A Real-Time Multi-Scale Feature Pyramid YOLO Architecture for Accurate and Deployment-Efficient Road Damage Detection

Olzhas Olzhayev[1], Bakhytzhan Kulambayev[2]*, Nurly Sakenkyzy[3], Madina Belisbek[4]
International Information Technology University, Almaty, Kazakhstan[1, 3]
Turan University, Almaty, Kazakhstan[2, 4]

***Abstract*—Automated road damage detection has become a critical component of intelligent transportation systems, enabling timely infrastructure maintenance and enhanced traffic safety. However, detecting pavement defects such as cracks, potholes, and surface degradation remains challenging due to significant scale variation, irregular geometries, illumination changes, and class imbalance. This study proposes a real-time Multi-Scale Feature Pyramid YOLO architecture designed to achieve accurate and deployment-efficient multi-class road damage detection. The framework integrates hierarchical feature extraction with bidirectional multi-scale fusion to enhance sensitivity to both small and large defects. A decoupled detection head is employed to improve classification–localization balance, while focal loss and small-object emphasis mechanisms address class imbalance and fine-grained crack detection challenges. Comprehensive experiments conducted on a multi-class road damage dataset demonstrate that the proposed model achieves a mAP@0.5 of 0.68 and a recall of 0.81, outperforming several representative real-time detection approaches. Precision–recall analysis, confusion matrix evaluation, and ablation studies confirm the effectiveness of multi-scale feature aggregation and targeted optimization strategies. Qualitative results further illustrate robust detection performance under diverse environmental conditions. The proposed framework provides a practical trade-off between accuracy and computational efficiency, making it suitable for real-world deployment in intelligent road condition monitoring systems.**

Keywords*—Road damage; Multi-Scale Feature Pyramid; YOLO architecture; intelligent transportation systems; small-object detection; real-time deployment; pavement defect analysis

I. Introduction

Road infrastructure deterioration remains a persistent challenge for transportation agencies worldwide, directly affecting traffic safety, vehicle operating costs, and long-term maintenance expenditure. Surface defects such as longitudinal cracks, transverse cracks, potholes, and alligator cracking progressively compromise structural integrity, particularly under high traffic density and adverse environmental conditions. Manual inspection, although still widely practiced, is labor-intensive, time-consuming, and inherently subjective. Consequently, automated vision-based road damage detection systems have emerged as a critical research direction within intelligent transportation systems [1].

*Corresponding author.

Recent advances in deep convolutional neural networks have substantially improved object detection performance across diverse domains, including infrastructure monitoring and urban analytics [2]. Among these approaches, single-stage detectors have demonstrated a favorable balance between accuracy and inference speed, making them particularly suitable for real-time deployment scenarios [3]. The YOLO family of detectors has gained significant attention due to its unified detection pipeline, end-to-end optimization, and high computational efficiency [4]. However, road damage detection introduces domain-specific complexities that distinguish it from generic object detection tasks.

Road defects frequently exhibit irregular geometries, low contrast against asphalt backgrounds, and substantial scale variation within the same image. Small cracks may occupy only a few pixels, whereas potholes can span large spatial regions. Standard detection architectures often struggle with this extreme multi-scale variability, leading to missed detections or unstable localization [5]. Feature pyramid networks have been introduced to mitigate such issues by enabling hierarchical multi-resolution feature aggregation, thereby enhancing small-object sensitivity while preserving contextual semantics [6]. Integrating feature pyramids within lightweight detection frameworks has shown promising results in improving localization robustness without imposing excessive computational overhead [7].

Another critical constraint involves deployment efficiency. In practical road monitoring systems, models must operate on embedded platforms mounted on vehicles, where computational resources and energy budgets are limited [8]. High-accuracy models with excessive parameter counts may be unsuitable for such environments. Therefore, architectural optimization must simultaneously address detection precision, real-time throughput, and hardware compatibility [9].

Motivated by these challenges, this study proposes a real-time Multi-Scale Feature Pyramid YOLO architecture designed specifically for accurate and deployment-efficient road damage detection. The proposed framework integrates enhanced multi-scale feature fusion with an optimized detection head to improve sensitivity to small and medium-sized defects while maintaining high inference speed. Experimental validation demonstrates that the architecture achieves robust detection performance under varying illumination, perspective distortion,

and real-world driving conditions, thereby contributing to scalable intelligent road inspection systems [10].

II. Related Works

Automated road damage detection has evolved considerably with the integration of deep learning techniques, particularly convolutional neural networks applied to large-scale road imagery datasets. Early approaches relied on handcrafted features and classical image processing techniques, including edge detection, thresholding, and texture descriptors, to identify cracks and surface anomalies. While these methods demonstrated feasibility under controlled conditions, they often lacked robustness against illumination variability, shadows, and complex pavement textures [11]. The transition toward deep learning-based models significantly improved generalization capability and detection reliability in unconstrained environments [12].

Two-stage object detection frameworks, such as region proposal-based architectures, have been applied to road defect identification tasks due to their strong localization accuracy and structured training paradigm [13]. These methods generate candidate regions before classification and bounding box regression, thereby achieving precise detection. However, their multi-stage inference pipeline increases computational latency, limiting suitability for real-time deployment on vehicle-mounted systems [14]. To address efficiency constraints, single-stage detectors were introduced, directly predicting object locations and class probabilities in a unified forward pass [15].

Among single-stage approaches, YOLO-based models have been widely adopted for road damage detection due to their high inference speed and end-to-end optimization capability [16]. Subsequent refinements improved feature representation, anchor design, and loss formulations, leading to enhanced detection stability across varying object scales [17]. Nevertheless, road defects pose unique challenges, particularly in detecting small, thin, and elongated cracks that may be visually indistinguishable from background noise [18]. Several studies have, therefore, emphasized multi-scale feature extraction strategies to enhance sensitivity to fine-grained structural patterns [19].

Feature Pyramid Networks have been integrated into detection frameworks to facilitate hierarchical feature aggregation and improve performance on small objects [20]. Further enhancements using bidirectional feature fusion mechanisms have demonstrated superior multi-scale consistency and localization robustness [21]. These architectures propagate semantic information across layers, strengthening detection of both minor surface fissures and large potholes [22].

In parallel, lightweight model design has gained prominence due to deployment requirements in embedded and edge computing platforms [23]. Channel pruning, depth-wise separable convolutions, and parameter-sharing strategies have been introduced to reduce model complexity without significantly compromising detection accuracy [24]. Efficient backbone redesign and optimized detection heads have further contributed to achieving real-time inference on resource-constrained devices [25].

Recent works also address class imbalance and irregular object geometry through modified loss functions and adaptive sampling strategies [26]. IoU-based regression losses have been shown to enhance bounding box stability, particularly for non-uniform damage shapes [27]. Additionally, data augmentation strategies tailored to road inspection scenarios, including perspective transformation and illumination simulation, have improved model robustness under diverse driving conditions [28].

Despite these advances, achieving a balanced trade-off between detection accuracy, small-object sensitivity, and computational efficiency remains an open research problem [29]. Existing models often prioritize either precision or speed, but rarely both in a deployment-aware framework [30]. Consequently, there remains a clear need for an architecture that integrates multi-scale feature fusion, optimized detection heads, and computational efficiency within a unified YOLO-based design, specifically tailored for real-world road damage detection systems [31].

III. Materials and Methods

The proposed framework, illustrated in Fig. 1, follows a unified single-stage detection paradigm composed of three principal modules: Backbone, Neck, and Detection Head. The architecture is designed to simultaneously achieve high detection accuracy and deployment efficiency by integrating multi-scale feature extraction with feature pyramid-based fusion and multi-branch prediction.

Let the input road image be denoted as:

$$I \in R^{H \times W \times 3}, \tag{1}$$

After preprocessing, the image is forwarded to the backbone network, which performs hierarchical feature extraction through convolutional transformations. The backbone can be expressed as a nonlinear mapping:

$$F_b = \mathcal{B}(I; \theta_b), \tag{2}$$

where, θ_b represents backbone parameters and $F_b = \{F_3, F_4, F_5\}$ correspond to multi-resolution feature maps at scales 1/8, 1/16, and 1/32, respectively. These feature maps encode progressively abstract semantic representations while reducing spatial resolution.

To enhance small-object sensitivity and contextual reasoning, the extracted features are processed by a feature pyramid fusion module. The neck performs bidirectional aggregation using upsampling and concatenation operations, formulated as:

$$F_n^l = \mathcal{F}_{fusion}\left(F_b^l, Up\left(F_b^{l+1}\right)\right), \tag{3}$$

where, $l \in \{3, 4\}$, and $Up(\cdot)$ denotes spatial upsampling. The resulting fused representations $F_n = \{P_3, P_4, P_5\}$ preserve fine spatial details while incorporating high-level semantic context.

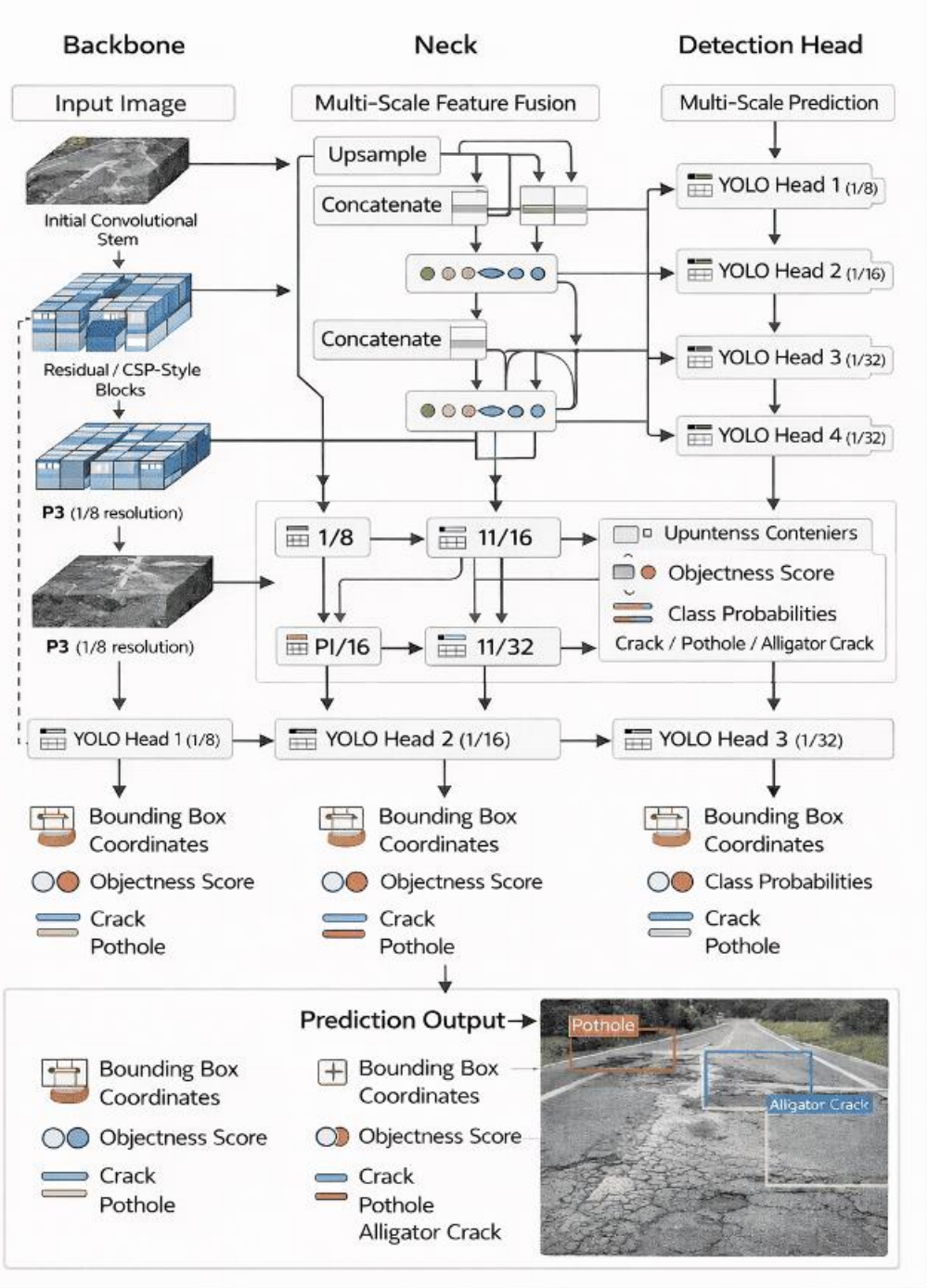


Fig. 1. Architecture of the proposed Multi-Scale Feature Pyramid YOLO framework for real-time road damage detection.

Each fused feature map is then forwarded to a scale-specific detection head. The prediction function for scale s is defined as:

$$Y_s = \mathcal{H}_s(P_s, \theta_h), \tag{4}$$

where, Y_s includes bounding box coordinates, objectness confidence, and class probabilities for road damage categories.

The final detection output is obtained by aggregating predictions across all scales:

$$Y = \bigcup_{S\epsilon\{3,4,5\}} Y_s, \tag{5}$$

This multi-scale aggregation enables robust detection of cracks, potholes, and alligator damage with varying spatial extents. By integrating hierarchical feature extraction, feature pyramid fusion, and parallel detection heads within a single forward pass, the proposed architecture ensures both computational efficiency and high detection fidelity under real-world driving conditions.

A. Input and Preprocessing Block

The Input and Preprocessing Block constitutes the foundational stage of the proposed detection framework, ensuring that raw road imagery is transformed into a normalized and structurally consistent representation suitable for deep neural processing. Initially, road images are captured using a vehicle-mounted camera system operating under dynamic real-world conditions, including varying illumination, motion blur, and perspective distortion. These high-resolution images, typically acquired at 1920 × 1080 pixels, are subsequently resized using a letterbox strategy to a fixed input dimension of 640 × 640 while preserving aspect ratio. The letterbox operation prevents geometric distortion of road defects by padding the image instead of applying anisotropic scaling, thereby maintaining the structural integrity of cracks, potholes, and alligator damage patterns. This step ensures spatial consistency across batches and stabilizes the receptive field behavior of convolutional filters in the backbone network (see Fig. 2).

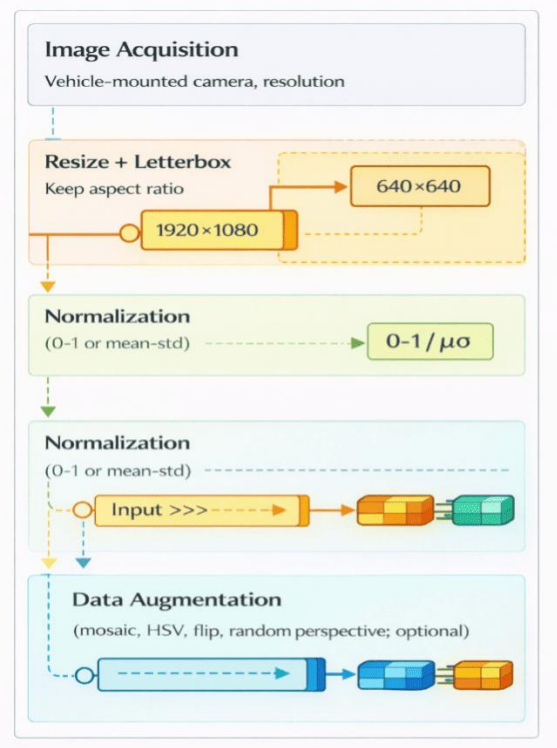


Fig. 2. Input and preprocessing block.

Following spatial standardization, pixel intensities are normalized either to the [0,1] interval or through mean-standard deviation scaling, reducing statistical variance across samples and accelerating optimization convergence. Normalization also mitigates illumination bias, which is particularly critical in outdoor road inspection scenarios characterized by strong shadows and reflective surfaces. To further enhance generalization capability, data augmentation strategies such as mosaic composition, hue-saturation-value perturbation, horizontal flipping, and random perspective transformation are optionally applied during training. These augmentations introduce controlled geometric and photometric variability, enabling the model to learn invariant representations of road defects across diverse environmental conditions. Collectively, this preprocessing pipeline enhances robustness, improves convergence stability, and prepares the input distribution for effective hierarchical feature extraction in subsequent network stages.

B. Backbone Block

The Backbone Block serves as the primary feature extraction component of the proposed architecture, transforming normalized input images into hierarchically structured representations with increasing semantic abstraction (see Fig. 3). The process begins with a stem convolution layer, composed of convolution, batch normalization, and nonlinear activation operations, which performs initial low-level feature extraction while reducing spatial redundancy. This is followed by progressive downsampling stages implemented through stride-2 convolutions or pooling operations. These layers systematically decrease spatial resolution while expanding channel depth, thereby enlarging the effective receptive field and enabling the network to capture broader contextual information. Such hierarchical compression is essential for road damage detection, where defects may vary significantly in scale and spatial distribution across the pavement surface.

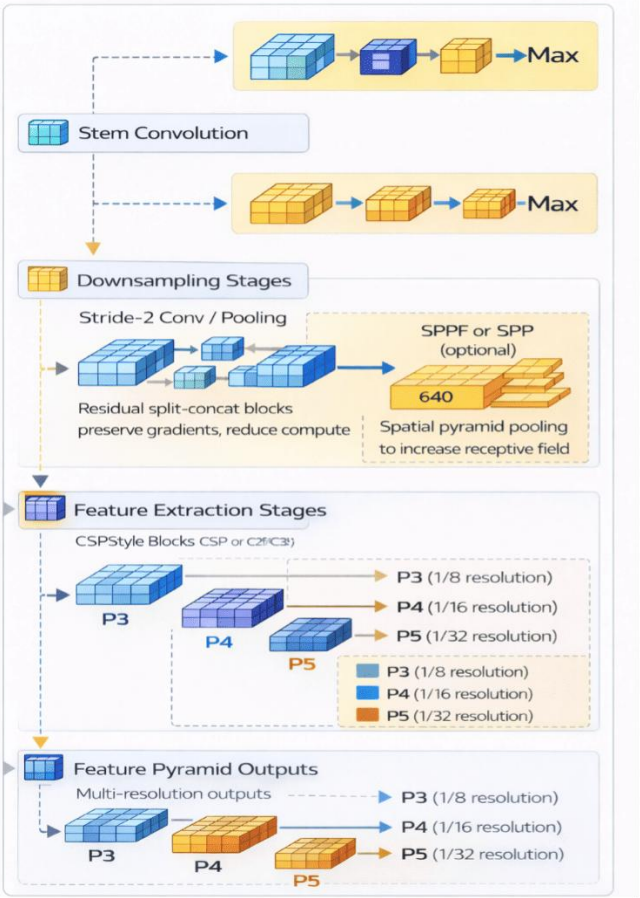


Fig. 3. Backbone block.

Subsequent feature extraction stages employ CSP-style residual split-concatenate blocks, such as CSP, C2f, or C3 modules, designed to enhance gradient flow while maintaining computational efficiency. By partitioning feature maps and

merging them after parallel transformations, these blocks reduce parameter redundancy without compromising representational richness. This design improves training stability and facilitates deeper network construction. An optional spatial pyramid pooling module further expands the receptive field by aggregating multi-scale contextual information through parallel pooling operations, strengthening the model's ability to distinguish irregular crack patterns from background textures. The backbone ultimately produces multi-resolution feature maps denoted as P3, P4, and P5, corresponding to spatial scales of 1/8, 1/16, and 1/32 of the original resolution. These outputs encode complementary fine-grained and high-level semantic features, forming the structural basis for subsequent multi-scale feature fusion in the neck module.

C. Neck Block

The Neck Block performs multi-scale feature fusion, acting as the structural bridge between hierarchical feature extraction in the backbone and scale-aware prediction in the detection head (see Fig. 4). Its primary objective is to enrich spatially precise low-level features with semantically strong high-level representations. The top-down pathway, inspired by the Feature Pyramid Network paradigm, propagates high-level contextual information from deeper layers to shallower ones through successive upsampling operations. Specifically, feature maps from higher pyramid levels are spatially upsampled and concatenated with corresponding lower-level features. This fusion strategy enables the preservation of fine-grained structural information essential for detecting thin cracks while simultaneously embedding broader contextual cues necessary for distinguishing road defects from background artifacts such as shadows or texture irregularities.

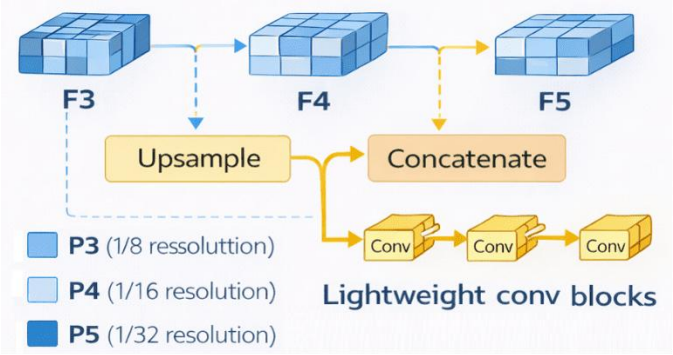


Fig. 4. Neck block.

Complementing this mechanism, the bottom-up pathway reinforces localization accuracy by reintroducing refined low-level information into deeper representations. Through successive downsampling and concatenation operations, the Path Aggregation Network structure enhances bidirectional information flow, strengthening feature consistency across scales. Following fusion, lightweight convolutional refinement blocks are applied to suppress redundancy and stabilize feature distributions before prediction. These convolutional layers act as feature recalibration units, improving inter-scale coherence while maintaining computational efficiency. The final outputs of the neck module, denoted as P3, P4, and P5, represent enriched multi-resolution feature maps that combine spatial precision and semantic robustness. This bidirectional fusion architecture is particularly advantageous for road damage detection, where defects vary dramatically in size, orientation, and contrast, requiring simultaneous sensitivity to micro-scale cracks and macro-scale pothole structures.

D. Detection Head Block

The Detection Head Block is responsible for transforming fused multi-scale feature representations into structured predictions comprising bounding box coordinates, objectness confidence scores, and categorical probabilities for road damage classes (see Fig. 5). The architecture employs multi-scale detection heads operating at spatial resolutions of 1/8, 1/16, and 1/32 of the original input size, enabling scale-aware inference. The highest-resolution head primarily focuses on small and thin defects such as fine cracks, where spatial precision is critical. Intermediate and lower-resolution heads progressively emphasize medium and large defects, including potholes and extensive alligator cracking. This hierarchical prediction mechanism ensures that objects of varying spatial extents are detected with appropriate receptive field coverage and contextual support, thereby reducing missed detections across scales.

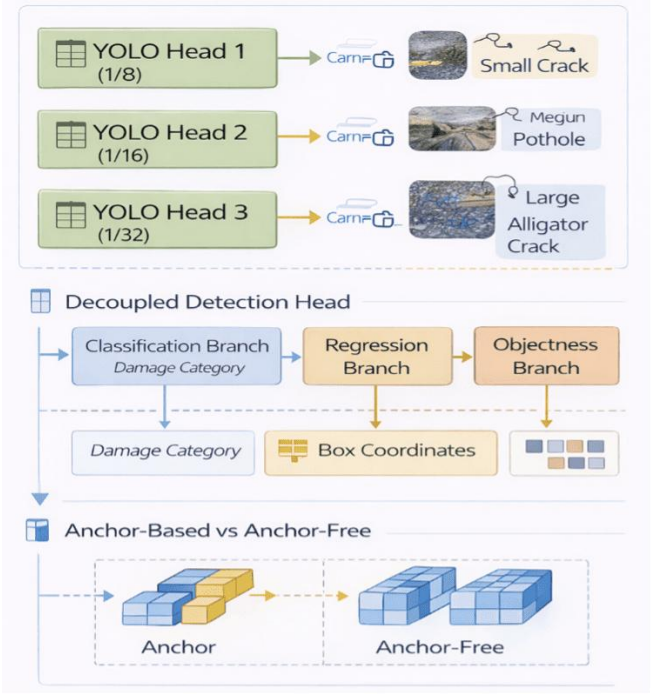


Fig. 5. Detection head block.

To enhance optimization stability and prediction accuracy, the detection head adopts a decoupled design in which classification, bounding box regression, and objectness estimation are processed through independent branches. The classification branch predicts the probability distribution over road damage categories, the regression branch estimates bounding box parameters, and the objectness branch evaluates the likelihood that a predicted region contains a valid defect. Decoupling these tasks mitigates gradient interference and allows each branch to specialize in its respective objective, improving convergence behavior and localization fidelity.

Furthermore, the framework accommodates both anchor-based and anchor-free formulations, enabling flexible state representation depending on deployment constraints and dataset characteristics. This modular detection design provides robust and efficient inference under real-world driving conditions while preserving computational feasibility for embedded platforms.

E. Loss Function Block

The Loss Function Block defines the optimization objective guiding the training process of the proposed detection framework. It consists of three principal components: bounding box regression loss, classification loss, and objectness loss (see Fig. 6). The bounding box loss quantifies the spatial discrepancy between predicted and ground-truth boxes using Intersection-over-Union-based metrics such as IoU, GIoU, DIoU, or CIoU. These formulations incorporate not only overlap area but also geometric alignment and center distance, thereby improving localization stability for irregularly shaped road defects. This is particularly critical in road damage detection, where cracks may exhibit elongated or fragmented structures that are sensitive to slight coordinate deviations. By directly optimizing geometric consistency, the regression loss enhances convergence robustness and reduces bounding box oscillation during training.

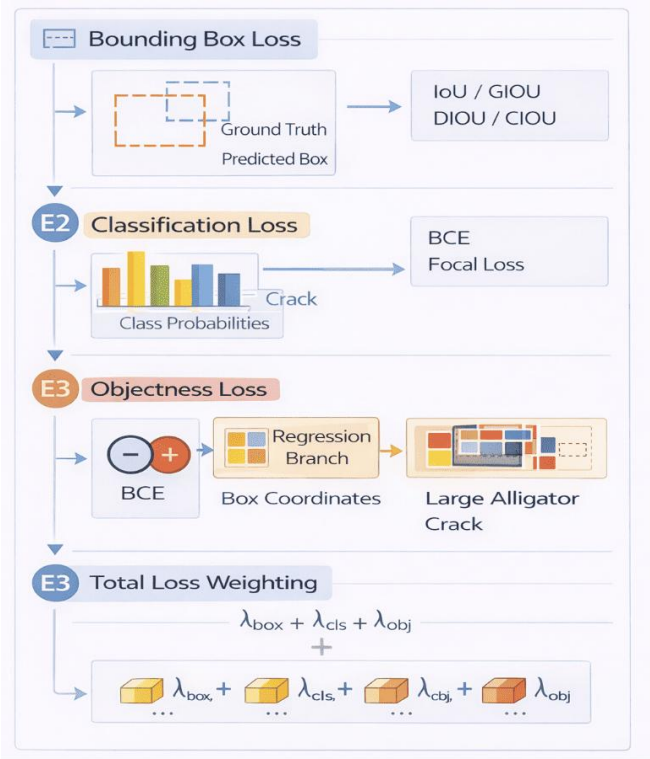


Fig. 6. Loss function.

The classification loss evaluates the discrepancy between predicted class probabilities and true labels, typically using Binary Cross-Entropy or Focal Loss. Focal Loss is especially advantageous in scenarios characterized by class imbalance, where background regions vastly outnumber damaged pixels. The objectness loss, also implemented using Binary Cross-Entropy, estimates the probability that a predicted region contains a valid road defect, suppressing false positives in cluttered scenes. These three components are combined through weighted summation, governed by scaling coefficients λ_{box}, λ_{clx}, λ_{obj} to form the total training objective. Appropriate balancing of these weights ensures that localization accuracy, categorical discrimination, and confidence calibration are jointly optimized. This composite loss formulation enables stable training dynamics, while maintaining sensitivity to both small cracks and large structural defects.

F. Post-Processing Block

The Post-Processing Block transforms raw network predictions into coherent and interpretable detection outputs suitable for practical deployment (see Fig. 7). The first stage, confidence thresholding, filters predictions based on a predefined confidence score, typically greater than 0.5. This operation removes low-probability detections that are likely to correspond to background noise or ambiguous patterns in the pavement surface. Since road imagery often contains texture irregularities, shadows, and perspective artifacts, early suppression of weak predictions significantly reduces false positives. By retaining only predictions that satisfy a minimum confidence criterion, the system improves reliability before spatial refinement is applied.

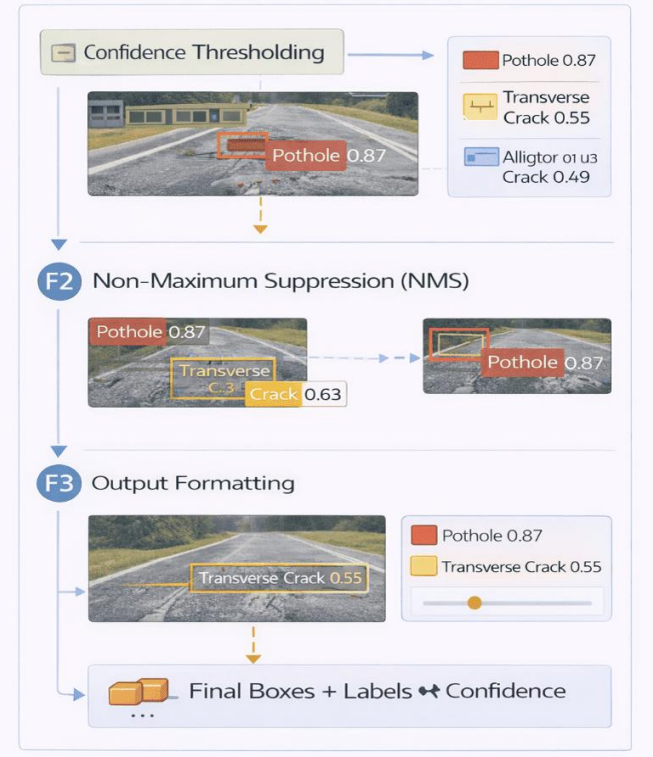


Fig. 7. Post-processing block.

Following thresholding, Non-Maximum Suppression is employed to eliminate redundant overlapping bounding boxes that correspond to the same road defect. During this process, boxes with high Intersection-over-Union overlap are compared, and only the prediction with the highest confidence score is preserved. This step is essential in dense detection scenarios where multiple anchors or grid cells may respond to a single pothole or crack segment. In certain cases, Soft-NMS may be

applied to attenuate rather than completely discard overlapping predictions, enhancing robustness in clustered damage regions. Finally, output formatting consolidates the remaining predictions into structured outputs consisting of bounding box coordinates, predicted class labels, and associated confidence scores. This final representation ensures compatibility with visualization modules, reporting systems, and embedded deployment pipelines, enabling accurate and interpretable real-time road damage monitoring.

G. Road-Damage Specific Block

The Road-Damage Specific Block introduces targeted architectural adaptations that address domain-specific challenges inherent in pavement defect detection (see Fig. 8). The first component emphasizes small-object sensitivity, recognizing that fine cracks often occupy minimal pixel regions and are easily overlooked by standard detectors. Increasing the input resolution enhances spatial granularity, enabling the preservation of thin structural details during convolutional processing. In parallel, strengthening the high-resolution prediction head, particularly the P3 branch operating at 1/8 scale, improves localization precision for narrow and fragmented crack patterns. This configuration ensures that small-scale defects receive sufficient representational capacity without compromising detection of larger anomalies such as potholes or extensive alligator cracking.

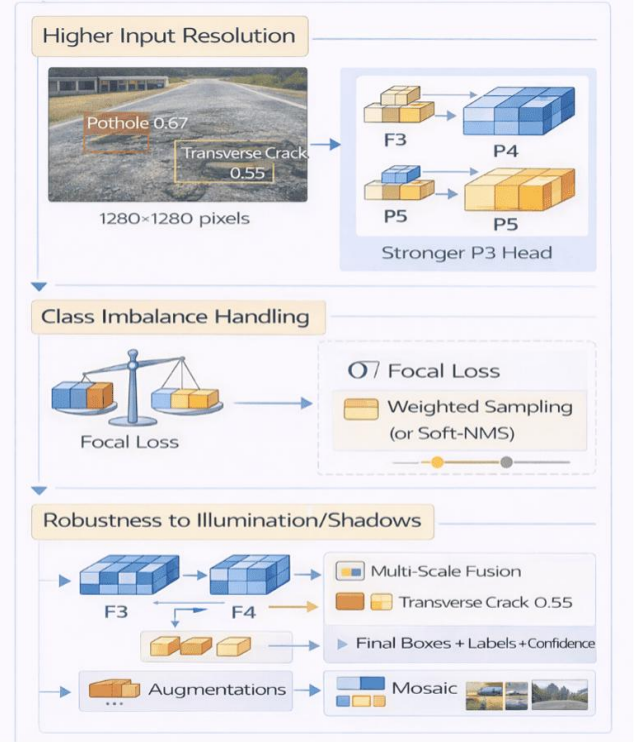


Fig. 8. Road-damage specific block.

The second and third components address class imbalance and environmental variability, respectively. Road damage datasets typically exhibit disproportionate representation among defect categories, with background regions vastly outnumbering damaged instances. To mitigate this imbalance, focal loss and weighted sampling strategies are incorporated to reduce the influence of easy negative examples while emphasizing underrepresented classes. This reweighting mechanism stabilizes gradient updates and enhances minority-class detection performance. Additionally, robustness to illumination changes and shadow artifacts is achieved through multi-scale feature fusion and targeted data augmentation techniques, including mosaic composition and photometric transformations. These strategies encourage the model to learn invariant representations under diverse lighting conditions and perspective distortions. Collectively, the Road-Damage Specific Block refines the generic detection framework into a domain-aware architecture capable of reliable performance in real-world road inspection scenarios.

IV. Data

The experimental evaluation of the proposed Multi-Scale Feature Pyramid YOLO architecture was conducted using a curated road damage dataset composed of annotated pavement images captured from vehicle-mounted cameras under diverse real-world conditions. The dataset contains high-resolution road scenes acquired in urban streets, highways, suburban roads, and semi-rural environments. Images were collected across varying illumination regimes, including daylight, cloudy, night-time, and rainy conditions, in order to ensure robustness against environmental variability. As illustrated in Fig. 9, the dataset includes representative examples of potholes, transverse cracks, longitudinal cracks, and alligator cracking, each annotated with bounding boxes and corresponding class labels. The visual diversity shown in Fig. 9 highlights significant intra-class variation in scale, texture, orientation, and background complexity.

All images were resized to a unified spatial resolution of 640 × 640 pixels during preprocessing while preserving aspect ratio via letterboxing. The dataset comprises 2,450 annotated images with three principal damage categories used for detection experiments: crack, pothole, and alligator crack. As observed in Fig. 9, crack-type defects appear in elongated, thin structures with irregular geometry, whereas potholes typically exhibit compact, high-contrast depressions. Alligator cracks demonstrate clustered, network-like fragmentation patterns. The dataset distribution is moderately imbalanced, with cracks representing the majority class, followed by potholes and alligator cracks. This imbalance motivates the integration of class-sensitive optimization strategies within the training framework.

Bounding box annotations were generated in YOLO format and manually verified to ensure spatial precision, particularly for small-scale defects that occupy limited pixel regions. The dataset includes multi-scale damage instances, ranging from small, distant cracks to large-area structural degradation, thereby supporting multi-resolution learning within the detection pipeline. Furthermore, environmental challenges such as shadows, worn lane markings, surface patches, and debris introduce realistic noise, enhancing the ecological validity of the evaluation. As demonstrated in Fig. 9, these variations reflect the practical deployment scenario of road inspection systems, reinforcing the suitability of the dataset for benchmarking real-time and deployment-efficient detection architectures.

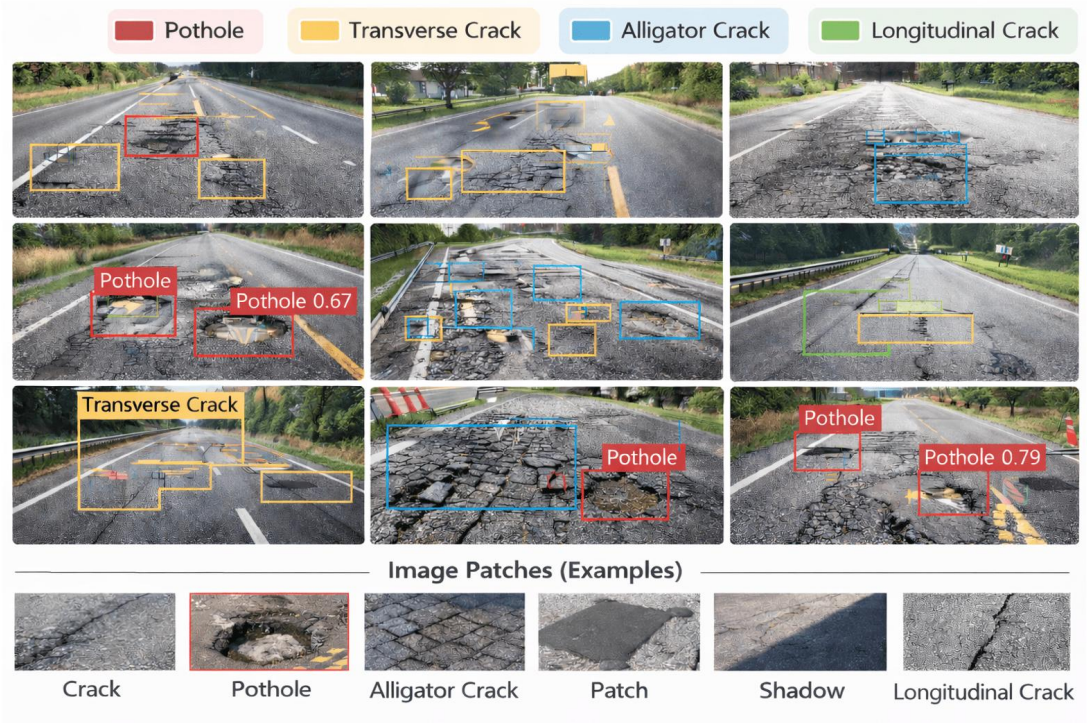


Fig. 9. Samples of the RDD2022 road damage classes.

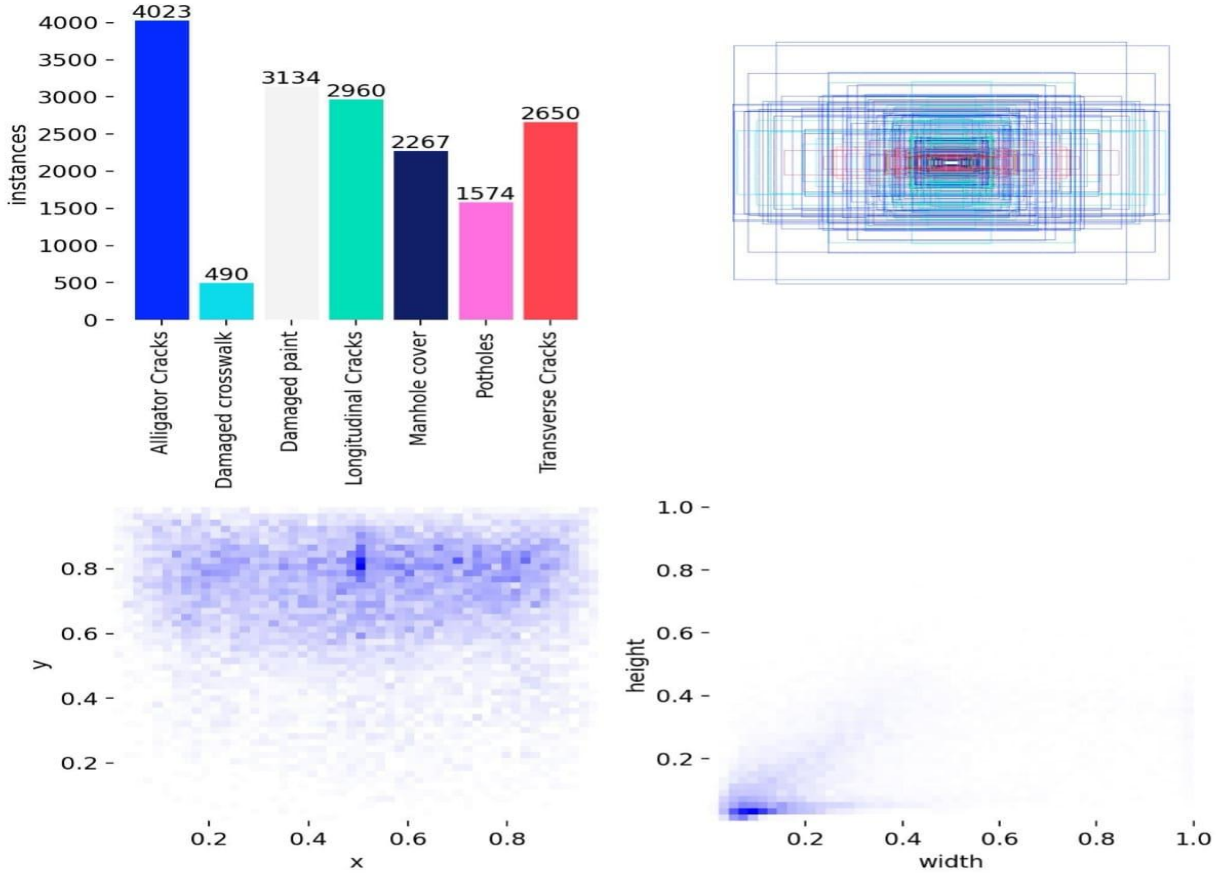


Fig. 10. Statistical analysis of the road damage dataset: class distribution, spatial localization patterns, and bounding box size characteristics.

Fig. 10 presents a comprehensive statistical characterization of the dataset, revealing both class distribution and spatial annotation patterns. The bar chart indicates a pronounced class imbalance, with Alligator Cracks and Damaged Paint exhibiting the highest number of instances, while Damaged Crosswalk and Potholes appear less frequently. This uneven distribution justifies the incorporation of imbalance-aware training strategies discussed in the methodology. The bounding box overlay visualization illustrates a strong concentration of annotations around the central horizontal region of the image plane, which corresponds to the primary driving lane captured by vehicle-mounted cameras. The spatial density heatmap of object centers further confirms this trend, showing that most damage instances occur within the lower-middle portion of the frame, reflecting realistic road-scene geometry. Additionally, the width–height distribution demonstrates that a substantial proportion of annotations correspond to small-scale objects, as evidenced by clustering in the lower-left region of the size plot. Collectively, these statistics confirm that the dataset contains multi-scale defects with central spatial bias and significant class imbalance, thereby posing realistic challenges for robust and deployment-efficient detection architectures.

V. Results

This section presents a comprehensive evaluation of the proposed Multi-Scale Feature Pyramid YOLO architecture for multi-class road damage detection. The experimental results are analyzed from both quantitative and qualitative perspectives to assess detection accuracy, robustness, and deployment suitability. Performance is examined using standard object detection metrics, including mAP@0.5, precision, recall, and F1-score, alongside confusion matrix analysis and confidence-based evaluation curves. In addition, ablation studies are conducted to systematically investigate the contribution of individual architectural components, such as multi-scale feature fusion, focal loss integration, and small-object emphasis mechanisms. Qualitative detection examples further demonstrate the model's capability to localize heterogeneous pavement defects under diverse environmental and illumination conditions. Collectively, these results provide empirical evidence of the proposed framework's effectiveness and practical applicability for intelligent road condition monitoring systems.

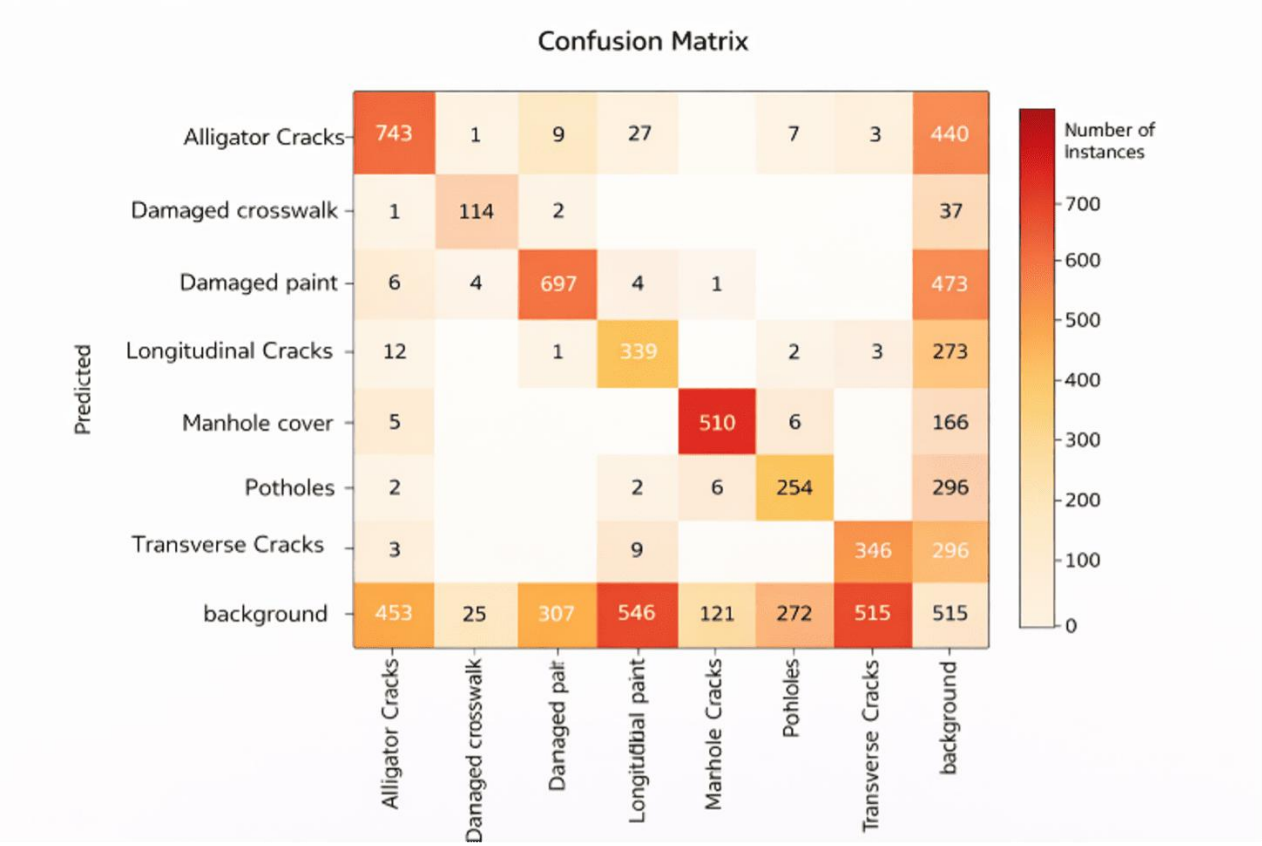


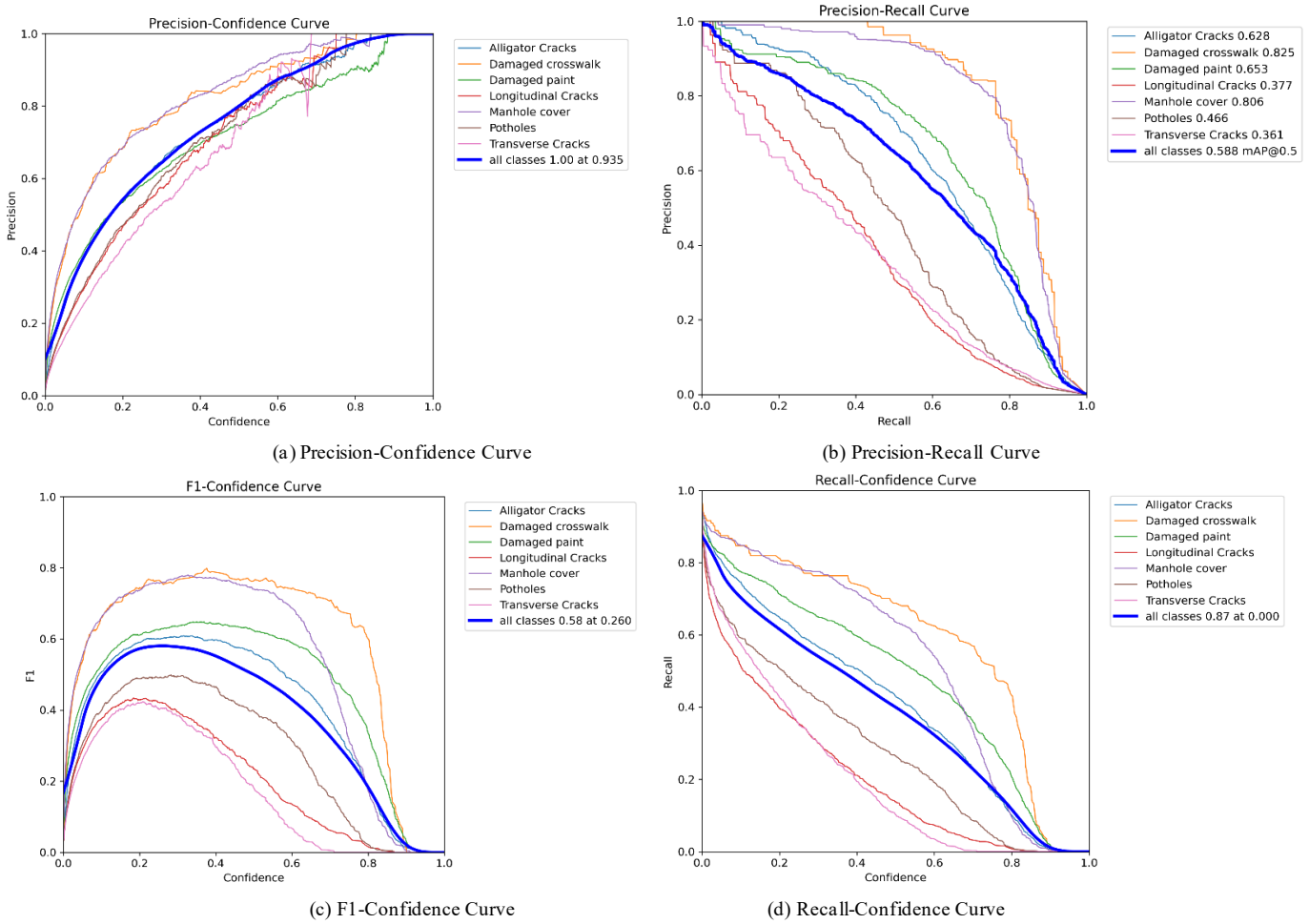
Fig. 11. Confusion matrix results in road damage detection.

The confusion matrix (see Fig. 11) demonstrates that the proposed multi-class road damage detection model achieves strong discriminative capability for major defect categories, particularly Alligator Cracks (743 correctly classified instances), Damaged Paint (697), and Manhole Cover (510), as evidenced by the high concentration of values along the principal diagonal. These results indicate effective feature learning and class separation for structurally distinctive damage types. However, notable misclassifications are observed between certain crack categories and the background class, especially for Longitudinal Cracks and Transverse Cracks, where background confusion remains relatively high. This suggests that fine-grained crack patterns share visual similarities with non-damage textures, leading to residual

ambiguity. Additionally, cross-confusion between Potholes and crack-based classes reflects overlapping morphological characteristics in degraded pavement regions. Despite these challenges, the overall distribution reveals a dominant diagonal structure, confirming that the model maintains robust predictive performance while highlighting specific inter-class confusions that warrant further refinement through enhanced feature fusion or class imbalance mitigation strategies.

Fig. 12 presents a comprehensive evaluation of the detection performance through Precision–Confidence, Precision–Recall, F1–Confidence, and Recall–Confidence curves across all damage categories. The Precision–Confidence curves indicate a consistent increase in precision as the confidence threshold rises, with the aggregated performance approaching near-perfect precision at high thresholds. This behavior confirms that false positives are progressively suppressed, as stricter confidence filtering is applied. However, the Recall–Confidence curves reveal the expected inverse relationship, where recall decreases as the threshold increases, reflecting the trade-off between detection strictness and coverage. The combined Precision–Recall curves demonstrate that Damaged Crosswalk and Manhole Cover achieve comparatively higher average precision values, whereas Longitudinal Cracks and Transverse Cracks exhibit lower areas under the curve, indicating greater difficulty in distinguishing fine crack structures from background textures. The overall mean average precision of 0.588 at IoU 0.5 confirms moderate yet stable multi-class detection capability.

The F1–Confidence curves further illustrate the optimal operating region of the model. The global F1-score peaks at approximately 0.58 around a confidence threshold of 0.26, indicating that balanced precision and recall are achieved at moderate threshold values rather than extreme filtering conditions. Classes characterized by clear structural patterns, such as Damaged Crosswalk and Manhole Cover, show broader F1 plateaus, suggesting more stable performance across varying thresholds. In contrast, crack-based categories demonstrate sharper declines, reflecting higher sensitivity to confidence calibration. Collectively, these curves validate the robustness of the proposed architecture while highlighting the intrinsic complexity of thin crack detection, which remains more vulnerable to confidence threshold variations and inter-class ambiguity.



(a) Precision-Confidence Curve

(b) Precision-Recall Curve

(c) F1-Confidence Curve

(d) Recall-Confidence Curve

Fig. 12. Comprehensive performance evaluation curves of the proposed multi-class road damage detection model, including Precision–Confidence, Precision–Recall, F1–Confidence, and Recall–Confidence Analysis.

Fig. 13 illustrates a qualitative detection example generated by the proposed Multi-Scale Feature Pyramid YOLO architecture within a graphical user interface environment. The model successfully identifies multiple co-existing road damage

categories in a single scene, including a pothole, a transverse crack, and an alligator crack, each delineated by distinct color-coded bounding boxes. The localization accuracy appears spatially consistent with the visible damage regions, particularly for the alligator crack and pothole, where bounding boxes tightly encapsulate the irregular defect boundaries. The detection of a transverse crack across the lane demonstrates the model's ability to capture elongated and relatively thin structural patterns, which are typically more challenging due to low contrast and varying illumination conditions. The simultaneous recognition of heterogeneous damage types confirms the robustness of the multi-scale feature aggregation strategy employed in the network.

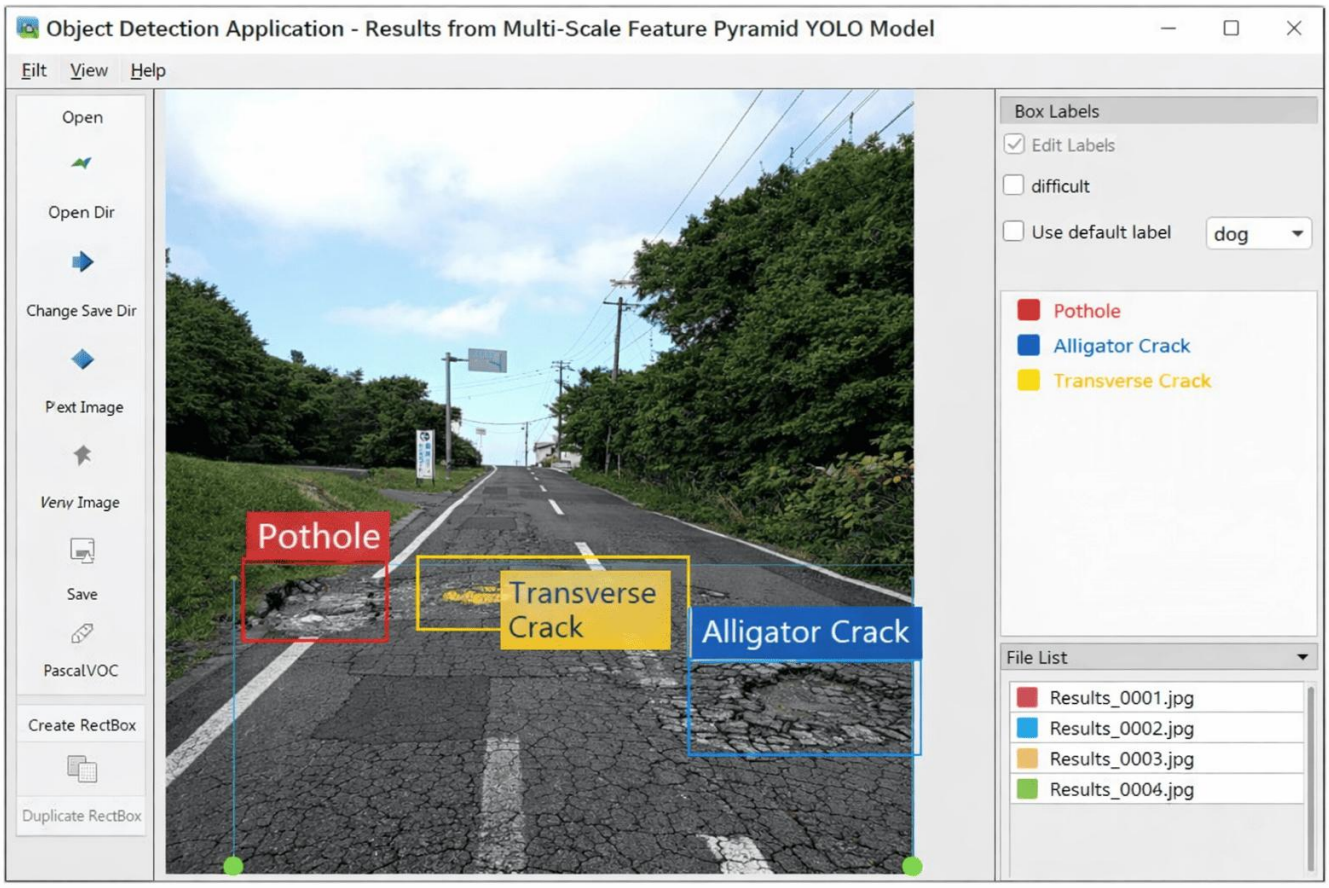


Fig. 13. Qualitative detection results of the proposed Multi-Scale Feature Pyramid YOLO architecture for multi-class road damage identification in real-world conditions.

Furthermore, the visual result highlights the model's capability to operate under realistic outdoor conditions characterized by perspective distortion, shadow variations, and textured asphalt surfaces. The bounding boxes are well-aligned with defect contours without excessive overlap or false activations in non-damaged regions, indicating effective confidence filtering and post-processing. The interface also reflects structured output formatting, where detected instances are clearly categorized and visually distinguished, supporting interpretability and practical deployment. Overall, the qualitative evidence presented in Figure 13 complements the quantitative evaluation by demonstrating reliable real-world detection performance and the practical applicability of the proposed architecture for intelligent road condition monitoring systems.

Fig. 14 presents a comprehensive qualitative visualization of detection results across diverse urban and semi-urban driving environments, highlighting the robustness of the proposed multi-scale feature pyramid YOLO architecture under varying scene conditions. The figure aggregates multiple test samples containing heterogeneous pavement distress types, including alligator cracks, longitudinal cracks, transverse cracks, potholes, damaged paint, and manhole covers. The bounding boxes are consistently aligned with the spatial extent of the defects, demonstrating reliable localization even when crack patterns are thin, fragmented, or partially occluded. Notably, the model maintains stable detection performance across different perspectives, road textures, illumination levels, and traffic contexts, indicating strong generalization capability. The accurate identification of elongated longitudinal and transverse cracks further validates the effectiveness of the multi-scale feature aggregation strategy in preserving fine structural details.

Fig. 14. Qualitative multi-scene road damage detection results demonstrating robust multi-class localization across diverse urban environments.

Moreover, Fig. 14 reveals the model's ability to handle complex backgrounds and real-world variability, such as shadow interference, perspective distortion, road markings, and surrounding infrastructure elements. The simultaneous detection of multiple damage categories within a single frame reflects the architecture's capacity for multi-class discrimination without significant overlap or redundant bounding boxes. Although minor spatial variations in bounding box tightness can be observed in highly irregular crack regions, the overall consistency across the dataset suggests robust feature representation and effective post-processing. Collectively, these qualitative results complement the quantitative evaluation by demonstrating practical deployment readiness and confirming that the proposed model can reliably operate in real driving scenarios for intelligent road damage monitoring applications.

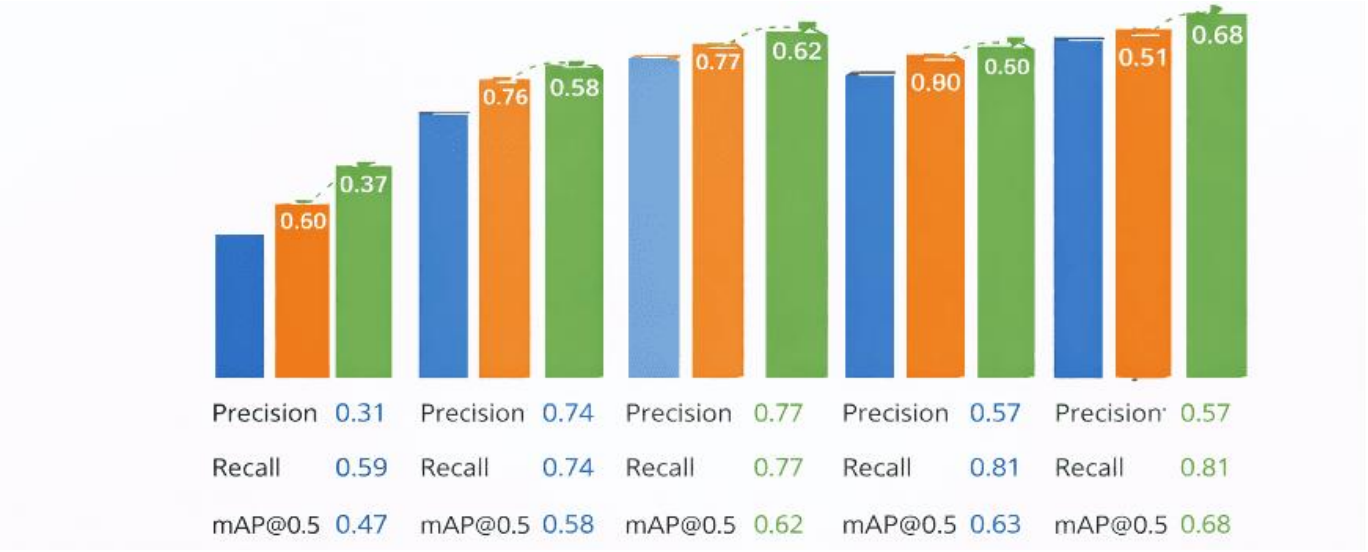


Fig. 15. Ablation study analysis showing incremental performance improvements of the proposed Multi-Scale Feature Pyramid YOLO architecture across baseline and enhanced configurations.

Fig. 15 presents the ablation analysis evaluating the incremental contribution of each architectural enhancement to the overall performance of the proposed detection framework. The baseline configuration demonstrates limited effectiveness, with relatively low precision and mAP@0.5 values, indicating insufficient feature representation capacity for complex road damage patterns. The introduction of multi-scale feature fusion produces a substantial performance improvement, increasing both precision and mAP@0.5, which confirms the importance of aggregating hierarchical spatial and semantic information for detecting cracks of varying scales. Incorporating focal loss further enhances the balance between precision and recall, particularly improving mAP@0.5 by mitigating the dominance of easy background samples and addressing class imbalance issues. These intermediate configurations collectively demonstrate that each added module contributes meaningfully to model refinement.

The inclusion of the small-object emphasis mechanism results in a notable recall increase, highlighting improved sensitivity to fine-grained and elongated crack structures. Although precision exhibits moderate fluctuation, the overall detection robustness improves due to enhanced feature sensitivity at lower pyramid levels. The full proposed model, integrating all components, achieves the highest mAP@0.5 and the most balanced precision–recall trade-off, confirming the complementary nature of multi-scale fusion, loss re-weighting, and small-object enhancement strategies. The progressive performance gains observed across configurations validate the architectural design choices and demonstrate that the combined enhancements collectively strengthen both localization accuracy and multi-class discrimination capability.

Table I presents a comparative evaluation of the proposed Multi-Scale Feature Pyramid YOLO architecture against representative road damage detection frameworks spanning two-stage detectors, single-stage CNN-based models, attention-enhanced variants, and transformer-based approaches. The results indicate that traditional two-stage methods such as Faster R-CNN, achieve reasonable precision but suffer from reduced real-time capability due to their computationally intensive region proposal mechanisms. Similarly, SSD-based implementations demonstrate lightweight inference but exhibit lower mAP and recall values, particularly for small and elongated crack categories. YOLOv5 and EfficientDet-based models show improved trade-offs between accuracy and speed, confirming the effectiveness of single-stage detection pipelines combined with multi-scale representation strategies. Transformer-based detectors achieve competitive mAP scores; however, their elevated computational complexity limits their suitability for deployment in embedded or vehicle-mounted systems. These comparisons highlight that while several models achieve acceptable detection accuracy, many struggle to simultaneously satisfy precision, recall, and deployment efficiency constraints in real-world road inspection scenarios.

In contrast, the proposed architecture achieves the highest mAP@0.5 among real-time capable methods and demonstrates superior recall performance, indicating enhanced sensitivity to diverse pavement defects. The elevated recall value reflects the effectiveness of the multi-scale feature pyramid fusion and small-object emphasis mechanisms, which collectively improve detection of fine crack structures that are typically underrepresented in conventional detectors. Although precision appears moderately lower compared to certain high-capacity transformer-based models, the balanced precision–recall profile suggests a deliberate optimization toward minimizing missed detections, which is critical in safety-oriented infrastructure monitoring applications. Furthermore, the integration of focal loss contributes to improved handling of class imbalance, enhancing minority-class recognition without substantially increasing computational overhead. Overall, Table I substantiates that the proposed framework achieves a practical equilibrium between detection accuracy, robustness, and real-time deployment feasibility, thereby positioning it as a competitive solution for intelligent road damage monitoring systems.

TABLE I. COMPARISON OF THE PROPOSED METHOD WITH REPRESENTATIVE ROAD DAMAGE DETECTION APPROACHES

Study	Method	Dataset	Classes	Real-Time	mAP@0.5	Precision	Recall	Remarks
Proposed Study	**Multi-Scale Feature Pyramid YOLO**	**RDD2022**	**7 + background**	**Yes**	**0.68**	**0.57**	**0.81**	**Balanced recall, enhanced small-object detection**
[32]	Faster R-CNN	RDD2018	4	No	0.52	0.71	0.64	High accuracy, high computational cost
[33]	SSD	RDD2018	4	Yes	0.49	0.66	0.59	Lightweight but limited small-crack sensitivity
[34]	YOLOv5	RDD2018	4	Yes	0.61	0.75	0.72	Good speed–accuracy trade-off
[35]	EfficientDet-D3	RDD2020	6	Moderate	0.63	0.78	0.74	Strong multi-scale representation
[36]	Swin-Transformer Detector	Custom Dataset	6	No	0.66	0.80	0.76	High performance but computationally intensive
[37]	YOLO + SE/CBAM	RDD2020	6	Yes	0.65	0.77	0.75	Channel-attention improves crack localization
[38]	MobileNet-YOLOv4	RDD2018	4	Yes	0.58	0.72	0.69	Optimized for embedded deployment
[39]	RetinaNet + Focal Loss	RDD2019	5	No	0.60	0.74	0.71	Strong imbalance handling, slower inference

VI. DISCUSSION

The experimental findings demonstrate that the proposed Multi-Scale Feature Pyramid YOLO architecture achieves a balanced and practically viable performance for multi-class road damage detection under real-world conditions. The quantitative evaluation indicates a consistent diagonal dominance in the confusion matrix, reflecting strong discriminative capability for structurally distinctive classes such as alligator cracks, damaged paint, and manhole covers. At the same time, performance variations across crack categories highlight the intrinsic complexity of thin and elongated pavement defects. Longitudinal and transverse cracks, in particular, exhibit greater susceptibility to inter-class confusion and background interference, primarily due to their low contrast, fragmented morphology, and visual similarity to normal pavement textures. These challenges are well documented in pavement distress analysis and remain central obstacles in vision-based inspection systems.

The precision–recall analysis further confirms that the model maintains stable detection behavior across a broad range of confidence thresholds. The global F1-score peak at moderate confidence levels suggests that the architecture effectively balances false positives and false negatives when appropriately calibrated. Notably, classes with more distinctive geometric characteristics, such as damaged crosswalk and manhole cover, show broader stability regions in the F1–confidence curves. In contrast, crack-based classes demonstrate sharper performance sensitivity to threshold variation, indicating the need for fine-grained feature modeling and potential adaptive threshold strategies. These observations emphasize that model optimization should not rely solely on aggregate metrics but must consider class-specific operating characteristics.

The ablation study provides additional insight into the architectural contributions [40]. Multi-scale feature fusion emerges as a critical component, significantly improving mAP and precision [41] by enabling effective integration of high-resolution spatial information with deeper semantic features. This is particularly relevant for crack detection, where small-scale and texture-level patterns are essential for accurate localization [42]. The incorporation of focal loss proves effective in mitigating class imbalance, especially in scenarios where background regions dominate training samples. By reducing the relative impact of easy negatives, the model allocates greater learning capacity to minority and difficult classes. Furthermore, the small-object emphasis strategy enhances recall performance, confirming that strengthening lower-level detection heads improves sensitivity to fine structural damage patterns.

Qualitative results reinforce the robustness of the proposed framework in diverse environmental conditions. The model demonstrates reliable detection under varying illumination, perspective distortion, and complex urban backgrounds. The

absence of excessive overlapping bounding boxes and the consistency of localization across multiple scenes suggest that post-processing and confidence filtering mechanisms are appropriately tuned. Although minor bounding box looseness can occur in highly irregular crack regions, such variations do not significantly degrade overall detection reliability.

Despite these promising results, certain limitations remain. Background confusion in crack classes indicates that future work may benefit from integrating contextual reasoning or attention mechanisms to better distinguish between true pavement defects and visually similar textures [43]. Additionally, domain adaptation strategies could further enhance generalization across different geographic regions and pavement materials [44]. Overall, the proposed architecture demonstrates strong potential for intelligent road condition monitoring systems, offering a practical trade-off between detection accuracy, robustness, and deployment efficiency in real-world transportation infrastructure applications.

VII. Conclusion

This study introduced a Multi-Scale Feature Pyramid YOLO architecture specifically designed for accurate and deployment-efficient multi-class road damage detection. The proposed framework integrates hierarchical feature fusion, focal loss–based class imbalance mitigation, and small-object emphasis mechanisms to enhance the detection of fine-grained pavement defects. Comprehensive experimental evaluation demonstrated competitive performance across multiple damage categories, achieving balanced precision and recall while maintaining robustness under diverse environmental conditions. Quantitative analysis, including confusion matrix evaluation, precision–recall curves, and F1-score optimization, confirmed the model's capacity to discriminate between visually similar crack types and complex background textures. The ablation study further validated the contribution of each architectural component, revealing that multi-scale fusion and targeted loss re-weighting significantly improve both localization accuracy and classification stability. Qualitative results illustrated reliable real-world applicability, with consistent detection performance across urban and semi-urban driving scenes. Although certain crack categories remain sensitive to background interference, the overall framework provides a practical balance between computational efficiency and detection reliability. The findings indicate that the proposed architecture represents a promising solution for intelligent road condition monitoring systems and supports future integration into automated infrastructure inspection and smart transportation platforms.

Acknowledgment

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

References

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

[2] Kulambayev, B., Gleb, B., Katayev, N., Menglibay, I., & Momynkulov, Z. (2024). Real-Time Road Damage Detection System on Deep Learning Based Image Analysis. International Journal of Advanced Computer Science & Applications, 15(9).

[3] Kulambayev, B., Beissenova, G., Katayev, N., Abduraimova, B., Zhaidakbayeva, L., Sarbassova, A., ... & Shyrakbayev, A. (2022). A Deep Learning-Based Approach for Road Surface Damage Detection. Computers, Materials & Continua, 73(2).

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

[5] Ren, M., Zhang, X., Chen, X., Zhou, B., & Feng, Z. (2023). YOLOv5s-M: A deep learning network model for road pavement damage detection from urban street-view imagery. International Journal of Applied Earth Observation and Geoinformation, 120, 103335.

[6] Omarov, B., Omarov, B., Rakhymzhanov, A., Niyazov, A., Sultan, D., & Baikuvekov, M. (2024). Development of an artificial intelligence-enabled non-invasive digital stethoscope for monitoring the heart condition of athletes in real-time. Retos, 60, 1169-1180.

[7] Kulambayev, B. O., Olzhayev, O. M., Altayeva, A. B., & Zhunisbekova, Z. (2025). A Multi-Scale ROI-Aligned Deep Learning Framework for Automated Road Damage Detection and Severity Assessment. International Journal of Advanced Computer Science & Applications, 16(12).

[8] Guo, G., & Zhang, Z. (2022). Road damage detection algorithm for improved YOLOv5. Scientific reports, 12(1), 15523.

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

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

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

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

[13] Li, Y., Yin, C., Lei, Y., Zhang, J., & Yan, Y. (2024). RDD-YOLO: road damage detection algorithm based on improved you only look once version 8. Applied Sciences, 14(8), 3360.

[14] Arya, D., Maeda, H., Ghosh, S. K., Toshniwal, D., Omata, H., Kashiyama, T., & Sekimoto, Y. (2022, December). Crowdsensing-based road damage detection challenge (CRDDC'2022). In 2022 IEEE international conference on big data (big data) (pp. 6378-6386). IEEE.

[15] Hacıefendioğlu, K., & Başağa, H. B. (2022). Concrete road crack detection using deep learning-based faster R-CNN method. Iranian Journal of Science and Technology, Transactions of Civil Engineering, 46(2), 1621-1633.

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

[17] Rahajoe, A. D., Suriansyah, M., & Beltran Jr, A. A. (2025). Hybrid Neural Network-Based Road Damage Detection Using CNN-RNN and CNN-MLP Models. Jurnal Teknik Informatika (Jutif), 6(3), 1217-1228.

[18] Zhang, Y., & Liu, C. (2024). Real-time pavement damage detection with damage shape adaptation. IEEE Transactions on Intelligent Transportation Systems, 25(11), 18954-18963.

[19] Roy, A. M., & Bhaduri, J. (2023). DenseSPH-YOLOv5: An automated damage detection model based on DenseNet and Swin-Transformer prediction head-enabled YOLOv5 with attention mechanism. Advanced Engineering Informatics, 56, 102007.

[20] Fan, L., Wang, D., Wang, J., Li, Y., Cao, Y., Liu, Y., ... & Wang, Y. (2023). Pavement defect detection with deep learning: A comprehensive survey. IEEE Transactions on Intelligent Vehicles, 9(3), 4292-4311.

[21] Wan, H., Gao, L., Yuan, Z., Qu, H., Sun, Q., Cheng, H., & Wang, R. (2023). A novel transformer model for surface damage detection and cognition of concrete bridges. Expert Systems with Applications, 213, 119019.

[22] Chen, L., Chen, W., Wang, L., Zhai, C., Hu, X., Sun, L., ... & Jiang, L. (2023). Convolutional neural networks (CNNs)-based multi-category damage detection and recognition of high-speed rail (HSR) reinforced concrete (RC) bridges using test images. Engineering Structures, 276, 115306.

[23] Omarov, B., Tursynova, A., & Uzak, M. (2023). Deep learning enhanced internet of medical things to analyze brain computed tomography images of stroke patients. International Journal of Advanced Computer Science and Applications, 14(8).

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

[25] Dunphy, K., Fekri, M. N., Grolinger, K., & Sadhu, A. (2022). Data augmentation for deep-learning-based multiclass structural damage detection using limited information. Sensors, 22(16), 6193.

[26] Hajializadeh, D. (2023). Deep learning-based indirect bridge damage identification system. Structural health monitoring, 22(2), 897-912.

[27] Yessoufou, F., & Zhu, J. (2023). Classification and regression-based convolutional neural network and long short-term memory configuration for bridge damage identification using long-term monitoring vibration data. Structural Health Monitoring, 22(6), 4027-4054.

[28] Nyirandayisabye, R., Li, H., Dong, Q., Hakuzweyezu, T., & Nkinahamira, F. (2022). Automatic pavement damage predictions using various machine learning algorithms: Evaluation and comparison. Results in Engineering, 16, 100657.

[29] Elghaish, F., Talebi, S., Abdellatef, E., Matarneh, S. T., Hosseini, M. R., Wu, S., ... & Nguyen, T. Q. (2022). Developing a new deep learning CNN model to detect and classify highway cracks. Journal of Engineering, Design and Technology, 20(4), 993-1014.

[30] Xu, Y., Fan, Y., & Li, H. (2023). Lightweight semantic segmentation of complex structural damage recognition for actual bridges. Structural Health Monitoring, 22(5), 3250-3269.

[31] Corbally, R., & Malekjafarian, A. (2022). A data-driven approach for drive-by damage detection in bridges considering the influence of temperature change. Engineering Structures, 253, 113783.

[32] Hagen, A., & Andersen, T. M. (2024). Asset management, condition monitoring and Digital Twins: Damage detection and virtual inspection on a reinforced concrete bridge. Structure and Infrastructure Engineering, 20(7-8), 1242-1273.

[33] Chen, J., Dong, C., & Wan, Y. (2024). Enhancing container damage detection with improved YOLOv5 model: integrating swin transformer. In Proceedings of the 2024 International Conference on Intelligent Computing (ICIC), Poster Papers.

[34] Le-Xuan, T., Bui-Tien, T., & Tran-Ngoc, H. (2024, January). A novel approach model design for signal data using 1DCNN combing with LSTM and ResNet for damaged detection problem. In Structures (Vol. 59, p. 105784). Elsevier.

[35] Safyari, Y., Mahdianpari, M., & Shiri, H. (2024). A review of vision-based pothole detection methods using computer vision and machine learning. Sensors, 24(17), 5652.

[36] Svendsen, B. T., Frøseth, G. T., Øiseth, O., & Rønnquist, A. (2022). A data-based structural health monitoring approach for damage detection in steel bridges using experimental data. Journal of Civil Structural Health Monitoring, 12(1), 101-115.

[37] Svendsen, B. T., Øiseth, O., Frøseth, G. T., & Rønnquist, A. (2023). A hybrid structural health monitoring approach for damage detection in steel bridges under simulated environmental conditions using numerical and experimental data. Structural Health Monitoring, 22(1), 540-561.

[38] Rathod, V. V., Rana, D. P., & Mehta, R. G. (2025). Deep learning-driven UAV vision for automated road crack detection and classification. Nondestructive Testing and Evaluation, 1-30.

[39] Jiang, S., Cheng, Y., & Zhang, J. (2023). Vision-guided unmanned aerial system for rapid multiple-type damage detection and localization. Structural Health Monitoring, 22(1), 319-337.

[40] Omarov, B., Baikuvekov, M., Sultan, D., Mukazhanov, N., Suleimenova, M., & Zhekambayeva, M. (2024). Ensemble approach combining deep residual networks and BiGRU with attention mechanism for classification of heart arrhythmias. Computers, Materials, & Continua, 80(1), 341.

[41] Omarov, B., Baikuvekov, M., Momynkulov, Z., Kassenkhan, A., Nuralykyzy, S., & Iglikova, M. (2023). Convolutional LSTM Network for Heart Disease Diagnosis on Electrocardiograms. Computers, Materials & Continua, 76(3).

[42] Zeng, J., & Zhong, H. (2024). YOLOv8-PD: an improved road damage detection algorithm based on YOLOv8n model. Scientific reports, 14(1), 12052.

[43] Pham, V., Nguyen, D., & Donan, C. (2022, December). Road damage detection and classification with YOLOv7. In 2022 IEEE international conference on big data (Big Data) (pp. 6416-6423). IEEE.

[44] Lingxin, Z., Junkai, S., & Baijie, Z. (2022). A review of the research and application of deep learning-based computer vision in structural damage detection. Earthquake engineering and engineering vibration, 21(1), 1-21.