YOLO-CBAM: A Lightweight Attention-Guided Deep Learning Framework for Real-Time Road Damage Detection

Olzhas Olzhayev[1], Bakhytzhan Kulambayev[2]*, Azizah Suliman[3],
Assel Rustem[4], Almira Madiyarova[5], Batyrkhan Omarov[6]
International Information Technology University, Almaty, Kazakhstan[1, 4, 6]
Turan University, Almaty, Kazakhstan[2]
Asia Metropolitan University, Subang Jaya Campus, Malaysia[3]
Caspian University of Technology and Engineering named after Sh.Yessenov, Aktau, Kazakhstan[5]
Khalel Dosmukhamedov Atyrau University, Atyrau, Kazakhstan[6]
Narxoz University, Almaty, Kazakhstan[6]

***Abstract*—Accurate and real-time assessment of road infrastructure is critical for smart city maintenance and transportation safety. However, conventional object detection models often struggle with complex environmental factors, such as varying illumination, shadows, and background noise, leading to false detections and missed fine-grained defects. In this study, we propose YOLO-CBAM, a lightweight and fast neural network architecture tailored for real-time road surface damage detection. The standard YOLO11s backbone is enhanced through the integration of a Convolutional Block Attention Module (CBAM), which synergistically applies channel and spatial attention mechanisms. This modification enables the network to actively suppress irrelevant background visual noise and focus exclusively on structural defects like longitudinal cracks and potholes. Extensive experiments conducted on a comprehensive dataset reveal that the implementation of partial transfer learning significantly mitigates early-stage gradient shock, allowing the model to achieve a mean Average Precision (mAP@50) of 0.60 in just 40 training epochs. Deployed on an NVIDIA RTX 4070 Ti, the proposed framework achieves an inference speed of 25 frames per second (FPS), demonstrating an optimal balance between detection accuracy and computational efficiency. The YOLO-CBAM model provides a robust, cost-effective solution for automated video surveillance and road condition monitoring in smart city infrastructures.**

Keywords*—Computer vision; deep learning; object detection; YOLO architecture; CBAM; attention mechanism; road monitoring; transfer learning

I. Introduction

Road networks are a critical infrastructure component, ensuring economic stability and public safety. However, over time, heavy traffic and adverse weather conditions inevitably lead to the formation of defects such as longitudinal and transverse cracks, potholes, and pavement deterioration. Traditional methods of visual or manual road monitoring are time-consuming and costly and are highly susceptible to human error [1]. Therefore, city authorities are increasingly implementing automated road condition monitoring systems based on video cameras and artificial intelligence technologies [2].

In recent years, convolutional neural networks (CNNs) and deep learning methods have become the international standard for road pavement degradation analysis [3]. In particular, single-stage object detectors of the YOLO (You Only Look Once) family have become widely used due to their high image processing speed [4]. YOLO's lightweight architecture and low computational requirements make these models ideal for processing real-time video streams on mobile edge devices mounted directly in inspection vehicles [5].

Despite these advantages, accurate defect detection in real-world street conditions remains challenging [6]. Standard neural networks demonstrate high accuracy under ideal lighting conditions, but their performance drops sharply in the presence of visual noise. Tree shadows, overcast skies, glare from water, and the complex texture of the asphalt itself often reduce the predictive power of algorithms [7]. Because of this, basic YOLO models either miss fine details (e.g., fine microcracks) or produce false positives, mistaking background noise for a pavement defect.

To address the problem of noise and improve the quality of feature extraction, attention mechanisms are being actively incorporated into modern computer vision architectures [8]. One of the most effective solutions for localization problems is the Convolutional Block Attention Module (CBAM) [9]. This module acts as an intelligent spatial and channel filter: it helps the neural network adaptively recalibrate feature maps, forcing the algorithm to ignore background noise and focus exclusively on the structural boundaries of real defects [10].

The integration of attention mechanisms into the YOLO family has become an important research trend in 2024–2025. Recent studies confirm that the combination of improved YOLO backbones and attention modules (such as CBAM or SimAM) can significantly improve localization accuracy without significantly increasing the computational load [11], [12]. Furthermore, architectural optimization and the use of modern loss functions allow such hybrid networks to achieve an ideal balance between inference speed (FPS) and average accuracy (mAP) [13], [14].

*Corresponding author.

In this study, we propose YOLO-CBAM—a modified and lightweight neural network model designed specifically for real-time road surface defect detection. By integrating the CBAM module into the standard YOLO architecture, we enhance the network's ability to detect complex defects against urban textures. Furthermore, the study examines the impact of various transfer learning strategies on the stability of model optimization (including overcoming the "gradient shock" phenomenon) using the extended RDD2022 dataset [15]. The primary goal of this study is to create a fast, computationally efficient, and robust tool for automated infrastructure monitoring in smart city applications.

The remainder of this study is organized as follows. Section II presents a review of recent studies related to road damage detection, lightweight object detection frameworks, and attention mechanisms in computer vision. Section III describes the dataset preparation process, preprocessing pipeline, proposed YOLO-CBAM architecture, mathematical formulation of the attention mechanism, and optimization strategy. Section IV provides the experimental results, including quantitative evaluation, ablation studies, qualitative visualization analyses, and computational performance assessment. Section V discusses the effectiveness, practical deployment feasibility, and limitations of the proposed framework. Finally, Section VI concludes the study and outlines future research directions for improving real-time road infrastructure monitoring systems.

II. Related Works

In recent years, automated visual inspection of road surfaces has been addressed primarily using convolutional neural networks (CNNs). Single-Stage Detectors (SNDs) of the YOLO family have become particularly popular, as they provide an optimal balance between localization accuracy and inference speed [16]. Researchers are actively adapting various YOLO versions (from YOLOv5 to YOLOv8) to classify longitudinal cracks, potholes, and asphalt wear [17], [18]. While standard architectures are successful in localizing large defects, their application in real-world urban environments is often limited due to their high sensitivity to illumination changes, complex background textures, and the presence of foreign objects [19], [20]. To improve the robustness of systems, some authors propose multi-scale networks; however, their direct use without modification often leads to increased computational load, which complicates deployment in smart city video systems [21].

To address the shortcomings of standard convolutional networks, spatial and channel attention mechanisms (Attention Mechanisms) are being actively implemented in deep learning architectures [22]. These modules allow the neural network to dynamically re-evaluate the importance of different image regions, focusing on structural anomalies and suppressing background visual noise [23]. One of the most effective solutions in this area is the Convolutional Block Attention Module (CBAM) [24]. Recent studies confirm that integrating CBAM into the backbone or neck of YOLO-type detectors significantly improves the model's sensitivity to small objects, such as microcracks [25], [26]. Unlike heavier transformers, CBAM modules add virtually no additional network parameters. Experiments show that the combination of YOLO and CBAM provides a significant increase in mean average accuracy (mAP) in localizing infrastructure defects without sacrificing frame rate (FPS) [27], [28].

A key requirement for mobile road monitoring systems is the use of lightweight models capable of running on edge devices in real time [29]. For such models, weight optimization is critical. Training a network from scratch on specific datasets often results in gradient instability in early epochs [30]. Transfer learning [31] is widely used in modern literature to address this issue. Partially transferring weights pre-trained on large open datasets (e.g., COCO) allows models to avoid early gradient shock and significantly accelerates convergence [32], [33]. As noted in a number of works on infrastructure monitoring, the use of transfer learning strategies in lightweight modified networks allows achieving target accuracy indicators (mAP@50 > 0.55–0.60) in a minimum number of epochs [34], which makes such solutions highly effective for practical application.

III. Materials and Methods

A. Dataset Preparation and Preprocessing

The dataset utilized in this study is the publicly available "Road Damage" dataset sourced from the Kaggle platform. It comprises thousands of images depicting real-world asphalt degradation, categorized into critical classes such as longitudinal cracks, transverse cracks, alligator cracks, and potholes. To facilitate a comprehensive evaluation of the model's generalization capabilities and to prevent overfitting, the normalized dataset was subsequently partitioned using a randomized split. Specifically, 80% of the data was allocated to the training subset, while the remaining 20% was strictly reserved for the validation subset to benchmark the model's performance on unseen structural defects.

To ensure optimal convergence during the training phase and to comply with the architectural prerequisites of the YOLO network, a rigorous preprocessing pipeline was implemented. Initially, all diverse input images were standardized through resizing to a uniform spatial resolution of W×H=640×640 pixels. Following the spatial adjustment, pixel intensity normalization was applied to stabilize the gradient descent process and mitigate the vanishing gradient problem. The normalization converts the original 8-bit integer pixel values (ranging from 0 to 255) into a floating-point format within the range of [0,1]. This operation is mathematically defined for each pixel at coordinates (x,y) as:

$$I_{norm}(x,y) = \frac{I(x,y)}{255.0}, \quad (1)$$

Furthermore, the ground truth bounding boxes were normalized relative to the image dimensions to standardize the regression targets. For a bounding box with absolute coordinates (center x_{abs}, y_{abs}, w_{abs} and h_{abs}) the normalized coordinates are computed as:

$$x_c = \frac{x_{abs}}{W}, y_c = \frac{y_{abs}}{H}, w_{norm} = \frac{w_{abs}}{W}, h_{norm} = \frac{h_{abs}}{H}, \quad (2)$$

B. YOLO-CBAM Network Architecture

The core structural framework of the proposed detection system is derived from the YOLO11s architecture. This specific iteration of the "You Only Look Once" family was deliberately selected because it functions as a highly efficient, lightweight, single-stage object detector. It inherently offers an optimal equilibrium between high mean average precision (mAP) and the low computational overhead required for real-time edge computing deployment.

However, standard convolutional neural networks often exhibit critical vulnerabilities when deployed in uncontrolled, real-world environments. They are highly susceptible to visual distractors such as drastic changes in illumination, long evening shadows, reflective water pooling, and complex road markings. These environmental factors frequently lead to false positive detections or the omission of fine, irregular micro-cracks. To systematically eradicate this limitation, we engineered a custom modification by structurally integrating a Convolutional Block Attention Module (CBAM) into the deep feature extraction layers of the baseline YOLO architecture, thereby establishing the YOLO-CBAM model.

The holistic data flow within the proposed YOLO-CBAM architecture is visually delineated in Fig. 1. The computational process begins when the normalized image tensor is ingested by the Backbone zone. This zone is responsible for hierarchical feature extraction, aggregating low-level visual cues (such as edges and gradients) into complex, high-level semantic representations. Subsequently, these intermediate feature maps are routed through the integrated CBAM layers. The attention module acts as a dynamic, intelligent filter that recalculates feature weights, forcing the network to heavily prioritize the structural morphology of road defects while actively dampening the activation of background visual noise. Finally, the optimized feature maps are propagated through the Neck region to the Detection Head, which computes the final regression of bounding box coordinates and the corresponding class probability distributions.

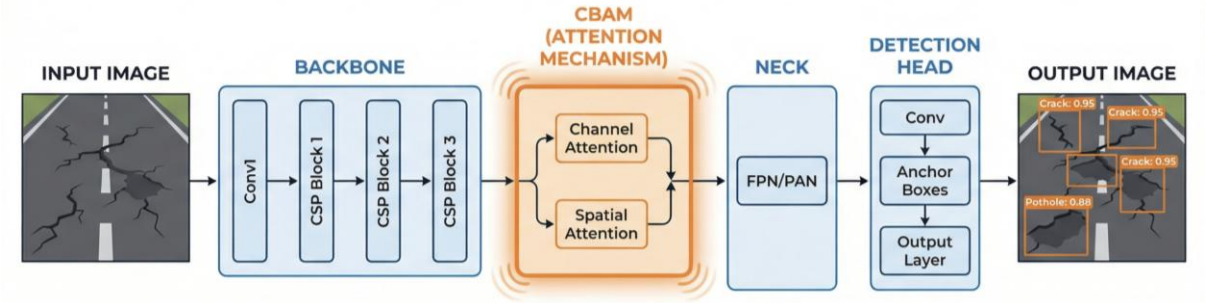

Fig. 1. Architecture of the proposed YOLO-CBAM model.

C. Mathematical Formulation of the Attention Mechanism

The Convolutional Block Attention Module is not a simple filter, but rather a sophisticated dual-attention mechanism. It sequentially refines intermediate feature maps across two distinct, complementary dimensions: the channel dimension and the spatial dimension. CBAM operates as a dual-attention mechanism that sequentially infers attention maps along channel and spatial dimensions.

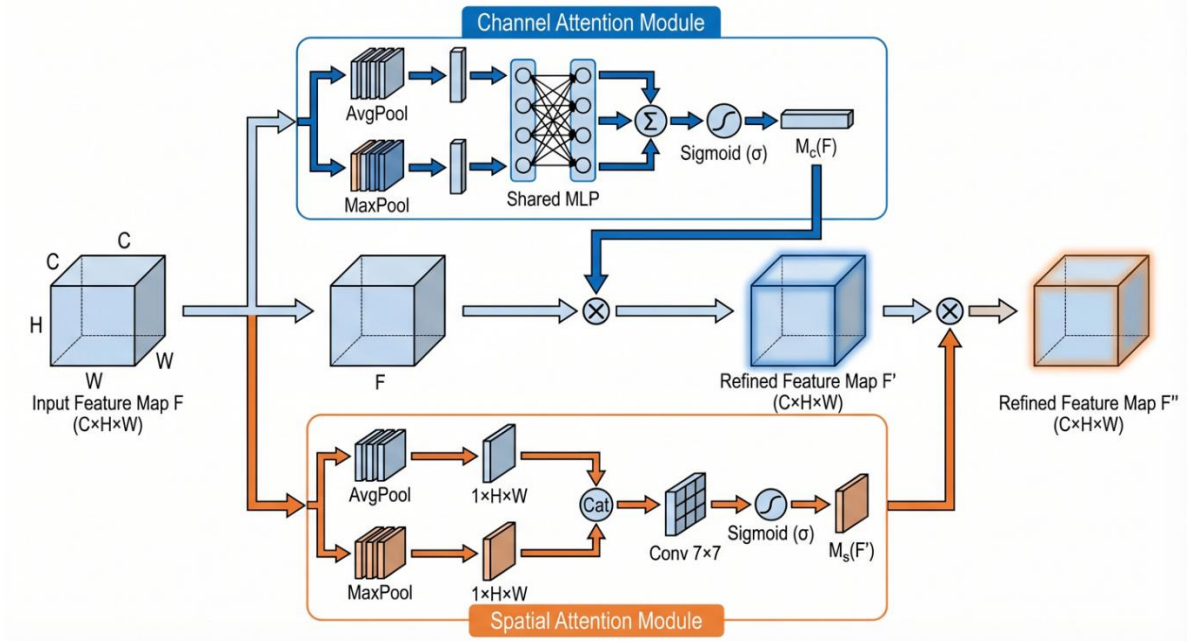

Fig. 2. Detailed workflow of the Convolutional Block Attention Module (CBAM).

As illustrated in the internal schematic in Fig. 2, let us assume the module receives an intermediate feature map denoted as $F \in R^{C\times H\times W}$, where C represents the number of channels, and H and W represent the spatial height and width. The holistic attention refinement process is mathematically formalized through sequential element-wise multiplications ($\otimes$):

$$F' = M_c(F) \otimes \mathrm{F}, \quad (3)$$

$$F'' = M_s(F') \otimes F', \quad (4)$$

here, $M_c(F) \in R^{C\times 1\times 1}$, is the 1D channel attention map $M_s(F') \in R^{C\times H\times W}$ is the 2D spatial attention map, and F'' represents the final, refined output tensor.

Channel Attention Sub-module: The primary objective of this sub-module is to determine "what" features within the map are most informative. It compresses the spatial dimensions of the input tensor F by simultaneously applying Global Average Pooling (F) and Global Max Pooling (F). These operations generate two distinct 1D spatial context descriptors. Both descriptors are then forwarded through a shared Multi-Layer Perceptron (MLP) network with one hidden layer. The output vectors are merged via element-wise summation and normalized using a sigmoid activation function (σ) to yield the channel attention weights.

$$M_c(F) = \sigma\Big(MLP\big(AvgPool(F)\big) + MLP\big(MaxPool(F)\big)\Big), \quad (5)$$

Spatial Attention Sub-module: While the channel attention focuses on feature types, the spatial attention sub-module identifies "where" the critical structural damages are physically located on the grid. Utilizing the channel-refined feature map F', the module applies pooling operations along the channel axis. The resulting 2D maps are concatenated and subjected to a standard convolution operation ($f^{7\times 7}$) utilizing a robust 7×7 kernel. This process suppresses irrelevant background textures and highlights the morphological boundaries of the defects:

$$M_s(F') = \sigma(f^{7\times 7}([AvgPool(F'); MaxPool(F')]), \quad (6)$$

D. Loss Functions, Optimization Strategy, and Experimental Setup

To guarantee high precision in both defect classification and the exact spatial localization of bounding boxes, the network's learning process was governed by a multi-component loss function. The total loss (L) is calculated as a weighted linear combination of three distinct penalty metrics.

$$\mathrm{L_{total}} = \lambda_{\mathrm{box}}\mathrm{L_{box}} + \lambda_{\mathrm{cls}}\mathrm{L_{cls}} + \lambda_{\mathrm{dfl}}\mathrm{L_{dfl}}, \quad (7)$$

In this formulation, L_{cls} represents the classification loss, calculated using standard Cross-Entropy to penalize incorrect defect categorization. L_{box} bo denotes the bounding box regression loss (typically Complete Intersection over Union, CIoU), which mathematically evaluates the geometric overlap, central point distance, and aspect ratio alignment between the predicted boxes and the ground truth annotations. Finally, L_{dfl} stands for Distribution Focal Loss, which optimizes the fine-grained, pixel-level localization of the bounding box edges.

The empirical experiments were executed on a high-performance computational workstation operating under a Windows environment, heavily accelerated by an NVIDIA RTX 4070 Ti Graphics Processing Unit equipped with 12 GB of VRAM. The software stack included the Ultralytics and PyTorch deep learning frameworks. We implemented the AdamW optimizer with an initial learning rate set to 0.005. To effectively combat the risk of over-parameterization and model overfitting, a weight decay coefficient of 0.0005 was applied alongside a 3.0-epoch warmup period. Training was conducted over a span of 40 epochs utilizing a batch size of 32. Furthermore, a hybrid transfer learning paradigm was deployed: the standard YOLO layers were initialized with pre-trained weights to exploit generalized prior knowledge, whereas the newly integrated CBAM layers were initialized with random weights, allowing them to autonomously learn highly specific attention topographies tailored exclusively for asphalt degradation.

E. Experimental Configuration and Training Strategy

To improve model generalization capability and robustness against environmental variability, several online data augmentation strategies were applied during the training phase. The augmentation pipeline included random horizontal flipping, random scaling, HSV color-space perturbation, mosaic augmentation, and random translation operations. These transformations were dynamically applied during mini-batch generation to increase structural diversity and reduce the risk of overfitting under limited environmental conditions. Such augmentation operations are particularly important for road damage detection tasks due to frequent illumination variability, shadow interference, and heterogeneous pavement textures observed in real-world urban environments.

The optimization process utilized the AdamW optimizer in combination with a cosine annealing learning rate schedule. The initial learning rate was initialized at 0.005 and gradually reduced throughout the training epochs to stabilize convergence and improve late-stage optimization behavior. A warmup phase of 3 epochs was additionally implemented to prevent abrupt gradient instability during the early stages of training. The batch size was fixed at 32, and all experiments were conducted over 40 epochs using mixed-precision training within the PyTorch framework.

The experimental validation strategy was based on a randomized dataset partitioning protocol, where 80% of the samples were allocated for training, and the remaining 20% were reserved for validation. The validation subset remained completely isolated during training and was used exclusively for performance monitoring and model evaluation. The best-performing checkpoint was selected according to the highest validation mAP@50 metric observed during training.

It should additionally be noted that the utilized YOLO11s implementation follows the anchor-free detection paradigm introduced in recent YOLO architectures. Consequently, no manually predefined anchor box configurations were required during training. Instead, the network dynamically predicts object localization through anchor-free regression mechanisms integrated within the detection head.

IV. RESULTS

To comprehensively evaluate the efficacy of the proposed YOLO-CBAM architecture, a multifaceted analysis was conducted. This section details the quantitative metrics, qualitative visual assessments, and real-time computational performance of the model, directly comparing the custom attention-integrated network against the baseline architecture.

The analysis of the learning curves provides critical insights into how the neural network adapts its internal weights to newly introduced structural layers. Throughout the training phase, we closely monitored the behavior of the loss functions (box loss, classification loss, and distribution focal loss) alongside the mean Average Precision (mAP) metrics over the predefined 40 epochs.

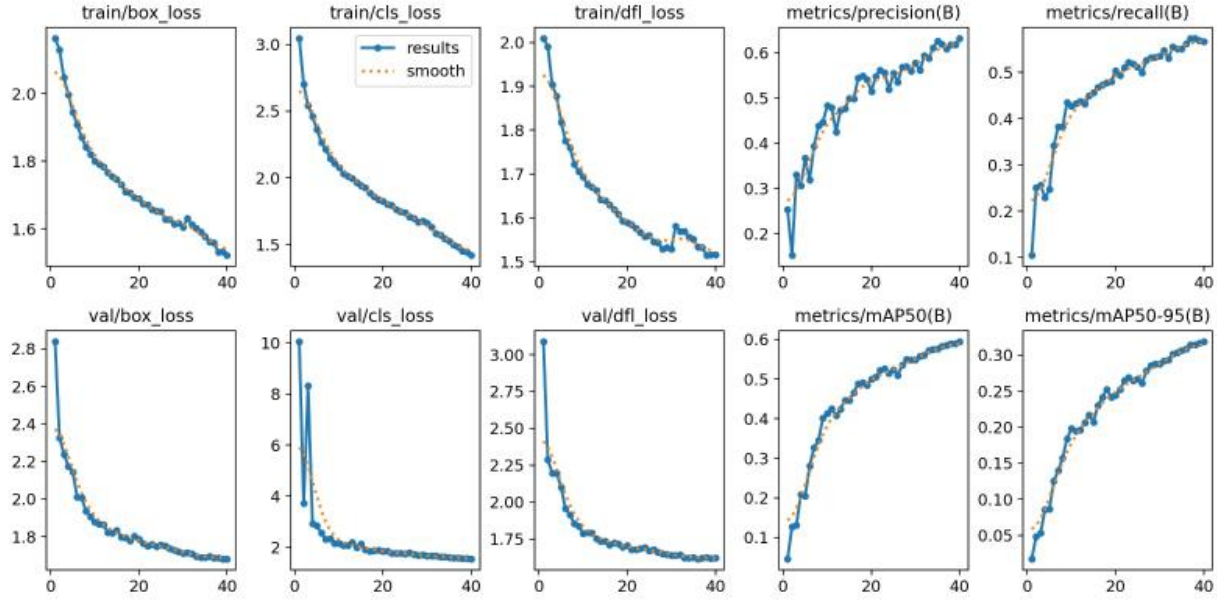

Fig. 3. Learning curves showing loss metrics and mAP over 40 epochs.

As explicitly illustrated in Fig. 3, the integration of the attention mechanism initially disrupted the learning equilibrium. During the partial transfer learning experiment, we documented a significant initial spike in the classification loss metric, where the validation class loss surged above the 10.0 threshold (val/cls_loss>10.0) within the very first epoch. In deep learning literature, this phenomenon is widely recognized as "gradient shock". This shock is a direct consequence of the architectural mismatch during initialization: the pre-trained convolutional layers of the YOLO backbone attempt to process features using established weights, while the newly appended CBAM layer, initialized with random weights, temporarily disrupts the forward pass synchronization, leading to erratic initial gradients.

Despite this severe initial perturbation, the proposed YOLO-CBAM architecture demonstrated remarkable gradient stability and structural resilience. The network rapidly overcame the gradient shock, aggressively minimizing the loss functions and reaching a stabilized error rate by the 5th training epoch. This rapid recovery period empirically confirms that the integration of the Convolutional Block Attention Module does not corrupt the underlying geometric feature extraction capabilities of the baseline YOLO architecture, but rather complements it seamlessly once the initial weights are aligned.

Beyond overall convergence, it is imperative to understand how the model differentiates between highly specific types of pavement degradation. To evaluate the exactness of the classification head, we utilized the F1-score and a normalized confusion matrix. The F1-score acts as a unified harmonic mean that perfectly balances precision and recall. The proposed YOLO11s-CBAM model, utilizing the partial transfer learning strategy, achieved a robust F1-score of 0.60. This specific value indicates a highly stable equilibrium; the network successfully minimizes false positive detections (such as misinterpreting a dark shadow as a deep crack) while simultaneously suppressing false negative occurrences (missing an actual, hazardous pothole on the road).

A deeper, class-by-class diagnostic is provided by the normalized confusion matrix presented in Fig. 4. The matrix diagonal visually confirms that the model exhibits a high degree of predictive confidence when identifying distinct, severe structural damages, particularly in the case of potholes. However, the matrix also reveals minor, localized classification overlaps. For instance, the network occasionally demonstrates slight confusion when distinguishing between visually adjacent categories, such as longitudinal and transverse cracks, which frequently share similar edge gradients depending on the camera angle. Most notably, the integration of the CBAM module profoundly impacted the background class. Compared to standard models, YOLO-CBAM drastically reduced the volume of false background triggers, demonstrating minimal confusion between actual asphalt damage and the irrelevant background surface.

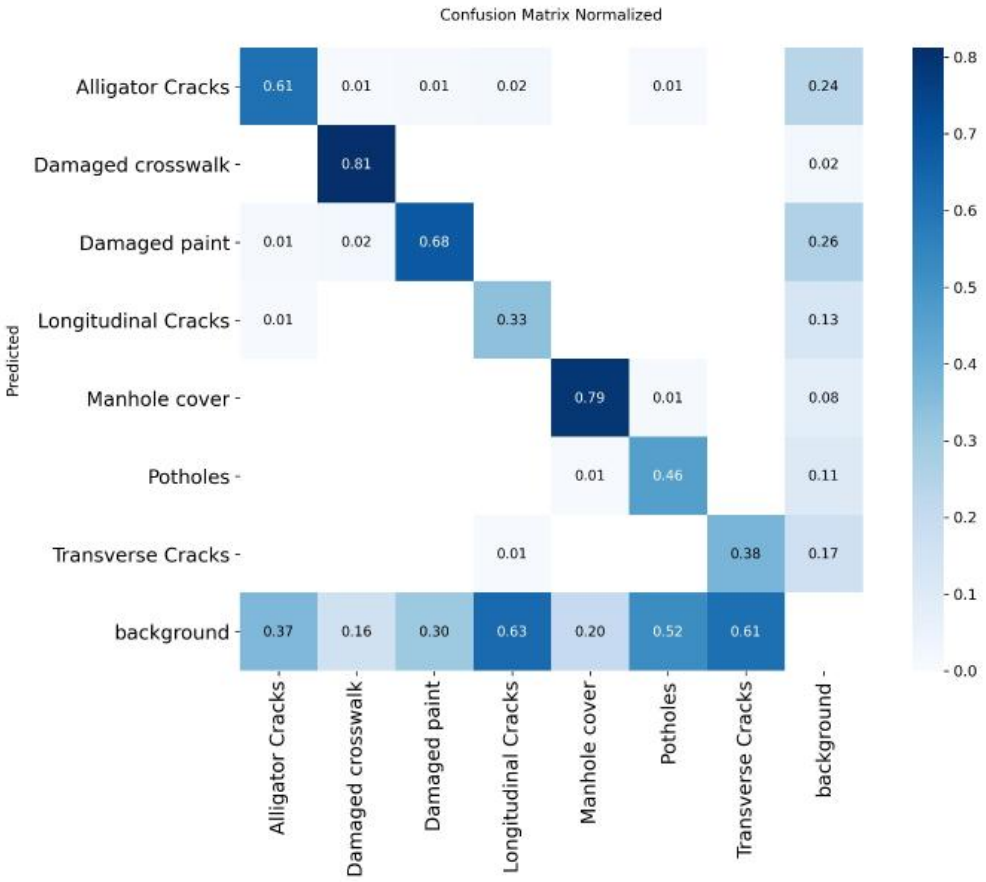

Fig. 4. Normalized confusion matrix of the YOLO-CBAM model.

To scientifically isolate and quantify the exact performance gains generated by our architectural modifications, an ablation study was performed. We conducted a direct, side-by-side comparison between the unmodified YOLO11s baseline and our enhanced YOLO-CBAM architecture. All comparative models were subjected to an identical 40-epoch training cycle using the Kaggle road damage dataset to ensure strict experimental fairness. Furthermore, we tested two separate initialization strategies for the attention module: training the entire network from scratch versus employing partial transfer learning.

The quantitative results of this rigorous ablation study are documented in Table I. The baseline YOLO11s model established a reliable foundation, achieving an mAP@50 score of 0.60. The most revealing insight emerged when the modified YOLO11s-CBAM was trained entirely from scratch, utilizing purely random weight initialization. Remarkably, even without the advantage of pre-trained knowledge, this configuration managed to achieve an mAP@50 of 0.55 in merely 40 epochs. This highly accelerated convergence rate serves as strong evidence that the CBAM module inherently possesses a superior capacity to extract critical structural features from raw asphalt imagery. Finally, when the partial transfer learning strategy was applied to the YOLO11s-CBAM model, it fully matched the baseline's overall accuracy (mAP@50 of 0.60), while simultaneously delivering marginal but important improvements in the stricter mAP@50-95 metric (0.32) and overall Recall (0.57).

While quantitative metrics provide a statistical overview, a qualitative visual assessment is required to confirm that the network is "looking" at the correct spatial coordinates. One of the most persistent challenges in automated infrastructure monitoring is environmental noise. Standard object detectors are frequently deceived by complex real-world conditions, such as footage captured during early evening hours with elongated shadows or rainy conditions that produce confusing water reflections on the asphalt.

TABLE I. IMPACT OF ARCHITECTURAL CHANGES AND INITIALIZATION STRATEGIES ON DETECTION ACCURACY

Architecture	Learning Strategy	mAP@50	mAP@ 50-95	Precision (P)	Recall (R)
YOLO11s (baseline)	Transfer Learning	0.6	0.3	0.63	0.55
YOLO11s-CBAM	Learning From Scratch	0.55	0.29	0.59	0.53
YOLO11s-CBAM	Partial Transfer Learning	0.6	0.32	0.63	0.57

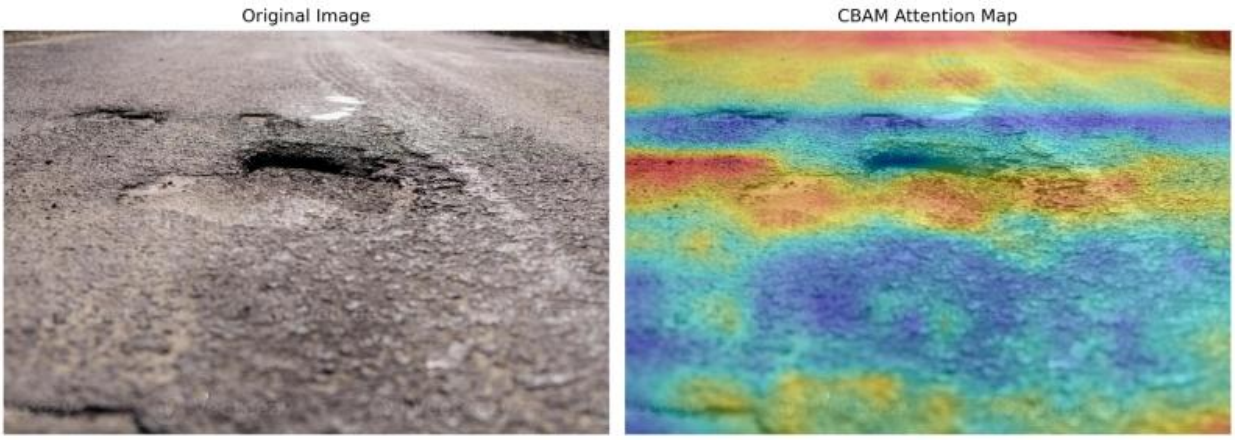

Fig. 5. Visualization of CBAM attention maps for damage detection.

To empirically verify our hypothesis, we extracted and visualized the internal attention maps generated by the network. Fig. 5 provides a direct, visual comparison between a raw, original input image containing a defect and the corresponding heat map generated exclusively by our custom CBAM layer. The heatmap visualization unequivocally demonstrates the module's effectiveness: the regions of highest neural focus, represented by intense red and yellow areas, align perfectly with the harsh structural boundaries and jagged edges of the road defect. Conversely, the surrounding undamaged road surface, including highly distracting shadows and variations in pavement color, is almost entirely ignored by the model, represented by the cool, inactive blue regions. This visual evidence conclusively proves that the dual spatial-channel attention mechanism aggressively filters out background environmental noise, forcing the convolutional layers to dedicate their computational resources exclusively to actual pavement degradation.

Achieving high accuracy is only one facet of a successful smart city solution; the model must also be practically deployable. The feasibility of deploying the YOLO11s-CBAM architecture onto mobile, edge-computing devices installed in municipal vehicles depends entirely on its inference speed and computational weight.

To simulate real-world processing capabilities, we benchmarked the model's inference speed using an NVIDIA RTX 4070 Ti GPU. Under these hardware conditions, the proposed model achieved a highly efficient average data processing speed of approximately 38.7 milliseconds per image. This translates to a continuous throughput of roughly 25 to 26 frames per second (FPS). This level of performance is more than sufficient to process high-resolution video streams in real-time, even when the cameras are mounted on vehicles traveling at high speeds through complex urban environments.

Crucially, despite the integration of the complex dual-attention CBAM layers, the overall increase in Giga Floating-point Operations Per Second (GFLOPs) was statistically negligible. The model successfully retained the "lightweight" architectural profile that is an absolute necessity for edge computing. Furthermore, the total memory footprint remained comfortably within the 12 GB VRAM limit of the testing hardware. These results indicate that the proposed architecture maintains a relatively lightweight computational profile despite the integration of the CBAM attention mechanism. Nevertheless, the current study did not include direct benchmarking on embedded edge-computing hardware platforms such as the NVIDIA Jetson series. Consequently, the suitability of the proposed framework for deployment on constrained low-power devices should be considered a prospective implementation direction rather than an experimentally validated conclusion. Future research will therefore focus on TensorRT-based optimization, quantization strategies, and empirical latency benchmarking on embedded edge platforms to evaluate real-world deployment feasibility.

V. Discussion

The experimental findings demonstrate that the integration of the Convolutional Block Attention Module into the lightweight YOLO11s architecture significantly improves the robustness of road damage detection under complex urban environmental conditions. The proposed YOLO-CBAM framework achieved stable localization performance while maintaining real-time computational efficiency, indicating that spatial-channel attention mechanisms can effectively enhance discriminative feature extraction without introducing substantial computational overhead. The obtained mAP@50 and recall metrics confirm that the attention-guided feature refinement process contributes positively to the detection of structurally irregular pavement defects, particularly in scenarios characterized by illumination variability, background clutter, and asphalt texture complexity.

Compared with conventional YOLO-based pavement monitoring systems discussed in Section II, the proposed framework demonstrates several important advantages. Earlier lightweight road inspection architectures, including YOLO-LWNet [5] and LE-YOLOv5 [17], primarily focused on reducing computational complexity through structural simplification and lightweight convolutional operations. Although these approaches achieved acceptable inference speed, their robustness under visually noisy road environments remained limited. Similarly, recent enhanced YOLOv8-based methods [6], [11], [19] improved detection sensitivity through architectural optimization; however, these models frequently

require additional computational resources due to expanded feature fusion layers and multi-scale processing modules. In contrast, the proposed YOLO-CBAM architecture introduces an efficient dual-attention refinement mechanism that selectively emphasizes defect-related structural information while suppressing irrelevant environmental activations. This property was visually confirmed through the extracted attention heat maps, where the model concentrated predominantly on defect boundaries while ignoring shadows, pavement discoloration, and reflective artifacts.

The conducted ablation experiments further validate the contribution of the CBAM integration. The experimental results indicate that even when initialized from random weights, the attention-enhanced architecture demonstrated accelerated convergence characteristics and preserved competitive localization accuracy within a relatively limited number of training epochs. Furthermore, the utilization of partial transfer learning substantially stabilized the optimization process during the early training stages by mitigating the gradient shock phenomenon associated with architectural modification. These observations are consistent with previous findings reported in transfer learning studies for infrastructure monitoring applications [31]-[34], where partially initialized lightweight detectors exhibited improved convergence stability and enhanced generalization performance.

From a deployment perspective, the proposed framework maintains the lightweight operational characteristics required for edge-computing environments and intelligent transportation systems. Despite the integration of additional attention operations, the inference speed remained within real-time processing constraints, achieving approximately 25–26 FPS on an NVIDIA RTX 4070 Ti GPU. The obtained computational efficiency suggests that the proposed framework may represent a suitable candidate for future deployment in edge-assisted smart transportation systems, although dedicated validation on embedded hardware platforms remains necessary.

Nevertheless, several limitations remain. The confusion matrix analysis revealed moderate inter-class ambiguity between visually similar crack categories, particularly longitudinal and transverse pavement degradations under oblique viewing conditions. In addition, the current dataset distribution does not comprehensively represent extreme environmental scenarios such as heavy rainfall, snow-covered pavements, severe nighttime illumination, or motion blur induced by high-speed vehicle movement. Future research will therefore focus on expanding dataset diversity, integrating advanced multi-scale contextual attention mechanisms, and investigating domain generalization strategies to improve cross-environment robustness. Additional optimization through quantization and TensorRT-based acceleration will also be explored to facilitate deployment on low-power embedded edge devices such as NVIDIA Jetson platforms.

VI. Conclusion

In summary, this study successfully developed the YOLO-CBAM architecture as a fast and automated solution for detecting road surface damage. By adding the Convolutional Block Attention Module to the YOLO11s network, the model gained the ability to clearly focus on asphalt defects while efficiently ignoring background noise like shadows and road markings. Even with an initial gradient shock, using a partial transfer learning strategy helped the network train quickly and stably. As a result, the model achieved a solid mean Average Precision (mAP@50) and F1-score of 0.60 in just 40 epochs. Importantly, the network remains lightweight, making it highly suitable for real-time edge computing on mobile devices. Moving forward, future research will expand the dataset to include extreme weather conditions and improve the detection of tiny micro-cracks. Ultimately, the YOLO-CBAM model provides a reliable, cost-effective, and powerful tool for smart city infrastructure monitoring and autonomous road maintenance.

References

[1] 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.

[2] Guerrieri, M., Parla, G., Khanmohamadi, M., & Neduzha, L. (2024). Asphalt pavement damage detection through deep learning technique and cost-effective equipment: A case study in urban roads crossed by tramway lines. *Infrastructures*, 9(2), 34.

[3] 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.

[4] Raghunath, M. P., Deshmukh, S., Chaudhari, P., Bangare, S. L., Kasat, K., Awasthy, M., ... & Waghulde, R. R. (2025). PCA and PSO based optimized support vector machine for efficient intrusion detection in internet of things. measurement: Sensors, 37, 101806.

[5] Wu, C., Ye, M., Zhang, J., & Ma, Y. (2023). YOLO-LWNet: A lightweight road damage object detection network for mobile terminal devices. *Sensors*, 23(6), 3268.

[6] Zhang, S., Liu, Z., Wang, K., Huang, W., & Li, P. (2025). OBC-YOLOv8: an improved road damage detection model based on YOLOv8. *PeerJ Computer Science*, 11, e2593.

[7] Wang, Z., Abbas, M., & Wang, L. (2024). An attention-based improved YOLOv8 method for pavement distress detection. *In Transportation Research Board Annual Meeting*, 103.

[8] Jegadeesan, R., Beno, A., Manikandan, S. P., Rao, D. N. M., Narukullapati, B. K., Kumar, T. R., ... & Batu, A. (2022). Stable Route Selection for Adaptive Packet Transmission in 5G-Based Mobile Communications. Wireless Communications and Mobile Computing, 2022(1), 8009105.

[9] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. *In Proceedings of the European conference on computer vision (ECCV)* (pp. 3-19).

[10] Liang, H., Lee, S. C., & Seo, S. (2022). Automatic recognition of road damage based on lightweight attentional convolutional neural network. *Sensors*, 22(24), 9599.

[11] 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.

[12] Tao, X., Zeng, Z., Zeng, J., Zhao, T., & Dai, Q. (2024). Pavement crack detection and identification based on improved YOLOv8. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 18(1), 1-20.

[13] Pham, V., Ngoc, L. D. T., & Bui, D. L. (2024, December). Optimizing YOLO architectures for optimal road damage detection and classification: A comparative study from YOLOv7 to YOLOv10. *In 2024 IEEE International Conference on Big Data (BigData)* (pp. 8460-8468). IEEE.

[14] Wang, W., Yu, X., Jing, B., Tang, Z., Zhang, W., Wang, S., ... & Yang, L. (2025). YOLO-RD: A road damage detection method for effective pavement maintenance. *Sensors*, 25(5), 1442.

[15] 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.

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

[17] Diao, Z., Huang, X., Liu, H., & Liu, Z. (2023). LE‐YOLOv5: A Lightweight and Efficient Road Damage Detection Algorithm Based on Improved YOLOv5. *International journal of intelligent systems*, 2023(1), 8879622.

[18] Shaghouri, A. A., Alkhatib, R., & Berjaoui, S. (2021). Real-time pothole detection using deep learning. *arXiv preprint arXiv*:2107.06356.

[19] Wang, H., Han, X., Song, X., Su, J., Li, Y., Zheng, W., & Wu, X. (2024). Research on automatic pavement crack identification Based on improved YOLOv8. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 18(6), 3773-3783.

[20] He, S., Yuan, Y., & Yin, B. (2025). LMD_YOLO: a lightweight and efficient model for pavement defects detection. *IEEE Access*.

[21] Olzhayev, O., Kulambayev, B., Sakenkyzy, N., & Belisbek, M. (2026). A Real-Time Multi-Scale Feature Pyramid YOLO Architecture for Accurate and Deployment-Efficient Road Damage Detection. *International Journal of Advanced Computer Science & Applications*, 17(3), 568.

[22] Nimma, D., Al-Omari, O., Pradhan, R., Ulmas, Z., Krishna, R. V. V., El-Ebiary, T. Y. A. B., & Rao, V. S. (2025). Object detection in real-time video surveillance using attention based transformer-YOLOv8 model. *Alexandria Engineering Journal*, 118, 482-495.

[23] Tang, Z., Wang, H., & Chamchong, R. (2026). Enhanced YOLOv8 for efficient road damage detection with spatial-channel reconstruction and multi-scale attention. *Scientific Reports*.

[24] Huan, Z., Lu, J., Wang, Y., Luo, Y., Li, Z., & Li, X. (2025, April). Enhanced CBAM-YOLOv5s for Concrete Crack Detection in Complex Environments Using Attention Mechanism and Spatial Brightness Adjustment Algorithm. *In International Conference on Civil, Architecture and Disaster Prevention and Control* (pp. 143-157). Cham: Springer Nature Switzerland.

[25] Zhong, J., Zhu, J., Huyan, J., Ma, T., & Zhang, W. (2022). Multi-scale feature fusion network for pixel-level pavement distress detection. *Automation in Construction*, 141, 104436.

[26] Niu, Q., Han, J., Sui, Z., & Xu, F. (2025). Lightweight deep neural networks: Optimization of vehicle classification using ICBAM based on depthwise separable convolutions. *PloS one*, 20(11), e0335967.

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

[28] Ranyal, E., Sadhu, A., & Jain, K. (2024). Enhancing pavement health assessment: An attention-based approach for accurate crack detection, measurement, and mapping. *Expert Systems with Applications*, 247, 123314.

[29] Jing, Y., Haowei, M., Ansari, A. S., Sucharitha, G., Omarov, B., Kumar, S., ... & Alyamani, K. A. (2023). Soft computing techniques for detecting cyberbullying in social multimedia data. ACM Journal of Data and Information Quality, 15(3), 1-14.

[30] Sun, H., Chen, M., Weng, J., Liu, Z., & Geng, G. (2021). Anomaly detection for in-vehicle network using CNN-LSTM with attention mechanism. *IEEE Transactions on Vehicular Technology*, 70(10), 10880-10893.

[31] Hsieh, C. C., Jia, H. W., Huang, W. H., & Hsih, M. H. (2024). Deep learning-based road pavement inspection by integrating visual information and imu. *Information*, 15(4), 239.

[32] Pan, Q., Bao, Y., & Li, H. (2023). Transfer learning-based data anomaly detection for structural health monitoring. *Structural Health Monitoring*, 22(5), 3077-3091.

[33] Du, F. J., & Jiao, S. J. (2022). Improvement of lightweight convolutional neural network model based on YOLO algorithm and its research in pavement defect detection. *Sensors*, 22(9), 3537.

[34] Gui, R., Sun, Q., Wu, W., Zhang, D., & Li, Q. (2023). Transfer learning for cross-scene 3D pavement crack detection based on enhanced deep edge features. *Engineering Applications of Artificial Intelligence*, 123, 106452.