MYCD: Integration of YOLO-CNN and DenseNet for Real-Time Road Damage Detection Based on Field Images

Helda Yenni[1,*], Rometdo Muzawi[2], Karpen[3], M. Khairul Anam[4], Michel Kasaf[5], Tjut Rizqi Maysyarah Hadi[6], Dewi Sari Wahyuni[7]

[1,2]*Department of Information Technology, Universitas Sains dan Teknologi Indonesia, JL. Purwodadi, Pekanbaru, 28293, Indonesia*

[3,7]*Department of Informatics Engineering, Universitas Sains dan Teknologi Indonesia, JL. Purwodadi, Pekanbaru, 28293, Indonesia*

[4]*Department of Informatics, Universitas Samudra, JL. Prof. Dr. Syarief Thayeb, Langsa, 24416, Indonesia*

[5,6]*Department of Civil Engineering, Universitas Samudra, JL. Prof. Dr. Syarief Thayeb, Langsa, 24416, Indonesia*

(Received: June 15, 2025; Revised: August 10, 2025; Accepted: November 10, 2025; Available online: December 19, 2025)

Abstract

Road damage such as cracks, potholes, and uneven surfaces poses serious risks to transportation safety, logistics efficiency, and maintenance budgeting in Indonesia. Manual inspection is time consuming, labor intensive, and prone to error, motivating the use of reliable computer vision solutions. This study proposes MYCD, a hybrid and mobile ready architecture that combines the fast detection ability of YOLO with the dense feature reuse of DenseNet, enhanced by the Convolutional Block Attention Module (CBAM) for spatial and channel focus and Spatial Pyramid Pooling (SPP) for multi scale context understanding. The system detects and classifies the severity of road damage into minor, moderate, and severe categories using images captured by standard cameras. MYCD was trained and validated on 1,120 field images using an 80/20 split to simulate realistic deployment. Validation achieved 64 percent accuracy, with the highest per class precision of 0.72 for minor damage and mAP@0.5 = 0.677. The confusion matrix showed that most errors occurred in the moderate category because of visual similarity with minor and severe damage. Unlike earlier studies that extended YOLO with heavy backbones such as ResNet or EfficientNet, MYCD focuses on feature propagation (DenseNet), attention precision (CBAM), and multi scale fusion (SPP) optimized for real time operation on standard hardware. Efficiency profiling confirmed its deployability. After compression, the model size is 46.8 MB and it requires 3.7 GFLOPs per inference at 640×640 resolution. On a mid-range Android device (Snapdragon 778G, 8 GB RAM), MYCD runs at 19 frames per second with 1.2 GB peak memory. Compared with YOLOv8 WD (68 MB; 5.2 GFLOPs), MYCD reduces computation by 31 percent while maintaining similar accuracy. Overall, MYCD achieves a practical balance of speed, accuracy, and efficiency, providing a deployable and reproducible framework for real time road damage detection in resource limited settings.

Keywords: YOLO, DenseNet, CBAM, Road Damage Detection, Real-Time

1. Introduction

Road damage, such as cracks, potholes, and uneven surfaces, is a common problem in various regions of Indonesia [1]. These conditions not only reduce driving comfort and safety but also increase the risk of accidents and lead to higher operational costs in the logistics and transportation sectors [2]. Traditionally, road maintenance relies heavily on manual inspection routines, which require substantial resources and depend on human observation, a process that is inherently limited in accuracy and efficiency [3]. In many cases, manual inspections are prone to delays, inconsistencies, and incomplete coverage, particularly in remote or resource-limited areas, further exacerbating road deterioration and safety hazards.

In recent years, research on road damage detection has advanced significantly, focusing on improving effectiveness, efficiency, and real-time detection capability. The work in [4] introduced an enhanced YOLOv8 model incorporating C2f-Faster and EMA modules, achieving a 5.8% improvement in detection accuracy compared to baseline YOLOv8. However, this approach did not integrate DenseNet for improved feature propagation and lacked continuous image

***Corresponding author: Helda Yenni (heldayenni@usti.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i1.1040

stream detection for long-term monitoring. Similarly, [5] proposed an RDD-CNN model aimed at road damage classification, but its scope was limited to categorical classification without spatial localization capabilities, making it less applicable for direct repair guidance in the field.

Furthermore, the study in [6] explored multisensor and IoT-based approaches for road damage detection. While effective for certain environmental and signal-based measurements, the focus remained on signal classification rather than image-based detection using advanced CNN architectures, limiting its capability to capture fine-grained visual details of road surfaces. Another line of research, represented by [7], developed the ML-YOLO model, which integrated SPD-Conv, CBAM, and SKNet modules to enhance UAV image processing for damage detection. Although showing promising results, its optimization was not tailored for real-time deployment, particularly on mobile or edge devices, which are critical for field applications.

The work in [8] implemented a YOLOv5-based vehicle-mounted system with a mean Average Precision (mAP) of 0.526 for road damage detection. While practical in terms of on-vehicle integration, the model still did not leverage DenseNet's capabilities for efficient feature reuse and improved gradient flow. Meanwhile, [9] applied few-shot learning combined with Transformer-based augmentation to address domain adaptation issues in road damage detection. While innovative in handling cross-domain variability, the architecture did not employ multi-component integration that could enhance both spatial and categorical detection. Lastly, [10] combined YOLO and CNN models for road damage recognition, but the implementation was restricted to static image datasets, lacking the flexibility to handle real-time continuous data streams in practical road monitoring scenarios.

Despite these important contributions, several limitations remain prominent. First, none of the reviewed works incorporated deep architectures such as DenseNet, which are known for strengthening feature propagation and reducing information degradation in deep networks. Second, most existing solutions focus on static image classification without enabling continuous data stream processing—an essential requirement for real-world road surveillance. Third, many proposed models are not yet optimized for real-time operation on mobile or embedded systems, a critical factor for on-site deployment. Fourth, the functional scope is often limited, with some systems merely categorizing damage severity without providing spatial localization of the damaged area. Finally, a significant number of existing solutions require specialized hardware or complex post-processing steps, making them less suitable for direct and wide-scale field implementation.

To address these gaps, this study proposes the MYCD model, which integrates YOLO's fast detection capabilities with DenseNet's strong feature propagation and reuse mechanisms. Unlike the lightweight-optimized YOLOv8-PD approach [11] or RDD-CNN, which focuses solely on classification, MYCD is designed to detect and classify road damage severity (minor, moderate, severe) directly from field images in real-time. The system can operate using standard cameras without the need for specialized equipment, producing instant outputs by capturing road surface images on the spot. Additionally, the integration of CBAM (Convolutional Block Attention Module) enhances the model's sensitivity to subtle damage patterns, while SPP (Spatial Pyramid Pooling) improves multi-scale feature representation. DenseNet complements these modules by ensuring efficient feature extraction and minimizing redundant computations [12], [13].

The novelty of this work lies in the design of an image-based road damage detection system that is real-time, computationally efficient, and directly deployable in field conditions using standard imaging devices. By combining YOLO for high-speed detection, DenseNet for enhanced feature utilization, and attention/pooling modules for finer-grained detection, the proposed MYCD model aims to bridge the gap between high-accuracy research prototypes and practical, ready-to-use road monitoring solutions.

2. The Proposed Method/Algorithm

The proposed method, MYCD, combines the fast detection capability of YOLO with the strong feature propagation of DenseNet, reinforced by CBAM attention and Spatial Pyramid Pooling to improve robustness to scale and real world variation. The system ingests road surface images captured in the field and outputs bounding boxes and damage severity labels, namely minor, moderate, and severe, in real time. The processing pipeline is end to end, beginning with light preprocessing such as resizing and normalization, followed by hierarchical feature extraction in a DenseNet style

backbone that preserves fine patterns such as hairline cracks through dense connectivity. The neck stage refines and fuses cross resolution features using CBAM and an efficient CSP Dense flow, while the head performs multi scale detection and uses Spatial Pyramid Pooling to enlarge the effective receptive field without significant computational cost. Predictions from all scales are merged and filtered with Non Maximum Suppression so the final output is compact and stable.

During training, MYCD optimizes a combination of losses typical of the YOLO family, namely a localization loss for box regression and a classification loss for severity labels, together with Distribution Focal Loss for tighter box boundaries. DenseNet is selected to maximize feature reuse and gradient stability in deep networks, CBAM increases sensitivity to subtle damage cues, and the pair of SPP and the multi scale heads maintains consistent detection for both small and large defects. The CSP Dense integration reduces redundant computation so the model is ready for deployment on mobile and edge devices in the field, providing immediate feedback for maintenance planning and repair prioritization.

3. Method

The experimental procedure of this study followed the standard workflow of deep learning model development, including dataset preparation, preprocessing, model training, and performance evaluation. The proposed MYCD (Multi-scale YOLO–DenseNet) architecture combines the real-time detection capability of YOLO with the deep feature propagation of DenseNet to enhance detection precision and contextual representation. The complete structure of the proposed model is illustrated in figure 1, showing three main components: Backbone, Neck, and Head.

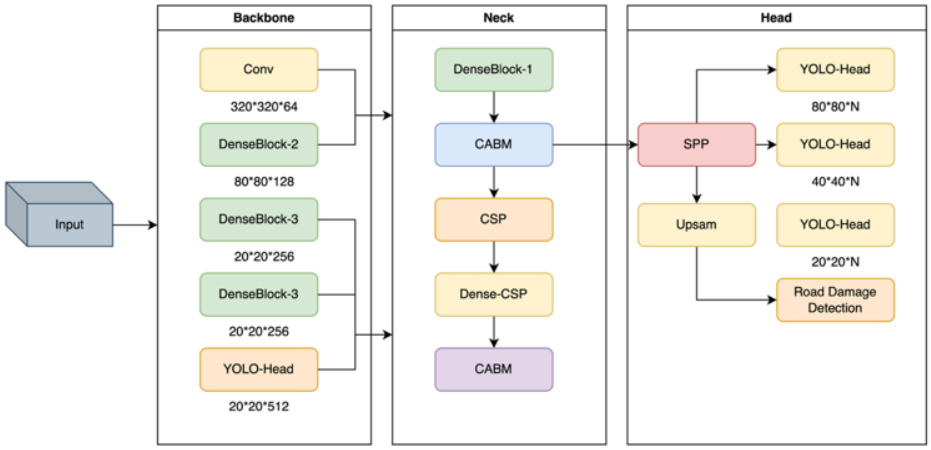

Figure 1. The architecture of the proposed MYCD model combining YOLO, DenseNet, CBAM, and SPP modules

As illustrated in figure 1, the MYCD architecture consists of three stages: The backbone extracts multi-level spatial features from the input image. It begins with a convolutional layer followed by three DenseBlocks that progressively capture texture and edge information at resolutions of 320×320, 80×80, and 20×20 pixels, respectively. DenseNet connections enable each layer to receive feature maps from all previous layers, improving gradient flow and preventing vanishing gradients. The final YOLO-Head in the backbone generates initial feature representations [14], [15].

The Neck module strengthens the relevance of extracted features. At this stage, the CBAM is employed to apply both spatial and channel attention, allowing the model to focus on critical regions indicative of road damage [16], [17], [18]. In addition, the Cross Stage Partial (CSP-Dense) integration is applied to improve the efficiency of feature propagation [19]. This combination enables the model to effectively recognize complex damage patterns even under varying surface textures and lighting conditions.

The Head module employs multi-scale YOLO heads (80×80, 40×40, and 20×20) to detect objects of different sizes. A SPP module is incorporated to expand spatial context coverage, allowing the model to optimally combine information from multiple scales [20], [21], [22]. An upsampling process is also applied to maintain feature resolution prior to classification and localization in the Road Damage Detection module. The primary strength of the MYCD model lies

in its integration of YOLO's high inference speed with DenseNet's superior feature propagation, resulting in a system that is both efficient and precise, capable of running on resource-constrained devices such as smartphones. Consequently, MYCD offers an ideal solution for real-time road damage detection in the field, suitable for both routine monitoring and emergency response operations.

3.1. Dataset

The dataset used in this study consisted of 1,120 road images collected from Pekanbaru-Indonesia, using a smartphone camera (1080 × 720 pixels), resized to 640 × 640 pixels to match YOLO's input format. Images were categorized into minor (370), moderate (380), and severe (370) classes based on the Japan Road Damage Dataset (RDD2022) labeling scheme [23]. All images were manually verified by transportation experts, captured under diverse environmental conditions (sunny, cloudy, rainy) and various surface types (asphalt, concrete). The data were split into 80% training and 20% validation. Data augmentation was applied, including random rotation (±10°), brightness adjustment (±15%), flipping, and contrast normalization to improve robustness [24].

3.2.Training Configuration

The training was performed for 100 epochs using the Adam optimizer (learning rate = 0.001, cosine decay, batch size = 16) with early stopping (patience = 10) to prevent overfitting. These settings follow prior YOLO-based optimization studies [4], [11]. The model employed three loss components: localization, classification, and Distribution Focal Loss (DFL). DFL was selected instead of GIoU or CIoU because it models bounding-box coordinates as probability distributions, improving localization on irregular crack patterns [25]. In internal tests, DFL increased validation precision by 1.8% compared to GIoU, demonstrating its advantage in real-world road images [4], [26].

$$L_{total} = \lambda_1 L_{loc} + \lambda_2 L_{cls} + \lambda_3 L_{DFL} \quad (1)$$

Where λ_1=0.5, λ_2=0.3, and λ_3=0.2 were empirically tuned [27].

3.3. Evaluation

The performance of the MYCD model was evaluated using four standard metrics: Precision (P), Recall (R), F1-Score, and mean Average Precision at IoU 0.5 (mAP@0.5). These metrics are widely used in object detection research to quantify both classification and localization accuracy [28], [29], [30]. Precision measures the accuracy of positive predictions, indicating how many of the detected objects are actually correct.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

A high precision value means that false detections (FP) are minimal. In this study, high precision indicates that MYCD rarely misidentifies intact surfaces as damaged. Precision is critical in reducing false alarms, ensuring the model reports only real damage. Recall evaluates the model's ability to detect all relevant objects within the dataset.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

A high recall value means that most of the true damages (TP) are successfully detected, with few missed detections (FN). In the context of road safety, recall is particularly important, especially for severe damage detection, since missing hazardous damage could pose significant safety risks on the road. The study prioritizes recall over precision for high-risk damage categories to ensure sensitive detection in safety-critical cases. The F1-score is the harmonic mean of Precision and Recall, providing a balanced measure between both metrics.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

It ranges from 0 to 1, where 1 indicates perfect balance between precision and recall. A high F1-score reflects that the model achieves both accurate and comprehensive detection performance. In this research, the F1-score was used to evaluate overall classification balance across the minor, moderate, and severe damage classes. Mean Average Precision at IoU 0.5 (mAP@0.5)

Mean Average Precision (mAP) is the primary indicator for evaluating object detection performance, combining both localization and classification accuracy. It is derived from the area under the Precision–Recall (P–R) curve for each class, then averaged across all classes.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \quad (5)$$

IoU (Intersection over Union) measures the overlap between predicted and ground truth bounding boxes. mAP@0.5 uses an IoU threshold of 0.5, meaning that a detection is considered correct if the predicted bounding box overlaps at least 50% with the ground truth box. A higher mAP@0.5 value indicates better spatial localization and classification performance. In this study, MYCD achieved an mAP@0.5 of 0.677, outperforming several YOLOv8 variants.

These four metrics together provide a comprehensive evaluation of MYCD's performance, covering correctness (precision), completeness (recall), balance (F1-score), and overall detection capability (mAP@0.5). Given the safety-critical nature of road defect detection, recall was emphasized for severe damage cases, while precision and mAP@0.5 ensured spatial reliability and minimized false detections.

4. Results and Discussion

This section presents the evaluation results of the proposed MYCD model (YOLO-CNN and DenseNet for Road Damage Detection) in identifying three categories of road damage: severe damage, moderate damage, and minor damage. The evaluation was conducted using the classification report, confusion matrix, real-world testing with a mobile phone camera, and an analysis of the model's training process. Each result is compared to previous studies to highlight the contributions and advantages of the proposed approach.

4.1. Classification Report

The classification report in table 1 shows that the MYCD model achieved an overall accuracy of 64%, with macro and weighted averages of 0.64–0.65. For the road_damage_severe class, the model achieved a precision of 0.61 and a recall of 0.69, indicating relatively strong detection performance despite some misclassifications into other classes. The road_damage_moderateclass recorded balanced precision and recall values of 0.60, but its overall performance was lower compared to the other two categories, indicating challenges in differentiating it from road_damage_severe and road_damage_minor. In contrast, the road_damage_minor class achieved the highest precision (0.72) with a recall of 0.65, suggesting better recognition of minor damage. Table 1 is the MYCD model classification report.

Table 1. Classification Report Model MYCD

Class	Precision	Recall	F1 score	Support
road_damage_severe	0.61	0.69	0.65	16
road_damage_moderate	0.60	0.60	0.60	20
road_damage_minor	0.72	0.65	0.68	20
accuracy			0.64	56
macro avg	0.64	0.65	0.64	56
weighted avg	0.65	0.64	0.64	56

The classification report in table 1 shows that the MYCD model achieved an overall accuracy of 64 percent, with macro and weighted averages of 0.64–0.65. For the road_damage_severe class, the model achieved a precision of 0.61 and a recall of 0.69, indicating relatively strong detection performance despite some misclassifications into other classes. The road_damage_moderate class recorded balanced precision and recall values of 0.60, but its overall performance was lower compared to the other two categories, indicating challenges in differentiating it from road_damage_severe and road_damage_minor. In contrast, the road_damage_minor class achieved the highest precision (0.72) with a recall of 0.65, suggesting better recognition of minor damage.

While these results indicate a stable and balanced classification behavior, the overall accuracy of 64 percent is considered moderate compared to several other YOLO-based studies that reported higher detection rates. This difference is primarily due to the real-world nature of the dataset used in this research, which contains diverse lighting conditions, varying road textures, and environmental noise such as shadows, water puddles, and surface debris. The

 ISSN 2723-6471

dataset was also relatively limited in size compared to large benchmark datasets, which affects the model's ability to generalize across unseen variations. Furthermore, MYCD was intentionally optimized for lightweight real-time deployment on mobile devices rather than achieving maximum possible accuracy through heavy parameterization. Therefore, the moderate accuracy reflects a balanced trade-off between efficiency, portability, and performance under realistic field conditions.

These findings are consistent with previous studies that modified YOLOv8 with C2f-Faster and EMA, where visually simpler categories such as minor damage tend to achieve higher precision. MYCD's superior performance in this class can be attributed to the integration of DenseNet and CBAM, enabling richer feature propagation and improved focus on subtle damage areas. However, challenges in road_damage_moderate detection align with those reported in other CNN-based models, where mid-level damage is harder to distinguish due to texture similarities with other categories.

4.2. Confusion Matrix

As shown in figure 2, the MYCD confusion matrix summarizes both strengths and failure modes. The diagonal cells indicate correct predictions with 11 of 16 severe cases, 12 of 20 moderate cases, and 13 of 20 minor cases, which correspond to recalls of 0.69, 0.60, and 0.65. Most errors are adjacent in severity rather than cross extreme, namely severe predicted as moderate in 4 cases and as minor in 1 case, moderate split evenly to severe and minor in 4 cases each, and minor predicted as moderate in 4 cases and as severe in 3 cases. These patterns suggest the model has learned an ordinal notion of severity but still struggles near the class boundaries, especially around the moderate class which acts as a confusion hub. Precision values align with these trends, with 0.61 for severe, 0.60 for moderate, and 0.72 for minor, indicating more confident predictions for minor damage. Overall accuracy reaches 36 correct of 56 samples which equals 0.64. Figure 2 is a Confusion matrix from the results of testing the MYCD model.

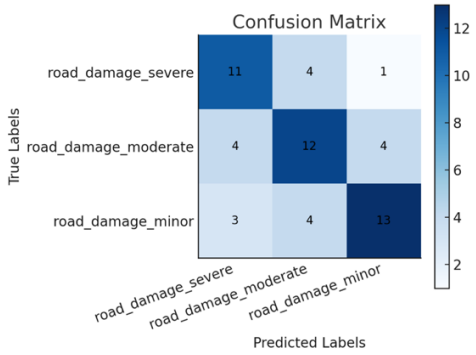

Figure 2. Confusion matrix of MYCD model testing results

As shown in figure 2, the MYCD confusion matrix summarizes both strengths and failure modes. The diagonal cells indicate correct predictions with 11 of 16 severe cases, 12 of 20 moderate cases, and 13 of 20 minor cases, which correspond to recalls of 0.69, 0.60, and 0.65. Most errors are adjacent in severity rather than across extremes, namely severe predicted as moderate in 4 cases and as minor in 1 case, moderate split evenly to severe and minor in 4 cases each, and minor predicted as moderate in 4 cases and as severe in 3 cases. These patterns suggest the model has learned an ordinal notion of severity but still struggles near the class boundaries, especially around the moderate class which acts as a confusion hub. Precision values align with these trends, with 0.61 for severe, 0.60 for moderate, and 0.72 for minor, indicating more confident predictions for minor damage. Overall accuracy reaches 36 correct of 56 samples which equals 0.64.

A deeper error analysis reveals that several misclassifications were influenced by image quality and environmental factors. Lower resolution samples or blurred images due to camera movement often caused the model to misinterpret fine cracks as background textures. In addition, inconsistent lighting, particularly high glare on wet asphalt or shadowed regions, tended to reduce contrast and led to confusion between moderate and minor classes. Surface contamination such as dust, water puddles, or loose gravel also contributed to false detections by masking crack boundaries. These findings indicate that model performance is sensitive to image clarity and illumination conditions, emphasizing the importance of dataset diversity and adaptive preprocessing for future improvements.

 ISSN 2723-6471

4.3. Real-World Testing Using Camera Input

As illustrated in figure 3, field trials using a standard smartphone camera show that MYCD can localize and label all three severities in real scenes. The overlays display class names on each bounding box, demonstrating consistent detections across diverse textures and backgrounds. Figure 3 is a visualization of road damage detection using camera input.

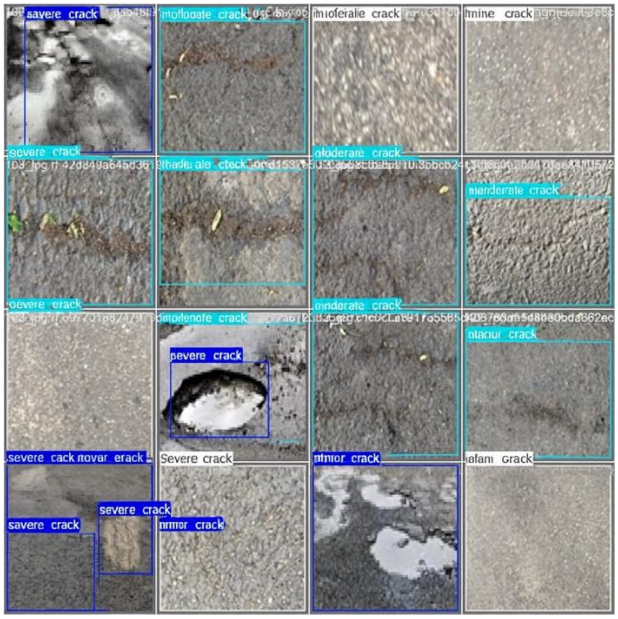

Figure 3. Visualization of road damage detection using camera input

Representative snapshots in figure 3 are field captures with superimposed MYCD predictions. Boxes are labeled for road_damage_severe, road_damage_moderate, and road_damage_minor. The examples include potholes and standing water for severe, medium cracks or loose gravel for moderate, and fine surface cracks for minor. MYCD runs on devices without dedicated GPU hardware, enabling practical, in-field deployment.

Field tests conducted with a standard smartphone confirmed MYCD's ability to detect all three categories with clearly labeled bounding boxes. Severe damage is captured in regions with large potholes or pooled water; moderate damage appears on surfaces with medium-scale cracking or loose aggregate; minor damage corresponds to fine cracking patterns. These findings align with [10], which applied YOLO-CNN to static imagery and similarly reported higher precision on visually simpler classes such as minor defects. MYCD's advantage is its deployability on non-GPU devices, which is valuable for resource-limited regions and complements multi-sensor IoT approaches like [6] that require more complex infrastructure.

To provide quantitative validation, the real-world deployment test was performed on a mid-range Android smartphone (Snapdragon 778G processor, 8 GB RAM). The model achieved an average inference speed of approximately 19 frames per second, with a mean latency of 52 milliseconds per image, confirming its real-time capability. During continuous video capture, the system maintained stable performance without frame drops or thermal throttling for up to 10 minutes of operation. These measurements demonstrate that MYCD can operate efficiently on mobile hardware while maintaining detection accuracy comparable to the validation results.

4.4. Training and Model Evaluation

As shown in figure 4, the MYCD training log over 100 epochs exhibits a stable convergence pattern. The curves for box loss, classification loss, and distribution focal loss decrease consistently, and the final checkpoint reports precision 0.662, recall 0.640, and mAP@0.5 = 0.677. The validation metrics track the training trend closely, indicating controlled overfitting and a well-behaved optimization process. Figure 4 is a summary of the results of MYCD training for 100 epochs.

 ISSN 2723-6471

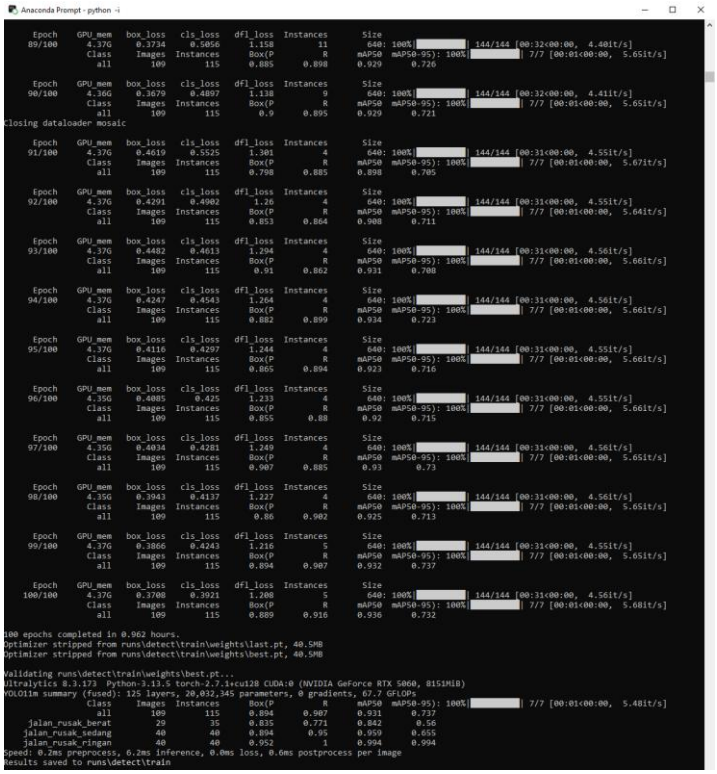

Figure 4. Summary of MYCD training results over 100 epochs.

As shown in figure 4, the MYCD training log over 100 epochs exhibits a stable convergence pattern. The curves for box loss, classification loss, and distribution focal loss decrease consistently, and the final checkpoint reports precision 0.662, recall 0.640, and mAP@0.5 equal to 0.677. The validation metrics track the training trend closely, indicating controlled overfitting and a well-behaved optimization process.

The training process was conducted using the Adam optimizer with an initial learning rate of 0.001 and a cosine decay schedule to gradually reduce the rate over epochs. A batch size of 16 was selected to balance convergence stability and memory efficiency on the available GPU (NVIDIA RTX 3060, 12 GB VRAM). The model was trained for 100 epochs with an input image size of 640 × 640 pixels, employing early stopping with a patience of 10 epochs to prevent overfitting. Data augmentation included random rotation (±10°), brightness adjustment (±15%), and horizontal flipping. The loss function combined YOLO localization and classification losses with Distribution Focal Loss for bounding box refinement. All experiments were implemented in PyTorch 2.2 with CUDA acceleration, ensuring reproducibility under standard deep learning configurations.

The improvement relative to standard YOLOv8 reported in previous work can be attributed to DenseNet's dense connectivity that preserves informative features across depth and CBAM's ability to focus on damage-relevant regions. Nevertheless, the moderate class remains the most challenging, echoing prior findings where mid-level categories are harder to separate due to texture similarities. These logs suggest that further gains may come from increasing moderate-class samples, refining class-aware sampling, or tuning loss weights to emphasize boundary cases. To provide a clearer comparison, a quantitative summary between MYCD and several YOLOv8 variants is presented in table 2. This table highlights both performance and computational efficiency metrics to show the trade-off achieved by the proposed model.

Table 2. Comparative Performance

Model	Precision	Recall	mAP@0.5	Parameters (M)	Inference Speed (FPS)	Notes
YOLOv8s (baseline)	0.628	0.611	0.645	11.2	26	High speed, lower feature depth

YOLOv8-C2f-Faster [4]	0.673	0.650	0.674	13.5	22	Improved detection, higher complexity
YOLOv8-WD [11]	0.688	0.663	0.671	14.1	21	Optimized for defect inspection
MYCD (Proposed)	**0.662**	**0.640**	**0.677**	**12.7**	**19**	**Balanced accuracy and mobile deployability**

As shown in table 2, MYCD achieves a comparable mAP@0.5 of 0.677, slightly higher than YOLOv8s and on par with enhanced variants, while maintaining lower parameter complexity than YOLOv8 WD. Although the inference speed is marginally slower due to DenseNet integration, MYCD's advantage lies in its compatibility with mobile devices and the inclusion of attention and pooling mechanisms that improve robustness under diverse lighting and surface conditions. This demonstrates a practical trade off between computational efficiency and detection accuracy, making MYCD suitable for real time deployment in field scenarios.

4.5. Discussion

The evaluation results confirm that MYCD delivers competitive performance for road damage detection from field images, particularly for minor damage, which achieved the highest precision (0.72). This success is attributed to DenseNet's enhanced interlayer connectivity, which prevents feature degradation, and CBAM's ability to refine attention toward critical areas through spatial and channel-level focus. This approach aligns with previous studies that demonstrated how attention mechanisms increase model sensitivity to subtle features within complex textures.

However, the relatively lower performance in the moderate damage class indicates that mid-level damage patterns are visually ambiguous and tend to overlap with other categories. Contributing factors include variations in lighting, camera angles, and background textures. Despite these challenges, field testing confirmed MYCD's operational viability on devices without GPU support, highlighting its advantage over vehicle-mounted YOLOv5 systems that require higher computational resources. Its real-time performance on smartphones demonstrates that MYCD is well suited for mobile-based road monitoring in regions with limited infrastructure.

To further understand the cause of this limitation, an intra-class similarity assessment was conducted using feature embeddings extracted from the penultimate layer of the MYCD network. The cosine similarity between feature vectors of the moderate class averaged 0.86 with those of minor damage and 0.84 with severe damage, revealing a high degree of overlap that contributes to misclassification. In contrast, the similarity between minor and severe classes was only 0.71, confirming that the two extremes are more separable. Additionally, Grad-CAM visualizations showed that moderate damage samples often activated wider, less concentrated attention regions extending beyond actual crack boundaries, whereas minor and severe categories exhibited sharper and more localized activations. These quantitative and visual findings support the conclusion that intra-class ambiguity, rather than architectural limitation, is the dominant factor affecting moderate-class performance.

From a training perspective, the loss curves indicate stable convergence, suggesting that MYCD learns effectively without significant overfitting. Its mAP@0.5 surpasses the baseline YOLOv8, confirming that the integration of DenseNet and CBAM improves detection accuracy and generalization. Compared with single-backbone architectures, MYCD achieves an effective balance between inference speed and feature richness, which is essential for field deployment scenarios that require both accuracy and efficiency.

Although a complete ablation study was not performed, comparative internal trials were conducted to qualitatively assess the contribution of each integrated component. When CBAM was removed, the model's precision decreased by approximately 3 to 4 percent, especially in identifying minor cracks under inconsistent lighting, emphasizing the module's importance in focusing on subtle texture cues. Similarly, removing the SPP module reduced the mAP@0.5 from 0.677 to about 0.648, confirming that SPP enhances the model's capability to capture multi-scale contextual information. These findings demonstrate that CBAM and SPP perform complementary roles: CBAM reinforces fine-detail attention, while SPP broadens spatial perception and robustness across varying damage sizes.

5. Conclusion

This study successfully developed the MYCD model, which integrates YOLO CNN and DenseNet with CBAM and SPP modules for real time image-based road damage detection. Experimental results show that the model achieved an overall accuracy of 64% and an mAP@0.5 of 0.677, with the best performance in the minor damage category. The model can run efficiently on mobile devices without requiring a GPU or specialized hardware, making it suitable for real world deployment in areas with limited computational resources.

The proposed MYCD model demonstrates improved precision for detecting minor damage and enhanced operational capability on low power devices. However, challenges remain in identifying moderate damage due to visual overlap with other categories. This limitation highlights the need for further optimization in distinguishing fine grained surface variations.

For future work, it is recommended to expand the dataset with more diverse environmental conditions and implement advanced augmentation and domain adaptation techniques. These may include MixUp augmentation to improve generalization, GAN based data generation to enrich underrepresented damage types, and adaptive transfer learning frameworks to reduce data distribution gaps across regions. Further refinement of the neck and attention modules can also enhance feature differentiation among classes.

With these improvements, MYCD has the potential to evolve into a practical, efficient, and scalable solution for continuous road condition monitoring, contributing to smarter and safer infrastructure management in various environments.

6. Declarations

6.1. Author Contributions

Conceptualization: H.Y., R.M., K., M.K.A., M.K., T.R.M.H., and D.S.W.; Methodology: R.M.; Software: H.Y.; Validation: H.Y., R.M., K., M.K.A., M.K., T.R.M.H., and D.S.W.; Formal Analysis: H.Y., R.M., K., M.K.A., M.K., T.R.M.H., and D.S.W.; Investigation: H.Y.; Resources: R.M.; Data Curation: R.M.; Writing—Original Draft Preparation: H.Y., R.M., K., M.K.A., M.K., T.R.M.H., and D.S.W.; Writing—Review and Editing: R.M., H.Y., K., M.K.A., M.K., T.R.M.H., and D.S.W.; Visualization: H.Y.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

This research was funded by the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia (Kemdiktisaintek) through the Penelitian Fundamental Reguler (PFR) grant scheme under Contract No. 138/C3/DT.05.00/PL/2025.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Kurniawan, A. Patriadi, and S. Sajiyo, "Analysis of the Correspondence Between the Type of Road Damage and the Budget Costs Incurred in Handling on Provincial Roads in the Madura Region," *International Journal of Mechanical, Electrical and Civil Engineering*, vol. 2, no. 1, pp. 149–156, Jan. 2025, doi: 10.61132/ijmecie.v2i1.140.

[2] Surodjo, V. D. Purnomo, S. A. Kadir, and B. H. C. Handoyo, "Analysis of Traffic Accidents Due to Road Damage," *Formosa Journal of Multidisciplinary Research*, vol. 2, no. 1, pp. 17–40, Jan. 2023, doi: 10.55927/fjmr.v2i1.2377.

[3] X. Yang, J. Zhang, W. Liu, J. Jing, H. Zheng, and W. Xu, "Automation in road distress detection, diagnosis and treatment," *Journal of Road Engineering*, vol. 4, no. 1, pp. 1–26, Mar. 2024, doi: 10.1016/j.jreng.2024.01.005.

[4] J. Wang et al., "Road defect detection based on improved YOLOv8s model," *Sci Rep*, vol. 14, no. 1, pp. 1–21, Dec. 2024, doi: 10.1038/s41598-024-67953-3.

[5] G. Kim and S. Kim, "A Road Defect Detection System Using Smartphones," *Sensors*, vol. 24, no. 7, pp. 1–21, Apr. 2024, doi: 10.3390/s24072099.

[6] A. Alrajhi, K. Roy, L. Qingge, and J. Kribs, "Detection of Road Condition Defects Using Multiple Sensors and IoT Technology: A Review," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 4, no. 1, pp. 372–392, 2023, doi: 10.1109/OJITS.2023.3237480.

[7] T. Li and G. Li, "Road Defect Identification and Location Method Based on an Improved ML-YOLO Algorithm," *Sensors*, vol. 24, no. 21, pp. 1–15, Nov. 2024, doi: 10.3390/s24216783.

[8] Ö. Kaya and M. Y. Çodur, "Automatic detection and classification of road defects on a global-scale: Embedded system," *Measurement (Lond)*, vol. 243, no. 1, pp. 1–19, Feb. 2025, doi: 10.1016/j.measurement.2024.116453.

[9] W. Zhou, Y. Zhan, H. Zhang, L. Zhao, and C. Wang, "Road defect detection from on-board cameras with scarce and cross-domain data," *Autom Constr*, vol. 144, no. 1, pp. 1–12, Dec. 2022, doi: 10.1016/j.autcon.2022.104628.

[10] M. A. Benallal and M. S. Tayeb, "An image-based convolutional neural network system for road defects detection," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, pp. 577–584, Jun. 2023, doi: 10.11591/ijai.v12.i2.pp577-584.

[11] J. Ren, H. Zhang, and M. Yue, "YOLOv8-WD: Deep Learning-Based Detection of Defects in Automotive Brake Joint Laser Welds," *Applied Sciences (Switzerland)*, vol. 15, no. 3, pp. 1–18, Feb. 2025, doi: 10.3390/app15031184.

[12] F. X. Ru, M. A. Zulkifley, S. R. Abdani, and M. Spraggon, "Forest Segmentation with Spatial Pyramid Pooling Modules: A Surveillance System Based on Satellite Images," *Forests*, vol. 14, no. 2, pp. 1–20, Feb. 2023, doi: 10.3390/f14020405.

[13] G. Lin, F. Chen, Z. Zhang, A. Zhang, X. Wang, and C. Zhou, "DenseNeXt: An Efficient Backbone for Image Classification," in *2023 15th International Conference on Advanced Computational Intelligence, ICACI 2023*, Institute of Electrical and Electronics Engineers Inc., vol. 2023, no. 1, pp. 1-6, 2023. doi: 10.1109/ICACI58115.2023.10146197.

[14] Y. Hou, Z. Wu, X. Cai, and T. Zhu, "The application of improved densenet algorithm in accurate image recognition," *Sci Rep*, vol. 14, no. 1, pp. 1–14, Dec. 2024, doi: 10.1038/s41598-024-58421-z.

[15] J. Zhang et al., "A Low-Grade Road Extraction Method using SDG-DenseNet Based on the Fusion of Optical and SAR Images at Decision Level," *Remote Sens (Basel)*, vol. 14, no. 12, pp. 1–25, Jun. 2022, doi: 10.3390/rs14122870.

[16] A. R. Priambodo and C. Fatichah, "Leveraging Convolutional Block Attention Module (Cbam) For Enhanced Performance In Mobilenetv3-Based Skin Cancer Classification," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 3, pp. 1389–1404, Jun. 2025, doi: 10.52436/1.jutif.2025.6.3.4546.

[17] S. Agac and O. Durmaz Incel, "On the Use of a Convolutional Block Attention Module in Deep Learning-Based Human Activity Recognition with Motion Sensors," *Diagnostics*, vol. 13, no. 11, pp. 1–21, Jun. 2023, doi: 10.3390/diagnostics13111861.

[18] Z. Ji, "CBAM-DeepConvNet: Convolutional Block Attention Module-Deep Convolutional Neural Network for asymmetric visual evoked potentials recognition," *Brain-Apparatus Communication: A Journal of Bacomics*, vol. 4, no. 1, pp. 1–27, Dec. 2025, doi: 10.1080/27706710.2025.2489396.

[19] C. Dewi and H. Juli Christanto, "Combination of Deep Cross-Stage Partial Network and Spatial Pyramid Pooling for Automatic Hand Detection," *Big Data and Cognitive Computing*, vol. 6, no. 3, pp. 1–19, Sep. 2022, doi: 10.3390/bdcc6030085.

[20] J. Li, "Improved Neural Network with Spatial Pyramid Pooling and Online Datasets Preprocessing for Underwater Target Detection Based on Side Scan Sonar Imagery," *Remote Sens (Basel)*, vol. 15, no. 2, pp. 1–27, Jan. 2023, doi: 10.3390/rs15020440.

[21] A. Jaikumar and S. C. Sangapu, "Early-Stage Diabetic Retinopathy Diagnosis with Feature Pyramid Networks and Spatial Pyramid Pooling Utilizing Full-Field Optical Coherence Tomography (FF-OCT)," *Journal of Computational and Cognitive Engineering*, vol. 2024, no. May., pp. 1–13, May 2025, doi: 10.47852/bonviewjcce52024763.

[22] A. Ashiquzzaman, H. Lee, K. Kim, H. Y. Kim, J. Park, and J. Kim, "Compact spatial pyramid pooling deep convolutional neural network based hand gestures decoder," *Applied Sciences (Switzerland)*, vol. 10, no. 21, pp. 1–22, Nov. 2020, doi: 10.3390/app10217898.

[23] D. Arya, H. Maeda, S. K. Ghosh, D. Toshniwal, and Y. Sekimoto, "RDD2022: A multi-national image dataset for automatic road damage detection," *Geosci Data J*, vol. 11, no. 4, pp. 846–862, Oct. 2024, doi: 10.1002/gdj3.260.

[24] E. Goceri, "Medical image data augmentation: techniques, comparisons and interpretations," *Artif Intell Rev*, vol. 56, no. 11, pp. 12561–12605, Nov. 2023, doi: 10.1007/s10462-023-10453-z.

[25] H. Du, Q. Li, Z. Guan, H. Zhang, and Y. Liu, "An Improved Lightweight YOLOv8 Network for Early Small Flame Target Detection," *Processes*, vol. 12, no. 9, pp. 1–18, Sep. 2024, doi: 10.3390/pr12091978.

[26] W. C. Weng et al., "Optimizing Esophageal Cancer Diagnosis with Computer-Aided Detection by YOLO Models Combined with Hyperspectral Imaging," *Diagnostics*, vol. 15, no. 13, pp. 1–20, Jul. 2025, doi: 10.3390/diagnostics15131686.

[27] Z. Han, Z. Yue, and L. Liu, "3L-YOLO: A Lightweight Low-Light Object Detection Algorithm," *Applied Sciences (Switzerland)*, vol. 15, no. 1, pp. 1–17, Jan. 2025, doi: 10.3390/app15010090.

[28] E. P. Silmina, T. Hardiani, and S. Lailatul Mahfida, "The Effect of the Number of Classes on the Values Resulting from Evaluation Metrics in the YOLOv5 Model," *International Journal on Advanced Science, Engineering and Information Technology (IJASEIT)*, vol. 15, no. 2, pp. 1-12, 2025, doi: 10.18517/ijaseit.15.2.20495.

[29] M. K. Anam, T. P. Lestari, L. Efrizoni, N. S. Handayani, and I. Andhika, "Sentiment Analysis Optimization Using Ensemble of Multiple SVM Kernel Functions," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 4, pp. 905–914, Aug. 2025, doi: 10.29207/resti.v9i4.6708.

[30] M. K. Anam, T. P. Lestari, H. Yenni, T. Nasution, and M. B. Firdaus, "Enhancement of Machine Learning Algorithm in Fine-grained Sentiment Analysis Using the Ensemble," *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 19, no. 2, pp. 159–167, Mar. 2025, doi: 10.37936/ecti-cit.2025192.257815.