IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 May 2025, accepted 21 May 2025, date of publication 26 May 2025, date of current version 12 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3573651

RESEARCH ARTICLE

Computationally Enhanced UAV-Based Real-Time Pothole Detection Using YOLOv7-C3ECA-DSA Algorithm

SITI FAIRUZ MAT RADZI[1], MOHD AMIRUDDIN ABD RAHMAN[1], (Member, IEEE), MUHAMMAD KHAIRUL ADIB MUHAMMAD YUSOF[1], NURIN SYAZWINA MOHD HANIFF[1], AND ROMI FADILLAH RAHMAT[2], (Member, IEEE)

[1]Department of Physics, Faculty of Science, Universiti Putra Malaysia (UPM), Serdang 43400, Malaysia
[2]Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan 20155, Indonesia

Corresponding authors: Mohd Amiruddin Abd Rahman (mohdamir@upm.edu.my) and Romi Fadillah Rahmat (romi.fadillah@usu.ac.id)

This work was supported by Selangor State Government through Geran Penyelidikan Negeri Selangor Fund under Grant SUK/GPNS/2023/TSE/02.

ABSTRACT Road deterioration due to potholes has a significant impact on traffic safety and infrastructure maintenance, highlighting the need for detection systems that combine precision with real-time capabilities. Although YOLO-based algorithms have been widely adopted for their speed and efficiency in object detection, achieving a balance between high accuracy and low inference time remains a challenge, particularly in scenarios involving small objects and complex features. This study introduces YOLOv7-C3ECA-DSA, an improved YOLOv7 architecture designed to address these limitations. The model incorporates Cross-Stage Enhanced Channel Attention (C3ECA) blocks in the backbone network and Depthwise Shuffle Attention (DSA) in the detection head to enhance feature learning and boundary detection, achieving high detection accuracy and real-time inference capabilities. The experimental results demonstrate that YOLOv7-C3ECA-DSA achieves an mAP0.5 of 85.3% with an inference time of 10.9 ms, outperforming the prior methods. The proposed model also performs well under adverse vision conditions such as night and rain. However, the model performance may be limited in severe environmental conditions, such as low-contrast surfaces or heavily obscured potholes. Despite these limitations, the research significantly advances real-time pothole detection by striking an optimal balance between computational efficiency and detection accuracy. The findings underscore the effectiveness of YOLOv7-C3ECA-DSA in addressing practical challenges in real-time infrastructure monitoring, making it suitable for scalable deployment in autonomous vehicle systems and road maintenance applications.

INDEX TERMS Pothole detection, deep learning, YOLOv7, C3ECA, improved shuffle attention.

I. INTRODUCTION

Road transportation is crucial for the economic development of a country as it facilitates the connection between communities and businesses, thus providing access to education, employment, healthcare, and social services [1]. However, due to economic expansion and technical advancements in recent times, it contributes to an adverse effect on the quality of the transportation infrastructure. As road conditions worsen, the capability of traffic to endure and recover from disruptions deteriorates [2]. A significant issue faced by drivers is the rising number of potholes on roadways. These not only cause damage to the vehicles, but they also present a danger to the safety of the drivers. Potholes are a prevalent kind of road deterioration caused by various factors, including inadequate maintenance and service, environmental and climatic conditions including rainfall and snowfall, and human-induced strain on road surfaces. These concerns provide significant challenges due to their frequent unexpectedness and the need for ongoing

The associate editor coordinating the review of this manuscript and approving it for publication was Sudhakar Radhakrishnan.

monitoring to guarantee the road infrastructure [3] remains in optimal shape.

A system capable of detecting and notifying maintenance services or other drivers of the presence of potholes could significantly improve the condition of the roads and ensure the safety of motorists. Nevertheless, relevant authorities can easily identify which potholes offer the highest risk and need to be fixed as soon as possible [4]. Currently, the recognition of potholes has primarily depended on sensor-based systems or manual inspection, both of which have constraints in terms of scalability, accuracy, and real-time responsiveness. Performing manual inspections is costly and labor-intensive. Apart from that, it is an unreliable and impractical approach for efficiently surveying extensive road networks. The expense of deploying and maintaining sensor-based systems increases depending on the size of road network coverage, and their efficacy can be hindered by various environmental conditions. To address these challenges, deep learning-based models have emerged as a promising approach for real-time detection of various-sized objects like potholes in complex road environments.

Deep learning-based target detection can be divided into two distinct groups based on the detection method. One of which consists of two-stage target detection algorithms, demonstrated as R-CNN [5], Fast-RCNN [6], Faster-RCNN [7], and Mask R-CNN [8]. This algorithmic class creates a predetermined frame for the target based on the algorithm and subsequently categorizes and predicts the frame using the network layer. In certain instances, it exhibits superior detection accuracy. Nevertheless, the training and detection speed of the system is slow, proving that it is impractical for real-time applications in certain cases, such as the detection of potholes.

Another group is one-stage target detection algorithms, demonstrated by SSD [9], and YOLO series algorithms (e.g., YOLO [10], YOLO 9000 [11], YOLOv3 [12], YOLOv4 [13], and YOLOv7 [14]). These algorithms directly produce the predicted classification probabilities and coordinate values of the detected targets, resulting in faster performance compared to the two-stage algorithm. Although a two-stage algorithm usually offers superior accuracy, particularly in complex scenes with small or overlapping objects, its additional processing steps reduce efficiency for real-time pothole detection. Nevertheless, potholes come with their own set of challenges, where they vary significantly in size and shape, which complicates detection for single-stage models that rely on predefined anchor boxes or features. Environmental factors such as shadows, lighting variations, and road markings can obscure potholes on diverse road surfaces, making it difficult to differentiate potholes from their surroundings.

Consequently, in this study, we propose an enhanced single-stage detection with improved handling of complex features and environmental variability, the YOLOv7-C3ECA-DSA, which is built on YOLOv7 architecture to achieve higher accuracy and robustness in real-time pothole detection. Hence, the research put forward some improvements, as described in the following contribution:

- We combine a three-convolutional-layer (C3) module with Efficient Channel Attention (ECA) in the final block of the backbone to allow the model to capture more depth and discriminative features while maintaining computational efficiency. This significantly enhances the backbone's capability to handle complex feature interactions quickly, which is vital for tasks like object detection in real-time applications.
- We incorporate Depthwise-Convolution into the shuffle attention mechanism at the head of YOLOv7, which improves the model's efficiency in recalibrating feature maps at the final detection stage. This design allows for better spatial and channel attention, ensuring that critical features are highlighted just before making final predictions, leading to improved object detection accuracy.
- We integrate YOLOv7-C3ECA-DSA architecture in the Efficient Layer Aggregation Network (ELAN) modules of the backbone. By integrating Cross-Stage Enhanced Channel Attention (C3ECA) and depthwise shuffle attention, the architecture enhances the representation of multiscale features and minimizes information loss during feature extraction. This improvement significantly strengthens the model's capability to detect and localize the potholes.

II. RELATED WORK

A. IMAGE PROCESSING FOR POTHOLE DETECTION

Traditional 2D image processing techniques, such as thresholding, segmentation, and contour extraction, have been extensively utilized to detect potholes and other road damage. Methods like Otsu's thresholding [15], histogram-based thresholding [16], and morphological operations [17] have been employed to identify damaged regions and extract pothole contours to improve the accuracy of road detection. However, these approaches often faced challenges in handling complex textures and varying environmental conditions, including lighting and shadows. Completing the 2D image-based approaches, researchers have also explored the use of 3D point cloud data for pothole detection [18]. These innovative 3D techniques leverage geometric modeling and segmentation algorithms that incorporate surface normal information, clustering, and region-growing methods. This enables enhanced accuracy and reliability in identifying road defects by incorporating depth data to refine the geometric modeling and labeling processes. However, the computational complexity and specialized hardware requirements associated with 3D point cloud techniques pose limitations to their widespread scalability.

The emergence of machine learning and deep learning has revolutionized the field of pothole detection, enabling automated and accurate identification using annotated datasets

and neural networks. Researchers have applied deep convolutional neural networks (CNN) [19] to various computer vision tasks, such as image classification, object detection, and segmentation, to identify potholes. Studies have evaluated the performance of pre-trained models like ResNet50, InceptionV2, and VGG19 on the classification of potholes in images of muddy and highway roads, with the VGG19 model demonstrating superior accuracy, precision, and recall [20]. Building upon these foundational methods, object detection algorithms such as Faster R-CNN and Single Shot Detector (SDD) have provided more efficient and automated solutions for pothole detection. Faster R-CNN is a two-stage detection model that uses a Region Proposal Network (RPN) to generate object candidates, followed by classification and bounding box regression. The authors utilized an improved Faster R-CNN model enhanced with an attention mechanism for crack recognition, achieving high detection accuracy. While the modifications improved the recognition of complex crack patterns, the model's computational demands remain a challenge for real-time applications, especially when processing large-scale datasets or operating in low-resource environments. In contrast, SSD is a single-stage detection model that directly predicts object categories and bounding boxes, offering faster inference times. Andika et al. [21] utilized the SSD MobileNet algorithm, incorporating image enhancement techniques to improve the smoothness and sharpness of images. This approach resulted in a model accuracy of 75% with image enhancement, compared to 73% without it. While the addition of image enhancement led to a modest improvement in accuracy, the overall precision remains relatively modest, especially when compared to more complex detection methods.

The YOLO (You Only Look Once) family of algorithms has significantly advanced the field of pothole detection by achieving a balance between accuracy and speed. Researchers experimented with various versions of YOLO algorithms, specifically v3, v4, v5, v7, v8, and v9, to categorize road damage, including potholes. These base YOLO models are often optimized to improve detection accuracy and precision. However, the base versions typically struggle to achieve high mean Average Precision (mAP), as they may fail to recognize fine-grained characteristics of objects, particularly those with complex textures or structures like potholes. For example, Zhu et al. [22] developed the unmanned aerial vehicle (UAV) asphalt pavement distress database (UAPD) and attained a suboptimal mean Average Precision (mAP) of 56.62% using YOLOv3, indicating limitations to achieving higher detection accuracy. Building on this, Heo et al. [4] introduced a deep metric learning-based approach with an optimized SPFPN-YOLOv4 tiny model to address 10 categories of distress, encompassing 4 types of cracks, seals, and potholes. This method attains a classification precision of 87.28%.

Lightweight networks are persistently evolving, with the YOLO series undergoing constant improvements and modifications to address both efficiency and precision. YOLOv5 [23] has markedly reduced in size relative to YOLOv4 [24]. Numerous researchers combined YOLOv5 with lightweight networks to develop algorithmic models that are more proficient in detecting pavement distress. The lightweight adaptations of YOLOv5, which utilize the suitable initial anchors CIoU (Complete-IoU) Loss [25] and DIoU-NMS (Distance IoU – Non-Maximum Suppression) [26], can achieve real-time detection while maintaining comparable precision. Wan et al. [27] improved the YOLOv5 backbone network by using ShuffleNet V2 and integrating an efficient channel attention (ECA) module. This alteration led to a 28.8% reduction in the model's size while retaining the model's precision. Chu et al. [28] presented an improved Ghost-YOLOv5s detection technique, characterized by a computational cost of 2.2 G FLOPs and an inference time of 72 ms. This method primarily aims to improve detection efficiency via the feature fusion procedure.

Recent advancements in pothole detection leverage the capabilities of the latest YOLO object detection algorithms, YOLOv8 and YOLOv9, to achieve enhanced accuracy and efficiency in identifying road damage. Bhavana et al. [29] introduced the YOLOv8-based POT-YOLO model that targets the detection of various pothole types, including cracks, oil stains, patches, and debris, by preprocessing video frames with the Contrast Stretching Adaptive Gaussian Filter to mitigate distortions, followed by the application of a Sobel edge detector for accurate pothole localization. This approach demonstrates exceptional performance, with POT-YOLO achieving an accuracy of 99.10%, a precision of 97.6%, a recall of 93.52%, and an F1-score of 90.2%, outperforming other models such as Faster R-CNN, SSD, and Mask R-CNN. Youwai et al. [30] introduced the YOLO9tr model, developed based on the YOLOv9 architecture, which incorporates a partial attention block that drastically enhances feature extraction and attention capabilities. YOLO9tr achieves high precision; however, its high computational requirements for inference render it unsuitable for real-time applications. Nonetheless, this study primarily identified fractures and depressions. Although these studies have improved detection performance, they mostly focus on a wide range of pavement issues, including cracks and fractures, instead of solely focusing on potholes.

Our proposed YOLOv7-C3ECA-DSA method overcomes these challenges by leveraging deep learning, which is less affected by environmental noise. Its integration of advanced attention mechanisms (C3ECA and DSA) enables it to better extract features, even in complex and varied road conditions, ensuring superior pothole detection under diverse environments. This approach also significantly reduces computational demands.

B. METHODS OF CAPTURING POTHOLE IMAGES

Researchers have been utilizing methods to capture images of potholes; a few methods that have been used include video analysis, LiDAR, smartphone applications, and unmanned aerial vehicles (UAVs). Smartphones have also emerged as a practical and cost-effective tool for pothole detection,

with researchers developing detection models that leverage the device's camera and sensors [32]. These models employ machine learning algorithms trained on extensive road image datasets to improve detection accuracy. Kim further highlights the value of low-cost pothole detection using smartphone sensors, showcasing their potential in detecting road anomalies like potholes [33]. The integration of smartphone sensors, such as accelerometers, can further enhance the detection process by providing additional data on road conditions [34]. However, smartphones have some limitations, such as lower resolution and precision, especially when identifying smaller or partially obscured potholes [35]. The effectiveness of smartphone-based systems is also highly dependent on environmental conditions, including lighting, weather, and camera positioning. Moreover, smartphones mounted in vehicles can only capture data from the vehicle's perspective, which can limit their ability to cover extensive road networks or provide a comprehensive view of the road surface [36].

Wang et al. [37] have illustrated the effectiveness of LiDAR in accurately detecting and segmenting potholes, which is crucial for developing targeted maintenance strategies. The combination of LiDAR with other imaging techniques can further enhance detection capabilities, providing comprehensive data for road management systems. However, LiDAR systems come with significant drawbacks. They are expensive to acquire and maintain, which makes them less practical for large-scale deployment, especially for routine pothole monitoring over extended road networks. LiDAR data also requires specialized software and expertise to process, which can be both time-consuming and computationally expensive [38].

Video analysis, particularly from vehicle-mounted cameras, has also been explored as a method for real-time pothole detection. This approach can be integrated with intelligent transportation systems to provide alerts and updates on road conditions in real-time. Additionally, deep learning models, such as YOLO, have been employed to process video feeds for pothole detection, demonstrating high accuracy in identifying road defects [39], [40]. However, vehicle-mounted video systems are not without their limitations. The camera's field of view is often limited, meaning it may miss important areas of the road surface, especially if potholes are located off to the side or not directly in the vehicle's path [45]. Additionally, video captured from a moving vehicle can suffer from motion blur, particularly at high speeds or during sharp turns, which can impair the ability to detect small or subtle road defects [46].

UAVs have emerged as a prominent tool due to their ability to capture high-resolution imagery and video of road surfaces from diverse perspectives, enabling detailed analysis and accurate pothole identification. Nomqupu et al. [31] emphasized the advantages of UAV imagery, particularly in detecting small features such as potholes, which are often overlooked by traditional sensors. The high spatial resolution provided by UAVs facilitates detailed analysis and accurate identification of potholes, making them a suitable option for road monitoring. Furthermore, studies have demonstrated that UAVs can effectively utilize multispectral sensors to enhance pothole detection through advanced image processing techniques [31]. Hence, in our study, we utilized UAV imagery for pothole detection due to its ability to overcome the limitations of other methods, such as the low resolution of smartphone cameras, the high cost and complexity of LiDAR systems, and the restricted field of view and environmental sensitivity of vehicle-mounted video analysis. UAVs offer high-resolution imagery from diverse perspectives, enabling more accurate and detailed analysis, particularly in detecting small potholes often overlooked by traditional sensors.

C. C3ECA AND SHUFFLE ATTENTION IN PREVIOUS DETECTION SYSTEMS

The integration of attention mechanisms, specifically C3ECA and shuffle attention, has significantly advanced detection systems in various domains, including wildlife monitoring [41], defect detection, and object recognition. These mechanisms enhance the model's ability to focus on relevant features while reducing computational complexity, which is crucial for real-time applications.

The C3ECA module, an advanced attention mechanism, has proven to significantly enhance detection accuracy within the YOLOv5 architecture. Liang's work demonstrates that by incorporating multiple attention modules, including Squeeze-and-Excitation (SE), the C3ECA-YOLOv5l algorithm achieves enhanced detection rates for bird species near substations, addressing the critical need for safety in power systems [42]. This demonstrated improvement in the detection accuracy model's performance in real-time scenarios, which is essential for practical applications. Ren et al. further explored the application of C3ECA in the context of steel surface defect detection. Their improvements in YOLOv5 architecture by integrating the C3ECA mechanism resulted in notable improvements in the model's performance metrics, including recall and mean Average Precision (mAP) [43]. Moreover, the model's computational complexity was reduced, with a smaller parameter count and file size, making it particularly well-suited for deployment in resource-constrained environments. This optimization facilitates real-time applications without compromising accuracy or efficiency.

Meanwhile, shuffle attention has emerged as a powerful tool in various detection frameworks, including YOLOv7 and YOLOv5. The shuffle attention mechanism has been integrated into the YOLOv7 backbone to enhance feature extraction capabilities, leading to improved detection accuracy [44]. Additionally, the YOLOv5-SA-FC model, which incorporates shuffle attention, has shown significant improvements in pig detection and counting, highlighting its effectiveness in agricultural applications [47]. The mechanism works by efficiently combining channel and spatial attention, allowing the model to focus on the most

relevant features while maintaining computational efficiency. Moreover, the use of shuffle attention has been extended to underwater image processing, where it aids in enhancing the model's attention to detecting objects in challenging environments. Research by [48] indicates that incorporating shuffle attention into the YOLOv5s algorithm significantly improves the model's performance in detecting fish in blurred underwater scenes. This adaptability across different contexts underscores the versatility of shuffle attention in various detection tasks.

Combining C3ECA and Shuffle Attention offers a sophisticated mechanism to enhance feature representation in neural networks. By effectively utilizing both channel attention and spatial conditions, the resultant framework can significantly improve performance on various computer vision tasks. Future research should focus on optimizing this integration to suit specific application needs and exploring its adaptability to different deep learning contexts.

D. OBJECT DETECTION USING YOLOv7 VARIATION

Recently, significant research and associated efforts have been devoted to the object detection algorithm YOLOv7, with a primary emphasis on ongoing enhancements and optimizations aimed to improve the algorithm's object detection accuracy, efficiency, and overall lightweight design specifically in pavement engineering. YOLOv7 stands out by offering several key advantages over its predecessors. These include significantly higher accuracy, faster inference speeds, and enhanced efficiency in terms of computational resources. Notably, YOLOv7 achieves state-of-the-art performance, delivering an average precision (AP) of 56.8%, while reducing parameters and computational complexity. By introducing innovations such as the introduction of compound model scaling, the use of Extended Efficient Layer Aggregation Networks (E-ELAN), and a trainable bag-of-freebies, it further contributes to its superior performance. [14]

Rojas et al. [49] utilized the base YOLOv7 model to detect low-level nuclear waste and obtain the highest accuracy compared to other state-of-the-art methods, including OWL-ViT, STEGO, and SD Mask R-CNN. The author later explained that the ability for YOLOv7 to perform instance segmentation was particularly useful for this task. This allowed the model to generate precise pixel masks around objects, which was critical for accurately identifying different types of waste in a cluttered environment.

However, Komori et al. [50] explored and evaluated the design of YOLOv7 integrated with attention modules, including a base model with simple operations of convolutions and sigmoid activation, Squeeze and Excitation (SE), and Convolutional Block Attention Module (CBAM), where each of the three types of attention modules is implemented in the head, in-between, and backbone of YOLOv7. The attention-enhanced YOLOv7 models, particularly the backbone-attention model with CBAM, significantly improved the detection of urinary sediment particles compared to both the original YOLOv7 and other state-of-the-art models. The improvements in recall and mAP highlight the effectiveness of using attention mechanisms to enhance feature extraction and reduce background noise in medical image detection tasks.

TABLE 1. Experimental configuration.

Parameter	Configuration
CUDA	11.8
Programming Language	Python 9.3
GPU	GeForce RTX 4070
CPU	Intel® Core™ i7-14700F

TABLE 2. Training configuration.

Training Parameter	Value
Batch size	16
Epoch	100
Initial learning rate	0.01
Recurrent learning rate	0.1
Image Size	416 x 416

He et al. [51] utilized Bidirectional Feature Pyramid Network (BiFPN) and Ghost Convolution modules, which help to enhance feature fusion across different scales and reduce the number of parameters by generating more feature maps using fewer operations. The authors applied knowledge distillation, transferring knowledge from a larger teacher model to a smaller student model to improve detection accuracy while maintaining a lightweight structure. The lightweight YOLOv7-BiFPN-G model significantly reduced the number of parameters to 7.4M, enabling real-time road crack detection on mobile devices while maintaining comparable performance in precision and mAP when compared to larger models. The system developed using YOLOv7-BiFPN-G demonstrates the model's effectiveness for efficient and cost-effective road crack detection.

Ning et al. [52] proposed YOLOv7-RDD, which utilized YOLOv7 integrated with Efficient Lightweight Aggregation Network (EDC), spatial feature pyramid structure (SPPCSPD), and SimAM attention to characterize a large dataset comprising 12 distinct types of pavement distress, namely longitudinal cracks, transverse cracks, net cracking, potholes, zebra crossing loss, marking loss, manhole sinkage, block patch, strip patch, swelling, strip patch failure, and manhole break. The dataset consisted of 2013 images with 5039 distress tags. The proposed YOLOv7-RDD model yielded an mAP@0.5 of 89.5%, which shows improvement from the best base YOLOv7 version (YOLOv7-tiny) by 2.5%.

III. METHODOLOGY

A. DATASET AND EXPERIMENTAL ENVIRONMENT

The experimental configurations are listed in Table 1. The specific settings of the hyperparameters of the network training are listed in Table 2.

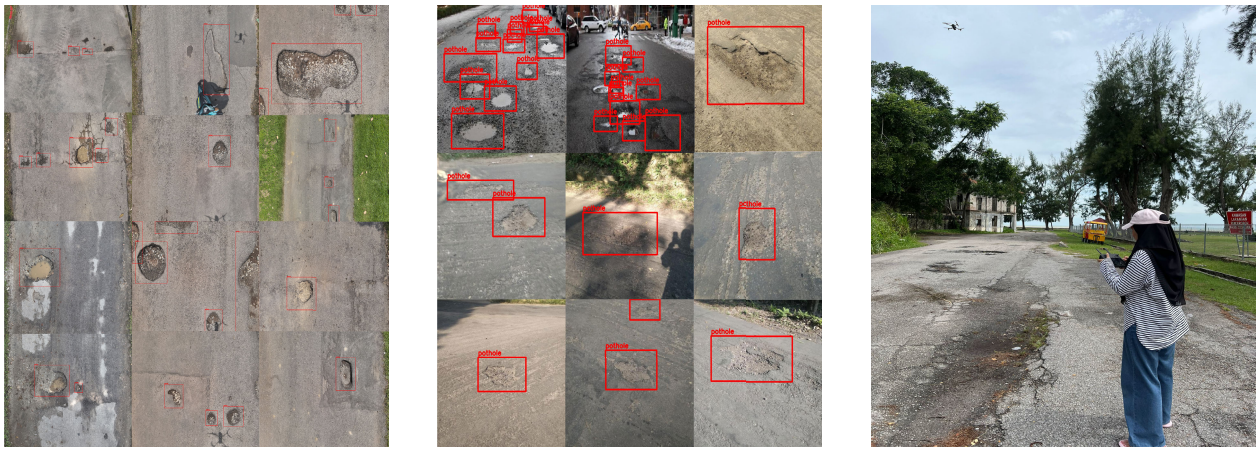


FIGURE 1. (a) Dataset from [61] with tags. (b) DJI Mavic 3 Enterprise Drone with tags. (c) Drone operation to capture pothole.

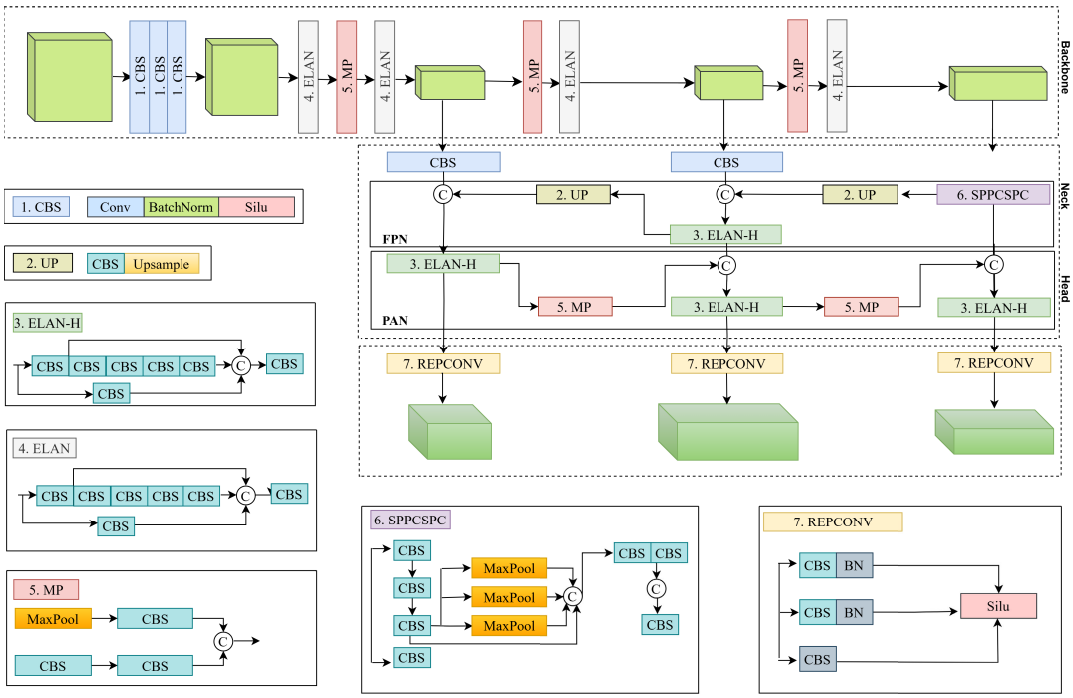


FIGURE 2. Original architecture of YOLOv7.

To verify the generalization of the model, two datasets are selected for training and validation of the model. Dataset 1 (Figure 1(a)) consists of 100 images of potholes using a DJI Mavic 3 Enterprise drone capturing the images of Selangor State Road of B11, B13, and B17 routes (Figure 1(c)). Dataset 2 (Figure 1(b)) consists of 665 datasets of road potholes taken from [61]. In this study, we combine those two datasets and divide them into training and validation sets according to a ratio of 70:30.

B. YOLOV7 NETWORK STRUCTURE

YOLOv7 is a single-stage target detection network with fast detection speed, high accuracy, and easy training and deployment. The network model structure for YOLOv7 is divided into three parts: backbone, neck, and head, as shown in Figure 2.

The backbone network extracts the multiscale features of the input image and generates a multiscale feature map, which is then passed as input to the neck network. The main function of the neck network is to facilitate the feature fusion, where the Feature Pyramid Network (FPN) [14] structure transfers richer, robust semantic features from the deeper layers to the shallower layers, thereby strengthening the entire pyramid and improving semantic representation across multiple scales. The Path Aggregation Network (PAN) structure [53] in the neck network transfers

stronger positional information from shallow to deep layers, improving localization across scales. Both FPN and PAN boost the representational ability of the network and distribute multiscale learning across 3 detection networks. The head then combines enhanced feature data to perform object detection and classification.

YOLOv7 consists of several modules, including CBS, MP, ELAN, ELAN-H, UPSample, SPPCSPC, and RepConv [54]. The CBS module integrates standard convolution, batch normalization, and SiLU activation. The MP module is responsible for downsampling and includes a maximum pooling layer and a CBS module. The ELAN module, with two branches, serves as an efficient feature aggregation network, enhancing the network's ability to learn features by managing both short and long gradient paths.

The ELAN-H module functions similarly to the ELAN module but differs in that it uses four convolutional layers for output, while ELAN uses six. The UPSample module performs upsampling using nearest neighbor interpolation. The SPPCSPC module broadens the receptive field, enabling the algorithm to manage varying image resolutions via maximum pooling at diverse receptive field dimensions. The RepConv module behaves differently during training and deployment, with training summing outputs from three branches, while deployment reparameterizes the parameters into a single main branch.

C. PROPOSED YOLOv7-C3ECA-DSA

The YOLOv7 backbone architecture downsamples the input image five times using the ELAN and MP1 modules, which can lead to the loss of semantic information, particularly for small target defects. The original ELAN feature extraction module struggles to effectively capture global contextual features, resulting in weaker performance and potential missed detections of small targets. To address this, the modified network of YOLOv7, YOLOv7-C3ECA-DSA (Figure 3), enhances the final ELAN modules by incorporating C3ECA in the backbone. This improvement enhances the ability of the network to capture both spatial and channel-specific features, focusing more accurately on important regions. The attention mechanism also embeds coordinated information, boosting the detection and localization of small objects, especially in cluttered or dense scenes, leading to improved feature representation and target recognition.

The original neck network in YOLOv7 struggles to effectively capture both spatial and channel-specific features, leading to suboptimal feature extraction and challenges in representing important details in complex scenes. This limitation hinders the learning ability of deeper layers, potentially causing neurons to stagnate and fail to adapt to new information. Consequently, the head network receives poorly refined features, which reduces the accuracy and quality of predictions. To address this, the study replaces the original RepConv in the YOLOv7 neck with the improved Depthwise-Shuffle Attention module (DSA). This update enhances the attention of the network to spatial and channel features, improves the flow of feature maps across layers, and strengthens feature learning by integrating Depthwise Convolution, reducing overfitting, and enhancing performance in complex environments.

The enhanced attention mechanisms in the improved neck architecture ensure that small objects, which might lose semantic information in deeper layers, are accurately detected. While the original YOLOv7 neck could handle multi-scale feature integration, the shuffle attention method enhances feature extraction across different dimensions, leading to more precise predictions, particularly in challenging scenarios.

1) C3ECA BLOCK

The study integrated an attention-based mechanism into the YOLOv7 architecture to enhance feature extraction and pothole detection capabilities. To preserve the compact design of the model, the implementation of attention modules prioritized improving performance without escalating complexity.

The C3 module, which is a key component of the YOLOv5 architecture, is built upon the CSPNet framework. This module provides robust feature extraction capabilities while addressing the issue of gradient duplication within the backbone network. Given the advantages offered by the C3 module, we incorporated it into our network design [55]. While the C3 module enhances feature extraction, we observed the limitations in its ability to handle occluded or small objects effectively. The ECA module is a lightweight attention mechanism that incorporates channelization from the Squeeze-and-Excitation Network. Unlike SENet, which uses complex, fully connected layers, ECA uses a simple one-dimensional convolution to capture relationships between channels without increasing the computational costs. Hence, ECA is more effective at distinguishing target features from complex backgrounds without adding any computational costs.

By combining ECA into the C3 structure shown in Figure 4, we integrate the C3ECA module, which addresses the limitations of each component when used individually, resulting in a more effective module for detecting potholes. The C3 structure uses a split-and-merge approach that captures features at different scales, making it versatile in extracting detailed information. Meanwhile, ECA helps in highlighting the most important channels and improving the ability of the network to accurately detect features even in complex situations, such as distinguishing potholes from road textures.

In our proposed architecture, we positioned the C3ECA module as the final block in the backbone of YOLOv7. At this stage, features contain high-level information crucial for object detection, especially for small or partially hidden objects. The final layers in the backbone are responsible for extracting details essential for bounding box regression and classification. C3ECA refines the most important features for

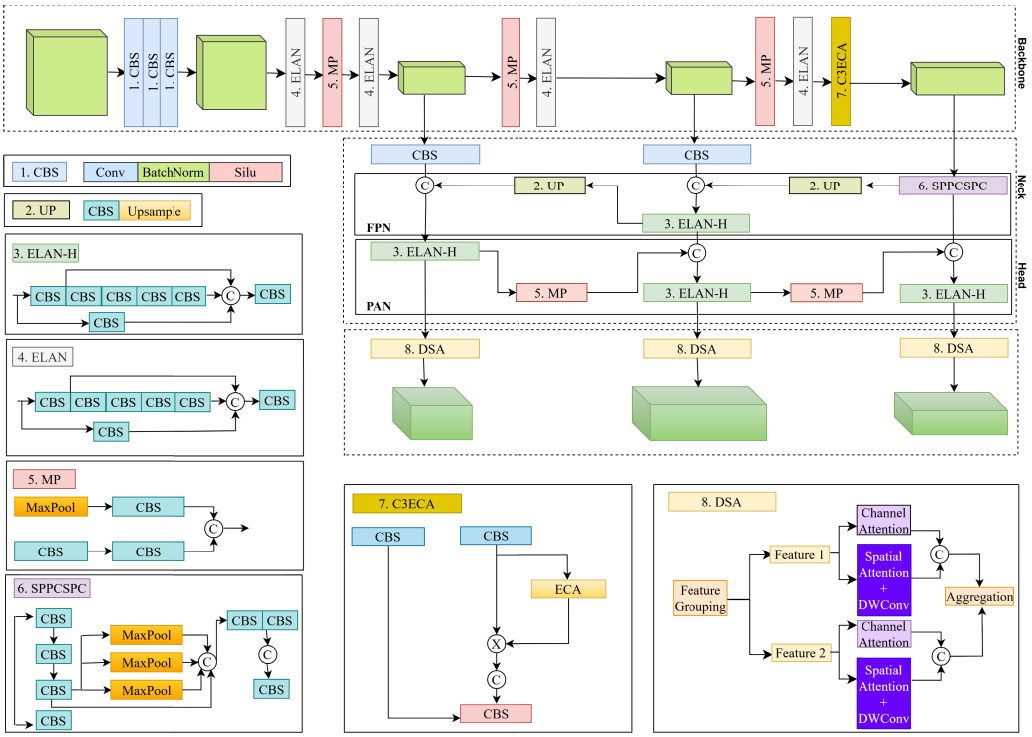


FIGURE 3. Proposed architecture of YOLOv7-C3ECA-DSA.

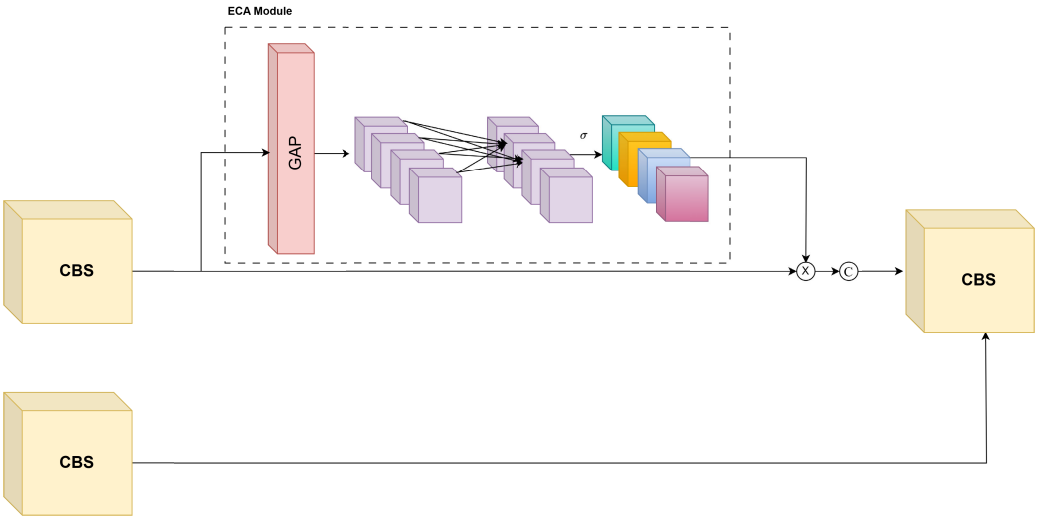


FIGURE 4. C3ECA module structure.

detection by improving positional accuracy and ensuring the network focuses on the details.

2) IMPROVED DEPTHWISE-SHUFFLE ATTENTION (DSA) MODULE

We introduced depthwise convolution into the shuffle attention module within the head network to enhance the efficiency of pothole detection. We positioned the DSA before the detection head by replacing RepConv to enhance feature quality right before prediction. This allows the model to better focus on the most relevant spatial and channel-specific features, thereby increasing the accuracy of object localization and classification. By refining features at this final stage, the model is better equipped to detect subtle and partially obscured potholes, which enhances overall detection performance without adding excessive computational cost. This is because depthwise convolution [56] and shuffle attention [57] are both designed to be lightweight and efficient.

In computer vision, channel attention and spatial attention are the two predominant mechanisms. Channel attention captures dependencies between feature channels, while

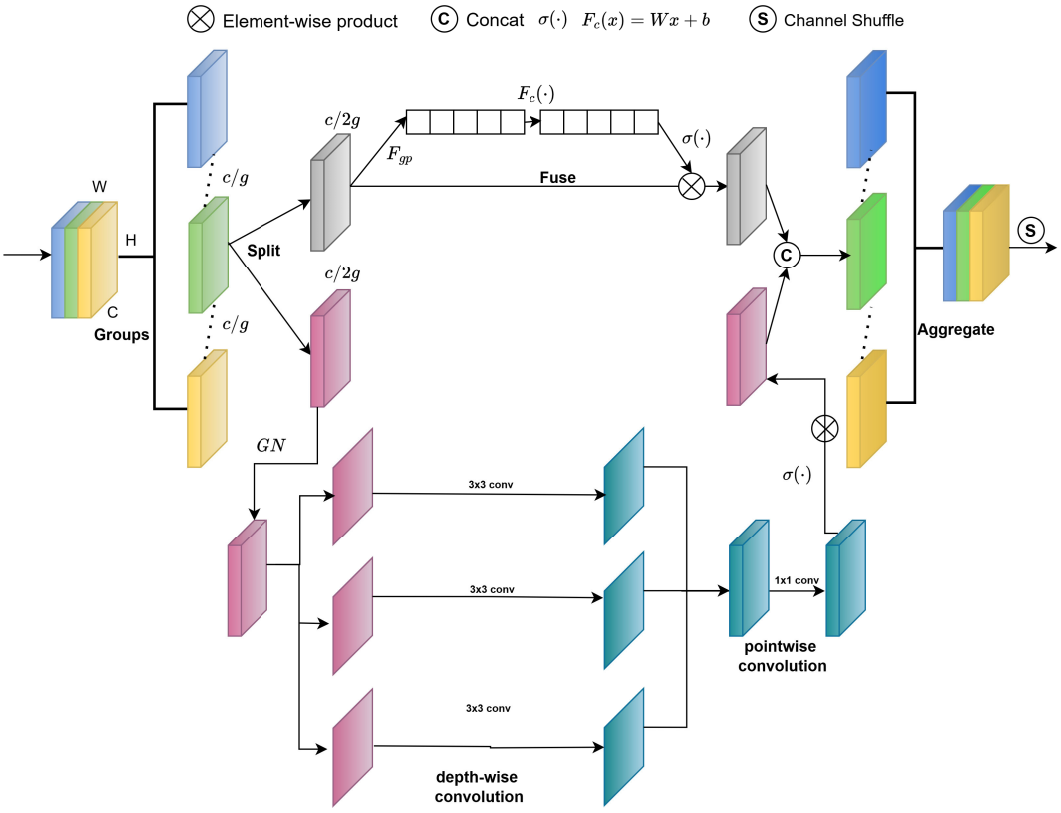


FIGURE 5. Integrated Depthwise Convolution in Shuffle Attention.

spatial attention extracts relationships between pixels. Both mechanisms utilize distinct strategies for feature aggregation, transformation, and activation to refine the feature extraction process. While combining these two attention mechanisms enhances performance, it also increases computational complexity [57].

The SA module offers a lightweight and efficient solution by grouping the input channels and applying the shuffle unit to perform both channel and spatial attention on sub-features. Additionally, the module employs attention masks for each position, reducing noise and enhancing meaningful semantic information. The SA module operates in three steps: feature grouping, mixed attention, and feature aggregation, as depicted in Figure 5.

In the feature grouping phase, the input feature tensor $X \in \mathbb{R}^{C\times H\times W}$ is divided into G groups along the channel dimension, generating sub-feature tensors $X_k \in X^{\left(\frac{C}{G}\times H\times W\right)}$, where each sub-feature learns specific semantic information during training. This corresponds to the leftmost part of Figure 5. Each sub-feature X_k is further split into two branches X_{k1} and $X_{k2} \in \mathbb{R}^{\left(\frac{C}{2G}\times H\times W\right)}$. The green branch handles channel attention, while the yellow branch focuses on spatial attention, enabling the capture of both semantic and positional information.

In the channel attention phase, the channel attention mechanism in the SA module presents an enhanced approach to mitigate the high parameter count observed in SE channel attention [58]. It commences by applying global average pooling to generate channel-wise data, which consolidates and averages the feature maps across each channel. This averaged value serves as a proxy for the feature information of the channel. The resulting data exhibits a dimension of $\mathbb{R}^{\left(\frac{C}{2G}\times 1\times 1\right)}$, as expressed by the following formula:

$$s = F_{\text{gp}}\left(X_{k1}\right) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{k1}(i,j) \tag{1}$$

Next, the operation is carried out via the fully connected layer, and the sigmoid activation function is employed to generate attention weights corresponding to each individual channel:

$$X'_{k1} = \sigma\left(F_c\left(s\right)\right) \bullet X_{k1} = \sigma\left(W_1 s + b_1\right) \bullet X_{k1} \tag{2}$$

The operation $F_c(\bullet)$ is defined by multiplying the input s with a matrix $W_1 \in \mathbb{R}^{\left(\frac{C}{2G}\times 1\times 1\right)}$, followed by adding a bias term $b_1 \in \mathbb{R}^{\left(\frac{C}{2G}\times 1\times 1\right)}$. This operation scales and transforms s as described in equation 2.

In the spatial attention phase, Group Norm [59] is first applied to the input feature X_{k2}. In our proposed method, we add a Depthwise Convolution after normalization to refine spatial features and improve feature representation,

as formulated in the following equation:

$$[Y_{i,j} = \sum_{k=-1}^{1} \sum_{l=-1}^{1} \left(X_{\lfloor \frac{c}{2G} \rfloor, i+k, j+l} \times W_{\lfloor \frac{c}{2G} \rfloor, k, l} \right) \quad (3)$$

where i, j are spatial positions in the feature map, k, l are kernel positions, c is the channel index, G is the group parameter, X is the input tensor, and W is the weight.

The transformation $F_c(\bullet)$ is then applied, with parameters W_2 and b_2, both of size $\mathbb{R}^{\left(\frac{C}{2G} \times 1 \times 1\right)}$. The formula for spatial attention with Depthwise Convolution is:

$$X'_{k2} = \sigma\left(W_2 \bullet (\mathrm{GN}(X_{k2}))\, \mathrm{DepWiseConv} + b_2\right) \bullet X_{k2} \quad (4)$$

Finally, the outputs of the channel and spatial attention branches are concatenated as follows:

$$[X'_{k1} = \left[X'_{k1}, X'_{k2}\right] \in \mathbb{R}^{\left(\frac{C}{G} \times H \times W\right)} \quad (5)$$

This maintains the same dimensions as the input to the group.

In the final aggregation step, channel shuffle operations are employed to enable information exchange between different groups. This results in an attention map that retains the same dimensions as the original input feature map X. This process is depicted in the rightmost section of Figure 5.

IV. RESULT AND DISCUSSION

A. COMPARISON WITH OTHER RELATED WORK MODELS

The proposed YOLOv7-C3ECA-DSA architecture was evaluated against a range of detection models, including conventional two-stage detectors, Faster RCNN-Resnet50 and Faster RCNN-SqueezeNet, various versions of lightweight YOLO, which are YOLOv3-Tiny, YOLOv5n, YOLOv8n, YOLOv9-tiny, YOLOv10n, YOLO11n, and the original YOLOv7, Realtime Detection Transformer (RT-DETR), and the proposed method by the previous author, YOLOv9tr and EMTTA-n. This comparison is to verify the accuracy and validity of our proposed model, YOLOv7-C3ECA-DSA. The comparison of mAP0.5, mAP0.75, mAP0.5-0.95, F1-score, inference time, and parameter are shown in Table 3.

In this study, the YOLOv7 model was selected as the base model due to its superior detection capability compared to newer architectures like YOLOv8 and YOLOv11 in our application scope of pothole detection. It achieved the highest accuracy across various evaluation criteria, such as mAP0.5, mAP0.5-0.95, and mAP0.75, with only a minor 0.7% decrease in F1-score compared to other detection models, such as RT-DETR. Although YOLOv7 has a slightly higher computational cost in terms of parameters and inference time, its advanced architectural components, such as ELAN and CBS layers, greatly enhance feature extraction and localization capabilities. These modules significantly enhance feature extraction and localization capabilities, despite the increased computational complexity. Hence, we developed our proposed model based on the strong foundation of YOLOv7 architecture.

In comparison to conventional detectors, our proposed model demonstrated a significant improvement over Faster R-CNN variants. The model achieved a notable increment of 18.7% and 51.6% in mAP0.5 compared to the ResNet50 and SqueezeNet implementations, respectively. Although RCNN-SqueezeNet was designed as a lightweight model with fewer parameters, it exhibited the lowest accuracy among all evaluated models. While the more complex architecture of RCNN-ResNet50 promoted better performance over RCNN-SqueezeNet, it was still suboptimal compared to the proposed model.

Our proposed model exhibited superior performance compared to the original lightweight YOLO architectures, including YOLOv3-tiny, YOLOv5n, YOLOv8n, YOLOv7-tiny, YOLOv9-tiny, YOLOv10n, and YOLOv11n, as well as the base YOLOv7 model. Notably, YOLOv7-C3ECA-DSA achieved the highest mAP0.5, demonstrating enhanced pothole detection accuracy with a minimum improvement of 1.43% in mAP0.5 compared to the YOLO versions. This suggests improved capabilities in accurately identifying true positive targets with confidence scores exceeding 0.5.

Recent specialized object detection architectures tailored for pavement distress recognition, such as EMTTA-n [60] and YOLOv9tr [30], exhibited notable performance gains within their respective frameworks. EMTTA-n incorporated Channel Attention and SE attention mechanisms, resulting in a 10% improvement over YOLOv5n. Likewise, YOLOv9tr implemented Partial Selective Attention, leading to a 17.38% improvement over the base YOLOv9-tiny model. Nevertheless, our proposed YOLOv7-C3ECA-DSA model surpassed these specialized architectures, demonstrating a 0.24% higher mAP0.5 metric in comparison to YOLOv9tr. In comparison to RDD-YOLO [64], a recent variant designed to address pavement distress detection, the YOLOv7-C3ECA-DSA model still outperforms, despite RDD-YOLO achieving a much lower inference time of 3.8 ms. RDD-YOLO integrates the Efficient Lightweight Aggregation Network (EDC) and Spatial Feature Pyramid Structure (SPPC-SPD), providing a solid mAP@0.5 of 0.731, but this comes at the cost of a significant drop in accuracy. YOLOv7-C3ECA-DSA, by contrast, achieves a higher mAP0.5 of 0.853, representing an 11.68% increase in detection accuracy over RDD-YOLO. While RDD-YOLO boasts a faster inference time, the YOLOv7-C3ECA-DSA model only incurs a 7.3% increase in inference time (10.9 ms compared to 3.8 ms), making it a highly competitive option for real-time applications without compromising accuracy.

Compared to RT-DETR, which utilizes an end-to-end transformer-based architecture to balance inference speed and accuracy, our proposed model demonstrated superior performance in both metrics by achieving a 2.28% higher mAP0.5 while reducing inference time by 27.3%. This improvement is particularly noteworthy given that the design of RT-DETR focuses on maintaining efficiency through simplified detection pipelines and reduced computational complexity.

To further highlight the overall superiority of YOLOv7-C3ECA-DSA compared to other models, each of the model's

TABLE 3. Comparison with other detection models.

	mAP0.50		mAP0.50-0.95		F1-Score		Inference Time (ms)	
	Value	Rank	Value	Rank	Value	Rank	Value	Rank
Faster RCNN-Resnet50	0.707	13	0.468	7	0.500	15	465	16
Faster RCNN-SqueezeNet	0.503	16	0.218	16	0.258	16	341	15
YOLOv3-tiny [62]	0.753	8	0.437	13	0.718	7	2.5	4
YOLOv5n [62]	0.765	7	0.464	8	0.708	8	2.8	5
YOLOv7-tiny	0.602	15	0.309	15	0.630	14	2.4	3
YOLOv8n	0.726	11	0.441	12	0.681	12	2.1	1
YOLOv9-tiny	0.725	12	0.456	10	0.684	11	4.6	8
YOLOv10n	0.694	14	0.433	14	0.658	13	3.3	6
YOLO11n	0.744	9	0.463	9	0.691	10	2.3	2
YOLOv9tr [30]	0.851	2	0.581	1	0.791	4	68.6	14
RDD-YOLO [64]	0.731	10	0.449	11	0.705	9	3.8	7
EMTTA-n [60]	0.842	3	0.555	5	0.790	5	43.1	13
RT-DETR-Pothole [64]	0.834	5	0.565	2	0.800	3	15.0	12
YOLOv7	0.841	4	0.549	6	0.803	2	10.3	10
YOLOv7-C3ECA-DSA	0.853	1	0.557	4	0.823	1	10.9	11

	Parameter (M)		GFLOPs		FPS		Average Rank	
	Value	Rank	Value	Rank	Value	Rank	Value	Rank
Faster RCNN-Resnet50	41.0	13	4.0	15	2.15	16	14.0	13
Faster RCNN-SqueezeNet	29.0	9	0.8	16	2.93	15	15.1	14
YOLOv3-tiny [62]	9.8	7	14.5	6	400.00	4	7.4	7
YOLOv5n [62]	2.1	2	5.8	13	357.14	5	6.9	3
YOLOv7-tiny	6.0	8	13.0	7	416.67	3	9.3	11
YOLOv8n	2.6	5	6.8	9	476.19	1	7.3	6
YOLOv9-tiny	1.7	1	6.4	11	217.39	8	8.7	10
YOLOv10n	2.6	5	8.2	8	303.03	6	9.4	12
YOLO11n	2.5	3	6.3	12	434.78	2	6.7	2
YOLOv9tr [30]	10.0	11	40.2	4	14.58	14	7.1	5
RDD-YOLO [64]	2.55	4	6.6	10	263.16	7	8.3	8
EMTTA-n [60]	2.8	7	4.1	14	23.20	13	8.6	9
RT-DETR-Pothole [63]	32.8	14	108.0	1	66.67	12	7.0	4
YOLOv7	36.4	15	103.0	2	97.09	10	7.0	4
YOLOv7-C3ECA-DSA	32.6	13	93.7	3	91.74	11	6.3	1

parameters was ranked as shown in Table 3. By aggregating rankings across all evaluated criteria, this metric offers a comprehensive perspective on model performance. The proposed YOLOv7-C3ECA-DSA achieved the lowest average rank of 4.8, outperforming all other detection models, including EMTTA-n (5.0), RT-DETR (5.2), and YOLOv9tr (5.4). This result underscores the balanced design of YOLOv7-C3ECA-DSA, which excels in achieving both high accuracy and computational efficiency.

B. TRADEOFF BETWEEN SPEED AND ACCURACY

The Faster RCNN variants, though comprehensive in their approach, exhibited prohibitively high inference times due to their architectural complexity. Consequently, these models appeared as outliers and were not included in the scatter plot visualization, as their significantly higher computational cost rendered them impractical for real-time applications.

Most YOLO models exhibited optimal inference times but were unable to surpass the mAP0.5 threshold of 0.8, with YOLOv7 proving to be a notable exception. Although specialized approaches such as EMTTA-n and YOLOv9tr attained high accuracy levels with mAP0.5 of 0.842 and 0.851, respectively, their significantly increased inference times, 43.1 and 68.6 ms, rendered them impractical for real-time applications, despite their impressive detection capabilities.

When factoring in computational efficiency metrics such as GFLOPs, FPS, and parameters, YOLOv7-C3ECA-DSA emerges as the superior model for real-time pothole detection. YOLOv7-C3ECA-DSA not only achieves a higher mAP0.5 of 0.853, but it does so with 32.6 GFLOPs, which is 19.1% more efficient compared to YOLOv9tr's 40.2 GFLOPs, allowing for faster processing. In terms of FPS, YOLOv7-C3ECA-DSA leads with 91.74 FPS, representing a 525.6% improvement over YOLOv9tr (which achieves 14.58 FPS) and a 29.9% increase compared to EMTTA-n (which achieves 23.2 FPS). Furthermore, with only 32.6 million parameters, YOLOv7-C3ECA-DSA is more lightweight compared to YOLOv9tr (which has 40.2 million parameters), resulting in faster deployment and less memory consumption, crucial for real-time applications. Despite the slightly increased inference time (10.9 ms)

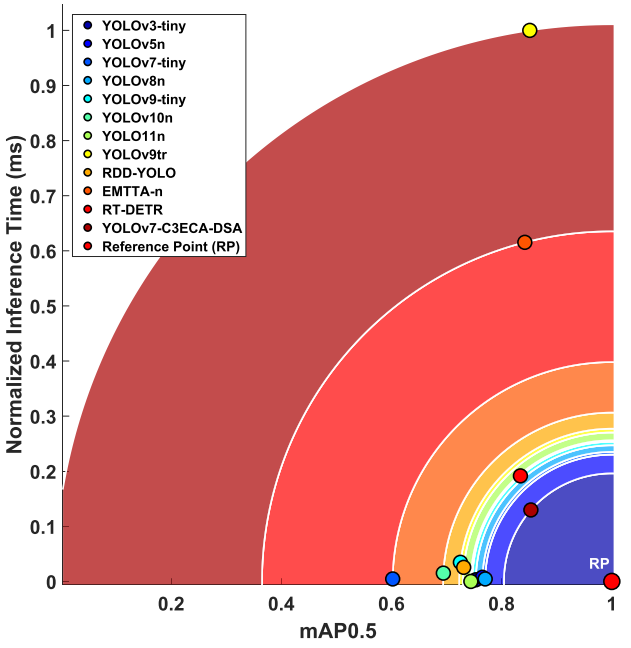


FIGURE 6. mAP0.5 vs normalized inference time plot of different models.

compared to RDD-YOLO (3.8 ms), YOLOv7-C3ECA-DSA compensates for this with an 11.68% increase in mAP0.5, making it the ideal solution for balancing accuracy and speed.

The contour plot, in Figure 6, further visualizes the trade-off between mAP0.5 and normalized inference time, with the ideal reference point (RP) representing the optimal performance, defined as a normalized inference time of 0 ms and mAP0.5 of 1. The normalized inference times were calculated using the following formula:

$$t_{normalized} = \frac{t - t_{min}}{t_{max} - t_{min}} \tag{6}$$

where t is the respective model's inference time, t_{max} is the highest inference time among the compared models (e.g., 465 ms for Faster RCNN-Resnet50), and t_{min} is the lowest inference time among the compared models (e.g., 2.3 ms for YOLOv11n).

The Euclidean distance was calculated to quantify the proximity of each model to this reference point. As depicted in the graph, YOLOv7-C3ECA-DSA is positioned closest to the reference point, with the shortest Euclidean distance among all models evaluated. This highlights the balanced optimization of accuracy and inference time. In contrast, models like EMTTA-n and YOLOv9tr, while achieving relatively high mAP0.5 scores, are further from the ideal performance due to their higher inference times.

The strong performance of our proposed model, YOLOv7-C3ECA-DSA, is due to key architectural improvements that enhance feature extraction and computational efficiency. The C3 block, adapted from YOLOv5, combines spatial and channel information to improve feature representation, while the addition of ECA allows adaptive, channel-wise recalibration to emphasize important features. This combination strengthens the ability of the model to detect objects in complex scenes. Moreover, the use of depthwise convolution in shuffle attention minimizes complexity, while the shuffle mechanism improves cross-channel interaction, optimizing spatial and depthwise feature utilization. Together, these innovations enable YOLOv7-C3ECA-DSA to achieve high detection accuracy with manageable inference times, making it effective and efficient for pothole detection.

C. COMPARISON FOR ADDITIONAL DATA POTHOLE DETECTION IN ADVERSE VISION CONDITIONS

To validate the robustness of YOLOv7-C3ECA-DSA under challenging conditions, additional tests were conducted using rainy and nighttime datasets. Comparisons were focused on YOLOv5n, YOLOv7, YOLOv11n, and RT-DETR models, which were among the top five ranked performers in previous evaluations. The detailed performance results for these conditions are summarized in Table 4.

In rainy weather conditions, the proposed YOLOv7-C3ECA-DSA model demonstrated superior performance compared to the original YOLO versions, namely YOLOv5n, YOLOv7, and YOLOv11n, achieving at least a 10% improvement in mAP0.5 over these models. However, a slight decrease of 2.2% in mAP0.5 was observed when compared to YOLOv9tr, indicating a minor trade-off in performance against the latest architecture.

Under nighttime conditions, the YOLOv7-C3ECA-DSA model continued to show strong performance, with an mAP0.5 improvement of at least 5.4% compared to YOLOv5n, YOLOv7, and YOLOv11n. Nevertheless, when compared to YOLOv9tr, a noticeable decrease of 9.0% in mAP0.5 was recorded, suggesting that YOLOv9tr is more robust under extremely low-light scenarios but at the cost of higher complexity.

In terms of inference time, the proposed YOLOv7-C3ECA-DSA model exhibited slightly higher latency compared to YOLOv5n, YOLOv11n, and YOLOv7, which is expected given the performance enhancements achieved. However, the inference time and GFLOPs of YOLOv7-C3ECA-DSA remain significantly lower than those of RT-DETR, which suffers from substantial computational overhead despite showing inferior detection performance in adverse weather conditions. This limitation aligns with observations made by Liu et al. [65], who reported that while RT-DETR performs adequately under standard scenarios, it struggles with degraded visibility environments such as fog, rain, and nighttime, leading to reduced target detection accuracy. Although YOLOv9tr achieved a slightly higher detection performance, it incurred the highest inference time among all models tested (33.0 ms), representing a considerable increase relative to the proposed YOLOv7-C3ECA-DSA model, thereby impacting its suitability for real-time applications.

To further understand the behavior of the proposed YOLOv7-C3ECA-DSA model under challenging conditions, a qualitative visual comparison was conducted against

TABLE 4. Performance comparison under rainy and nighttime conditions.

	Rainy Condition				Nighttime Condition			
Model	**mAP0.5**	**F1-Score**	**Inference Time (ms)**	**GFLOPs**	**mAP0.5**	**F1-Score**	**Inference Time (ms)**	**GFLOPs**
YOLOv5n [62]	0.431	0.527	6.3	5.8	0.456	0.601	6.5	5.8
YOLO11n	0.493	0.577	5.9	6.3	0.488	0.608	6.2	6.3
RT-DETR-Pothole [63]	0.475	0.494	12.4	103.4	0.495	0.606	14.2	108.0
YOLOv9tr [30]	0.545	0.452	33.0	40.2	0.631	0.660	50.7	40.2
YOLOv7	0.505	0.589	9.6	103.0	0.513	0.649	11.6	103.0
YOLOv7-C3ECA-DSA	0.523	0.610	11.5	93.7	0.541	0.654	12.6	103.2

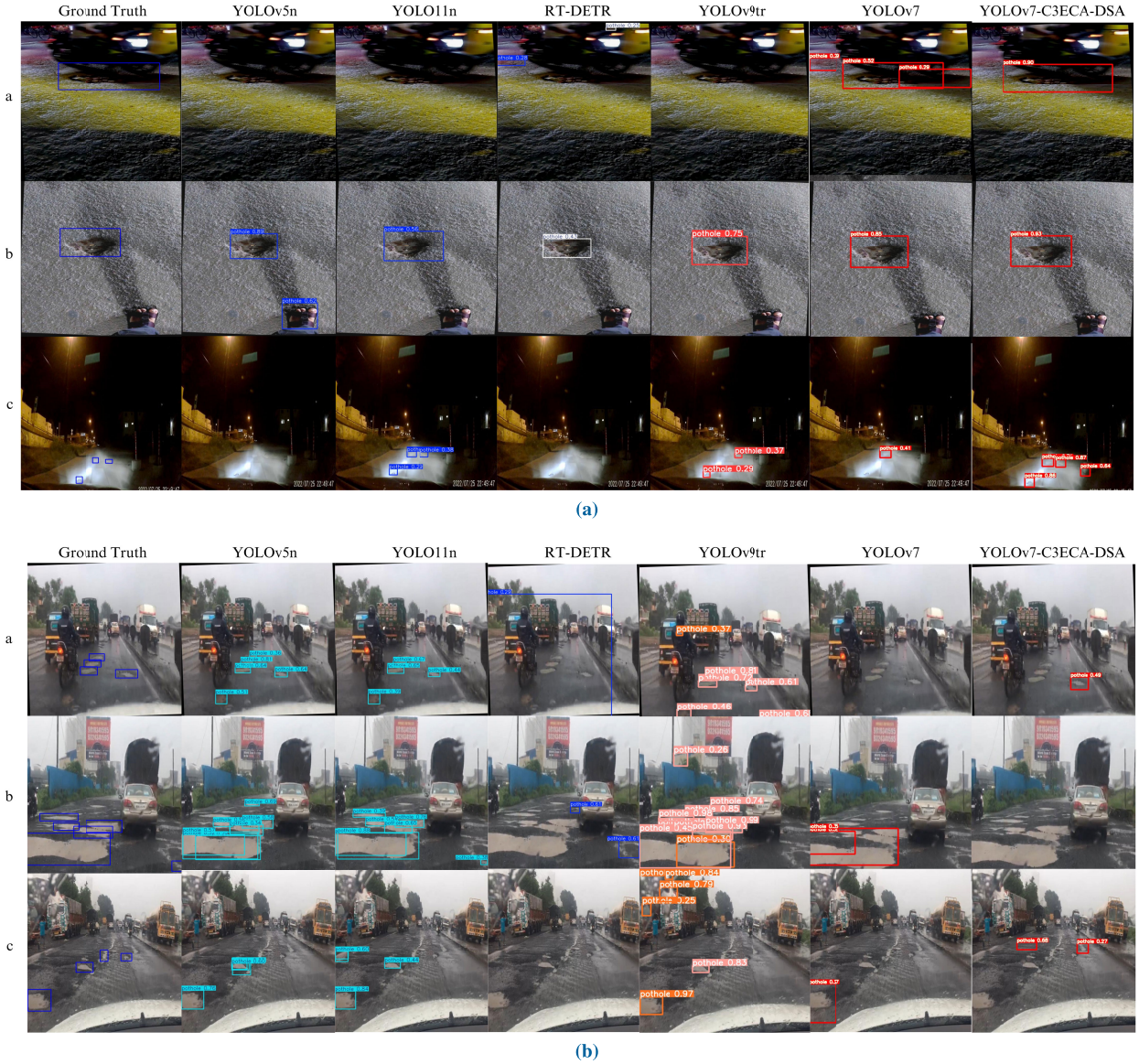


FIGURE 7. Detection performance comparison under adverse vision conditions: (a) Nighttime and (b) Rainy.

several baseline models, including YOLOv5n, YOLOv7, YOLOv11n, RT-DETR, and YOLOv9tr, using nighttime datasets. The visualization results are presented in Figure 7, and detailed observations are discussed below.

Overall, the YOLOv7-C3ECA-DSA model demonstrated more precise and reliable pothole detection compared to the baseline models. It consistently produced tighter and cleaner bounding boxes with fewer false positives,

outperforming YOLOv5n and YOLOv11n, which exhibited loose detections and high false positive rates, and YOLOv7, which occasionally overpredicted. Compared to RT-DETR, YOLOv7-C3ECA-DSA showed superior object separation, avoiding the merging of multiple potholes into single large detections, although minor underfitting and confidence drops were observed under shadow conditions. In comparison with YOLOv9tr, YOLOv7-C3ECA-DSA adopted a more conservative detection strategy, maintaining tighter bounding boxes and lower false positive rates, at the slight cost of missing very small or distant potholes under extreme lighting. These findings highlight the model's strong balance between precision and robustness in complex nighttime driving environments.

To further evaluate the detection performance under complex rainy conditions, a qualitative visual comparison was conducted across several baseline models, including YOLOv5n, YOLOv7, YOLOv11n, RT-DETR, and YOLOv9tr, as illustrated in Figure 7b. The proposed YOLOv7-C3ECA-DSA model consistently demonstrated superior localization and cleaner detections compared to YOLOv5n, YOLOv7, and YOLOv11n, with tighter bounding boxes and fewer false positives, although slight underprediction of smaller potholes was observed. Compared to RT-DETR, YOLOv7-C3ECA-DSA exhibited better object separation, avoiding the merging of multiple potholes into large bounding boxes. When assessed against YOLOv9tr, YOLOv7-C3ECA-DSA adopted a more conservative detection strategy, producing fewer but more accurate detections, whereas YOLOv9tr, despite detecting more potholes, tended to generate redundant detections and a higher rate of false positives. Overall, YOLOv7-C3ECA-DSA achieved a favorable balance between detection precision and robustness, particularly in challenging and cluttered rainy environments.

D. ABLATION STUDY

This study investigates several enhancements incorporated into the YOLOv7 object detection model and conducts ablation analysis to assess the individual and improving efficiency of these modifications in Table 5.

The baseline YOLOv7 achieved an mAP0.5 of 0.841 and an F1-score of 0.806. When evaluated individually, both the C3 and ECA modules demonstrated a drop in their performance, with C3 reducing mAP0.5 by 8.52% and F1-score by 2.63%, while ECA decreased mAP0.5 by 2.18% and F1-score by 3.26%. Interestingly, the combination of C3 and ECA resulted in a modest performance enhancement of 0.5% from YOLOv7. This interaction can be attributed to the complementary nature of the modules—the ability of the C3 module to capture hierarchical features and the ECA module's feature prioritization capabilities, which work in parallel with the optimization of the model's performance.

The baseline and ablation variants were also ranked as in Table 5 to further highlight the overall superiority of YOLOv7-C3ECA-DSA in ablation studies. By aggregating the ranks of all evaluation criteria, YOLOv7-C3ECA-DSA achieved the lowest average rank of 2.25, demonstrating the best balance between accuracy, computational efficiency, and inference time. In comparison, the baseline YOLOv7 achieved an average rank of 3.0, while other variations, such as YOLOv7-C3 and YOLOv7-ECA, resulted in average ranks of 3.71 and 4.14, respectively. This consistent performance across metrics emphasizes the strength of the proposed modifications, with YOLOv7-C3ECA-DSA emerging as the optimal configuration.

YOLOv7-C3ECA-DSA further demonstrates its superiority by outperforming other variants in both GFLOPs and FPS. Despite incorporating additional complexity through the C3 and ECA modules, YOLOv7-C3ECA-DSA achieves the lowest GFLOPs of 93.7, making it the most computationally efficient model compared to all variants. Additionally, YOLOv7-C3ECA-DSA achieves an FPS of 91.74, which is significantly higher than YOLOv7-C3 (55.87 FPS) and YOLOv7-ECA (49.02 FPS). This optimal balance between accuracy, GFLOPs, and FPS is particularly important for real-time applications, such as UAV-based pothole detection, where high inference speed and low computational cost are crucial. YOLOv7-C3ECA-DSA stands out not only in terms of detection accuracy but also in its computational efficiency, making it the ideal choice for real-time, high-performance systems.

The C3 module effectively captures features at different scales and provides multi-scale spatial features, while the ECA component selectively prioritizes these features based on their relevance and refines channel relationships, allowing the model to have better feature representation and focus on the more informative aspects of the input. When integrating Shuffle Attention into the YOLOv7-C3ECA architecture, we observed a reduction in both inference time and parameters, with minimal performance trade-offs, which is a slight decrease of 0.36% in mAP0.5 and 0.48% in F1-score. This minor performance decrease is justified by the significant computational efficiency gained through the implementation of pointwise group convolution in Shuffle Attention, which reduces parameters and computations by processing partitioned input channels in separate groups, resulting in a more lightweight yet effective model.

E. EVALUATION OF YOLOv7, YOLOv7-C3ECA, AND YOLOv7-C3ECA-DSA

The precision-recall curve in Figure 8 illustrates the comparative performance of the baseline YOLOv7 model, YOLOv7-C3ECA, and the proposed YOLOv7-C3ECA-DSA model. From the curve, the YOLOv7-C3ECA-DSA model demonstrates better performance than the other models in most recall thresholds. The higher precision at equivalent recall levels, compared to YOLOv7-C3ECA and the baseline YOLOv7, suggests that the inclusion of Shuffle Attention and Depthwise convolution improves the model's ability to detect true positives while reducing false positives.

The YOLOv7-C3ECA model exhibits a moderate improvement over the baseline YOLOv7, especially at higher

TABLE 5. Result of ablation analyses on the dataset.

	mAP0.50		mAP0.50-0.95		F1-Score		Inference Time (ms)	
	Value	Rank	Value	Rank	Value	Rank	Value	Rank
YOLOv7	0.841	4	0.549	3	0.8063	2	10.3	3
YOLOv7-C3ECA	0.845	2	0.553	2	0.8023	4	20.4	6
YOLOv7-ECA	0.823	5	0.544	4	0.7808	6	17.9	5
YOLOv7-C3ECA-SA	0.842	3	0.541	5	0.8062	3	7.2	2
YOLOv7-C3	0.775	6	0.454	6	0.7856	5	6.7	1
YOLOv7-C3ECA-DSA	0.853	1	0.557	1	0.823	1	10.9	4

	Parameter (M)		GFLOPs		FPS		Average Rank	
	Value	Rank	Value	Rank	Value	Rank	Value	Rank
YOLOv7	36.4	4	103	4	97.09	3	3.29	3
YOLOv7-C3	38.8	5	107	2	49.02	6	3.86	5
YOLOv7-ECA	6.1	1	104.3	3	55.87	5	4.14	6
YOLOv7-C3ECA	32.6	2	94.3	5	138.89	2	3.14	2
YOLOv7-C3ECA-SA	60.6	6	148.3	1	149.25	1	3.71	4
YOLOv7-C3ECA-DSA	32.6	2	93.7	6	91.74	4	2.25	1

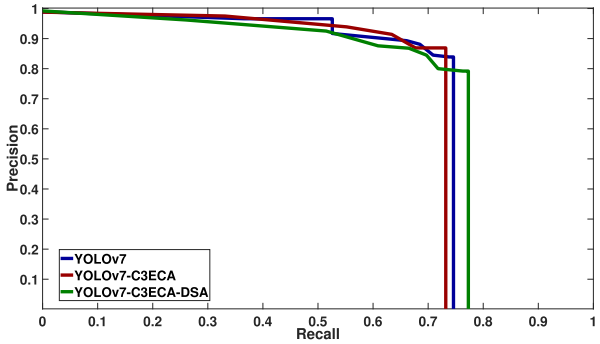


FIGURE 8. The precision-recall curve for YOLOv7, YOLOv7-C3ECA, and YOLOv7-C3ECA-DSA.

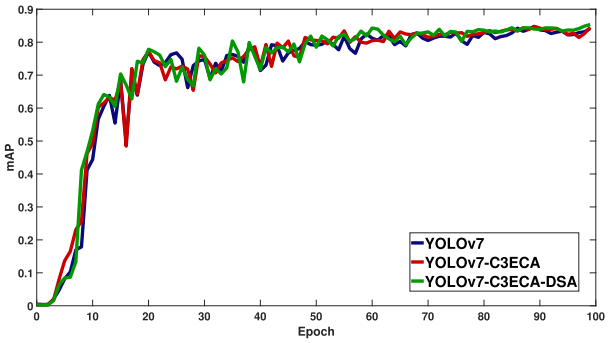


FIGURE 9. mAP vs epochs graph of YOLOv7, YOLOv7-C3ECA and YOLOv7-C3ECA-DSA.

recall levels. This improvement can be attributed to the inclusion of the C3 and ECA modules, which enhance feature extraction and prioritization. Conversely, the baseline YOLOv7 shows a decline in precision as recall increases, indicating its challenges in balancing these metrics at higher thresholds.

Our proposed model achieves a more balanced trade-off between precision and recall, maintaining relatively high precision at higher recall levels. This performance improvement can be linked to the synergy of the C3, ECA, and shuffle attention modules, which together enhance feature representation, spatial focus, and computational efficiency. These enhancements help the YOLOv7-C3ECA-DSA model to manage complex detection tasks with minimal compromises between precision and recall.

The graph in Figure 9 shows the mean Average Precision (mAP) performance across 100 epochs for YOLOv7, YOLOv7-C3ECA, and YOLOv7-C3ECA-DSA. The mAP measures the accuracy of object detection, and the progression across epochs provides an indication of how the models learn and stabilize during training.

All three models demonstrate a rapid increase in mAP during the initial 20 epochs, reflecting effective learning in the early stages. After this, the growth in mAP slows down, indicating that the models are approaching convergence. YOLOv7-C3ECA-DSA maintains slightly higher mAP values throughout training, reaching the highest mAP at the end of 100 epochs. YOLOv7-C3ECA achieves a slightly higher mAP than the baseline YOLOv7, but it falls short of YOLOv7-C3ECA-DSA in the later epochs. This suggests that the inclusion of shuffle attention in YOLOv7-C3ECA-DSA contributes to improved learning by enhancing feature extraction. The baseline YOLOv7 performs consistently but does not achieve the same final mAP as the modified models.

The mAP progression indicates that YOLOv7-C3ECA-DSA benefits from the integration of C3, ECA, and shuffle attention, which appear to improve the model's ability to learn and generalize over the training epochs.

Figures 10 and 11 show precision, recall, and F1-score plotted at different thresholds: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0.

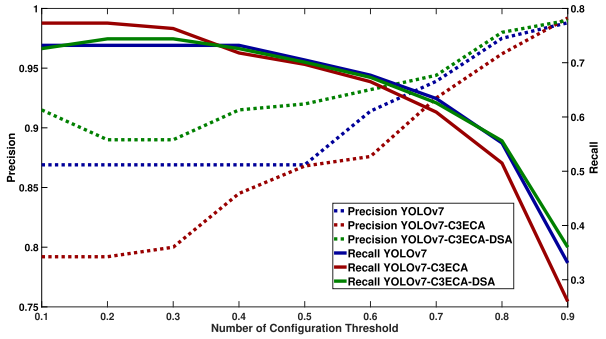


FIGURE 10. The precision and recall under different confidence thresholds in YOLOv7, YOLOv7-C3ECA and YOLOv7-C3ECA-DSA.

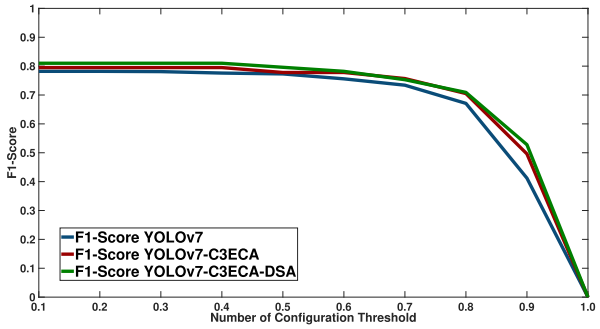


FIGURE 11. The F1-score under different confidence thresholds in YOLOv7, YOLOv7-C3ECA and YOLOv7-C3ECA-DSA.

Our proposed model demonstrated substantial improvements in precision across increasing confidence thresholds, particularly noteworthy in the range of 0.1 to 0.7. This suggests enhanced positive prediction abilities and more accurate identification of target objects. The YOLOv7-C3ECA model shows slightly better precision than the baseline YOLOv7 in the lower threshold ranges but is outperformed by YOLOv7-C3ECA-DSA as the threshold increases, demonstrating more consistent precision at high thresholds in the latter part. This consistent performance is indicative of the C3 and ECA modules working synergistically to optimize precision even under more stringent confidence conditions, enabling more reliable detection of objects as the confidence threshold rises. However, the recall values showed minor fluctuation between the models, implying similar capabilities in detecting all relevant instances. YOLOv7-C3ECA maintained a stable recall profile similar to YOLOv7, while YOLOv7-C3ECA-DSA displayed slightly higher recall values at moderate thresholds. This indicates that YOLOv7-C3ECA-DSA is better at maintaining a high detection rate, even when faced with moderate confidence thresholds, further improving its overall detection performance.

The evaluation of the F1 score across various confidence thresholds revealed the superior performance of the proposed model compared to the base YOLOv7 and YOLOv7-C3ECA models. YOLOv7-C3ECA demonstrates an intermediate performance between YOLOv7 and the proposed YOLOv7-C3ECA-DSA across all thresholds. While YOLOv7-C3ECA improves over the baseline YOLOv7, particularly in lower thresholds, its performance plateaus at higher thresholds, where YOLOv7-C3ECA-DSA continues to maintain a consistent edge. This consistent behavior highlights the additional benefit provided by the integration of shuffle attention in YOLOv7-C3ECA-DSA, which helps in balancing precision and recall. The incorporation of shuffle attention in YOLOv7-C3ECA-DSA not only improves detection accuracy but also enhances the F1-score, making it more effective in real-time object detection tasks where both precision and recall are equally important.

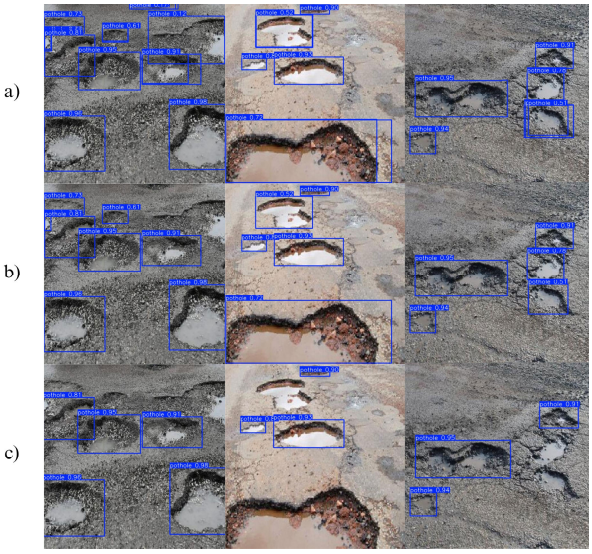
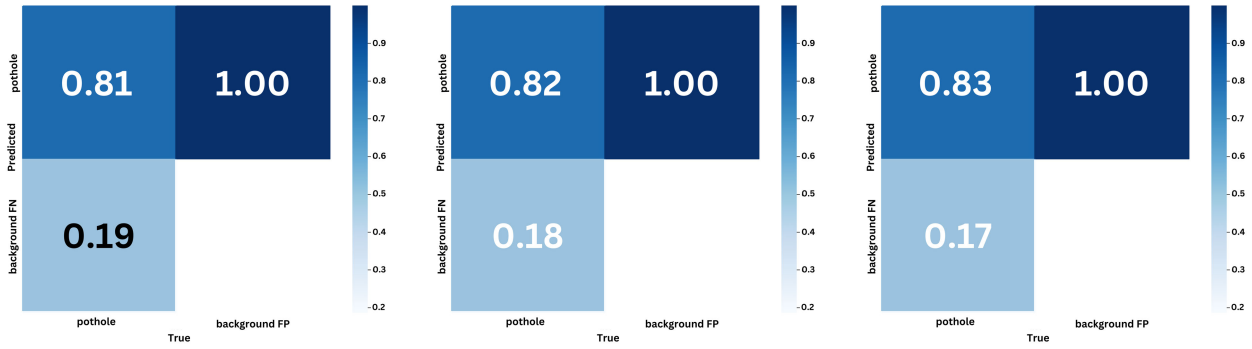


FIGURE 12. Visualization result at different confidence threshold level: a) 0.1 b) 0.5 and c) 1.0.

The confidence threshold significantly impacts object detection performance by controlling the trade-off between precision and recall. At a low threshold, as shown in Figure 12(a), the detection includes most of the object, resulting in a high recall, but includes many false positives, resulting in clutter visualization. At a medium threshold, as shown in Figure 12(b), it reduces false positives and overlapping detections while maintaining a balance between recall and precision. At a high threshold, as shown in Figure 12(c), the model keeps only the most confident detections, improving precision but reducing recall and potentially missing some objects.

Based on Figure 13(a), 13(b), and 13(c), analysis of the confusion matrices demonstrated that while both models encountered false positive detections and failed to identify some potholes, the YOLOv7-C3ECA-DSA architecture achieved a superior true positive rate. This suggests enhanced detection capability in comparison to the original YOLOv7 model and YOLOv7-C3ECA, although both continue to struggle in accurately differentiating between background and pothole. The YOLOv7-C3ECA model shows some improvement over YOLOv7 in reducing false positives. However, its performance is slightly behind



(a) (b) (c)

FIGURE 13. Comparison of the confusion matrix with (a) YOLOv7, (b) YOLOv7-C3ECA, and (c) YOLOv7-C3ECA-DSA.

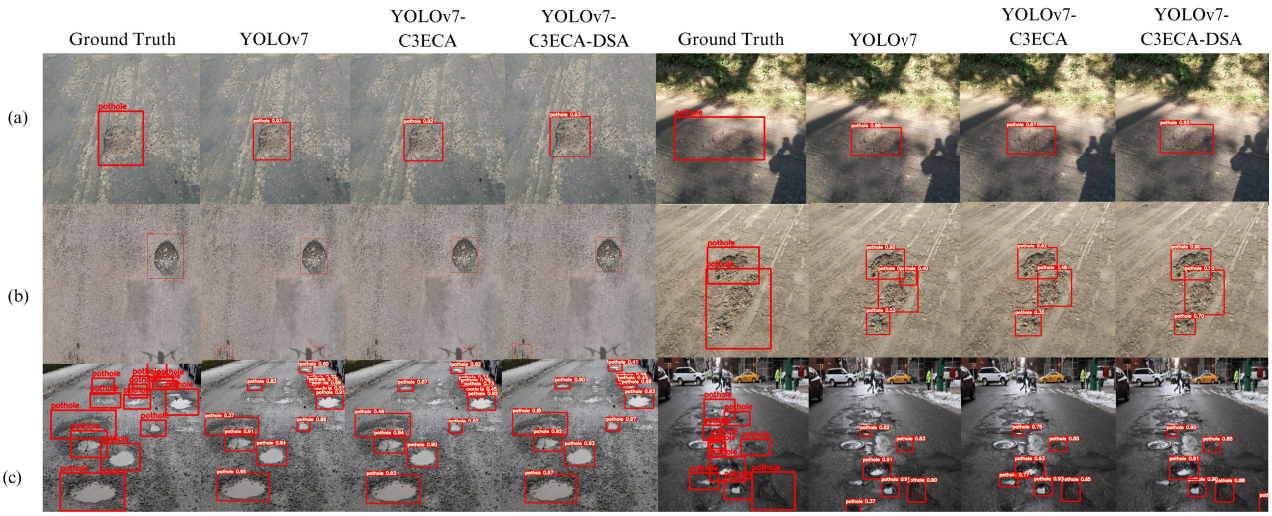


FIGURE 14. Visualization results of YOLOv7, YOLOv7-C3ECA and YOLOv7-C3ECA-DSA against ground truth.

YOLOv7-C3ECA-DSA, which benefits from additional enhancements that improve object differentiation and feature extraction.

The comparative analysis of detection performance across diverse scenarios reveals distinctive characteristics between the models. Figure 14(a) demonstrates that YOLOv7 detects the ground truth images but with lower confidence levels. YOLOv7-C3ECA shows some improvement in confidence due to its enhanced feature prioritization through the ECA module. The YOLOv7-C3ECA-DSA model further increases confidence levels, reflecting the impact of depthwise shuffle attention in refining feature selection. Figure 14(b) highlights the tendency of YOLOv7 and YOLOv7-C3ECA to overpredict, as evidenced by its detection of non-existent potholes and misclassification of drone shadows as potholes; nevertheless, our proposed model is able to correctly identify the pothole. The YOLOv7-C3ECA-DSA model builds on this improvement, which can be attributed to the depthwise shuffle attention mechanism, which facilitates capturing relevant features [66] from background noise; hence, it has clearer separation between potholes and road texture. Figure 14(c) illustrates a common limitation of all three models, where accurate detection is compromised by low contrast between potholes and surrounding asphalt, particularly in aged or weathered roads where degraded surfaces exhibit similar textural patterns to actual potholes. This limitation can lead to true negatives and interrupt its ability to detect potholes accurately. Despite YOLOv7 exhibiting a marginally higher overall mAP0.5, the proposed model demonstrates superior practical performance through reduced incidences of missed detections in pothole identification.

The comparative analysis of the prediction result across YOLOv7, YOLOv7-C3ECA and YOLOv7-C3ECA-DSA in Table 5 demostrates significant improvements and trade-offs in detection performance. YOLOv7-C3ECA improved the correctly predicted values by 10% over YOLOv7, while YOLOv7-C3ECA-DSA achieved a further 23.6% increment, reflecting the effectiveness of the C3, ECA, and

TABLE 6. Prediction results for YOLOv7, YOLOv7-C3ECA, and YOLOv7-C3ECA-DSA.

Models	Correctly Predicted	Overpredicted	Underpredicted
YOLOv7	65 (32.7%)	57 (28.6%)	77 (38.6%)
YOLOv7-C3ECA	85 (42.7%)	33 (16.6%)	81 (40.7%)
YOLOv7-C3ECA-DSA	112 (56.3%)	52 (26.1%)	35 (17.6%)

depthwise shuffle attention modules in enhancing detection accuracy. Overpredictions were reduced by 12% in YOLOv7-C3ECA and 2.5% in YOLOv7-C3ECA-DSA compared to YOLOv7, with YOLOv7-C3ECA-DSA provided a balanced approach. As for underpredictions, YOLOv7-C3ECA showed a 2.10% increase, indicating a slight trade-off in recall, whereas YOLOv7-C3ECA-DSA demostrated a substantial 21% reduction, highlighting its capability to improve both recall and precision effectively.

V. CONCLUSION

In this study, an improved YOLOv7-based pothole detection model is proposed to accurately identify potholes for real-time applications. The model integrates the C3 block, which provides multi-scale spatial features and captures hierarchical spatial relationships, with the ECA block, which refines channel relationships and emphasizes important channels within the YOLOv7 backbone. This synergistic combination significantly enhances relevant feature representations and improves pothole detection capabilities, as the hierarchical spatial information from the C3 block is effectively complemented by ECA's channel-wise feature discrimination. Moreover, the incorporation of a depthwise shuffle attention mechanism into the detection head further enhances feature mixing through efficient channel information exchange, facilitating improved spatial relationship modeling and feature discrimination. These enhancements enable the model to better recognize pothole shapes and sizes, achieving clearer separation between potholes and the surrounding road texture. Consequently, the YOLOv7-C3ECA-DSA architecture results in enhanced feature learning capabilities, making it particularly beneficial for object detection tasks. Experimental evaluations demonstrate that the YOLOv7-C3ECA-DSA model outperforms the baseline YOLOv7 by achieving a 1.43% improvement in mAP0.5, while maintaining a reasonable inference time of 10.7 ms, thereby meeting the requirements for high-accuracy, real-time detection.

While the YOLOv7-C3ECA-DSA model demonstrates significant improvements in detection accuracy and real-time performance, several limitations remain. First, although the model was evaluated under adverse conditions such as nighttime and rainy environments, the detection performance under these challenging scenarios was found to be lower than expected, indicating that further enhancements are needed to improve robustness in diverse and complex real-world settings. Second, the model, despite improved efficiency, still requires considerable computational resources compared to lightweight architectures, potentially limiting its deployment on ultra-low-power devices without further compression. Finally, while initial results are promising, broader validation across different geographical regions and road surface types is necessary to confirm the generalizability of the model.

Although the current study emphasizes improving the accuracy and efficiency of pothole detection for real-time implementation, several avenues for further enhancement remain. First, incorporating model compression and pruning techniques could reduce the computational cost, making the model even more efficient for real-time applications without sacrificing accuracy. Specifically, model compression methods such as weight quantization and knowledge distillation could reduce the model's memory footprint, making it more suitable for deployment on resource-constrained edge devices. Furthermore, pruning techniques, particularly structured pruning, could eliminate redundant or less critical connections, resulting in a leaner and faster model [67]. These techniques are essential for the practical deployment of the model on edge devices such as smartphones, embedded systems, and UAVs, where computational power and energy efficiency are critical considerations.

Additionally, future studies could explore the potential of self-supervised learning to further enhance feature learning capabilities. By leveraging large unlabeled datasets, self-supervised learning could improve the model's generalization ability, particularly in diverse environments where labeled data is scarce [68]. Moreover, incorporating more diverse datasets, including those captured from UAVs and other pothole sources, could further validate and enhance the proposed method's effectiveness across a broader range of real-world scenarios.

The findings of this study also have important practical implications for real-world pothole detection systems. The proposed YOLOv7-C3ECA-DSA model's ability to balance high detection accuracy with real-time processing capability makes it well-suited for deployment in smart city infrastructure, autonomous vehicles, and UAV-based road monitoring. By enhancing feature learning through hierarchical spatial and channel-wise attention mechanisms, the model is better equipped to detect potholes of varying shapes and sizes under challenging conditions. Furthermore, with future integration of model compression and pruning techniques, the architecture has the potential to be deployed efficiently on resource-constrained edge devices, enabling scalable, wide-area road condition monitoring systems.

REFERENCES

[1] P. Novotny and M. Janosikova, "Designating regional elements system in a critical infrastructure system in the context of the Czech republic," *Systems*, vol. 8, no. 2, p. 13, Apr. 2020, doi: 10.3390/systems8020013.

[2] M. Lacinák, "Resilience of the smart transport system–risks and aims," *Transp. Res. Proc.*, vol. 55, pp. 1635–1640, Jan. 2021, doi: 10.1016/j.trpro.2021.07.153.

[3] M. Cingel, M. Drliciak, and J. Celko, "Morning modal split model of economically active people in zilina region," *Transp. Res. Proc.*, vol. 55, pp. 1065–1072, Jan. 2021, doi: 10.1016/j.trpro.2021.07.077.

[4] D.-H. Heo, J.-Y. Choi, S.-B. Kim, T.-O. Tak, and S.-P. Zhang, "Image-based pothole detection using multi-scale feature network and risk assessment," *Electronics*, vol. 12, no. 4, p. 826, Feb. 2023, doi: 10.3390/electronics12040826.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Sep. 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[6] R. Girshick, "Fast R-CNN," 2015, *arXiv:1504.08083*.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," 2017, *arXiv:1703.06870*.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, Dec. 2015, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015, *arXiv:1506.02640*.

[11] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," 2016, *arXiv:1612.08242*.

[12] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[13] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[14] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475, doi: 10.1109/cvpr52729.2023.00721.

[15] T. Y. Goh, S. N. Basah, H. Yazid, M. J. A. Safar, and F. S. A. Saad, "Performance analysis of image thresholding: Otsu technique," *Measurement*, vol. 114, pp. 298–307, Jan. 2018, doi: 10.1016/j.measurement.2017.09.052.

[16] C. A. Glasbey, "An analysis of histogram-based thresholding algorithms," *CVGIP: Graph. Models Image Process.*, vol. 55, no. 6, pp. 532–537, Nov. 1993, doi: 10.1006/cgip.1993.1040.

[17] Y. O. Ouma and M. Hahn, "Pothole detection on asphalt pavements from 2D-colour pothole images using fuzzy c-means clustering and morphological reconstruction," *Autom. Construction*, vol. 83, pp. 196–211, Nov. 2017, doi: 10.1016/j.autcon.2017.08.017.

[18] S. N. Abd Mukti and K. N. Tahar, "Detection of potholes on road surfaces using photogrammetry and remote sensing methods (review)," *Scientific Tech. J. Inf. Technol., Mech. Opt.*, vol. 22, no. 3, pp. 459–471, Jun. 2022, doi: 10.17586/2226-1494-2022-22-3-459-471.

[19] M. Mahdianpari, B. Salehi, M. Rezaee, F. Mohammadimanesh, and Y. Zhang, "Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery," *Remote Sens.*, vol. 10, no. 7, p. 1119, Jul. 2018, doi: 10.3390/rs10071119.

[20] C. Saisree and U. Kumaran, "Pothole detection using deep learning classification method," *Proc. Comput. Sci.*, vol. 218, pp. 2143–2152, Jan. 2023, doi: 10.1016/j.procs.2023.01.190.

[21] F. Andika and Y. Bandung, "Road damage classification using SSD mobilenet with image enhancement," in *Proc. Int. Conf. Comput. Sci., Inf. Technol. Eng. (ICCoSITE)*. Bedford, MA, USA: Digital, Feb. 2023, pp. 540–545, doi: 10.1109/ICCoSITE57641.2023.10127763.

[22] J. Zhu, J. Zhong, T. Ma, X. Huang, W. Zhang, and Y. Zhou, "Pavement distress detection using convolutional neural networks with images captured via UAV," *Autom. Construction*, vol. 133, Jan. 2022, Art. no. 103991, doi: 10.1016/j.autcon.2021.103991.

[23] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2778–2788, doi: 10.1109/ICCVW54120.2021.00312.

[24] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," *Proc. Comput. Sci.*, vol. 199, pp. 1066–1073, Jan. 2022, doi: 10.1016/j.procs.2022.01.135.

[25] J. Gao, Y. Chen, Y. Wei, and J. Li, "Detection of specific building in remote sensing images using a novel YOLO-S-CIOU model. Case: gas station identification," *Sensors*, vol. 21, no. 4, p. 1375, Feb. 2021, doi: 10.3390/s21041375.

[26] S. Tan, G. Lu, Z. Jiang, and L. Huang, "Improved YOLOv5 network model and application in safety helmet detection," in *Proc. IEEE Int. Conf. Intell. Saf. Robot. (ISR)*, Mar. 2021, pp. 330–333, doi: 10.1109/ISR50024.2021.9419561.

[27] F. Wan, C. Sun, H. He, G. Lei, L. Xu, and T. Xiao, "YOLO-LRDD: A lightweight method for road damage detection based on improved YOLOv5s," *EURASIP J. Adv. Signal Process.*, vol. 2022, no. 1, pp. 1–15, Oct. 2022, doi: 10.1186/s13634-022-00931-x.

[28] Y. Chu, X. Xiang, Y. Wang, and B. Huang, "Pavement disease detection through improved YOLOv5s neural network," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, Oct. 2022, doi: 10.1155/2022/1969511.

[29] N. Bhavana, M. M. Kodabagi, B. M. Kumar, P. Ajay, N. Muthukumaran, and A. Ahilan, "POT-YOLO: Real-time road potholes detection using edge segmentation-based YOLO V8 network," *IEEE Sensors J.*, vol. 24, no. 15, pp. 24802–24809, Aug. 2024, doi: 10.1109/JSEN.2024.3399008.

[30] S. Youwai, A. Chaiyaphat, and P. Chaipetch, "YOLO9tr: A lightweight model for pavement damage detection utilizing a generalized efficient layer aggregation network and attention mechanism," *J. Real-Time Image Process.*, vol. 21, no. 5, p. 163, Oct. 2024, doi: 10.1007/s11554-024-01545-2.

[31] S. Nomqupu, A. Sali, A. Nyamugama, and N. Ndou, "Integrating sigmoid calibration function into entropy thresholding segmentation for enhanced recognition of potholes imaged using a UAV multispectral sensor," *Appl. Sci.*, vol. 14, no. 7, p. 2670, Mar. 2024, doi: 10.3390/app14072670.

[32] K. Ghafoor, "IoT and cloud based automated pothole detection model using extreme gradient boosting with texture descriptors," *Scalable Comput.*, vol. 24, no. 4, pp. 713–728, Nov. 2023, doi: 10.12694/scpe.v24i4.2176.

[33] G. Kim and S. Kim, "A road defect detection system using smartphones," *Sensors*, vol. 24, no. 7, p. 2099, Mar. 2024, doi: 10.3390/s24072099.

[34] F. Ozoglu and T. Gökgöz, "Detection of road potholes by applying convolutional neural network method based on road vibration data," *Sensors*, vol. 23, no. 22, p. 9023, Nov. 2023, doi: 10.3390/s23229023.

[35] D.-W. Jang and R.-H. Park, "Pothole detection using spatio-temporal saliency," *IET Intell. Transp. Syst.*, vol. 10, no. 9, pp. 605–612, Nov. 2016, doi: 10.1049/iet-its.2016.0006.

[36] S. Sattar, S. Li, and M. Chapman, "Road surface monitoring using smartphone sensors: A review," *Sensors*, vol. 18, no. 11, p. 3845, Nov. 2018, doi: 10.3390/s18113845.

[37] P. Wang, Y. Hu, Y. Dai, and M. Tian, "Asphalt pavement pothole detection and segmentation based on wavelet energy field," *Math. Problems Eng.*, vol. 2017, no. 1, Jan. 2017, Art. no. 1604130, doi: 10.1155/2017/1604130.

[38] E. Romero-Chambi, S. Villarroel-Quezada, E. Atencio, and F. Muñoz-La Rivera, "Analysis of optimal flight parameters of unmanned aerial vehicles (UAVs) for detecting potholes in pavements," *Appl. Sci.*, vol. 10, no. 12, p. 4157, Jun. 2020, doi: 10.3390/app10124157.

[39] M. H. Asad, S. Khaliq, M. H. Yousaf, M. O. Ullah, and A. Ahmad, "Pothole detection using deep learning: A real-time and AI-on-the-edge perspective," *Adv. Civil Eng.*, vol. 2022, no. 1, Jan. 2022, Art. no. 9221211, doi: 10.1155/2022/9221211.

[40] W. T. Mpofu, B. Ndlovu, S. Dube, and J. Mutengeni, "Pot-hole detection and reporting system using deep learning," in *Proc. 4th Asia Pacific Conference on Industrial Engineering and Operations Management*, 2023, pp. 1–12.

[41] X. Su, J. Zhang, Z. Ma, Y. Dong, J. Zi, N. Xu, H. Zhang, F. Xu, and F. Chen, "Identification of rare wildlife in the field environment based on the improved YOLOv5 model," *Remote Sens.*, vol. 16, no. 9, p. 1535, Apr. 2024, doi: 10.3390/rs16091535.

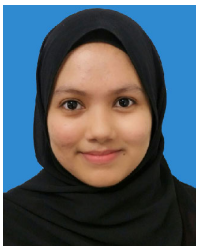
[42] Y. Liang, B. Wang, H. Huang, H. Pang, and X. Yue, "Bird detection algorithm incorporating attention mechanism," Res. Square, Durham, NC, USA, Tech. Rep. v1, Sep. 21, 2023, doi: 10.21203/rs.3.rs-3319901/v1.

[43] F. Ren, G. Wang, Z. Hu, M. Wu, and M. Devaraj, "Research on steel surface defect detection algorithm based on improved deep learning," *Int. J. Electr. Electron. Res.*, vol. 10, no. 4, pp. 1140–1145, Dec. 2022, doi: 10.37391/ijeer.100461.

[44] Y. Chen, "Insulator defect detection method upon fused attention mechanism and bidirectional feature fusion," *J. Phys., Conf. Ser.*, vol. 2632, no. 1, Nov. 2023, Art. no. 012013, doi: 10.1088/1742-6596/2632/1/012013.

[45] T. Fu, S. Zangenehpour, P. St-Aubin, L. Fu, and L. F. Miranda-Moreno, "Using microscopic video data measures for driver behavior analysis during adverse winter weather: Opportunities and challenges," *J. Mod. Transp.*, vol. 23, no. 2, pp. 81–92, Jun. 2015, doi: 10.1007/s40534-015-0073-3.

[46] Q. Guo, W. Feng, Z. Chen, R. Gao, L. Wan, and S. Wang, "Effects of blur and deblurring to visual object tracking," 2019, *arXiv:1908.07904*.
[47] W. Hao, L. Zhang, M. Han, K. Zhang, F. Li, G. Yang, and Z. Liu, "YOLOv5-SA-FC: A novel pig detection and counting method based on shuffle attention and focal complete intersection over union," *Animals*, vol. 13, no. 20, p. 3201, Oct. 2023, doi: 10.3390/ani13203201.
[48] F. Wu, Z. Cai, S. Fan, R. Song, L. Wang, and W. Cai, "Fish target detection in underwater blurred scenes based on improved YOLOv5," *IEEE Access*, vol. 11, pp. 122911–122925, 2023, doi: 10.1109/ACCESS.2023.3328940.
[49] A. Duani Rojas, L. Lagos, H. Upadhyay, J. Soni, and N. Prabakar, "AI-based detection and identification of low-level nuclear waste: A comparative analysis," *Neural Comput. Appl.*, vol. 36, no. 33, pp. 21061–21072, Nov. 2024, doi: 10.1007/s00521-024-10238-7.
[50] T. Komori, K. Sasaki, and T. Onoye, "Multi-class urinary sediment particles detection based on YOLOv7 with attention modules," *IEEE Access*, vol. 11, pp. 129753–129764, 2024, doi: 10.1109/ACCESS.2024.3448262.
[51] J. He, Y. Wang, Y. Wang, R. Li, D. Zhang, and Z. Zheng, "A lightweight road crack detection algorithm based on improved YOLOv7 model," *Signal, Image Video Process.*, vol. 18, no. S1, pp. 847–860, Aug. 2024, doi: 10.1007/s11760-024-03197-y.
[52] Z. Ning, H. Wang, S. Li, and Z. Xu, "YOLOv7-RDD: A lightweight efficient pavement distress detection model," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 6994–7003, Jul. 2024, doi: 10.1109/TITS.2023.3347034.
[53] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," 2018, *arXiv:1803.01534*.
[54] Z. Li and J. Li, "Automated surface defect detection for hot-pressed light guide plates based on GDA-YOLOv7," *IEEE Access*, vol. 11, pp. 142196–142213, 2023, doi: 10.1109/ACCESS.2023.3341928.
[55] H. Liu, F. Sun, J. Gu, and L. Deng, "SF-YOLOv5: A lightweight small object detection algorithm based on improved feature fusion mode," *Sensors*, vol. 22, no. 15, p. 5817, Aug. 2022, doi: 10.3390/s22155817.
[56] Y. He, C. Li, X. Li, and T. Bai, "A lightweight CNN based on axial depthwise convolution and hybrid attention for remote sensing image dehazing," *Remote Sens.*, vol. 16, no. 15, p. 2822, Jul. 2024, doi: 10.3390/rs16152822.
[57] Q.-L. Zhang and Y.-B. Yang, "SA-Net: Shuffle attention for deep convolutional neural networks," 2021, *arXiv:2102.00240*.
[58] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation networks," 2017, *arXiv:1709.01507*.
[59] Y. Wu and K. He, "Group normalization," 2018, *arXiv:1803.08494*.
[60] S. Wang, H. Jiao, X. Su, and Q. Yuan, "An ensemble learning approach with attention mechanism for detecting pavement distress and disaster-induced road damage," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 13667–13681, Oct. 2024, doi: 10.1109/TITS.2024.3391751.
[61] R. Atikur. (2022). *Annotated Potholes Image Dataset*. Accessed: Apr. 15, 2024. [Online]. Available: https://www.kaggle.com/chitholian/annotated-potholes-dataset
[62] T. Tamagusko and A. Ferreira, "Optimizing pothole detection in pavements: A comparative analysis of deep learning models," *Eng. Proc.*, vol. 36, no. 1, p. 11, Jun. 2023, doi: 10.3390/engproc2023036011.
[63] S. F. Mat Radzi, M. A. Abd Rahman, and M. L. A. Bin Mohamad, "RT-DETR-pothole: Lightweight real-time detection transformers for improved road pothole detection," in *Proc. 9th Int. Conf. Artif. Intell., Autom. Control Technol. (AIACT)*, 2025, pp. 1–7. [Online]. Available: https://www.researchgate.net/publication/392398955_RT-DETR-Pothole_Lightweight_Real-Time_Detection_Transformers_for_Improved_Road_Pothole_Detection#fullTextFileContent
[64] Y. Li, C. Yin, Y. Lei, J. Zhang, and Y. Yan, "RDD-YOLO: Road damage detection algorithm based on improved you only look once version 8," *Appl. Sci.*, vol. 14, no. 8, p. 3360, Apr. 2024, doi: 10.3390/app14083360.
[65] Z. Liu, C. Sun, and X. Wang, "DST-DETR: Image dehazing RT-DETR for safety helmet detection in foggy weather," *Sensors*, vol. 24, no. 14, p. 4628, Jul. 2024, doi: 10.3390/s24144628.
[66] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," 2017, *arXiv:1707.01083*.
[67] D. Yang, M. I. Solihin, Y. Zhao, B. Cai, C. Chen, A. A. Wijaya, C. K. Ang, and W. H. Lim, "Model compression for real-time object detection using rigorous gradation pruning," *IScience*, vol. 28, no. 1, Jan. 2025, Art. no. 111618, doi: 10.1016/j.isci.2024.111618.
[68] S. Gidaris, A. Bursuc, N. Komodakis, P. P. Pérez, and M. Cord, "Boosting few-shot visual learning with self-supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8058–8067, doi: 10.1109/ICCV.2019.00815.



SITI FAIRUZ MAT RADZI received the master's degree in science (physics) from Universiti Putra Malaysia, in 2022. Her research focuses on image processing, with particular expertise in developing and optimizing algorithms for object detection and classification in medical imaging. She has been actively involved in projects that integrate advanced image processing techniques with practical applications. Her academic interests include the development of efficient image processing pipelines, real-time detection systems, and the enhancement of machine learning architectures for improved accuracy and computational efficiency. Currently, she is exploring YOLO-based architectures to address real-world challenges in infrastructure monitoring and autonomous systems.

MOHD AMIRUDDIN ABD RAHMAN (Member, IEEE) received the B.Sc. degree in electrical engineering from Purdue University and the M.Sc. degree in sensor and instrumentation from UPM. He is currently an Associate Professor of data science with the Department of Physics, Universiti Putra Malaysia (UPM). He is the Deputy Director of the University's Research Management Centre. He is a Pioneer in co-tutelle Ph.D. program, having earned the first such degree from both The University of Sheffield and UPM. His research experience extends globally, including stints as an Invited Researcher with Alcatel-Lucent Bell Labs and a Postdoctoral Fellow with Indian Institute of Technology Kanpur. He has also played a pivotal role in national research initiatives, contributing significantly to the smart data and big data projects at the Ministry of Higher Education. He leads the Computational Intelligence Research Group at UPM and his research focuses on machine learning and artificial intelligence, with over 50 high-impact publications in the field. His areas of interest also include signal processing, pattern recognition, machine learning algorithms, electromagnetics-related computation and modeling, and RF and microwave-based sensor systems.

MUHAMMAD KHAIRUL ADIB MUHAMMAD YUSOF received the B.Sc. degree from Universiti Putra Malaysia (UPM), Seri Kembangan, Malaysia, in 2016, and the Ph.D. degree in space science from Universiti Kebangsaan Malaysia (UKM), Bangi, Malaysia, in 2021. He is currently a Senior Lecturer with the Department of Physics, UPM. His research interests include earthquake precursors, ground and space geomagnetic observations, signal processing, and computational physics.

NURIN SYAZWINA MOHD HANIFF received the B.S. degree (Hons.) in physics and the Ph.D. degree in medical physics from Universiti Putra Malaysia (UPM), in 2019. Her research interests include computed tomography, and machine learning classifiers.

ROMI FADILLAH RAHMAT (Member, IEEE) received the bachelor's and master's degrees from the Universiti Sains Malaysia and the Ph.D. degree from the Universitas Sumatera Utara. He is currently a Researcher and a Lecturer with the Universitas Sumatera Utara. He is the Vice Dean of the Faculty of Computer Science and Information Technology. His research interests include image processing, deep learning, brain-like computing, the Internet of Things, and mobile computing.