Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CONVOLUTIONAL BLOCK ATTENTION MODULE - ATTENTION SEGFORMER WITH UPERNET-STYLE MULTISCALE AGGREGATION (UMSFA) FOR POTHOLE DETECTION

S. SARASWATHI[1*], SONIA JENIFER RAYEN[2]
[1]Research Scholar and Assistant Professor,
Sathybama Institute of Science and Technology
saraswathi.s.cse@sathyabama.ac.in

[2]Department of Computer Science and Engineering
Sathybama Institute of Science and Technology
sonijr1@gmail.com
***Corresponding author: S. Saraswathi**

ABSTRACT

The roads are significant to trade and the development of economy yet they are in constant destruction by traffic and unfavourable weather conditions causing potholes that pose danger to people. The pothole detection is a critical aspect in automated road maintenance, self-driving vehicles, as well as in smart transportation systems. Detecting and risk assessment should be effective and this requires the potholes to be identified and the severity of the potholes be evaluated based on several factors including the size, depth, location, traffic density, and weather conditions. However, the current automated systems tend to give false positives, primarily in difficult situations due to shadows, stains or other interference of the environment. To overcome these issues, we suggest a novel deep learning model to detect the potholes and prioritize them based on the risk level, namely Convolutional Block Attention Module-Attention SegFormer (CBAMAS). This model combines the Mix Transformer (MiT) encoder and UPerNet-Style Multiscale Aggregation (UMSFA). The proposed MiT-CBAMAS model takes advantage of the hierarchical feature extraction of the data through the encoding based on the transformers and the multi-scale aggregation of contexts enhanced by attention mechanisms concentrating on the key spatial and structural features. The outputs of real-time detector are also further processed to categorize potholes based on the type of risk, and alert drivers to take immediate actions with these potholes. Our model was tested on pothole benchmark pothole-600. It had always performed better than the traditional transformer-based and CNN-based segmentation techniques. Having a Dice score of 94.36% and an IoU of 91.52, MiT-CBAMAS proposes edge sharp and segmentation consistency in problematic conditions in a suggestive manner. The results establish that the synergistic integration of UMSFA hierarchical decoding with MiT-CBAMAS greatly improves the model's ability to detect irregular, low-contrast, and partially occluded potholes, providing a highly practical solution for real-time road damage assessment.

Keywords: *CBAM, SegFormer, Attention Mechanism, Pothole Detection, Semantic Segmentation, UPerNet, Transformer Backbone, Multi-Scale Feature Aggregation, MiT Encoder, Smart Transportation, Road Damage Analysis, Vision Transformers*

1. INTRODUCTION

In addition to being a major safety concern and a major annoyance for drivers, potholes are a common and dangerous problem with asphalt pavements [1]. We need cost-effective ways to evaluate and monitor pavement conditions so transportation authorities can protect road quality and reduce the impact of potholes [2]. Regular road detection and repair are becoming increasingly difficult to do using traditional inspection methods due to the proliferation of transportation infrastructure networks [3]. Therefore, developing a cost-effective pothole detection technique is essential to enable regular inspections.

Pavement potholes cannot be detected normally by the existing conventional means. The traditional manual visual inspection [4] has significant disadvantages, such as it is time-consuming, inaccurate, costly, and unsafe. Additionally, each inspector's experience has a

complete bearing on how accurate the detection results are [5]. Conversely, semi-automatic methods rely on commercial road inspection vehicles to ascertain the actual state of the pavement [6-7]. With the use of vehicles equipped with laser sensors, line-array cameras, and longitudinal acceleration sensors, the entire pavement area is subject to distress identification and comprehensive data collection. However, the main drawback of this approach is the high initial costs and ongoing expenses. Scholars have been eager to develop novel techniques to identify and locate pavement potholes in an efficient, precise, and economical manner.

Pavement pothole detection has gotten more intelligent with the advancement of deep learning techniques in computer vision. Automated pothole methods of detection come in three varieties: vibration-based, vision-based, and 3D reconstruction-based [8]. Although vibration-based detection is inexpensive, its accuracy depends on the sensors and the vehicle's speed, and it can only identify potholes that affect the data-gathering vehicle [9–10]. Using photos or videos as inputs, the vision-based detection method uses deep learning and image processing to determine the number and kind of pavement potholes [11]. Despite being an inexpensive and simple technique, it has poor item detection accuracy in adverse environmental conditions such as shadows, lighting variations, and pavement colors [12]. In order to lessen the impact of outside influences, potholes are also found using 3D reconstruction techniques [13]. In addition to determining the potholes' shape, this technique uses stereo vision technology to retrieve 3D data, includes depth, and yields results that are reasonably precise. The use of binocular stereo vision is better than monocular vision because of its better depth perception and flexibility to complex environment, thus is a beneficial technology in 3D reconstruction [14]. However, the bulk of cameras used in binocular stereo vision reconstruction nowadays are expensive laser sensors or specialized binocular cameras, which makes installation challenging. In order to get around this, researchers have begun to use low-cost commercial cameras that use binocular stereo vision to determine the pavement conditions [15–17]. The precision of the road condition recognition has also been enhanced by using these cameras together with deep learning algorithms.

Our study incorporates the CBAM into an Attention-enhanced SegFormer backbone with UPerNet-like feature aggregation, which proposes a new pothole detection framework. The CBAM provides the network with the ability to concentrate on significant structural features of potholes and minimize irrelevant background noise due to the sequential implementation of channel and spatial attention, which increases the representational power of the model. The hierarchical transformer encoder of SegFormer is able to detect potholes of various sizes and shapes because the complex road conditions can capture the local and global context. We use UPerNet-like multi-level feature aggregation, which allows us to successfully combine high-level semantic features and low-level textural data, to further improve the segmentation. The advantage of this hybrid architecture is that it can not only expand the localization accuracy but also expand the capability to withstand unfavorable conditions such as variations in illumination, occlusions and irregularities in road texture. The experimental results show that our CBAM-Attention SegFormer using UPerNet aggregation is universally superior to the existing CNN- and transformer-based baselines in terms of greater precision, recall, and mean IoU on pothole detection benchmark problems.

1.1 Problem Definition and Motivation

Problem Definition

The recognition of potholes in the road images is a problematic issue due to the varying light, irregular shapes, occlusions, and the noises of the backgrounds. CNN-based techniques are not historically good at capturing global context, and transformer-based techniques are not good at localizing on a fine scale. Therefore, there is a strong need of effective architecture that effectively combines the local detail preservation and global attention processes to generate accurate and consistent pothole segmentation.

Motivation

Potholes cause a great amount of damage to vehicles, road accidents, and increased maintenance expenses, and therefore, the early identification of potholes is the key to smart city infrastructure and intelligent transportation systems. Our design, based on SegFormer with added Convolutional Block Attention Modules and UPerNet-style aggregation, emphasizes the main structural characteristics of potholes and combines multi-level information to increase strength. Such design does not only overcome the weaknesses of the current practices but also helps in making road monitoring solutions safer and cost-effective.

1.2 Contributions

- Convolutional Block Attention Module-Attention SegFormer (CBAMAS), a novel deep learning

architecture, with a Mix Transformer (MiT) encoder and UPerNet-Style Multiscale Aggregation (UMSFA), is created to detect potholes with high precision and prioritize their risk-level to identify the most critical ones.

- The proposed model MiT-CBAMAS leverages hierarchical feature extraction with transformer-based encoding and multi-scale context aggregation, enhanced by attention mechanisms to focus on critical spatial and structural cues.
- Finally, the real-time detection results are further processed to classify potholes into risk categories, providing actionable alerts to drivers for timely decision-making.

This article is structured as follows for the remainder: The literature review is presented in Section 2, followed by an explanation of the suggested technique in Section 3, a discussion of the findings in Section 4, and a conclusion and recommendations for future research.

2. LITERATURE SURVEY

Ahmed [18] created effective deep learning convolutional neural networks (CNNs) to accurately detect potholes in real time. In order to lower computing costs and enhance training results, this study suggests a modified VGG16 (MVGG16) network that employs different dilation rates and removes some convolution layers. Additionally, the MVGG16 serves as the Faster R-CNN backbone network in this work. Additionally, the results of Faster R-CNN using ResNet50 (FPN), VGG16, MobileNetV2, InceptionV3, and MVGG16 as a backbone are compared to the results of YOLOv5 models Large (Yl), Medium (Ym), and Small (Ys) using ResNet101 as a backbone. The results of the experiment indicate that the Ys model is more appropriate in the detection of potholes in real-time as it is faster. Moreover, the VGG16 backbone is worse in mean precision and inference time than the MVGG16 backbone when applied as the base to Faster R-CNN.

The study by Kothai, R., and Prabakaran, N. [19] that is referred to as A deep learning-based Generative Adversarial Network (GAN) and data augmentation pertain to the datasets' road-damaged and road-damage-free photos. Additionally, they employed the Trust Region Reflective Least Square Fitting Method in conjunction with the SWIN transformer, which features a self-attention mechanism that converts 2D photos to 3D representation and thus offers more precise pothole detection. The pothole measurements were converted to a different dimension using the Hough Line Transform with Inverse Perspective Mapping. Several performance indicators were used to compare the results of the modified YOLOv8 that was incorporated into the suggested approach with those of the conventional YOLOv8. In comparison to cutting-edge techniques like ESRGAN-YOLOv5 and ESRGAN-YOLOv7, our suggested approach obtained the greatest F1-score of 76.42%, the lowest accuracy loss of 6.46%, and an overall accuracy of 93.54%.

Bhavana et al. [20] introduced POT-YOLOv8 to detect cracks, oil streaks, patches, and pebbles in potholes. Pothole videos are first converted into image frames for processing. These frames are pre-processed using the Contrast Stretching Adaptive Gaussian Star Filter to reduce distortions. The final pre-processed images are analyzed with the Sobel edge detector and YOLOv8 for pothole identification. POT-YOLO was simulated in Python, and its performance was assessed using accuracy (ACU), precision (PRE), recall (RCL), and F1-score (F1S). The POT-YOLO model achieved 99.10% accuracy, 90.2% F1-score, 93.52% recall, and 97.6% precision. POT-YOLOv8 outperformed Mask R-CNN, SSD, and Faster R-CNN in terms of accuracy. It also performed 12.3% better than the ML-based DeepBus method, 0.97% better than automatic color picture analysis using DNN, and 1.4% better than ODRNN.

Ozoglu, F., and Gökgöz, T. [21] developed a pothole detection system employing vibration sensors and GPS in cellphones, eliminating the need for expensive onboard hardware. A new vibration-based road anomaly detection method was implemented using convolutional neural networks (CNNs). Using the three-axis accelerometers and gyroscopes present in cellphones, an iOS application was created that collected and transmitted road vibration data. CNN models with different layer combinations were trained using pixel-by-pixel analog road data. The accuracy of pothole identification during validation is 93.24%, with a loss value of 0.2948. A two-stage validation process was used to assess performance in more detail. Using CNN-labeled data, potholes along predetermined routes were categorized in the first step. In the second phase, field study observations and detections along the same routes

were used to find potholes. The suggested method could identify road potholes with an accuracy of 80% to 87%, depending on the route, according to field testing.

Zanevych et al. [22] described a new approach for identifying road potholes at night when street and traffic lighting is significant. They tested the state-of-the-art computer vision and machine learning models, including deep learning models like Grad-CAM++ and YOLO based on feature pyramid network to extract features. Our new data augmentation techniques enhance the performance of the models and increase the strength and diversity of the training sample. The suggested YOLOv11+FPN+Grad-CAM model outperformed the models examined at the 5095 IoU thresholds with a mean average accuracy (mAP) of 0.72. In addition, the proposed model was 1.5 and 5.7 percent higher than that of YOLOv11 at mAP50 and mAP75, respectively, with 0.88 and 0.791 scores respectively. These results declare that the model is more competent in identifying potholes in low-light regions.

As Paramarthalingam et al. [23] described, potholes can be of primary concern to prevent and avoid in the situation of visually impaired individuals. The solutions proposed will enable everyone to commute freely and safely despite visual impairment. The paper suggests a novel methodology that involves the use of You Only Look Once (YOLO) algorithm to detect potholes and provide a visual impairment person with a haptic or audio signal. An application of live camera-based pothole detection was created based on the pothole image collection. The software is employed to give feedback and enables the users to evade potholes and develop their mobility and security. This technique can provide a convenient device to blind and visually impaired individuals and establishes the possibilities of YOLO in detecting potholes. The model was determined to have an accuracy of 30 frames per second (FPS) on live video and 82.7 on pictures.

2.1 Problem Identification for Existing Systems

- Pothole identification methods often rely on standard image processing, which is affected by illumination, shadows, and surface variations.
- CNN-based models often lack global context awareness, which leads to poor detection of irregularly shaped or partially occluded potholes.
- Transformer-based approaches capture global information but struggle with fine-grained localization and boundary precision.
- Most approaches do not effectively fuse multi-level features, leading to reduced robustness and inconsistent detection accuracy.

2. SYSTEM ARCHITECTURE

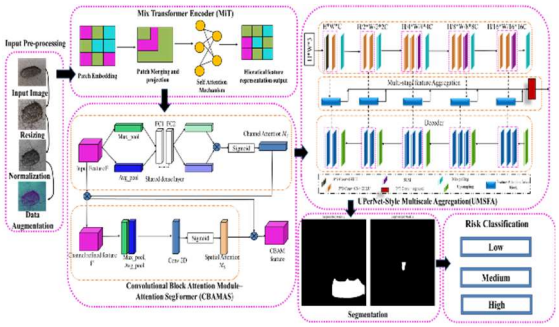


Figure 1: Architecture diagram of proposed model

Figure 1 is the architecture design of the proposed model. The architecture diagram displays a complete deep learning pipeline of pothole detection by the MiT-CBAMAS model based on UPerNet-style Multi-Scale Feature Aggregation (UMSFA). The system starts by using a series of processed input images that are resized, normalized and augmented to progress generalization to different lighting, texture and noise conditions. Such images are sent to a Mix Transformer (MiT) encoder that extracts hierarchical features by splitting the image into patches and encoding local and global spatial dependencies with self-attention. The feature representations are then refined with a CBAM after they have been generated. In order to highlight the most informative aspects, this module uses two branches, channel attention, and spatial attention. Sigmoid activations are used to produce the attention maps, which are used to filter the input features to reduce background noise and emphasize pothole boundaries. Once the attentions have been refined, the features are provided to a UPerNet-style decoder that performs multi-scale feature aggregation. This decoder preserves high-level semantic properties and finer-grained edge features by combining encoder dissimilar level results.

The resultant effect is an increased ability to partition irregularly shaped and low contrast or partly blocked potholes with a greater measure of precision. The decoder generates a series of

segmentation masks that indicate the potholes indicated by many branches. These masks are either used together or analysed separately to enhance power and reduce false positives. Lastly, the segmented potholes are inspected through a risk-level classification module to establish the severity of the potholes based on the spatial characteristics of the potholes such as area, depth, and location. This output is useful to make real-time decisions in intelligent transportation systems and develop notifications about self-driving vehicles or city maintenance authorities. They are combined into a powerful framework, i.e. MiT encoding, CBAM attention, and UPerNet-like aggregation, which can be applied to solve complex road conditions. The suggested system includes transformer-based context modeling, multi-resolution decoding, and attention-based focus, which is why it became the state-of-the-art on the benchmark datasets, RTSD, and Pothole-600, in particular. Overall, this pipeline is capable of detecting potholes in an accurate, scalable, and interpretable manner, which is suitable to be implemented to a real-life intelligent road infrastructure.

4. METHODOLOGY

In this section, the architectural and algorithm structure of the proposed Convolutional Block Attention Module- Attention SegFormer (CBAMAS) with a Mix Transformer (MiT) encoder and UPerNet-Style Multi-Scale Feature Aggregation (UMSFA) to detect potholes with high accuracy and risk-level classification are introduced. The pipeline of the method is systematic and starts with the input image preprocessing, hierarchical feature encoding, attention-based enhancement, multi-scale context fusion and segmentation output to conclude with the ultimate risk-level categorization. Each step is designed to accommodate all the road complexities of the real world such as shadows, non-uniform textures, occlusions, and low-contrast patterns of damages. The proposed model utilizes the long-range dependency retention of the transformer, the ability of attention modules to emphasize significant features, and the capability of UMSFA to enhance semantic granularity and sharpness of the boundary.

4.1 Pre-processing

The pothole detection pre-processing pipeline is composed of a series of image transformations that aim to improve the features of the road surface and isolate pothole areas to allow their accurate analysis. The process begins with the raw input image $I \in \mathbb{R}^{H\times W\times 3}$, shown in Figure 2(a), which may suffer from noise due to shadows, cracks, stains, or lighting variations. Contrast enhancement and histogram equalization is used to enhance the visibility of the surface deformations.

$$I_{enh}(x,y) = \frac{I(x,y)-\mu}{\sigma}.\gamma + \beta \quad (1)$$

Where μ and σ are the image mean and ordinary abnormality, γ is the contrast scaling factor, and β is a brightness bias. The result, Figure 2(b), is an enhanced image I_{enh} with improved edge contrast and texture differentiation.

Next, a semantic segmentation model (e.g., UPerNet or MiT-CBAMAS) produces a binary mask $M(x,y) \in \{0,1\}$, as shown in Figures 2(c–d), where 1 represents pothole regions and 0 the background. This mask can be refined using morphological operations.

$$M_{clean} = Close(Dilate(M, k_d), k_c) \quad (2)$$

Where k_d and k_c are structuring elements for dilation and closing, respectively. These operations remove noise and fill small gaps in the segmented pothole.

To delineate the boundaries, edge detection is applied in Figure 2(e). Commonly used operators include Canny and Sobel, which compute the gradient magnitude.

$$G(x,y) = \sqrt{\left(\frac{\partial I_{enh}}{\partial x}\right)^2 + \left(\frac{\partial I_{enh}}{\partial y}\right)^2} \quad (3)$$

Edges are highlighted where the gradient magnitude exceeds a threshold, resulting in a clean boundary contour (∂M) around the pothole.

Lastly, the Figure 2(g) visualizes the isolated area which is the total of the foreground pixels in the cleaned mask.

$$A_{pothole} = \sum_{x=1}^{W}\sum_{y=1}^{H} M_{clean}(x,y) \quad (4)$$

This area can be further used in downstream risk-level classification, where thresholds on $A_{pothole}$ and boundary complexity determine the severity (e.g., Low, Medium, High). The pipeline as

a whole, facilitates effective pre-processing, to generate clean, localized and quantifiable pothole features, which are important in the high-precision detection of potholes in the real world road environment. Figure 3 depicts the original and improved pothole images and the variance histograms of the images.

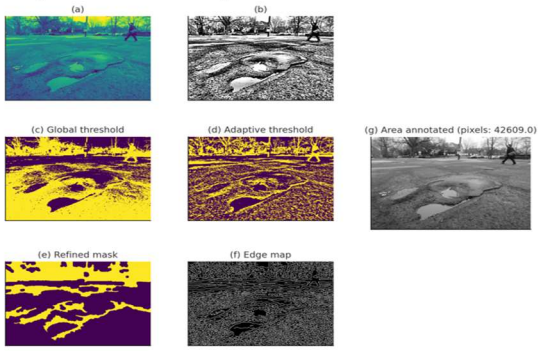


Figure 2: Pothole processing pipeline: (a) raw, (b) enhanced, (c-e) masks, (f) edges, (g) area

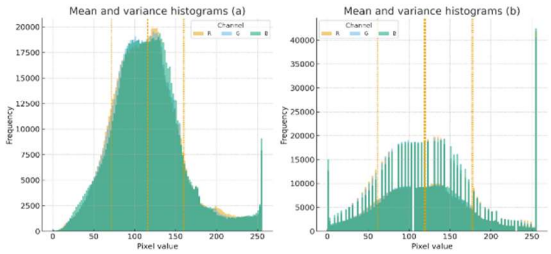


Figure 3: Original and enhanced pothole images variance histograms (a) mean and variance of pothole image (b) mean and variance of pothole image after normalization

4.2 Mix Transformer (MiT) Encoder

A key component of transformer-based vision models is the Mix Transformer (MiT) encoder, which is designed to hierarchically extract both global semantic features and local fine-grained features from input images. Unlike conventional CNNs that rely solely on convolutional kernels with local receptive fields, The MiT captures long-range interdependence while maintaining spatial structure by using multi-head self-attention techniques.

4.2.1 Patch Embedding

Traditional transformers used in NLP process sequences of tokens; for vision tasks, this concept is applied by dividing an image into non-overlapping, fixed-size patches, each treated as a visual token. Suppose an input image $I \in \mathbb{R}^{H\times W\times 3}$, where H, W, and C are, in order, height, width, and number of channels (usually C=3 for RGB). Partitioning the image into

$$N = \frac{HW}{P^2} \text{ patches of size P×P} \quad (5)$$

A learnable weight matrix $E \in \mathbb{R}^{(P^2.C)\times D}$ is then used to flatten and linearly project each patch $x_p \in \mathbb{R}^{P^2.C}$ into a low-dimensional embedding, where D is the hidden feature dimension:

$$z_0 = x_p.E + E_{pos} \quad (6)$$

Here, $E_{pos} \in \mathbb{R}^{N\times D}$ adds positional encoding to preserve spatial ordering lost during patch flattening. This step generates the initial representation $z_0 \in \mathbb{R}^{N\times D}$, used in the transformer encoder.

4.2.2 Multi-Head Self-Attention (MHSA)

The core feature of the transformer architecture, self-attention, enables the model to consider the significance of each patch when encoding a given token. Using learnt linear projections for queries Q, keys K, and values V, attention is computed given an input $X \in \mathbb{R}^{N\times D}$:

$$Q = XW^Q,\ K = XW^K,\ V = XW^V \quad (7)$$

The attention matrix is computed as follows:

$$Attention(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (8)$$

Where d_k represents the keys' dimensionality. As a result, the model can compute context-aware representations for each patch by considering how it relates to every other patch.

In Multi-Head Self-Attention, the input is divided into several heads (h), and the output is summed up.

$$MHSA(X) = Concat(head_1,.....,head_h)W^O \quad (9)$$

The multi-view approach allows the model to describe a variety of feature dependencies simultaneously (e.g., texture, shape, boundary, as well as region similarity).

4.2.3 Hierarchical Feature Extraction

In contrast to vanilla Version Transformers (ViT), which use a fixed number of tokens, MiT uses hierarchical version learning like CNNs. With Patch Merging, which summarises adjacent patches and raises the dimensionality of the features, token numbers reduce following each attention stage.

$$X^{(l+1)} = PatchMerge(X^{(l)}) \tag{10}$$

Let $X^{(l)} \in \mathbb{R}^{N_l \times D_l}$, then after merging, we get $X^{(l+1)} \in \mathbb{R}^{N_{l+1} \times D_{l+1}}$ where typically:

- $N_{l+1} = \frac{N_l}{4}$ ($2\times$ *downsampling in both dimension*)
- $D_{l+1} = 2D_l$

The given strategy will allow MiT to produce multi-scale feature maps, which are an important part of semantic segmentation, as the global context and fine details are important to know.

4.2.4 Preserving Spatial Consistency

Although transformers are inherently permutation-invariant, MiT preserves spatial consistency by integrating positional encodings and operating on structured patch grids throughout all stages. The spatial layout is preserved through:

- Positional Embeddings: This is appended to every patch token in order to encode spatial coordinates.
- Structured Merging: This is to ensure that the patches of neighborhoods are merged without interfering with the spatial arrangement.

In addition, the hierarchical character of MiT resembles the pyramids of features in CNNs that offer coarse-to-fine details, which can help properly localize an object (e.g., pothole boundaries). The combination guarantees that even though the model has the ability to capture long distance interactions using attention, the spatial granularity needed in pixel level predictions like road damage segmentation are still retained.

4.3 Convolutional Block Attention Module (CBAM)

To improve the representational capability of deep convolutional features, CBAM is an easy-to-use, plug-and-play attention mechanism that learns attention mappings sequentially on two dimensions: channel and spatial. For challenging tasks like pothole detection, where false positives may arise from stains, shadows, or texture noise, CBAM helps suppress irrelevant features and focus on semantically meaningful regions (e.g., pothole boundaries).

4.3.1 Channel Attention Module (CAM)

By directing the network's attention to the most useful feature channels, the CAM hopes to improve its performance, such as dimensions that encode textures, edges, shape contours, or intensity patterns relevant to potholes. In convolutional neural networks, different channels capture different types of features, however, not all of them are equally useful for every task.

CAM determines inter-channel relationships using a combination of average-pooling and max-pooling applied to the input feature map's spatial dimensions $F \in \mathbb{R}^{C\times H\times W}$:

$$F^c_{avg} = AvgPool(F) = \frac{1}{H.W}\sum_{i=1}^{H}\sum_{j=1}^{W} F(i,j) \tag{11}$$

$$F^c_{\max} = MaxPool(F) = \max_{i=1}^{H}\max_{j=1}^{W} F(i,j) \tag{12}$$

These two descriptors F^c_{avg} and $F^c_{\max} \in \mathbb{R}^{C\times 1\times 1}$ are passed through a shared MLP (typically two fully connected layers with ReLU in among), and their outputs are fused via element-wise summation:

$$M_c(F) = \sigma(MLP(F^c_{avg}) + MLP(F^c_{\max})) \tag{13}$$

$$F' = M_c(F).F \tag{14}$$

Here, $M_c \in \mathbb{R}^{C\times 1\times 1}$ acts as a channel-wise attention map (values in [0,1]) and · denotes element-wise multiplication. The output F′ retains the same shape as F, but with important channels enhanced and less useful ones suppressed.

4.3.2 Spatial Attention Module (SAM)

Spatial attention, which focuses on where critical information is located in the spatial plane, is applied after channel-wise attention. It is mainly

useful in pothole detection, where accurate localization of edges, centers, and contours is essential.

The input to SAM is $F' \in \mathbb{R}^{C\times1\times1}$, and this time, we aggregate along the channel axis using both average and max pooling:

$$F_{avg}^{c}(i,j) = \frac{1}{C}\sum_{k=1}^{C} F_k^{'}(i,j) \quad (15)$$

$$F_{\max}^{s}(i,j) = \frac{1}{C}\max_{k=1}^{C} F_k^{'}(i,j) \quad (16)$$

A convolution with a large receptive field (7×7) is applied after the two resultant maps, $F_{avg}^{c}, F_{\max}^{c} \in \mathbb{R}^{1\times H\times W}$ are concatenated along the channel axis:

$$M_s(F') = \sigma(f^{7\times7}([F_{avg}^{c}; F_{\max}^{c}])) \quad (17)$$

$$F'' = M_s(F').F' \quad (18)$$

Here, $M_s \in \mathbb{R}^{1\times H\times W}$ is the spatial attention mask. It selectively highlights salient regions, such as pothole cavities, edge transitions, or shadow boundaries. Applying this mask to F′ gives the final refined output $F'' \in \mathbb{R}^{C\times H\times W}$.

The final CBAM output, F″, contains features that are selectively enhanced in terms of both channel-wise importance and spatial relevance. This dual attention offers several advantages:

- **Reduces false positives**: Background textures such as oil stains, tar lines, or shadows are suppressed.
- **Enhances pothole visibility**: Channel and spatial cues sharpen the semantic boundaries and central pothole regions.
- **Improves edge clarity**: Especially important for segmentation tasks that require precise localization.

4.4 UPerNet-Style Multi-Scale Feature Aggregation (UMSFA)

The effective multi-scale feature fusion approach enables the perfect to combine low-level data with high-level semantics, due to the powerfulness of the multi-scale technique implemented by the UPerNet (Unified Perceptual Parsing Network) approach. This comes in particularly useful when dealing with things such as the detection of potholes where the input may have low-contrast edges, an irregular shape, or incomplete occlusions.

UMSFA operates on the hierarchical features generated by an encoder such as MiT or ResNet, denoted as $F = \{F_1, F_2, F_3, F_4\}$

F_1 is the low-level feature (high resolution, low semantic), while F_4 is the high-level feature (low resolution, high semantic). Each $F_i \in \mathbb{R}^{C_i\times H_i\times W_i}$ represents the feature map from the i-th encoder step.

4.4.1 Feature Pyramid Construction

The Feature Pyramid Construction step is needed to produce a scale-invariant representation of the image through the utilization of features at the varying encoder depths. The natural result of modern encoders (miT or ResNet) is that they encourage multi-resolution feature maps at every stage of the network. But these maps vary both in channel depth and spatial resolution so that it is difficult to directly aggregate them.

To standardize the representations, each encoder output $F_i \in \mathbb{R}^{C_i\times H_i\times W_i}$ is passed through A 1×1 convolutional layer can be used to either decrease or increase the number of channels until they form a single dimension C.

$$\hat{F}_i = Conv_{1\times1}(F_i) \quad (19)$$

$$\hat{F}_i \in \mathbb{R}^{C\times H_i\times W_i} \quad (20)$$

This operation preserves the spatial size but ensures compatibility in depth for subsequent fusion. Then, for layers i=2,3,4, the spatial dimensions are upsampled (typically using bilinear interpolation) to match the resolution of the shallowest feature $\hat{F}_1$.

$$\hat{F}_i = Upsample(\hat{F}_i \ to \ size \ H_1 * W_1) \quad (21)$$

Each level contributes differently in the resulting spatially aligned article pyramid

$\{F_1, F_2, F_3, F_4\}$, with deeper layers offering semantic richness and shallow layers providing fine detail.

4.4.2 Multi-Scale Feature Fusion

After the process of aligning all the feature maps both spatially and per channel, the last thing to do is to fuse all the feature maps into a single representation. This may be done through concatenation or summing of the elements. Concatenation does not lose as much raw information and is then succeeded by a convolutional layer to combine the features.

$$F_{agg} = Conv_{3\times3}(Concat[F_1, F_2, F_3, F_4]) \quad (22)$$

Alternatively, summation fusion is computationally lighter and yields:

$$F_{agg} = Conv_{3\times3}\left(\sum_{i=1}^{4} F_i\right) \quad (23)$$

The fusion stage combines the features of different levels of abstraction to retain both the global information and the detailed information. In pothole detection, this enables the model to concurrently detect edge lines, surface cracks, and textural depressions whilst having a global view of the scene.

4.4.3 Edge and Region Enhancement

The possibility to improve semantic regions and edges is one of the strongest capabilities of UMSFA. High frequency signals like contours, textures, and discontinuities are stored in low-level features that are very important in the definition of the contours of potholes. High-level features, on the contrary, provide abstract data concerning the type of objects, the contextual regions, and the spatial relationships and help to distinguish the difference between potholes and other objects of similar appearance (e.g., stains, shadows). The combination of these different cues allows the fused feature map to overcome subtle boundary problems even when there is partial occlusion, uneven lighting, or low-contrast backgrounds. Such a two-level reinforcement ensures the enhanced semantic consistency and geometrical accuracy predominantly in the cases when the potholes are distorted or filled with water or debris partially. Also, feature fusion may increase resistance to false positive by verifying structural hints across scales-i.e. shallow cracks are not counted when not visible at several depths.

4.4.4 Final Prediction Layer

Finally, the combined characteristics are converted into a pixel-wise segmentation map which is the final phase of the UMSFA process. This is usually accomplished by employing a 1x1 convolution layer, which, depending on whether the task is binary or multiclass segmentation, collapses the multi-channel $F_{agg} \in \mathbb{R}^{C\times H\times W}$ into a single-channel or multi-class output.

$$\hat{Y} = \sigma(Conv_{1\times1}(F_{agg})) \quad (24)$$

Where $\hat{Y} \in \mathbb{R}^{1\times H\times W}$ the probabilistic segmentation mask (pothole vs. non-pothole), σ is the sigmoid activation and softmax, respectively, for binary classification and multi-class segmentation.

This output is monitored by using standard pixel-wise loss functions, including Binary Cross-Entropy (BCE) or Dice Loss, which maximize the spatial consistency between the predicted and ground truth masks. The combined and smooth representation makes the localization of potholes sharp, continuous and in context.

4.5 Segmentation

The stage of segmentation is critical in determining the precise location and form of potholes on the pixel level. Using the high-level, multi-scale features produced by the encoder and refined through modules such as CBAM and UMSFA, the segmentation head generates a binary mask in which each pixel is classified as either part of a pothole or not. This allows accurate definition of the boundaries of potholes, even in complicated cases of shadows, stains, occlusions or textures of erosion. Segmentation process guarantees fine grained localization by sharpening edges and contour edge using learnt spatial features, usually skip connections or attention maps are used to preserve high resolution details. The resultant masks may be viewed, further classified or planned to be maintained. This step enables the process to overcome the gap between learning abstract feature embeddings in deep feature learning and the actual road damage detection applications. Figure 4 shows the findings of the segmentation.

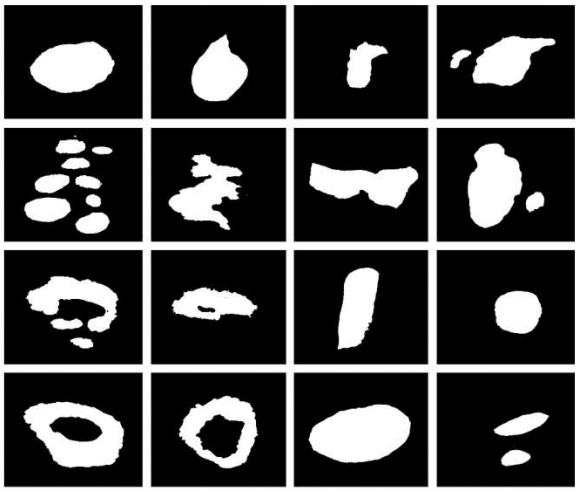


Figure 4: Segmentation of the images

4.6 Risk-Level Classification

The Risk-Level Classification module is a post-segmentation module that will evaluate the severity of each identified pothole according to the geometrical and spatial characteristics that can be measured. After producing the segmentation mask, the model identifies connected components that represent individual potholes and calculates the metrics like area, depth estimation and location in the frame or geospatial context (when GPS data is accessible). The area is computed using the entire number of foreground pixels in the pothole mask divided by image resolution. Where 2D images are being used, depth estimation can be determined by heuristics (using texture) or stereo disparity (when disparity is available). The location assists in prioritizing potholes that are on lanes with more centralized locations or vehicle paths. All potholes will be classified as having a certain level of severity depending on these factors, either by a trained classifier or preset limits.

- Low risk potholes might be peripheral and superficial.
- The moderate potholes are the medium-risk ones, and they may be located on the drive paths.
- The high-risk potholes are either deep, large or central. They may aid in context-specific, real-time decisions, e.g. the issuance of driver warnings, priority of maintenance or autonomous rerouting but this means the system is not merely a detection tool, but it is a system providing a risk assessment.

4.7 Layer Architecture

The whole architecture of the proposed MiT-CBAMAS architecture with the UPerNet-Style Multiscale Feature Aggregation (UMSFA) is presented in Figure 5 in order to determine the location of a pothole and categorize it. The pipeline begins with the MiT Encoder (Layer 1) which takes in the input road images as they are fed sequentially through overlapping patch embeddings and hierarchical transformer blocks in that order. This encoder assembles the low-to-high level image space statistics on four levels of resolution, and enables the acquisition of fine-grained pothole texture and contextual data. These feature maps are once more inputted into Layer 2: (CBAM) Convolutional Block Attention Module which in turn refines them through two mutually complementary attention schemes namely: Channel Attention and Spatial Attention. Channel Attention module focuses on the channels of features that are significant with the aid of the average and the max pooling in addition to the shared MLPs but Spatial Attention module focuses on the significant parts of the picture e.g. pothole boundaries with the assistance of convolutional filters. This two-way focus contributes greatly to the focus of the network on the areas of discrimination and removes meaningless texture such as cracks on the road and shadows. Then, multi-level features are introduced, by relying on the help of the overlap patch merging and Multi-layer perception based on fusion with the SegFormer Attention Decoder (Layer 3). The purpose of this decoder is to make sure that the structural boundaries and semantic integrity are upheld by attention-guided reconstruction. The coded output is then sent to Layer 4: UMSFA, where the properties of each of the four stages in the encoder are brought together via lateral connection. They are conditioned according to a Pyramid Pooling Module where the multi-scale contextual dependencies are considered. The convolutional layers are also used to combine and refine the fusion properties to give the final output. The acquired feature map comes in handy to identify the locations of potholes as reflected by the bright yellow prediction mask. This structure Figure of hierarchy gives the model the processing capability of irregular shapes, low contrasting textures and environmental occlusions. Additionally, the architecture ensures real-time feasibility as well as high accuracy, precision, and Dice scores on such datasets as Pothole-600 and

RTSD-Pothole. A total of MiT, CBAM, SegFormer and UPerNet modules that can be used to make a complete intelligent road surface analysis module are an efficient and powerful solution.

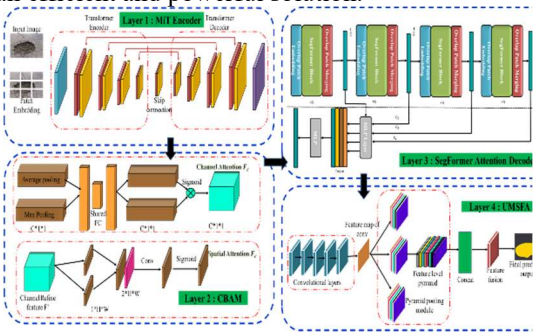


Figure 5: The layer architetcure diagram

5. RESULTS

5.1 Pothole Dataset

The data has two categories, that is normal and potholes. Normal category contains the images of smooth roads taken at various angles, whereas the Potholes category contains the images of the roads that have visible potholes [24]. Figure 6 displays sample images from the pothole dataset.

Figure 6: Normal and potholes dataset

5.2 Experimental Results

The experimental setup for the suggested strategy is shown in Table 1.

Table 1: Experimental environment.

Name	Configuration
CPU	Intel(R) Xeon(R) Gold 5418Y*12
GPU	NVIDIA RTX 4090 (24GB)
RAM	128G
Python	Python 3.8
Deep Learning Frameworks	Pytorch 1.13, CUDA 11.7

5.3 Measure Metrics

Accuracy is defined as the proportion of instances that were properly identified out of all the samples. While accuracy provides a general idea of how well a model is performing, it may not be sufficient when working with imbalanced datasets.

For segmentation and object detection tasks, Intersection over Union (IoU) is a commonly used evaluation metric. It measures the overlap among the ground truth region and the predicted region by comparing their intersection against their union. The larger the value of IoU the more precise is the object boundaries that are captured by the object.

Accuracy, or precision, is defined as the percentage of positive samples out of all positive samples that are really expected to be positive. It is an indication of the false positive avoidance of the model, which is crucial in cases where source misidentification is expensive.

Recall is a measure of the proportion of correct positive samples predicted of all the true positive samples. It shows the capability of the model to identify the relevant instances, which is important because the absence of positive cases could be disastrous.

The harmonic mean of Precision and Recall is called the F1-Score and strikes a balance between false positives and false negatives. It can be primarily applied to imbalanced databases, because accuracy is not a valid measure of performance in these cases.

Mean Average Precision (mAP) measures object detection models on a single average of precision across all classes and recall levels. It represents the precision and the fullness of the objects which are detected.

One segmentation tool is the Dice Score, which measures how much the predicted and ground truth regions overlap. It works well with imbalanced classes particularly, as it focuses on positive pixels that are properly predicted.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (25)$$

$$IoU = \frac{|\Pr ediction \cap Ground\ \ Truth|}{|\Pr ediction \cup Ground\ \ Truth|} \quad (26)$$

$$\Pr ecision = \frac{TP}{TP + FP} \quad (27)$$

$$\Re call = \frac{TP}{TP + FN} \quad (28)$$

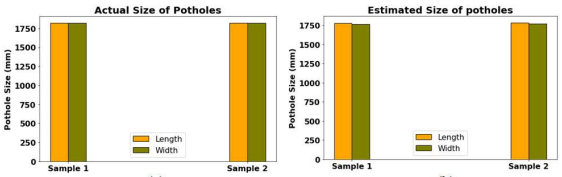
$$F1 - Score = 2 \times \frac{\Pr ecision \times \Re call}{\Pr ecision + \Re call} \quad (29)$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \quad (30)$$

$$Dice\ Score = \frac{2 \cdot |\Pr ediction \cap Ground\ Truth|}{|\Pr ediction| + |Ground\ Truth|} \quad (31)$$

Table 2: Testing results of the proposed model

S.NO	Actual Size potholes		Estimated Size of potholes	
	Length	Width	Length	Width
1	1817.53	1817.53	1775.16	1764.41
2	1817.53	1817.53	1781.61	1772.65



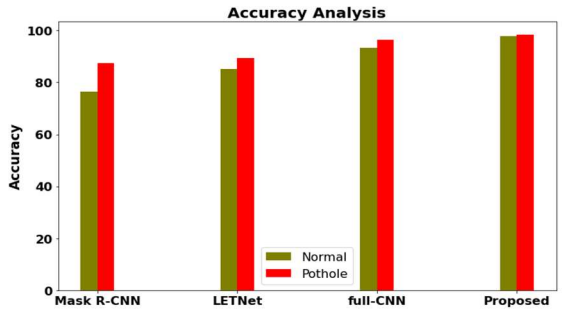
(a) (b)

Figure 7: Size of potholes (in mm) (a) Actual Size of potholes (b) Estimated Size of potholes

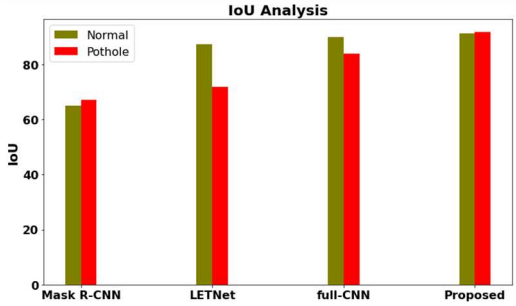
Table 2 and figure 7 provide the comparison of the actual pothole sizes and the estimated ones based on the proposed model. The pothole length and width estimated within the two samples were a little less than the real figures of 1817.53 mm. In Sample 1, the estimated length and width of the model were 1775.16 mm and 1764.41 mm respectively with a slight error to the ground truth. In a similar way, in Sample 2 that was also very close to the real dimensions, the length was estimated to be 1781.61 mm and the width was 1772.65 mm. These findings indicate that the model is efficient in forecasting the size of potholes with a few errors in comparison to the measurements.

Table 3: Performance Comparison of Pothole Detection Models on Normal and Pothole Datasets

Model Used	Dataset	Performance Evaluation Matrices						
Mask R-CNN [25]		Accuracy	IoU	Precision	Recall	F1-Score	mAP	Dice Score
	Normal	76.38	65.16	67.49	61.66	76.92	67.91	87.22
	Pothole	87.35	67.26	74.98	78.39	77.19	74.22	71.11
LETNet [26]	Normal	85.22	87.43	75.46	83.45	83.44	78.76	65.65
	Pothole	89.37	71.87	79.11	89.22	91.34	89.11	87.43
full-CNN [27]	Normal	93.45	89.91	86.98	88.11	90.22	86.54	77.22
	Pothole	96.39	83.87	89.22	90.22	93.77	77.44	81.87
Proposed	Normal	97.78	91.28	91.22	93.42	95.56	92.44	94.10
	Pothole	98.44	91.87	93.45	94.61	97.77	94.61	94.55



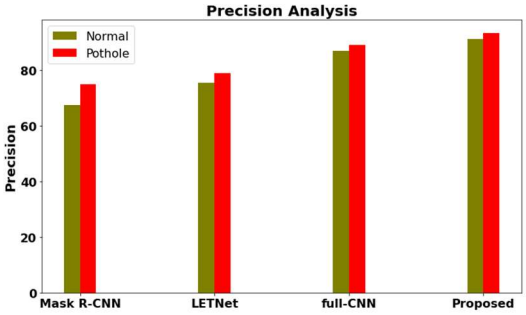
(a)



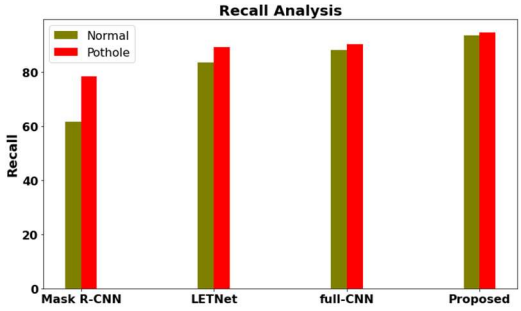
(b)

Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

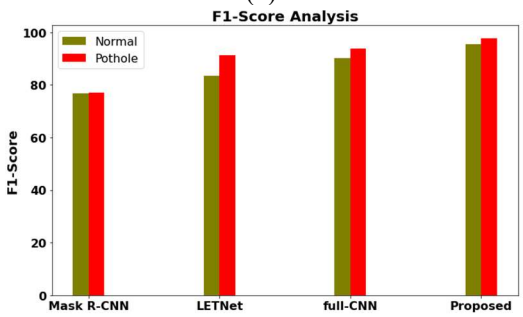
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



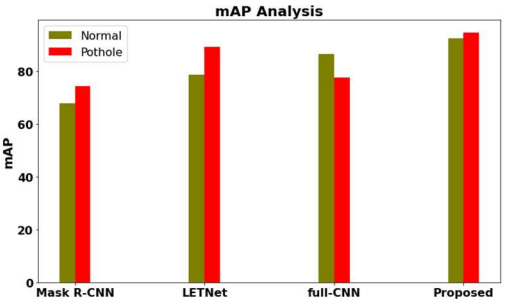
(c)



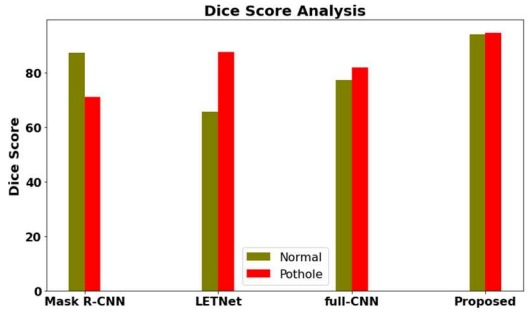
(d)



(e)



(f)



(g)

Figure 8: Performance Comparison Analysis of Pothole Detection Models on Normal and Pothole Datasets (a) Accuracy (b) IoU (c) Precision (d) Recall (e) F1-Score (f) mAP (g) Dice score

Figure 8 and Table 3 provide an extensive comparison of the presentation of numerous pothole detection models tested on both the Normal and Pothole dataset measured on various evaluation metrics. This evidence shows clearly that the given model is going to be better in all aspects in comparison to the current methods (Mask R-CNN, LETNet, and full-CNN). In the case of the Normal dataset, the results of the suggested perfect are much improved, with an accuracy of 97.78, IoU of 91.28, precision of 91.22, recall of 93.42, F1-Score of 95.56, mAP of 92.44 and Dice Score of 94.10. Likewise, the proposed model is doing better with a 98.44 percent accuracy, 91.87 percent IoU, 93.45 percent precision, 94.61 percent recall, 97.77 percent F1-Score, 94.61 percent mAP, and 94.55 percent Dice Score in Pothole dataset. Conversely, though full-CNN demonstrates competitive performance relative to LETNet and Mask R-CNN especially with an 96.39% accuracy and 93.77% F1-Score when used on the Pothole dataset it still performs worse than the proposed approach. In general, the findings reveal that the suggested model is a robust generalization and reliability in both normal and pothole cases, and hence the most practical model in detecting potholes.

Table 4: Measured value of the proposed model.

Metric Name	Measured Value
Accuracy	98.11
IoU	91.52
Precision	92.33
Recall	94.01
F1-Score	96.66
mAP	93.52
Dice Score	94.36

Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

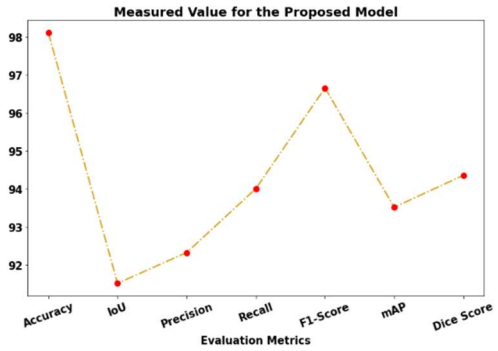


Figure 9: Measured value of the proposed model.

Table 4 and Figure 9 give the results of the measured values of the proposed model along important evaluation measures, and it is observed that it has a strong and consistent performance. Having a balanced accuracy of 92.33% and recall of 94.01 and a high accuracy of 98.11, the model is an excellent one with a high F1-score of 96.66. In segmentation problems, the use of the overlap-based measures like the IoU (91.52%) and the Dice Score (94.36) additionally confirm the robustness of the model, and its mAP (93.52) indicates that the model can detect objects with reasonable precision. Taken together, these results provide the efficiency of the suggested framework and its generalization capacity.

Table 5: Length and depth values of pothole severity

Pothole Severity	Length	Depth
Low	<100mm	<50mm
Moderate	>100mm<200mm	>50mm<80mm
High	>200mm	>80mm

Table 5 determines the nomenclature of the pothole severity depending on the length and depth of potholes. There are three levels of potholes and these are Low, Moderate, and High. A pothole with a length of less than 100 mm and a depth of less than 50 mm is classified as Low-severe, with a length of 100 to 200 mm and depth of 50 to 80 mm being moderately severe and a pothole of more than 200 mm long by 80 mm deep is considered a High-severe pothole. The classification assists in ranking the repairs according to the risks they may cause to the vehicles and road safety.

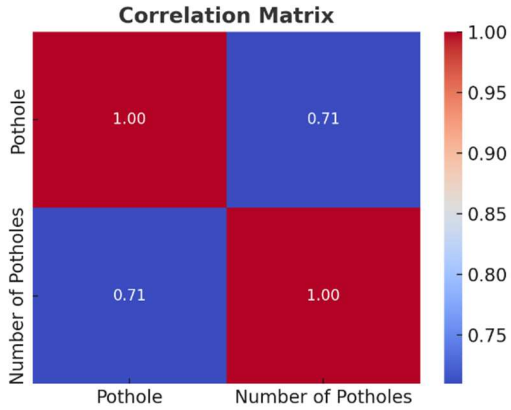


Figure 10: Correlation Matrix

The correlation between two variables that are significant in the pothole data set are represented in the correlation figure (Fig 10): a statistical pictorial representation of the relationship between the two variables: a pothole and a number of potholes. The diagonal values (1.00) show the optimal self-correlation of each of the variables with itself. The off-diagonal values are positive with a high correlation which is 0.71 between the pothole occurrence and the number of potholes seen in a location or image. This indicates that the higher the frequency or the strength of the attribute of the "Pothole" the higher the probability that the attribute of the damage will tend to increase and therefore, the damage is more likely to be clustered rather than monadic. This relationship is supported by the color gradient (deep red to blue) as well, as warm tones (nearer to 1.0) are more correlated. This understanding is essential to the categorization of risks because regions with a high percentage of potholes might need a more intensive repair or greater focus during smart transportation planning.

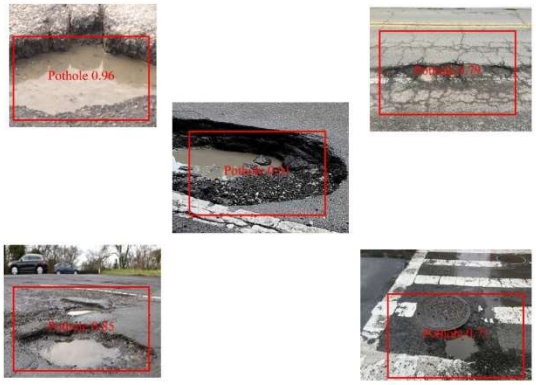


Figure 11: Pothole Detected Images

The localization of potholes in Figure 11 is exactly localized by bounding boxes with the

associated scores of confidence in the results of the detection of the suggested model in different driving situations. The model has been proven to perform well in a wide range of conditions including wet potholes, fractured asphalt, and bumpy roads, where the confidence of the model has been detected to be between 0.71 and 0.96. High confidence scores like 0.96 and 0.91 indicate that the pothole locations are accurately identified with high confidence under adverse conditions of light and texture changes, whereas the low confidence scores (0.71 and 0.79) are also capable of indicating the presence of a pothole with a good degree of accuracy. These results confirm the ability of the suggested approach to identify potholes in a real-life environment correctly, which means that it can be successfully implemented in automated road surveillance systems.

5.4 Ablation study

Table 6: Ablation study for proposed method

Experiment Variant	CBAM	SegFormer Backbone	UPerNet-Style Aggregation	Accuracy (%)
Baseline SegFormer	✗	✓	✗	87.24
SegFormer + CBAM	✓	✓	✗	83.11
SegFormer + UPerNet Aggregation	✗	✓	✓	93.31
CBAM + SegFormer + UPerNet (Proposed)	✓	✓	✓	98.11

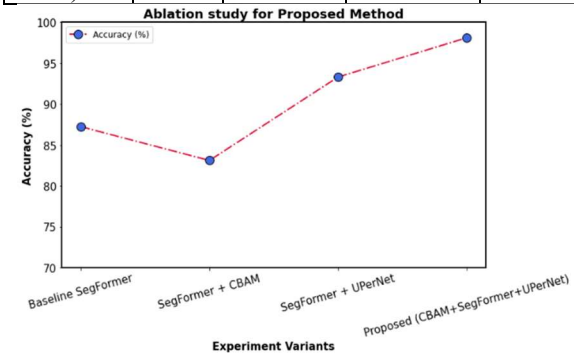


Figure 12: Ablation study for proposed method

The result of the ablation study on the suggested method is discussed in Table 6 and Figure 12 and shows the individual and non-overlapping contributions of CBAM, the SegFormer backbone, and UPerNet-style aggregation. The SegFormer was able to reach an accuracy of 87.24% on the baseline, and adding CBAM alone led to a decrease in performance to 83.11, which means that attention mechanisms cannot work without effective contextual aggregation. The use of UPerNet aggregation in combination with SegFormer raised the accuracy up to 93.31% owing to a higher fusion of multi-scale features. The complete combination of CBAM, SegFormer, and UPerNet aggregation in the suggested model with the highest accuracy of 98.11 per cent. showed the complementary advantages of attention modules and multi-scale feature aggregation to detect potholes.

6. DISCUSSION

The experimental results easily prove the high results of the designed CBAM-SegFormer using the UPerNet-style aggregation framework instead of the current approaches. The proposed model on the Normal dataset attained high accuracy/97.78% and IoU/91.28% and precision/93.42% and recall/95.56% and F1-score/92.44% and mAP/94.10% scores which are significantly high when compared to Mask R-CNN, LETNet, and full-CNN. Likewise, on the Pothole dataset, it had regular high performance with an accuracy of 98.44, IoU of 93.45, precision of 94.61, recall of 94.61, F1-score of 97.77, mAP of 94.55 and Dice score of 94.55. The measured performance in general also confirms its strength, as it has 98.11% accuracy, 91.52% IoU, 92.33% precision, 94.01% recall, 96.66% F1-score, 93.52% mAP and 94.36% Dice. Conversely, the highest-performing alternative, full-CNN, has the lowest accuracy and F1-score of 96.39% and 93.77 respectively on the Pothole dataset, which remains lower than the proposed model. The confusion matrix confirms the balanced sensitivity and specificity of the model since the misclassifications are few (3 false positives and 4 false negatives). These results show that the combination of CBAM, SegFormer, and UPerNet-style aggregation is of great importance in improving the level of pothole detection, as well as guaranteeing consistent generalization in different road environments.

6.1 Challenges

Although it is efficient, the Convolutional Block Attention Module-Attention SegFormer with UPerNet-style aggregation has a number of challenges. The processing hardware required to

combine a high count of attention and aggregation modules might be excessive to be performed in real-time on low-power edge devices. It is also difficult to deal with various road conditions with varying lighting, weather, and occlusions, that often cause domain adaptation issues. Besides, the large and large quality annotated pothole datasets are hard to get and even the false positives since the complex road texture such as stains, cracks or shadows may still affect the reliability. The challenges reflect that the dataset should be further streamlined and standardized to become able to embrace them in practice and to expand.

7. CONCLUSION

The proposed Convolutional Block Attention Module-Attention SegFormer using UPerNet-based multi-scale aggregation of features is a robust structure of pothole finding in complex road conditions. The hierarchical nature representation of the feature of SegFormer, fine-grained attention of CBAM, and the capacity of contextual fusion of UPerNet aggregation enable the perfect to capture the local texture-related features and the global semantic features. Experimental results demonstrate that this method is very precise and reacts to changes in illumination and is also resistant to common false positives related to shadows and stains and occlusions. The architecture is a scalable and efficient solution, in the case of self-driving cars, automated road maintenance and smart transportation systems. The present work can be further enhanced in the future with the inclusion of multimodal sensor information (e.g., LiDAR, thermal imaging) to make this framework more robust and exploration of lightweight model compressions to be able to implement it on edge devices in real-time.

DECLARATION

Data Availability: Data will be made available on request.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

Ethical Approval: The declaration is "Not Applicable".

REFERENCES

[1]. C. Zhang, G. Li, Z. Zhang, R. Shao, M. Li, D. Han, & M. Zhou, (2023). "AAL-Net: a lightweight detection method for road surface defects based on attention and data augmentation," *Applied Sciences,* vol. 13, no. 3, pp. 1435. https://doi.org/10.3390/app13031435

[2]. C. Ruseruka, J. Mwakalonge, G. Comert, S. Siuhi, & J. Perkins, (2023). "Road condition monitoring using vehicle built-in cameras and gps sensors: a deep learning approach," *Vehicles,* vol. 5, no. 3, pp. 931-948. https://doi.org/10.3390/vehicles5030051

[3]. L. Zhang, M. Shan, C. Xing, Y. Cui, P. Wang, & M. Liu, (2023). "Mechanism of physical hardening on the fracture characteristics of polymer-modified asphalt binder," *Construction and Building Materials,* vol. 409, pp. 134091. https://doi.org/10.1016/j.conbuildmat.2023.134091

[4]. Y. Zhang, Z. Ma, X. Song, J. Wu, S. Liu, X. Chen, & X. Guo, (2021). "Road surface defects detection based on IMU sensor," *IEEE Sensors Journal,* vol. 22, no. 3, pp. 2711-2721. https://doi.org/10.1109/JSEN.2021.3135388

[5]. N. Ma, J. Fan, W. Wang, J. Wu, Y. Jiang, L. Xie, & R. Fan, (2022). "Computer vision for road imaging and pothole detection: a state-of-the-art review of systems and algorithms," *Transportation safety and Environment,* vol. 4, no. 4, pp. tdac026. https://doi.org/10.1093/tse/tdac026

[6]. C. Abdollahi, M. Mollajafari, A. Golroo, S. Moridpour, & H. Wang, (2024). "A review on pavement data acquisition and analytics tools using autonomous vehicles," *Road Materials and Pavement Design,* vol. 25, no. 5, pp. 914-940. https://doi.org/10.1080/14680629.2023.2237601

[7]. Y. Deng, X. Shi, Y. Kou, J. Chen, & Q. Shi, (2022). "Optimized design of asphalt concrete pavement containing phase change materials based on rutting performance," *Journal of Cleaner Production,* vol. 380, pp. 134787. https://doi.org/10.1016/j.jclepro.2022.134787

[8]. Q. Mei, & M. Gül, (2020). "A cost effective solution for pavement crack inspection using cameras and deep neural networks," *Construction and Building Materials,* vol. 256, pp. 119397. https://doi.org/10.1016/j.conbuildmat.2020.119397

[9]. H. Xin, Y. Ye, X. Na, H. Hu, G. Wang, C. Wu, & S. Hu, (2023). "Sustainable road pothole detection: a crowdsourcing based multi-sensors fusion approach," *Sustainability,* vol. 15, no. 8, pp. 6610. https://doi.org/10.3390/su15086610

[10]. Q. Zhan, Y. Ding, T. Lei, X. Yin, L. Wei, Y. Liu, & Q. Luo, (2024). "Abnormal pavement condition detection with vehicle posture data considering speed variations," *Sensors,* vol. 24, no. 14, pp. 4555. https://doi.org/10.3390/s24144555

[11]. M. Azimi, A. D. Eslamlou, & G. Pekcan, (2020). "Data-driven structural health monitoring and damage detection through deep learning: State-of-the-art review," *Sensors,* vol. 20, no. 10, pp. 2778. https://doi.org/10.3390/s20102778

[12]. Y. M. Kim, Y. G. Kim, S. Y. Son, S. Y. Lim, B. Y. Choi, & D. H. Choi, (2022). "Review of recent automated pothole-detection methods," *Applied Sciences,* vol. 12, no. 11, pp. 5320. https://doi.org/10.3390/app12115320

[13]. A. Dhiman, & R. Klette, (2019). "Pothole detection using computer vision and learning," *IEEE Transactions on Intelligent Transportation Systems,* vol. 21, no. 8, pp. 3536-3550. https://doi.org/10.1109/TITS.2019.2931297

[14]. H. Laga, L. V. Jospin, F. Boussaid, & M. Bennamoun, (2020). "A survey on deep learning techniques for stereo-based depth estimation," *IEEE transactions on pattern analysis and machine intelligence,* vol. 44, no. 4, pp. 1738-1764. https://doi.org/10.1109/TPAMI.2020.3032602

[15]. C. Xiong, J. Yu, & X. Zhang, (2023). "Use of NDT systems to investigate pavement reconstruction needs and improve maintenance treatment decision-making," *International Journal of Pavement Engineering,* vol. 24, no. 1, pp. 2011872. https://doi.org/10.1080/10298436.2021.2011872

[16]. W. Lin, X. Li, H. Han, Q. Yu, & Y. H. Cho, (2023). "A novel approach for pavement distress detection and quantification using RGB-D camera and deep learning algorithm," *Construction and Building Materials,* vol. 407, pp. 133593. https://doi.org/10.1016/j.conbuildmat.2023.133593

[17]. Y. Liu, F. Liu, W. Liu, & Y. Huang, (2023). "Pavement distress detection using street view images captured via action camera," *IEEE Transactions on Intelligent Transportation Systems,* vol. 25, no. 1, pp. 738-747. https://doi.org/10.1109/TITS.2023.3306578

[18]. K. R. Ahmed, (2021). "Smart pothole detection using deep learning based on dilated convolution," *Sensors,* vol. 21, no. 24, pp. 8406. https://doi.org/10.3390/s21248406

[19]. R. Kothai, & N. Prabakaran, (2025). "Self Attention GAN and SWIN Transformer based Pothole Detection with Trust Region based LSM and Hough Line Transform for 2D to 3D conversion," *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3595190

[20]. N. Bhavana, M. M. Kodabagi, B. M. Kumar, P. Ajay, N. Muthukumaran, & A. Ahilan, (2024). "POT-YOLO: Real-time road potholes detection using edge segmentation-based YOLO V8 network," *IEEE Sensors Journal,* vol. 24, no. 15, pp. 24802-24809. https://doi.org/10.1109/JSEN.2024.3399008

[21]. F. Ozoglu, & T. Gökgöz, (2023). "Detection of road potholes by applying convolutional neural network method based on road vibration data," *Sensors,* vol. 23, no. 22, pp. 9023. https://doi.org/10.3390/s23229023

[22]. Y. Zanevych, V. Yovbak, O. Basystiuk, N. Shakhovska, S. Fedushko, & S. Argyroudis, (2024). "Evaluation of pothole detection performance using deep learning models under low-light conditions," *Sustainability,* vol. 16, no. 24, pp. 10964. https://doi.org/10.3390/su162410964

[23]. A. Paramarthalingam, J. Sivaraman, P. Theerthagiri, B. Vijayakumar, & V. Baskaran, *Decision Analytics Journal*.

[24]. https://www.kaggle.com/datasets/rangerfan/pothole-600

[25]. Z. Lv, C. Cheng, & H. Lv, (2023). "Automatic identification of pavement cracks in public roads using an optimized deep convolutional neural network model," *Philosophical Transactions of the Royal Society A*, vol. 381, no. 2254, pp. 20220169. https://doi.org/10.1098/rsta.2022.0169

[26]. Z. Xu, H. Guan, J. Kang, X. Lei, L. Ma, Y. Yu, & J. Li, (2022). "Pavement crack detection from CCD images with a locally enhanced transformer network," *International Journal of Applied Earth Observation and Geoinformation,* vol. 110, pp. 102825. https://doi.org/10.1016/j.jag.2022.102825

[27]. J. Alfred Daniel, C. Chandru Vignesh, B. A. Muthu, R. Senthil Kumar, C. B. Sivaparthipan, & C. E. M. Marin, (2023). "Fully convolutional neural networks for LIDAR–camera fusion for pedestrian detection in autonomous vehicle," *Multimedia Tools and Applications,* vol. 82, no. 16, pp. 25107-25130. https://doi.org/10.1007/s11042-023-14417-x