

Road surface defect segmentation using the U-Net deep learning model

Quang-Quy Tran, Thi-Lien Pham, Huy-Manh Trieu, Minh-Tien Dang, Anh-Kiet Lam, Quoc-Trung Ha

Department of Computer Science, Faculty of Information Technology, Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam

Article Info

Article history:

Received Sep 27, 2025

Revised Jul 31, 2026

Accepted Aug 12, 2026

Keywords:

Decoder

Encoder

Pothole dataset

Road surface

U-Net

ABSTRACT

The rapid expansion of transportation infrastructure in Vietnam has introduced new challenges in road maintenance, particularly in detecting and managing surface defects. Early identification and precise localization of these defects are essential to ensuring traffic safety and optimizing maintenance operations. To address these issues, this study proposes a U-Net-based deep learning model integrated with an internet of things (IoT) architecture for real-time road surface defect detection and monitoring. The proposed system operates on an embedded Raspberry Pi platform equipped with a camera, global positioning system (GPS) module, and auditory alert unit, enabling on-device inference with minimal latency. A mobile web application is also developed to visualize detected defects, record their geolocations, and support road authorities in timely maintenance planning. Experimental results on the pothole dataset demonstrate a training accuracy of 97.7%, while the model achieves a test accuracy of 89.49% on unseen data, confirming the robustness of the model under various environmental conditions. Unlike previous studies focusing solely on image-based detection, this work introduces a fully integrated IoT-driven ecosystem that bridges deep learning, embedded computing, and cloud-based visualization—enhancing automation, scalability, and real-time responsiveness in intelligent transportation systems.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Thi-Lien Pham

Department of Computer Science, Faculty of Information Technology

Thai Nguyen University of Information and Communication Technology

Thai Nguyen, Vietnam

Email: ptlien@ictu.edu.vn

1. INTRODUCTION

Road surface defect analysis is a fundamental component of transportation infrastructure management, as undetected or poorly assessed defects can significantly compromise traffic safety and increase maintenance costs. Conventional inspection methods, which mainly rely on manual surveys or vehicle-mounted visual inspections, are often labor-intensive, time-consuming, and highly dependent on human experience, making them difficult to scale for large and complex road networks. Consequently, automated vision-based approaches have attracted increasing attention, particularly with the rapid advancement of deep learning techniques.

Despite recent progress, road surface analysis remains a challenging task due to the irregular geometry of defects, complex pavement textures, and varying environmental conditions such as illumination changes, shadows, and weather effects [1], [2]. In addition, recent studies have shown that real-world

deployment scenarios introduce further challenges related to data diversity, robustness, and generalization capability [3], [4]. These challenges highlight the need for more reliable and scalable segmentation frameworks capable of supporting practical road infrastructure monitoring systems.

In recent years, object detection-based deep learning models have been widely adopted for pothole and road damage detection. Approaches based on single-stage detectors, such as the you only look once (YOLO) family, have shown strong real-time performance and satisfactory detection accuracy, making them suitable for deployment in intelligent transportation systems and large-scale monitoring scenarios [5]–[7]. Region-based detectors and hybrid frameworks have also been explored to improve robustness under challenging conditions [8]. Despite these advantages, detection-oriented models typically represent defects using bounding boxes, which only provide coarse localization and fail to capture the exact shape and spatial extent of damaged regions. This limitation restricts their usefulness in applications that require precise defect quantification, such as severity estimation and maintenance prioritization.

Semantic segmentation offers a more detailed representation by assigning class labels at the pixel level, enabling accurate delineation of defect boundaries. This fine-grained characterization is particularly important for road surface analysis, where defects often exhibit irregular shapes and gradual transitions with the surrounding pavement. Among segmentation architectures, U-Net has emerged as a widely adopted and effective model due to its symmetric encoder-decoder structure and skip connections, which allow the network to preserve spatial information while learning high-level semantic features [9]. Originally introduced for biomedical image segmentation, U-Net has since been successfully adapted to various surface defect and pavement crack segmentation tasks, demonstrating robustness across different imaging conditions [10]–[12].

Several recent studies have explored enhanced U-Net variants and hybrid segmentation models for road-related applications, incorporating residual connections, multi-scale feature fusion, or attention mechanisms to improve segmentation accuracy [11], [13]. These works highlight the suitability of segmentation-based approaches for capturing complex defect patterns that are difficult to represent using bounding boxes alone. At the same time, lightweight U-Net configurations remain computationally feasible for deployment on embedded and edge devices, which is a key requirement for real-time road monitoring systems [10].

Despite the growing literature on road defect detection and segmentation, existing studies generally follow two distinct directions. Detection-based approaches primarily emphasize real-time performance but often provide limited geometric detail, whereas segmentation-based methods tend to focus on accuracy evaluation under controlled or offline conditions, without sufficient consideration of system-level integration. In particular, the comparative advantages of U-Net-based segmentation over YOLO-style detection models in practical, internet of things (IoT)-enabled road monitoring scenarios remain insufficiently explored.

Moreover, only a limited number of studies have investigated how pixel-level segmentation can be effectively integrated with embedded hardware and location-aware monitoring systems to support real-world infrastructure maintenance. Recent research has demonstrated that deep learning-based road damage analysis models can achieve robust performance under challenging environmental conditions while remaining computationally efficient for deployment on mobile and edge devices [14], [15]. In addition, large-scale road damage datasets and practical inspection frameworks have shown strong potential for improving model generalization across multiple categories of pavement defects, further highlighting the applicability of deep learning-based approaches beyond isolated experimental settings [16], [17].

To address this research gap, the present study investigates the application of the U-Net architecture for pothole segmentation within an IoT-based real-time monitoring framework. By emphasizing pixel-level defect localization rather than bounding-box detection, the proposed approach aims to provide more informative representations of road surface damage while maintaining practical deployability on edge devices. This integration contributes to intelligent transportation infrastructure by supporting accurate defect assessment, real-time monitoring, and data-driven maintenance decision-making.

2. METHOD

2.1. U-Net model

This study adopts the U-Net architecture for image segmentation, as illustrated in Figure 1. The model follows a symmetric encoder-decoder design that enables effective learning of both spatial structures and semantic representations from input images [18]. At the bottleneck layer, high-level semantic features are

learned, which play a crucial role in distinguishing complex structures within the image. During decoding, spatial resolution is progressively restored through up-convolution operations, while skip connections are employed to preserve fine-grained spatial information from earlier layers. This design helps maintain structural consistency and improves boundary localization in the final segmentation results.

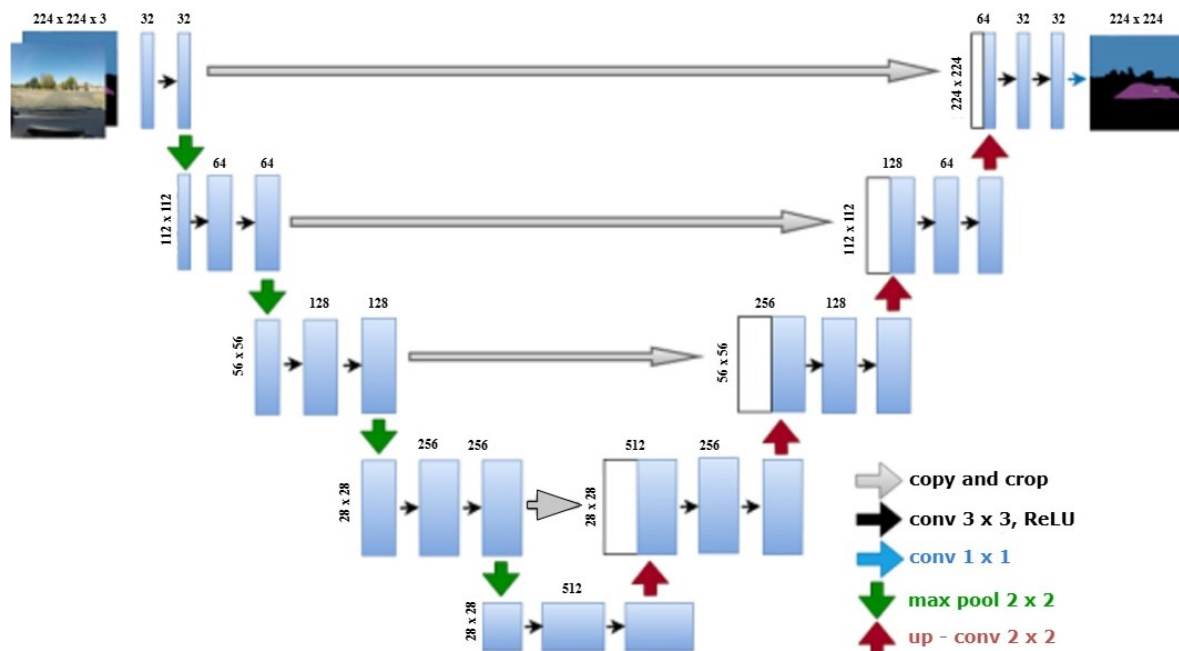


Figure 1. The overall architecture of the U-Net model [18]

Due to its balanced combination of representation capacity and structural efficiency, the U-Net architecture has been widely adopted in various segmentation tasks, particularly in road surface and pothole detection. Recent studies have shown that enhanced variants of U-Net can further improve segmentation robustness through adaptive training strategies and refined feature fusion mechanisms [19], [20]. In addition, hybrid architectures integrating transformer-based global context modeling have demonstrated strong capability in handling complex visual patterns encountered in real-world scenarios [21]. Overall, U-Net remains a reliable and effective backbone for accurate pixel-level segmentation in practical applications [9].

2.1.1. Encoder

The encoder is responsible for learning hierarchical feature representations by progressively extracting spatial and semantic information from the input image. It consists of repeated blocks of two convolutional layers with 3×3 kernels, each followed by a rectified linear unit (ReLU) activation function, and a 2×2 max-pooling operation with a stride of 2. This hierarchical downsampling process gradually reduces the spatial resolution of feature maps while simultaneously increasing their depth, allowing the network to capture increasingly abstract and discriminative features. Such a design has been proven effective for preserving contextual information while enhancing feature representation in semantic segmentation tasks [9], [10].

2.1.2. Decoder

The decoder aims to recover spatial information through transposed convolutional layers, also referred to as upsampling layers. After each upsampling step, the feature maps from the decoder are concatenated with the corresponding feature maps from the encoder via skip connections. This mechanism allows the model to combine low-level information (e.g., position, boundaries) with high-level semantic information, thereby enhancing segmentation quality, particularly around object edges. Finally, the output of the model is generated by a 1×1 convolutional layer with either a sigmoid or softmax activation function (depending on the number of segmentation classes), producing a pixel-wise probability distribution map over the input image. Figure 1

illustrates the overall architecture of the proposed segmentation framework, including the encoder–decoder structure and skip connection mechanism used for feature fusion.

Figure 2 illustrates the overall workflow of the proposed segmentation framework. The process begins with the input RGB image, which is first subjected to data preprocessing to ensure consistent representation. The processed data are then fed into the U-Net architecture, where hierarchical features are extracted through the encoder, refined at the bottleneck, and reconstructed by the decoder using skip connections. Finally, the model produces a pixel-wise segmentation mask that highlights road surface defects.

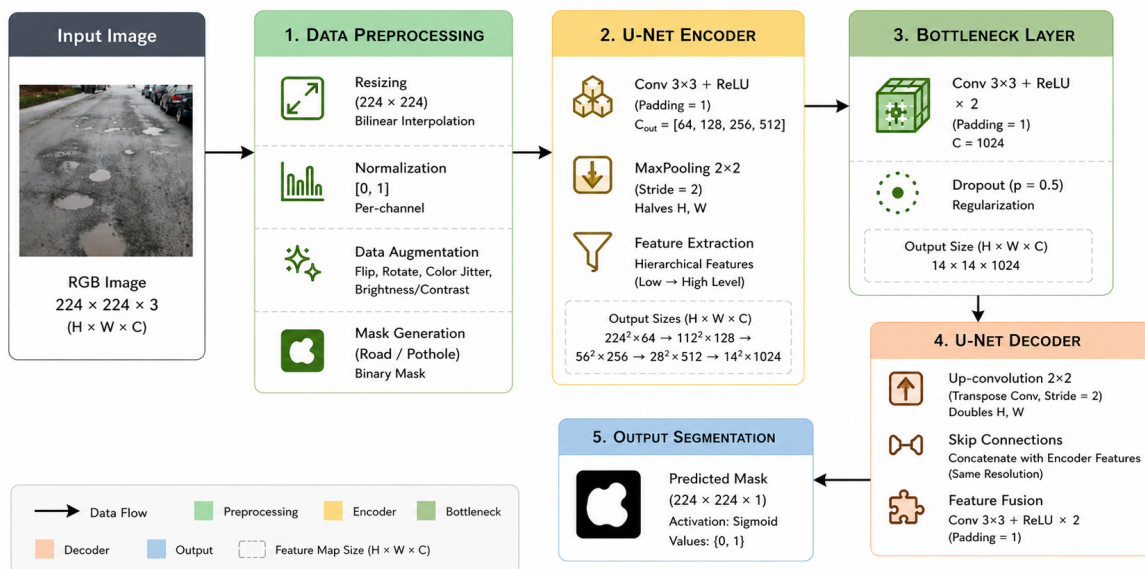


Figure 2. Workflow of the proposed segmentation framework

3. DATASET AND HYPERPARAMETERS

The dataset used in this study is the pothole image segmentation dataset, collected from Roboflow. It consists of a total of 780 annotated images, formatted according to the YOLOv8 specification for the tasks of pothole detection and segmentation. The images underwent preprocessing and data augmentation procedures to optimize the training of the models. Figures 3 and 4 illustrate examples of the dataset, where Figure 3 presents sample raw images and Figure 4 shows the corresponding labeled annotations.

To improve model robustness and generalization capability, several data augmentation techniques were applied during the training process. These augmentations included horizontal flipping, random rotation, brightness and contrast adjustment, Gaussian blur and noise injection, as well as random zooming and spatial shifting operations. Such transformations increase the diversity of training samples and can improve the robustness of deep learning models to variations in road appearance and imaging conditions [6].



Figure 3. Sample images from the pothole dataset under different road conditions

```

0 0.9897727281250001 0.4303030296875 0.9795454546875 0.40909090781249996 0.928409090625 0.375757575 0.871590909375 0.36969697031250004 0.8
0.38333333281250004 0.735227271875 0.384848484375 0.6829545453125 0.3772727265625 0.59431818125 0.35909090878125 0.515909090625
0.3545454546875 0.43181818125 0.3454545453125 0.34431818125 0.3424242421875 0.259090909375 0.3424242421875 0.177272728125 0.3515151515625
0.14318181875 0.360606059375 0.09318181875 0.36666666718749996 0.103409090625 0.4121212109375 0.164772728125 0.46363636406250003
0.17840909062500002 0.49696969687500003 0.177272728125 0.55 0.2613636359375 0.6121212109374999 0.41818181875000004 0.5909090921875
0.49431818125 0.5893939406250001 0.640909090625 0.5924242421875 0.6761363640625 0.5924242421875 0.7556818187500001 0.5878787890625
0.8102272718749999 0.5848484843749999 0.8613636359375001 0.5727272734375 0.9193181812500001 0.5681818187500001 0.9636363640625
0.5484848484375 1 0.46818181875000003 0.9897727281250001 0.4303030296875

```

Figure 4. Ground-truth segmentation masks corresponding to the input images, used for training and evaluating the segmentation model

The annotation files in the pothole image segmentation dataset strictly follow the YOLO labeling format, which is widely adopted for object detection and segmentation tasks. Each annotation corresponds to a single object instance and is represented by a five-element vector consisting of the class identifier, the normalized x- and y-coordinates of the bounding box center, and the normalized width and height of the bounding box. All spatial coordinates are normalized to the range [0, 1] with respect to the original image dimensions, ensuring scale invariance and robustness across images with varying resolutions.

In this dataset, the class_id value of 0 represents the pothole category, reflecting its single-class segmentation nature. The x_center and y_center values indicate the relative location of the object centroid, while the width and height parameters describe the spatial extent of the bounding box. This normalized representation enables efficient learning and stable convergence during training, as commonly adopted in modern object detection and segmentation frameworks [22], [23].

For instance, an annotation entry such as 0 0.89777 0.43083 0.97954 0.40909 indicates a pothole instance whose bounding box center is located at approximately 89.8% of the image width and 43.1% of the image height, with the bounding box spanning approximately 97.9% of the image width and 40.9% of the image height. This concise yet expressive representation facilitates efficient data parsing and enables consistent training and evaluation across large-scale datasets. Figure 4 illustrates the labeled data associated with each image, stored in .txt files. Each file contains a list of polygon coordinates used to generate masks for the segmentation task.

The objective of this study is to evaluate the performance of the U-Net model in pothole segmentation based on input images and their corresponding labels. To enhance the efficiency of model training, several preprocessing techniques were applied:

- i) Color space conversion: images were converted to the RGB color space, while masks were transformed into grayscale images.
- ii) Mask handling: if an image did not have an associated mask, the system automatically generated an empty mask (all pixel values set to 0).
- iii) Resizing: both images and masks were resized to a standard dimension of 224×224 pixels.
- iv) Data normalization: pixel values of both images and masks were scaled to the range [0, 1] to optimize the training process.
- v) Batch creation: images and masks were grouped into batches of a fixed size to improve training efficiency.
- vi) Data shuffling: after each epoch, the dataset was shuffled to prevent overfitting and to promote better generalization of the model.
- vii) Dataset splitting: the dataset was partitioned into 80% for the training set and 20% for the validation set, allowing for performance evaluation and detection of potential overfitting.

Those preprocessing steps enable the model to learn effectively from the data and enhance its ability to accurately detect potholes. During the deep learning model training process, the input data is not fed into the model all at once but is instead divided into smaller subsets called mini-batches. This partitioning allows the model to update its weights efficiently, reduces memory computation costs, and accelerates convergence during training. The size of each mini-batch, known as the batch size, determines the number of samples processed in each training iteration. Assuming the training dataset consists of a total of A samples and the batch size is set to N , the number of iterations required to complete one full pass over the dataset (i.e., one epoch) is calculated according to (1).

$$\text{Number of iterations per epoch} = \frac{A}{N} \quad (1)$$

To provide a clearer understanding of the network configuration and feature transformation process, Table 1 summarizes the detailed architecture of the proposed U-Net model. The table presents the encoder–decoder structure, including convolutional layers, pooling and upsampling operations, skip connections, output feature dimensions, and trainable parameters. This detailed representation helps illustrate how spatial and semantic information is progressively extracted, preserved, and reconstructed throughout the segmentation network.

Table 1. Detailed architecture of the proposed U-Net model

Stage	Layer (type)	Output shape	Parameters	Connected to
Input Layer				
1	InputLayer	$256 \times 256 \times 3$	0	–
Encoder (downsampling path)				
2	Conv2D + ReLU	$256 \times 256 \times 128$	3,584	InputLayer
3	Conv2D + ReLU	$256 \times 256 \times 128$	147,584	Conv2D_8
4	MaxPooling2D	$128 \times 128 \times 128$	0	Conv2D_9
5	Dropout	$128 \times 128 \times 128$	0	MaxPooling2D_3
6	Conv2D + ReLU	$128 \times 128 \times 256$	295,168	Dropout_3
7	Conv2D + ReLU	$128 \times 128 \times 256$	590,080	Conv2D_10
8	MaxPooling2D	$64 \times 64 \times 256$	0	Conv2D_11
9	Dropout	$64 \times 64 \times 256$	0	MaxPooling2D_4
10	Conv2D + ReLU	$64 \times 64 \times 512$	1,180,160	Dropout_4
11	Conv2D + ReLU	$64 \times 64 \times 512$	2,359,808	Conv2D_12
12	MaxPooling2D	$32 \times 32 \times 512$	0	Conv2D_13
13	Dropout	$32 \times 32 \times 512$	0	MaxPooling2D_5
Bottleneck layer				
14	Conv2D + ReLU	$32 \times 32 \times 512$	2,359,808	Dropout_5
15	Conv2D + ReLU	$32 \times 32 \times 512$	2,359,808	Conv2D_14
Decoder (upsampling path)				
16	Conv2DTranspose	$64 \times 64 \times 256$	1,179,904	Conv2D_15
17	Concatenate	$64 \times 64 \times 768$	0	Skip Connection
18	Dropout	$64 \times 64 \times 768$	0	Concatenate
19	Conv2D + ReLU	$64 \times 64 \times 256$	1,769,728	Dropout_6
20	Conv2D + ReLU	$64 \times 64 \times 256$	590,080	Conv2D_16
21	Conv2DTranspose	$128 \times 128 \times 128$	295,040	Conv2D_17
22	Concatenate	$128 \times 128 \times 384$	0	Skip Connection
23	Dropout	$128 \times 128 \times 384$	0	Concatenate_1
24	Conv2D + ReLU	$128 \times 128 \times 128$	442,496	Dropout_7
25	Conv2D + ReLU	$128 \times 128 \times 128$	147,584	Conv2D_18
26	Conv2DTranspose	$256 \times 256 \times 64$	73,792	Conv2D_19
27	Concatenate	$256 \times 256 \times 192$	0	Skip Connection
28	Dropout	$256 \times 256 \times 192$	0	Concatenate_2
29	Conv2D + ReLU	$256 \times 256 \times 64$	110,656	Dropout_8
30	Conv2D + ReLU	$256 \times 256 \times 64$	36,928	Conv2D_20
31	Conv2D (Output)	$256 \times 256 \times 2$	130	Conv2D_21

The hyperparameters adjusted for the proposed U-Net model are: i) batch size: 16, ii) epochs: 100, iii) learning rate: 0.0001, iv) validation split: 20% of the training set, v) patience: 10 (training stops if no improvement is observed after 10 consecutive epochs), vi) mode: accuracy (model performance is evaluated based on the accuracy metric, which will be discussed in the results and discussion section), and vii) min_delta: 0.005 (smallest improvement between epochs to reset the patience counter). The training configuration for the proposed U-Net model was carefully designed to achieve stable convergence and robust segmentation performance for pothole detection. A batch size of 16 was selected to balance graphics processing unit (GPU) memory efficiency and the preservation of fine-grained spatial information during feature learning. The model was trained for 100 epochs to ensure sufficient learning capacity while mitigating overfitting through an early stopping strategy with a patience value of 10 epochs and a minimum improvement threshold (min_delta) of 0.005.

In this study, sparse categorical crossentropy (SCC) was employed as the primary loss function for multi-class semantic segmentation. SCC is widely adopted in deep learning-based segmentation tasks because it efficiently handles integer-encoded class labels without requiring one-hot encoding, thereby reducing memory consumption and computational complexity. In addition, cross-entropy-based loss functions

have demonstrated strong optimization capability and stable gradient propagation in semantic segmentation frameworks [24], [25]. Compared with conventional categorical cross-entropy, SCC is particularly suitable for dense pixel-wise classification problems where each pixel belongs to a single semantic category. The mathematical formulation of SCC is expressed in (2).

$$L_{SCC} = -\frac{1}{N} \sum_{i=1}^N \log(p_{i,y_i}) \quad (2)$$

Where N denotes the total number of pixels, y_i represents the ground-truth class label of pixel i , and p_{i,y_i} denotes the predicted probability corresponding to the correct class label.

Given an input sample, let $\mathbf{z} = (z_1, z_2, \dots, z_C)$ denote the unnormalized logits produced by the model, where C is the total number of classes. The true label for the sample is represented by an integer $y \in \{1, 2, \dots, C\}$. First, the softmax function is applied to transform logits into a probability distribution over classes as in (3).

$$p_j = \frac{\exp(z_j)}{\sum_{k=1}^C \exp(z_k)} \quad \text{for } j = 1, 2, \dots, C \quad (3)$$

Where p_j denotes the predicted probability for class j .

The standard categorical crossentropy loss for a single sample is then defined as (4).

$$\mathcal{L} = -\sum_{j=1}^C y_j \log(p_j) \quad (4)$$

Where y_j is the one-hot encoding of the ground-truth class label. However, in SCC, the true label is an integer index, and the loss simplifies as in (5).

$$\mathcal{L} = -\log(p_y) \quad (5)$$

Substituting the softmax definition from (3) into (5), the SCC loss can be explicitly written in terms of logits as in (6).

$$\mathcal{L} = -\log\left(\frac{\exp(z_y)}{\sum_{k=1}^C \exp(z_k)}\right) = -z_y + \log\left(\sum_{k=1}^C \exp(z_k)\right) \quad (6)$$

Thus, in practice, modern deep learning frameworks implement SCC directly on logits without explicitly computing the softmax probabilities, utilizing the numerically stable formulation in (6). This avoids redundant computation and improves numerical stability, particularly when logits have large magnitudes.

Although the selected configuration may not represent an absolutely optimal setup for all tasks, the hyperparameters used in this study were fine-tuned based on practical experience with the objective of balancing accuracy and training efficiency. These settings contribute to enhancing the model's generalization ability across varying environmental conditions and datasets. For future work, hyperparameter optimization techniques such as grid search or random search could be employed to further improve performance and better adapt to specific datasets by discovering more effective combinations of training parameters.

4. RESULTS AND DISCUSSION

To evaluate the performance of the proposed U-Net segmentation model, the accuracy metric is employed as the primary evaluation criterion. Accuracy is a widely used indicator in classification and segmentation tasks, reflecting the proportion of correctly classified pixels over the total number of pixels in the output image. A higher accuracy value indicates that the model can more precisely predict the location and boundaries of the target objects to be segmented. The formula for calculating accuracy is presented in (7).

$$\text{Accuracy} = \frac{\sum_{i=1}^{H \times W} \mathbb{I}(y_i = \hat{y}_i)}{H \times W} \quad (7)$$

Where $\hat{y}_i$ denotes the model's predicted label at pixel i , y_i is the ground-truth label at pixel i , $H \times W$ represents the height and width of the image (i.e., the total number of pixels), and $\mathbb{I}(\hat{y}_i = y_i)$ is the indicator function, which equals 1 if the prediction is correct and 0 otherwise.

The training process over 100 epochs is illustrated in Figure 5. Figure 5(a) presents the evolution of the accuracy metric for both the training and validation datasets, while Figure 5(b) depicts the corresponding loss values across epochs for the training and validation datasets. Table 2 summarizes the quantitative performance of the proposed model across different evaluation metrics. The results demonstrate that the segmentation framework achieves high accuracy and robust generalization under varying road conditions. In particular, the proposed architecture provides improved boundary segmentation performance compared with baseline approaches.

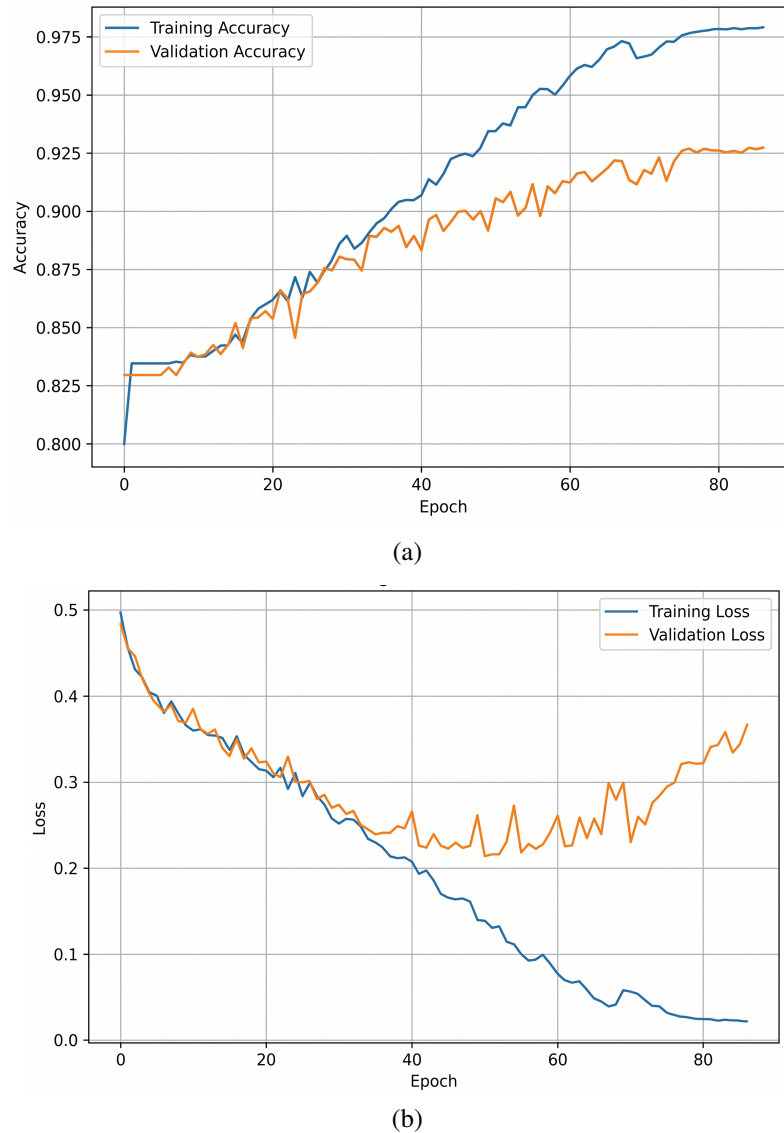


Figure 5. Training and validation results: (a) accuracy and (b) loss

Table 2. Results after training the U-Net model

Dataset	Loss	Accuracy
Training set	0.0284	0.9770
Validation set	0.2982	0.9267
Testing set	0.4629	0.8949

Figure 6 presents the qualitative segmentation results produced by the proposed model under different road surface conditions. Figure 6(a) shows examples of successful crack segmentation under normal lighting environments. Figure 6(b) illustrates segmentation performance in more challenging scenarios with shadows, uneven illumination, or complex pavement textures.



Figure 6. Results from the trained model: (a) original image and (b) predicted segmentation mask

Table 3 presents a comparative evaluation of the proposed U-Net framework with several state-of-the-art semantic segmentation models for pothole and pavement defect analysis. The comparison focuses on the loss functions, final loss values, and test accuracy obtained from different segmentation approaches. Experimental results show that SegFormer-B0 achieved the highest test accuracy of 90.1%, while the proposed U-Net obtained a competitive accuracy of 89.5% with the lowest final loss value of 0.23 among the evaluated deep learning models. In contrast, DeepLabV3+ achieved a slightly lower accuracy of 88.9% and a higher final loss of 0.2569, indicating less stable convergence during training. These findings demonstrate that the proposed U-Net provides a balanced trade-off between segmentation accuracy and optimization stability. Furthermore, the proposed U-Net achieved a test accuracy of 89.5%, which is comparable to the results reported for other segmentation architectures, including U-Net, regression-segmentation network (RS-Net), and MultiResUNet. It should also be noted that differences in datasets, loss functions, and evaluation protocols across studies may affect direct numerical comparison.

Table 3. Comparison of loss and accuracy for pothole segmentation models

Model	Dataset	Loss function	Final loss	Test accuracy (%)
Proposed U-Net	Roboflow pothole	Sparse categorical cross-entropy	0.23	89.5
DeepLabV3+	Roboflow pothole	Sparse categorical cross-entropy	0.2569	88.9
SegFormer-B0	Roboflow pothole	Sparse categorical cross-entropy	0.232	90.1
U-Net [12]	Pavement crack	Cross-entropy	Not reported	87.3
RS-Net [11]	Pavement defect	Binary cross-entropy + Dice	0.28	88.1
MultiResUNet [10]	Surface defect	Dice loss	Not reported	86.5

Through the experimental results, it can be observed that all evaluated segmentation models were capable of identifying pothole regions under realistic road surface conditions. Figure 6(a) illustrates the original road image containing potholes of varying shapes and sizes, while Figure 6(b) presents the corresponding segmentation masks generated by the trained model. The segmented regions successfully highlight damaged road areas using distinctive colors, demonstrating the ability of deep learning-based segmentation architectures to extract meaningful spatial and contextual features from complex road scenes.

However, the qualitative results also reveal several limitations. Certain pothole boundaries were partially misclassified or affected by minor segmentation noise, particularly in regions with unclear visual characteristics, shadows, or distant damaged areas located far from the image center. Although transformer-based models such as SegFormer-B0 achieved superior quantitative performance, the proposed U-Net framework still demonstrated promising applicability for pothole segmentation while maintaining

a relatively lightweight architecture and stable optimization behavior. These findings provide a useful foundation for future development of automated road monitoring and maintenance systems, particularly in smart transportation and intelligent infrastructure management applications.

5. CONCLUSION

This study examined the effectiveness of the U-Net architecture for pothole segmentation within an IoT-enabled real-time road monitoring framework. Experimental results demonstrate that the proposed model is capable of accurately delineating road surface defects, achieving a segmentation accuracy of approximately 89.49% on unseen test data, which indicates a good level of generalization under realistic conditions. By focusing on pixel-level segmentation rather than bounding-box detection, the proposed approach provides a more detailed representation of defect shapes and spatial extents, which is essential for assessing defect severity and supporting informed maintenance decisions. Moreover, the integration of the U-Net model with embedded hardware and location-aware monitoring highlights the practical applicability of the proposed system in intelligent transportation infrastructure, enabling timely detection and visualization of road surface damage. Future research will extend this work by evaluating the proposed framework on larger and more diverse datasets, incorporating additional segmentation metrics such as intersection over union (IoU) and Dice coefficient, and further optimizing the system for real-time deployment in large-scale road networks.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Quang-Quy Tran	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			
Thi-Lien Pham		✓	✓		✓	✓		✓	✓	✓	✓			
Huy-Manh Trieu		✓	✓	✓	✓	✓	✓	✓						
Minh-Tien Dang		✓				✓		✓	✓	✓				
Anh-Kiet Lam		✓				✓		✓	✓	✓				
Quoc-Trung Ha	✓		✓		✓			✓		✓				

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal Analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project Administration

Fu : Funding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The supporting data of this study are openly available in Roboflow at <https://universe.roboflow.com/gp-grakz/pothole-segmentation-g6hbh>.

REFERENCES

- [1] J. Yu, J. Jiang, S. Fichera, P. Paoletti, L. Layzell, D. Mehta, and S. Luo, "Road surface defect detection—from image-based to non-image-based: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 9, pp. 10581–10603, 2024, doi: 10.1007/s10845-019-01476-x.

- [2] D. Tabernik, S. Šela, J. Skvarč, and D. Skočaj, "Segmentation-based deep-learning approach for surface-defect detection," *Journal of Intelligent Manufacturing*, vol. 31, no. 3, pp. 759–776, 2020, doi: 10.1007/s10845-019-01476-x.
- [3] Y. Zanevych, V. Yovbak, O. Basystiuk, N. Shakhovska, S. Fedushko, and S. Argyroudis, "Evaluation of pothole detection performance using deep learning models under low-light conditions," *Sustainability*, vol. 16, no. 24, 2024, doi: 10.3390/su162410964.
- [4] L. Yang, J. Deng, H. Duan, and C. Yang, "An efficient semantic segmentation method for road crack based on EGA-UNet," *Scientific Reports*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-01983-3.
- [5] Y. K. Yik, N. E. Alias, Y. Yusof, and S. Isaak, "A real-time pothole detection based on deep learning approach," in *Journal of Physics: Conference Series*, vol. 1828, 2021, doi: 10.1088/1742-6596/1828/1/012001.
- [6] K. Gorro, E. Ranolo, L. Roble, and R. N. Santillan, "Road pothole detection using YOLOv8 with image augmentation," *Journal of Image and Graphics*, vol. 12, no. 4, pp. 417–426, 2024, doi: 10.18178/joig.12.4.417-426.
- [7] J. Zeng and H. Zhong, "YOLOv8-PD: An improved road damage detection algorithm based on YOLOv8n model," *Scientific reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-62933-z.
- [8] H. Tahir and E. S. Jung, "Comparative study on distributed lightweight deep learning models for road pothole detection," *Sensors*, vol. 23, no. 9, 2023, doi: 10.3390/s23094347.
- [9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention (MICCAI 2015)*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
- [10] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020, doi: 10.1016/j.neunet.2019.08.025.
- [11] L. Ali, H. AlJassmi, M. Swavaf, W. Khan, and F. Alnajjar, "Rs-net: Residual sharp U-Net architecture for pavement crack segmentation and severity assessment," *Journal of Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00981-y.
- [12] Q. Zhang *et al.*, "Improved U-net network asphalt pavement crack detection method," *PLoS ONE*, vol. 19, no. 5 May, 2024, doi: 10.1371/journal.pone.0300679.
- [13] I. Katsamenis *et al.*, "A deep learning framework for segmentation of road defects using ResUNet-a," in *17th International Conference on Pervasive Technologies Related to Assistive Environments*, 2024, pp. 449–455, doi: 10.1145/3652037.3663935.
- [14] S. Shim, J. Kim, S. W. Lee, and G. C. Cho, "Road damage detection using super-resolution and semi-supervised learning with generative adversarial network," *Automation in Construction*, vol. 135, 2022, doi: 10.1016/j.autcon.2022.104139.
- [15] C. Wu, M. Ye, J. Zhang, and Y. Ma, "YOLO-LWNet: a lightweight road damage object detection network for mobile terminal devices," *Sensors*, vol. 23, no. 6, 2023, doi: 10.3390/s23063268.
- [16] Y. Jiang, "Road damage detection and classification using deep neural networks," *Discover Applied Sciences*, vol. 6, no. 8, pp. 1127–1141, 2024, doi: 10.1007/s42452-024-06129-0.
- [17] S. K. Baduge *et al.*, "Artificial intelligence and smart vision for building and construction 4.0: machine and deep learning methods and applications," *Automation in Construction*, vol. 141, 2022, doi: 10.1016/j.autcon.2022.104440.
- [18] V. Pereira, S. Tamura, S. Hayamizu, and H. Fukai, "Semantic segmentation of paved road and pothole image using U-net architecture," in *2019 International Conference on Advanced Informatics: Concepts, Theory, and Applications, ICAICTA 2019*, 2019, pp. 1–4, doi: 10.1109/ICAICTA.2019.8904105.
- [19] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. M.-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021, doi: 10.1038/s41592-020-01008-z.
- [20] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: a nested U-net architecture for medical image segmentation," in *International workshop on deep learning in medical image analysis*, 2018, pp. 3–11, doi: 10.1007/978-3-030-00889-5_1.
- [21] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Computer vision*, Cham, Switzerland: Springer, 2022, pp. 205–218, doi: 10.1007/978-3-031-25066-8_9.
- [22] Q. M. U. Haq, S.-J. Ruan, M. A. Haq, S. Karam, J. L. Shieh, P. Chondro, D.-Q. Gao, "An incremental learning of YOLOv3 without catastrophic forgetting for smart city applications," *IEEE Consumer Electronics Magazine*, vol. 11, no. 5, pp. 56–63, 2022, doi: 10.1109/MCE.2021.3096376.
- [23] S. K. Jaiswal and R. Agrawal, "A comprehensive review of YOLOv5: advances in real-time object detection," *International Journal of Innovative Research in Computer Science and Technology*, vol. 12, no. 3, pp. 75–80, 2024, doi: 10.55524/ijirest.2024.12.3.12.
- [24] J. Terven, D. M. C.-Esparza, J. A. R.-González, A. R.-Pedraza, and E. A. C.-Urbola, "A comprehensive survey of loss functions and metrics in deep learning," *Artificial Intelligence Review*, vol. 58, no. 7, 2025, doi: 10.1007/s10462-025-11198-7.
- [25] M. Yeung, E. Sala, C. B. Schönlieb, and L. Rundo, "Unified focal loss: generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Computerized Medical Imaging and Graphics*, vol. 95, 2022, doi: 10.1016/j.compmedimag.2021.102026.

BIOGRAPHIES OF AUTHORS



Quang-Quy Tran    received his engineer's degree from Irkutsk National Research Technical University, Russian Federation, majoring in Computer Science, from 2007 to 2013. He then pursued his graduate studies at Kazan National Research Technological University, Russian Federation, specializing in Computer Science and Statistical Mathematics from 2014 to 2019. He successfully completed his Ph.D. dissertation and was awarded the doctoral degree in 2019 in the Russian Federation. Since 2020, he has been a lecturer at the Faculty of Information Technology, Thai Nguyen University of Information and Communication Technology, Vietnam. His main research interests include computer vision, computational mathematics, big data analytics, and queueing models in statistical mathematics. He can be contacted at email: tqquy@ictu.edu.vn.



Thi-Lien Pham    is a lecturer at the Department of Information Systems, Faculty of Information Technology, Thai Nguyen University of Information and Communication Technology, Vietnam. Her research interests include artificial intelligence (AI), AI agents, computer vision, and natural language processing (NLP). Her current research focuses on autonomous intelligent agents, multimodal AI, computer vision, and intelligent language technologies. She can be contacted at email: ptlien@ictu.edu.vn.



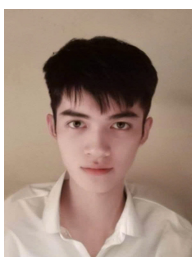
Huy-Manh Trieu    is currently pursuing the engineer's degree in Information Technology at Thai Nguyen University of Information and Communication Technology, Vietnam, from 2022 to 2027. His research interests include computer vision, machine learning, deep learning, NLP, generative AI, and AI agents. He is actively interested in intelligent systems and real-world AI applications, particularly in the areas of image analysis, intelligent automation, and smart infrastructure systems. He can be contacted at email: trieuhuymanh030@gmail.com.



Minh-Tien Dang    is currently a student at Thai Nguyen University of Information and Communication Technology, Vietnam. He is pursuing his studies in the field of information technology, with interests in image processing, computer vision, and AI. His academic work focuses on deep learning-based image segmentation and practical applications of intelligent systems in real-world problems. He can be contacted at email: tien1234d1@gmail.com.



Anh-Kiet Lam    is currently pursuing the engineer's degree in Information Technology at Thai Nguyen University of Information and Communication Technology, Vietnam, from 2022 to 2027. His primary research interests lie in the field of AI, with a strong focus on NLP, computer vision, and advanced agentic RAG architectures. He is deeply engaged in the development of intelligent systems and real-world AI applications, particularly specializing in cutting-edge projects such as automated resume analysis systems and virtual academic support assistants. Driven by a passion for innovation, he consistently seeks to explore practical AI implementations that enhance automated processes and optimize system workflows. He can be contacted at email: lilviet2004@gmail.com.



Quoc-Trung Ha    is currently pursuing the engineer's degree in Information Technology at Thai Nguyen University of Information and Communication Technology, Vietnam, from 2021 to 2026. His research interests include computer vision, machine learning, deep learning, NLP, generative AI, and AI agents. He is actively interested in intelligent systems and real-world AI applications, particularly in the areas of image analysis, intelligent automation, and smart infrastructure systems. He can be contacted at email: haquocchung775@gmail.com.