-Scale Detection

Road distresses exhibit large scale variation and irregular spatial distribution, where small cracks and large-area defects often appear simultaneously. In the original YOLOv7 head, predictions are performed independently on each feature level, which limits cross-scale interaction and weakens contextual modeling. Such a scale-isolated strategy is insufficient for road-distress detection that relies on the joint representation of multi-level semantics and global structural information.

To address this issue, the Dynamic Head (DyHead) [25] is introduced to replace the conventional detection head. DyHead organizes multi-level features into a unified representation and performs dynamic attention along the level, spatial, and channel dimensions, enabling adaptive feature selection for each prediction without increasing computational cost.

Given the input tensor $F \in \mathbb{R}^{L\times S\times C}$, dynamic aggregation is formulated as

$$F_{out} = \pi_C\left(\pi_S\left(\pi_L(F)\right)\right) \tag{3}$$

Where π_L, π_S, and π_C denote the level-, spatial-, and channel-wise attention functions, respectively. Level-wise attention models cross-scale dependencies, spatial-wise attention enhances structurally continuous regions such as cracks, and channel-wise attention performs task-aware feature recalibration.

By integrating DyHead, the detection head becomes scale-adaptive and context-aware, which improves the localization of small and irregular distresses and enhances robustness under complex road backgrounds.

4.1.5. Overall Detection Framework

The overall architecture of the proposed road-distress detection network is illustrated in Fig. 5. Based on YOLOv7, the framework is redesigned from three aspects, namely feature extraction, multi-scale feature fusion, and dynamic prediction, to improve the detection performance for distresses with large scale variation and complex morphology.

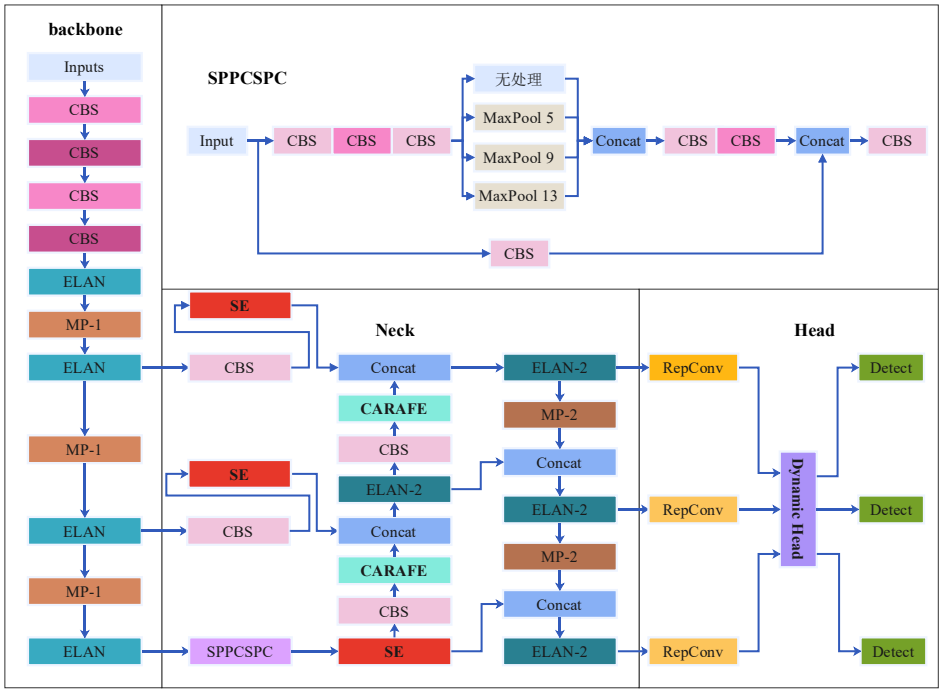

Figure 5. Improved YOLOv7 Network Architecture

First, to enhance the discriminative capability of the backbone, the SE attention modules are inserted after the SPPCSPC module and the convolutional layers at stages 54 and 66, which correspond to three feature-fusion nodes in the neck, i.e., two located before the Concat operations in the top–down pathway and one on the SPPCSPC branch prior to feature aggregation. This design enables the network to emphasize distress-related feature channels while suppressing redundant background responses, which is particularly beneficial for road scenes with strong texture interference.

Second, to improve the effectiveness of multi-scale feature aggregation, the conventional upsampling operation in the neck is replaced by the CARAFE operator. Through content-aware feature reassembly with a large receptive field, high-level semantic information can be adaptively propagated to high-resolution feature maps. This strengthens the representation of distresses with diverse sizes and irregular shapes and alleviates missed detections and misclassifications caused by insufficient feature fusion.

Finally, the original YOLOv7 detection head is replaced with the Dynamic Head to enable scale-adaptive and context-aware prediction. By modeling cross-level, spatial, and channel-wise dependencies in a unified manner, the detection head dynamically selects the most informative features for each target, thereby improving the localization accuracy and robustness for small and multi-scale distresses.

Through the above modifications, the proposed framework forms a coherent feature-processing pipeline: channel-wise feature recalibration in the backbone, content-adaptive multi-scale fusion in the neck, and dynamic feature selection in the head. This progressive refinement strategy significantly enhances the network's ability to perceive fine-grained textures, preserve structural continuity, and handle large scale variation in road-distress detection.

4.2. Distress Segmentation and Quantification

While object detection provides efficient localization of road distresses, bounding boxes are insufficient for accurately describing the spatial extent and geometric characteristics of surface damage. Fine-grained segmentation is therefore required to support quantitative analysis. In this study, a segmentation and quantification module is designed to extract crack and pothole regions at the pixel level and to derive meaningful geometric measurements for maintenance assessment.

4.2.1. MIResU-Net for Distress Segmentation

Road-distress segmentation is challenged by complex background interference and large variations in the scale and morphology of distress regions. Directly applying the conventional U-Net [26] often leads to blurred boundaries and incomplete extraction of fine structures. To address these issues, a multi-scale encoder–decoder network termed MIResU-Net is developed, as shown in Fig. 6.

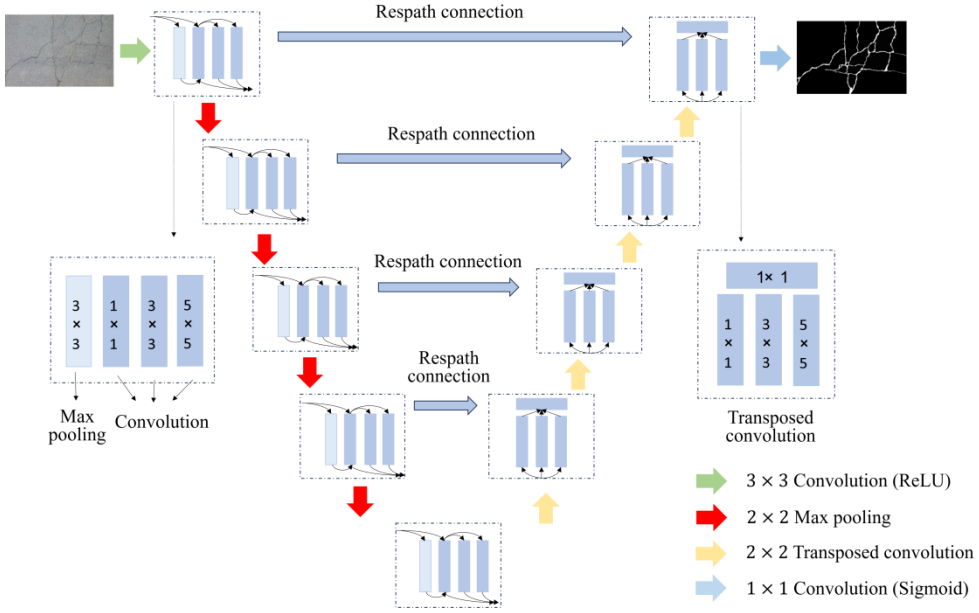

Figure 6. MIResU-Net network architecture

The proposed network follows a five-stage encoder–decoder architecture and enhances feature learning from three aspects.

First, the standard convolutions in the encoder are replaced by an M-Inception module, which employs parallel convolutional branches with different receptive fields. This multi-scale design enables the network to capture both global contextual information and local structural details of distresses while controlling the parameter size through 1×1 convolutions.

Second, the original skip connections are replaced by Respath modules to reduce the semantic gap between encoder and decoder features. By progressively refining shallow features before fusion, Respath improves boundary preservation and enhances the segmentation of thin and low-contrast distresses.

Third, a multi-scale decoding module is introduced to reconstruct high-resolution feature maps. Parallel convolutions with different kernel sizes are used to aggregate contextual information at multiple scales, which improves structural continuity and detail recovery in the final segmentation.

Considering the severe class imbalance between distress and background pixels, the weighted binary cross-entropy (WCE) loss is adopted to increase the penalty on false negatives and reduce missed detections of small distress regions.

Through multi-scale feature extraction, semantic-consistent feature fusion, and multi-scale reconstruction, MIResU-Net achieves more accurate and complete pixel-level segmentation of road distresses under complex road conditions.

4.2.2. Segmentation Output and Post-Processing

The output of MIResU-Net is a pixel-wise probability map indicating the likelihood of each pixel belonging to a distress region. A threshold operation is first applied to obtain the binary segmentation result.

To further improve the structural integrity and measurement reliability of the segmented distresses, several post-processing steps are performed. First, small isolated regions caused by noise are removed using connected-component analysis to suppress false positives in complex road backgrounds. Second, morphological operations are applied to fill small holes and smooth the boundaries, which enhances the continuity of crack structures. Finally, each connected distress region is labeled to enable subsequent quantitative analysis.

Based on the refined segmentation results, geometric attributes of distresses, such as area, length, and width, can be calculated at the pixel level. These quantitative indicators provide essential support for road-condition assessment and maintenance decision-making.

4.2.3. Quantitative Evaluation of Road Distresses

The objective of distress segmentation is to enable quantitative assessment of pavement damage, thereby providing reliable data support for road-condition evaluation and maintenance decision-making. Based on the pixel-level binary masks generated by MIResU-Net, the geometric parameters of different distress types are calculated through a calibration-based measurement strategy.

(1) Pixel-to-real-world calibration

To establish the mapping between image pixels and actual physical dimensions, a calibration object with a known length is captured using the same acquisition system as the road

images. Let M denote the real length (mm) of the calibration object and m the corresponding pixel length in the image. The pixel-to-length conversion factor k is defined as

$$k = \frac{M}{m} \tag{4}$$

In this study, a calibration object of 100 mm corresponds to 113 pixels in the image, yielding a scale factor of $k = 0.8850$ mm/pixel. This ensures consistent measurement accuracy under fixed imaging conditions.

(2) Quantification of linear cracks

For linear distresses, including longitudinal and transverse cracks, both length and width are measured. The crack width is estimated by computing the diameter of the maximum inscribed circle within the segmented crack region. This operation is performed using contour analysis, which provides the pixel width w. The corresponding real width W is calculated as

$$W = k \cdot w \tag{5}$$

The crack length is derived from the ratio between the total number of crack pixels n and the crack width:

$$l = \frac{n}{w} \tag{6}$$

Where l is the crack length in pixels. The real crack length L is then obtained by

$$L = k \cdot l \tag{7}$$

(3) Quantification of potholes and alligator cracks

For area-type distresses, such as potholes and alligator cracks, the quantitative indicator is the actual area. Let s denote the total number of pixels in the segmented distress region. The real area S is computed as.

$$S = k^2 s \tag{8}$$

Where k^2 represents the pixel-to-area conversion factor.

(4) Quantitative evaluation pipeline

The complete quantification process consists of three stages. First, the improved YOLOv7 model is used to identify the distress category. Second, MIResU-Net generates the pixel-level segmentation of the detected distress. Finally, the geometric parameters are calculated using the above measurement formulas.

Through this pipeline, the actual length, width, and area of different distress types can be obtained. These quantitative indicators provide a direct basis for pavement condition assessment and maintenance planning.

5. Experiments and Results

5.1. Experimental Setup

All experiments were conducted on a workstation equipped with an Intel Xeon Platinum 8352V CPU, 120 GB RAM, and an NVIDIA RTX 4090 GPU. The proposed framework was implemented in PyTorch.

For the detection task, the dataset was divided into training, validation, and test sets with a ratio of 7:2:1. The model was trained for 300 epochs using stochastic gradient descent with a momentum of 0.937 and a weight decay of 0.0005. The initial learning rate was set to 0.01 and the batch size was 16.

For the segmentation task, pixel-level annotations were used to train the network for 300 epochs with the Adam optimizer and a batch size of 8.

Detection performance was evaluated using Precision, Recall, and mAP@0.5, while segmentation performance was assessed using PA, IoU, and MIoU.

5.2. Road Distress Detection Results

5.2.1. Effect of Attention Mechanisms

To determine the most suitable attention strategy for road-distress detection, SE, CBAM, and BiFormer were embedded into YOLOv7 for comparison. As shown in Table 1, all attention mechanisms improve detection performance compared with the baseline. Among them, the SE module achieves the best balance between accuracy and computational cost, yielding the highest mAP with only a marginal increase in model complexity. In contrast, BiFormer significantly increases computational overhead without providing comparable accuracy gains. Therefore, SE was selected as the channel-attention module in the proposed detector.

Table 1. Detection performance of YOLOv7 variants on the test set

Method	P (%)	R (%)	$mAP_{0.5}$(%)	$mAP_{0.5:0.95}$(%)	GFLOPs	SizeMB
YOLOv7	85.70	81.28	84.56	60.45	99.0	70.9
YOLOv7-BiF	87.19	80.05	85.78	60.84	145.2	77.7
YOLOv7-CBAM	87.46	80.90	85.29	61.22	101.4	74.6
YOLOv7-SE	87.72	81.62	86.16	61.19	101.3	74.3

5.2.2. Ablation Study

To evaluate the contribution of each proposed component, ablation experiments were conducted by progressively introducing SE, CARAFE, and DyHead.

The results in Table 2 show that each module consistently improves the detection performance. SE enhances feature discrimination, CARAFE strengthens semantic propagation in the feature pyramid, and DyHead improves the detection of small-scale distresses. When all modules are combined, the proposed model achieves the highest Precision, Recall, and mAP, demonstrating the complementarity of the three components.

Table 2. Ablation Study

Configuration				P (%)	R (%)	$mAP_{0.5}$ (%)	GFLOPs	SizeMB
YOLOv7				85.70	81.28	84.56	99.0	70.9
YOLOv7	SE			87.72	81.62	86.16	101.3	74.3
YOLOv7		CARAFE		87.80	80.70	85.22	101.2	74.2
YOLOv7			DyHead	86.31	81.85	86.14	102.4	73.3
YOLOv7	SE	CARAFE		87.80	80.90	85.90	103.5	74.5
YOLOv7	SE	CARAFE	DyHead	88.60	81.66	86.75	103.7	74.7

5.2.3. Comparison with State-of-the-Art Detectors

Table 3. Comparison with state-of-the-art methods

Method	P (%)	R (%)	$mAP_{0.5}$ (%)	GFLOPs
Faster R-CNN	82.75	78.47	81.90	118.4
SSD	78.22	73.10	77.44	22.5
YOLOv5s	80.80	74.30	80.52	15.8
YOLOv7	85.70	81.48	84.56	99.0
Ours	88.60	81.66	86.75	103.7

The proposed detector was further compared with Faster R-CNN, SSD, YOLOv5s, and YOLOv7. As reported in Table 3, the proposed method achieves the best overall detection accuracy while maintaining moderate computational complexity.

5.2.4. Qualitative Analysis

The visual detection results are presented in Table 4. The proposed method shows superior performance in detecting small potholes, preserving crack continuity, and reducing missed detections under complex illumination and shadow interference. In contrast, the baseline YOLOv7 suffers from incomplete detection and false negatives in challenging scenarios.

Table 4. Ablation study for each class

	P (%)		R (%)		$mAP_{0.5}$ (%)	
	YOLOv7	Ours	YOLOv7	Ours	YOLOv7	Ours
Longitudinal crack	82.2	84.1	81.9	82.7	84.6	86.8
Transverse crack	82.8	88.7	79.9	80.6	84.6	86.0
Alligator crack	80.8	90.1	69.6	71.0	74.0	85.1
Pothole	92.5	92.9	91.2	91.3	92.4	92.9
Block crack	94.4	95.8	88.9	88.9	88.7	93.1
Repair	80.5	81.2	77.3	78.2	83.1	82.1

5.3. Road Distress Segmentation Results

5.3.1. Ablation Study of MIResU-Net

To verify the effectiveness of each architectural improvement, ablation experiments were conducted on the segmentation network. As shown in Table 5, each component contributes to the performance gain. The M-Inception module enhances multi-scale feature perception, the Respath connection reduces the semantic gap between encoder and decoder features, and the multi-scale decoding module improves structural recovery. The weighted loss further increases the sensitivity to small distress regions. With all modules integrated, MIResU-Net achieves the highest PA, IoU, and MIoU.

Table 5. Ablation study of MIResU-Net

Method	PA (%)	IoU (%)	MioU (%)
U-Net	85.09	77.82	76.00
MIU-Net	86.63	78.69	77.84
RU-Net	86.50	78.60	76.81
MU-Net	86.54	78.04	77.12
WU-Net	85.98	78.33	76.90
MIResU-Net	87.24	80.92	78.34

5.3.2. Comparison with Mainstream Segmentation Networks

The proposed method was compared with U-Net, FCN, SegNet, and ResUNet. The quantitative results in Table 6 show that MIResU-Net outperforms all comparison methods in all evaluation metrics. The qualitative segmentation results in Fig. 7 demonstrate that MIResU-Net produces more continuous crack structures, clearer pothole boundaries, and stronger robustness against complex backgrounds.

Table 6. Comparison with mainstream segmentation methods

method	PA (%)	IoU (%)	MioU (%)
U-Net	85.09	77.82	76.00
FCN	83.02	75.77	74.94
SegNet	84.41	76.50	75.72
ResUNet	86.06	78.26	77.00
MIResU-Net	87.24	80.92	78.34

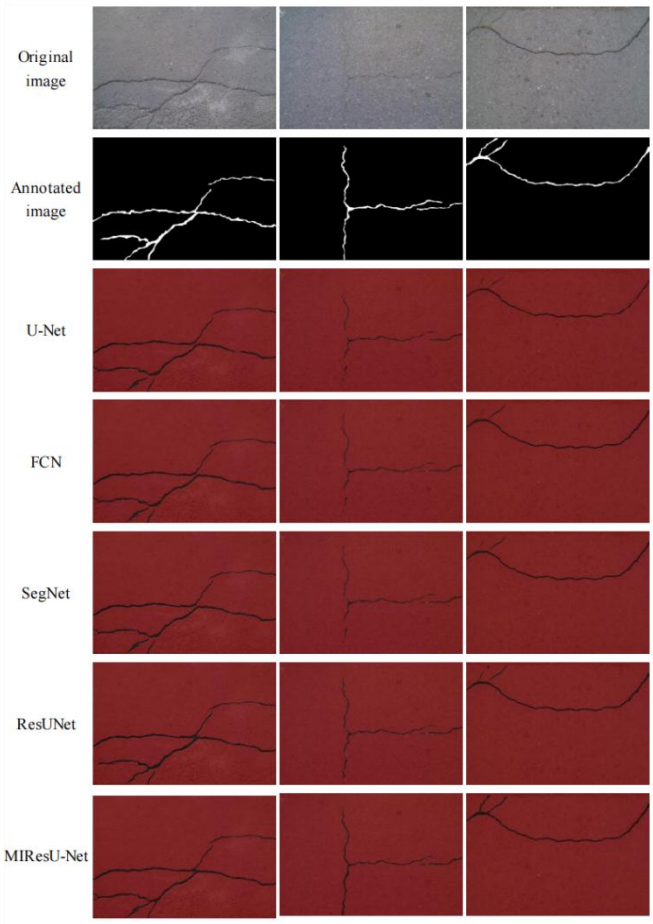

Figure 7. Visual comparison of segmentation results

5.4. Quantitative Evaluation Results

To obtain physically meaningful pavement-condition indicators, a calibration-based measurement was conducted to convert the segmentation outputs into real-world geometric parameters. A calibration object with a known length of 100 mm corresponded to 113 pixels in the image, resulting in a scale factor of 0.8850 mm/pixel. This scale factor enables reliable transformation from pixel-level measurements to actual dimensions.

For linear cracks, the width was estimated using the maximum inscribed circle of the segmented crack contour, as shown in Fig. 8. The crack length was then derived from the total number of crack pixels and the estimated width. The quantitative results for a representative longitudinal crack are summarized in Table 7, where the measured crack width and length are 2.93 mm and 456.14 mm, respectively.

For area-type distresses such as potholes, the actual area was calculated by counting the foreground pixels in the segmented region. A pothole containing 2433 pixels corresponds to a real area of 1905.59 mm^2. These results demonstrate that the proposed framework can provide metrically reliable and physically interpretable distress measurements for pavement-condition assessment.

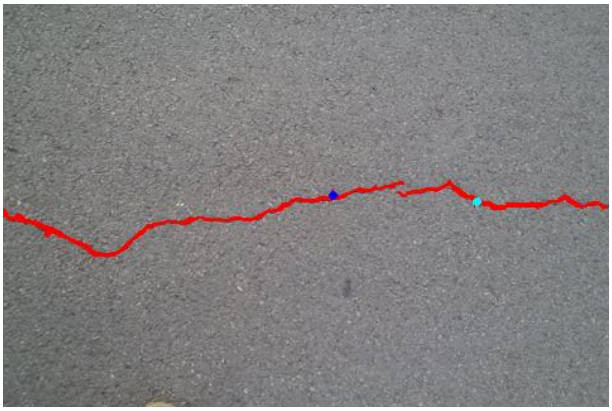

Figure 8. Visualization of crack-width estimation in pixels

Table 7. Quantitative evaluation of road distresses

Distress type	Pixel count	Pixel width	Pixel length	Distress width (mm)	Distress length (mm)
D00	1706	3.31	515.41	2.93	456.14

5.5. Discussion

The experimental results demonstrate that the proposed framework achieves consistent performance improvements across detection, segmentation, and quantification. For detection, the combination of SE, CARAFE, and DyHead enhances multi-scale feature representation with limited computational overhead. For segmentation, MIResU-Net improves semantic consistency and enables accurate extraction of thin and irregular distresses. For quantification, the calibration-based measurement converts segmentation outputs into physically interpretable parameters for pavement-condition assessment. Although the model introduces slightly higher complexity, future work will focus on model compression and real-time deployment on edge devices.

6. Conclusion

This study presents an integrated deep learning-based framework for road-distress detection, segmentation, and quantitative evaluation. An improved YOLOv7 model is developed for robust distress detection under complex road conditions, and MIResU-Net is introduced to extract crack and pothole regions at the pixel level. Based on the segmentation results, geometric parameters such as crack length and pothole area are further computed to enable objective pavement-condition assessment.

Experimental results on real-world road images demonstrate that the proposed framework achieves superior detection and segmentation performance and provides physically meaningful quantitative information for road-maintenance applications. Future work will focus on extending the framework to more distress categories and improving computational efficiency for large-scale and real-time deployment.

Acknowledgment

Supported by 'the Fundamental Research Funds for the Central Universities (2023MS136)'

References

[1] Haas, R. C. G., Hudson, W. R., & Zaniewski, J. P. (1994). Modern pavement management. Krieger Publishing Company.

[2] Chambon, S., & Moliard, J.-M. (2011). Automatic road pavement assessment with image processing: Review and comparison. International Journal of Geophysics, 2011, 989354. https://doi.org/10.1155/2011/989354

[3] Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T., & Omata, H. (2018). Road damage detection and classification using deep neural networks with smartphone images. Computer-Aided Civil and Infrastructure Engineering, 33(12), 1127–1141. https://doi.org/10.1111/mice.12387

[4] Arya, D., Maeda, H., Ghosh, S. K., Toshniwal, D., Mraz, A., Kashiyama, T., & Sekimoto, Y. (2021). Deep learning-based road damage detection and classification for multiple countries. Automation in Construction, 132, 103935. https://doi.org/10.1016/j.autcon.2021.103935

[5] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

[6] Chu, C., Wang, L., & Xiong, H. (2022). A review on pavement distress and structural defects detection and quantification technologies using imaging approaches. Journal of Traffic and Transportation Engineering, 9(2), 135–150. https://doi.org/10.1016/j.jtte.2021.04.007

[7] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 580–587. https://doi.org/10.1109/CVPR.2014.81

[8] Wang, W., Wu, B., Yang, S., & Wang, Z. (2018). Road damage detection and classification with Faster R-CNN. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 5220–5223). IEEE. Seattle, WA, United States. https://doi.org/10.1109/BigData.2018.8621987.

[9] Chen, Q., Gan, X., Huang, W., et al. (2020). Road damage detection and classification using Mask R-CNN with DenseNet backbone. Computer, Materials & Continua, 65(3), 2201–2215. https://doi.org/10.32604/cmc.2020.011108

[10] Yang, F., Yu, B., Zhao, J., et al. (2022). Bridge-bottom crack detection method based on improved YOLOv3. China Sciencepaper, 17(3), 252–259.

[11] Cao, M. T., Tran, Q. V., Nguyen, N. M., et al. (2020). Survey on performance of deep learning models for detecting road damages using multiple dashcam image resources. Advanced Engineering Informatics, 46, 101182. https://doi.org/10.1016/j.aei.2020.101182

[12] Luo, H., Jia, C., & Li, J. (2021). Highway pavement distress detection algorithm based on improved YOLOv4. Laser & Optoelectronics Progress, 58(14), 336–344.

[13] Feng, X., Xiao, L., Li, W., et al. (2020). Pavement crack detection and segmentation method based on improved deep learning fusion model. Mathematical Problems in Engineering, 2020, 1–22. https://doi.org/10.1155/2020/6413085

[14] K. C., R., & G., R. (2022). Road damage detection and classification using YOLOv5. In 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT) (pp. 489–494). IEEE. Kannur, India. https://doi.org/10.1109/ICICICT54557.2022.9917899

[15] Zhang, Y., Zuo, Z., Xu, X., et al. (2022). Road damage detection using UAV images based on multi-level attention mechanism. Automation in Construction, 138, 104264. https://doi.org/10.1016/j.autcon.2022.104264

[16] Wang, S., et al. (2022). An ensemble learning approach with multi-depth attention mechanism for road damage detection. In 2022 IEEE International Conference on Big Data (Big Data) (pp. 6439–6444). IEEE. Osaka, Japan. https://doi.org/10.1109/BigData55660.2022.10020445

[17] Wang, J., Gao, X., Liu, Z., & Wan, Y. (2023). GSC-YOLOv5: An algorithm based on improved attention mechanism for road crack detection. In 2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS) (pp. 1664–1671). IEEE. Xiangtan, China. https://doi.org/10.1109/DDCLS58216.2023.10132562

[18] Ma, H. (2022). Object surface defect detection based on deformable convolution and attention mechanism (Master's thesis). Yunnan University.

[19] Yang, L., He, H., & Liu, T. (2022). Road damage detection and classification based on multi-scale contextual features. In 2022 IEEE International Conference on Big Data (Big Data) (pp. 6445–6453). IEEE. Osaka, Japan. https://doi.org/10.1109/BigData55660.2022.10020446

[20] Tzutalin. (2015). LabelImg (Version 1.8.6) [Computer software]. GitHub. https://github.com/tzutalin/labelImg

[21] Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T., & Omata, H. (2018). Road damage detection and classification using deep neural networks with smartphone images. arXiv. https://arxiv.org/abs/1801.09454

[22] Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7464–7475. https://doi.org/10.1109/CVPR52729.2023.00721

[23] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7132–7141. https://doi.org/10.1109/CVPR.2018.00745

[24] Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C. C., & Lin, D. (2019). CARAFE: Content-aware reassembly of features. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 3007–3016. https://doi.org/10.1109/ICCV.2019.00310

[25] Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., & Zhang, L. (2021). Dynamic head: Unifying object detection heads with attentions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7369–7378. https://doi.org/10.1109/CVPR46437.2021.00729

[26] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28