Pothole and Manhole Cover Detection for Road Safety Systems

Duy-Linh Nguyen, Xuan-Thuy Vo, Adri Priadana, Jehwan Choi, and Kang-Hyun Jo
Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan, Korea
ndlinh301@mail.ulsan.ac.kr, xthuy@islab.ulsan.ac.kr, priadana@mail.ulsan.ac.kr, cjh1897@ulsan.ac.kr, acejo@ulsan.ac.kr

***Abstract*—Potholes and cracks on the road often appear during traffic operations. This is one of the main factors causing traffic accidents and is a major concern for vehicle owners. Early detection and repair of potholes and cracks is essential to ensure smooth traffic flow and avoid risks. This paper proposes a method to improve the YOLOv8s network in general object detection and applies it to detecting potholes and manhole covers to support automatic detection and road repair. The research focuses on replacing the original convolution operations in the backbone and neck modules with the new one called Receptive Field Coordinate Attention Convolution. This module uses the Group Convolution and Coordinate Attention mechanism to enhance feature extraction capabilities. The experimental results are conducted on the pothole and manhole cover detection dataset and reported using mean average precision. As a result, the proposed network achieves the best performance at 77.1% of mAP@0.5 and 36.8% of mAP@0.5:0.95 and demonstrates superiority over other networks.**

***Index Terms*—Coordinate Attention, Convolutional Neural Network (CNN), Group convolution, YOLOv8.**

I. Introduction

Road traffic is a type of transportation system that makes an important contribution to each country's economic and social development. Among them, pothole and manhole cover pose many challenges for maintenance agencies and are a great danger to all types of vehicles. These road defects are automatically generated during operation by many factors such as adverse weather, environment, heavy traffic load, and substandard maintenance [1]. Pothole and manhole cover not only reduce the experience of road users but also compromise safety, causing collisions and accidents that lead to vehicle damage and can cause death [2]. Usually, the collection and processing of road damage data is carried out manually or reported by road users through their experiences. Although these methods are simple to implement, they are ineffective, have high costs, and easily cause delays [3]. Today, the rapid development of sensor technology, machine learning, and artificial intelligence (AI) facilitate modern and more effective approaches to pothole detection and maintenance. The pothole and manhole cover detection systems utilize different sensor methods such as conventional cameras, infrared cameras, and LIDAR combined with machine learning algorithms to detect and evaluate road surface conditions more accurately [4]. Inspired by the vision-based approach, this paper proposes a technique to improve the YOLOv8s network architecture for pothole and manhole cover detection. Through a thorough analysis of each component in the YOLOv8s network architecture, this work replaces all standard convolution operations with a new convolution operation called Receptive Field Coordinate Attention Convolution (RFCAConv) inside the backbone and neck modules. The RFCAConv is a combination of lightweight convolution (Group convolution) and Coordinate Attention (CA) mechanisms to enhance feature extraction for each feature map level. Optimization of network parameters and computational complexity while still ensuring object detection efficiency holds great potential for deployment in real-time systems with low-computing and embedded devices.

The major contributions of this paper are as follows:

- Improves the YOLOv8s network architecture for pothole and manhole cover detection supporting road safety systems.
- The proposed approach achieves better performance than other methods on the pothole and manhole cover dataset.

The remainder of the manuscript is organized as follows: Section II presents the approaches related to pothole and manhole detection. Section III introduces the details of the proposed method. Section IV explains and analyzes the experiments. Finally, Section V concludes the issue and directs the future works.

II. Related Works

The related works section introduces the various methods applied in pothole and manhole cover detection. These techniques can be split into two groups: sensors-based and CNN-based techniques.

A. Sensors-based techniques.

Detecting and evaluating road surface conditions is a labor-intensive and costly task. Therefore, researchers focus on designing compact systems and ensuring accuracy using mobile sensors such as accelerometer [5], vibration sensor [6], LIDAR [7], and Stereo vision [8]. The accelerometer-based and vibration-based approaches achieve high accuracy and do not depend on vision. Still, they have low response speeds and require driving over potholes to detect and measure. The LIDAR-based approaches can detect objects in poor lighting and visibility conditions. Therefore, it is often used to detect potholes at night. The stereo image-based system can retrieve information about potholes' size, depth, and location. Despite its remarkably high accuracy, this method has disadvantages

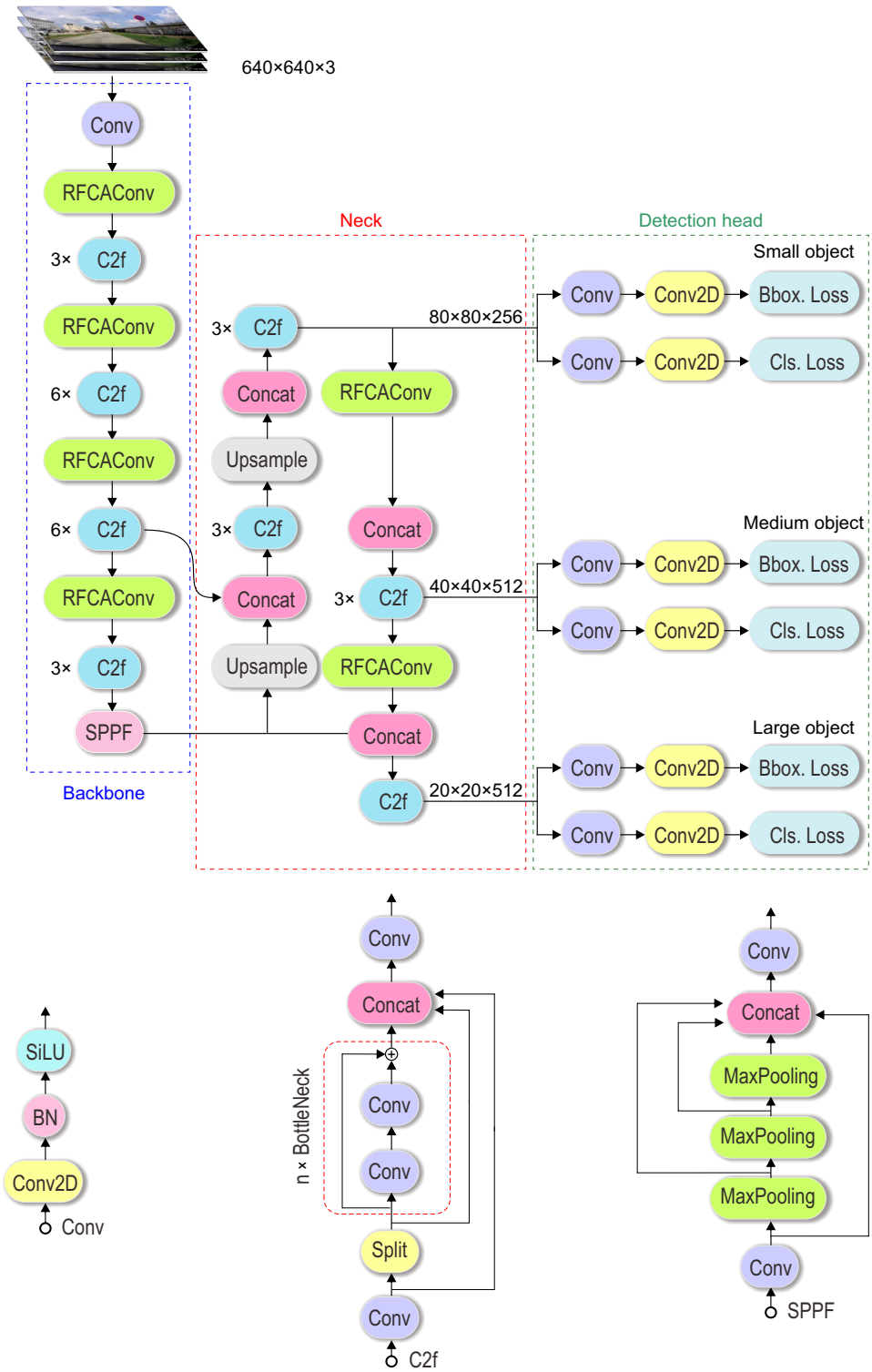

Fig. 1. The overall proposed network architecture and sub-modules.

such as computational complexity and sensitivity to movements and vibrations.

B. CNN-based techniques.

The authors in [9] evaluated the detection speed and precision of different semantic segmentation and object detection algorithms for the pothole detection task. The experimental results demonstrate that the performance of the YOLOv3 network architecture is superior to the YOLOv2 network architecture in both precision and speed. Similarly, the experiments in [10] also proposed a new dataset for pothole detection and improved the precision by integrating with Sparse R-CNN. Another study proposed a two-part location-aware convolutional neural network to detect potholes. In this method, a localization part is responsible for suggesting regions containing potholes and selecting the most relevant area. Then, these regions are recognized by another classification network. Later versions of the YOLO network family were also applied for pothole detection in research [11], including YOLOv4, YOLOv4 Tiny, and YOLOv5s. Experiments show that YOLOv4 and YOLOv4 Tiny achieve better performance than YOLOv5s. The work in [12] used and compared various pothole detection network architectures based on YOLOv5, YOLOR, and Faster R-CNN. These methods replace various backbone network architectures to create diverse comparisons.

III. Proposed Methodology

Fig. 1 shows the overall proposed network architecture. This network involves three modules: Backbone, Neck, and Detection head.

A. Proposed network architecture

This work thoroughly studies and assesses each component of the original YOLOv8 network architecture [13]. Based on that analysis, the research fine-tunes several blocks in the architecture of Backbone and Neck modules. More specifically, the Cross Stage Partial Bottleneck with two convolutions (C2f), the Spatial Pyramid Pooling Fast (SPPF), and the first Conv are reused. The remaining Conv blocks are replaced by an RFCAConv architecture [14]. The architecture of Conv, C2f, and SPPF are shown in the sub-module part of Fig. 1.

The backbone network consists of a Conv layer, followed by four identical blocks, and ends with the SPPF layer. The four intermediate blocks are built upon several C2f layers (ratio of 3, 3, 6, 6, and 3 times) and an RFCAConv layer. The architecture of the RFCAconv layer is shown in Fig. 2. The RFCAConv layer is inspired by the combination of the Coordinate Attention mechanism and standard convolution (Conv2D). RFCAConv is applied to solve the problem of sharing convolution kernel parameters and improve the feature extraction ability of the standard convolution operation. It uses lightweight convolutions (group convolution) to reduce network parameters and the Coordinate Attention mechanism helps the network capture the salient features. Suppose, $F \in R^{W\times H\times C}$ and $F' \in R^{W\times H\times C}$ are the input and output feature maps, respectively. The operating principle of RFCAConv is described as follows:

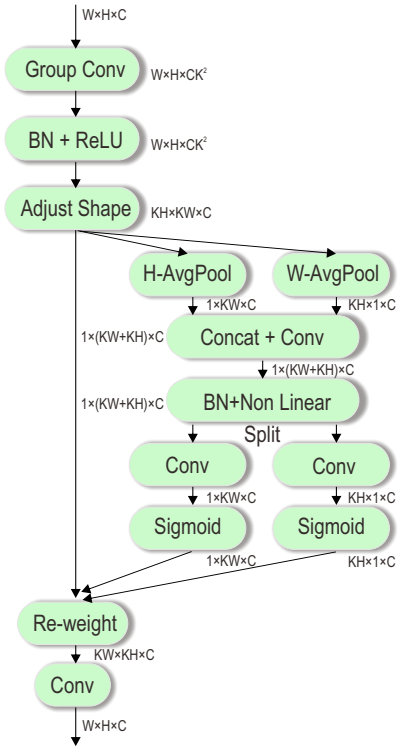

Fig. 2. The RFCAConv architecture.

$$F' = f^{3\times 3}(F_{Main} \times F_H \times F_W), \tag{1}$$

where $f^{i\times i}$ is the $i \times i$ Conv2D ($i = 1$ or $i = 3$). F_{Main} is the output feature map of the main branch. F_H and F_W are splitted from feature map F_{HW} of Coordinate Attention.

$$F_{Main} = Reshape(BN(ReLU(g^{3\times 3}(F)))), \tag{2}$$

$$F_{WH} = \delta([AVG_H(F_{Main}), AVG_W(F_{Main})]), \tag{3}$$

where δ denotes the non-linear activation function. The AVG_H and AVG_W are Average Pooling layers according to the H-axis (kernel size $(H, 1)$) and W-axis (kernel size $(1, W)$), respectively. The operation $[\cdot]$ describes the Concatenation layer.

$$F_H = \sigma(f^{1\times 1}(F'_H)), \tag{4}$$

$$F_W = \sigma(f^{1\times 1}(F'_W)), \tag{5}$$

in which, F'_H and F'_W are separated tensors from F_{HW}.

The Neck module reuses the Path Aggregation Network (PAN) architecture as in the original YOLOv8 network and replaces all Conv layers with the RFCAConv layers. The current feature maps are upsampled and aggregated with previous low-level feature maps from the Backbone module using Concatenation operations. Via this module, the feature maps will be enriched and generated with three scales for

small, medium, and large objects. These feature maps go through the Detection head module to identify the object.

The Detection head module also leverages the three detection heads from the original YOLOv8 with the decouple head and free-anchor technique. The output feature maps from the Neck module transfer to two sibling blocks. Each block consists of a Conv layer and standard convolution layer for bounding box regression (four coordinates of the box: x, y, h, w) and classification (number of classes: c) on three object levels. The Conv block is built based on a 1×1 standard convolution layer (Conv2D), a BN, and a ReLU activation function. Table I shows the details of the Detection head module.

TABLE I
THE DETAILS OF THE DETECTION HEAD MODULE.

Heads	Input	Anchor	Ouput	Object
1	$80 \times 80 \times 256$	Free	$80 \times 80 \times 4/80 \times 80 \times c$	Small
2	$40 \times 40 \times 512$	Free	$40 \times 40 \times 4/40 \times 40 \times c$	Medium
3	$20 \times 20 \times 512$	Free	$20 \times 20 \times 4/20 \times 20 \times c$	Large

B. Loss function

The loss function used in this paper is defined as follows:

$$\mathfrak{L} = \lambda_{Box}\mathfrak{L}_{Box} + \lambda_{DFL}\mathfrak{L}_{DFL} + \lambda_{Cls}\mathfrak{L}_{Cls}, \quad (6)$$

where the bounding box regression loss integrates $\mathfrak{L}_{Box}$ with $\mathfrak{L}_{DFL}$ and they apply the CIoU loss [15] and Distribution Focal Loss (DFL) [16], respectively. The classification loss $\mathfrak{L}_{Cls}$ is computed by the Binary Cross Entropy loss [17]. The λ_{Box}, λ_{Cls}, and λ_{dfl} are balancing parameters.

IV. EXPERIMENTS

A. Dataset

The pothole and manhole cover detection dataset [14] was taken from a specific road with damage in an industrial area of Žilina city, Slovakia with an image resolution of 1920×1080. The dataset includes 1,052 images under clear weather conditions (training set: 736 images, test set: 159 images, and validation set: 157 images) and four subsets under adverse weather conditions such as Rain, Sunset, Evening, and Night. The images are divided into two classes: Potholes and Manhole covers. A detailed description of this dataset is presented in Table II.

TABLE II
THE DETAILS OF EACH SUBSET IN THE POTHOLE AND MANHOLE COVER DETECTION DATASET.

Subset	Images	Instances	Potholes	Manhole covers
Clear	1,052	2,128	1,896	232
Rain	286	458	383	75
Sunset	201	404	364	40
Evening	250	339	286	53
Night	310	262	220	42

B. Experimental setup

Based on the YOLOv8 framework, this work uses the Python programming language and the Pytorch library to modify the architectures. The training and evaluation processes are conducted on a GeForce GTX 1080Ti 11GB GPU. The training phase applies the Stochastic Gradient Descent (SGD) optimization. The initial learning rate is set at 10^{-4} and then increase to 10^{-2}. The momentum is set at 0.937. The training process goes through 200 epochs with a batch size of 16. The balance parameters are set as follows: λ_{Box}=1.5, λ_{Cls}=0.5, and λ_{DFL}=1.5. Several data augmentation methods are used, such as translate, scale, flip, and mosaic. The input image size in the training phase is 640×640. For the inference phase, image sizes are 640×640 and 1080×1080, a batch size of 16, a confidence threshold = 0.5, and an IoU threshold = 0.5. The inference time is reported in milliseconds (ms).

C. Experimental results

The experiments evaluate the proposed network and compare it with the networks trained from scratch (YOLOv5s [18], YOLOv8s [13]), and several other networks in reference [14] with different input image resolutions. The comparison results are shown in Table III. For an input image size of 640×640, the proposed network achieves 74.2% of mAP@0.5 and 34.1% of mAP@0.5:0.95. This result is superior to other competitors except for YOLOv3-4 [14] on mAP@0.5 (0.5%↓) while network parameters and computational complexity are comparable to the YOLOv8s network. The speed of the proposed network is better than other networks and YOLOv5s (0.6 ms↓) and is comparable to YOLOv8s (0.4 ms↑). For an input image size of 1080×1080, the proposed network reaches 77.1% of mAP@0.5 and 36.8% of mAP@0.5:0.95. These results are largely better than other competitive networks except for YOLOv3-SPP [14] (2.0%↓) and YOLOv8s (0.7%↓). However, in terms of speed, the proposed network is lower than YOLOv8s (8.4 ms↑) network and can be compared to YOLOv5s (2.0 ms↑). The above results prove that choosing the input image size increases object detection accuracy but reduces the inference speed quite a lot. Therefore, choosing the appropriate input image size is necessary to build real-time systems on low-computing devices. Fig. 3 shows several qualitative results on the Potholes and Manhole covers dataset with different weather conditions (Clear, Evening, Night, Rain, and Sunset). This work also implements the comparison between the proposed network and YOLOv8s as presented in Fig. 4. The comparison proves that the proposed network is better than the YOLOv8s network architecture when detecting the potholes and manhole covers with long distances and large angles of the camera. The balance in the performance of the proposed network with acceptable computation complexity and network parameters allows the proposed model can be applied in real-time road safety systems. Nevertheless, the proposed model is still affected by several conditions that reduce the detection ability such as dense potholes, overlapping between potholes and manhole covers, distance from the

Fig. 3. The qualitative result on the validation set with various conditions.

objects to the camera, and the camera angles. Especially, the lighting conditions as the results are shown in Table IV.

D. Ablation studies

To evaluate the ability of the proposed module to combine with the components of the original architecture of the YOLOv8 network, this work conducted several ablation studies as shown in Table V. These experiments mainly focus on the feature extraction capabilities of the RFCAConv module with the first Conv, SPP, and SPPF modules. Replacing all Conv modules in the Backbone network reduces accuracy and increases network parameters and computational complexity. Combining the first Conv module and RFCAConv improves accuracy, network parameters, and computational complexity. Finally, the experiment achieved the best results using a combination of all three modules: the first Conv, RFCAConv, and SPPF. This is the selection of proposed network architecture that is trained, evaluated, and reported based on the Pothole and Manhole covers dataset.

V. Conclusion and Future Work

This study improves the YOLOv8s architecture network for pothole and manhole cover detection. The research focuses on rebuilding the backbone and neck modules using a combination of the first CONV, RFCAConv, and SPPF modules. With the balance of network parameters, computational complexity, and inference speed, the proposed network promises to be applied to real-time road safety systems with low-computing

TABLE III
THE COMPARISON RESULT BETWEEN THE PROPOSED NETWORK AND OTHER NETWORKS ON THE CLEAR WEATHER VALIDATION SET.

Models	Image size	Parameter	GFLOPs	Weight(MB)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Inf. time (ms)
YOLOv3-1 [14]	640×640	N/A	N/A	N/A	28.5	9.2	35.0
YOLOv3-2 [14]	640×640	N/A	N/A	N/A	56.3	20.2	35.0
YOLOv3-3 [14]	640×640	N/A	N/A	N/A	68.1	26.8	35.0
YOLOv3-4 [14]	640×640	N/A	N/A	N/A	**74.7**	31.4	35.0
YOLOv3-SPP [14]	640×640	N/A	N/A	N/A	71.1	28.6	36.0
YOLOv5s† [18]	640×640	7,050,367	15.4	14.3	67.6	25.3	2.8
YOLOv8s† [13]	640×640	11,126,358	28.4	22.5	73.7	34.0	**1.8**
Proposed method	640×640	**11,234,502**	**29.1**	**22.8**	**74.2**	**34.1**	**2.2**
YOLOv3 [14]	1080×1080	N/A	N/A	N/A	77.1	33.0	82.0
YOLOv3-SPP [14]	1080×1080	N/A	N/A	N/A	**79.1**	35.4	84.0
YOLOv5s† [18]	1080×1080	7,050,367	15.4	14.3	75.5	31.4	9.4
YOLOv8s† [13]	1080×1080	11,126,358	28.4	22.5	78.1	**37.5**	**3.0**
Proposed method	1080×1080	**11,234,502**	**29.1**	**22.8**	**77.1**	**36.8**	**11.4**

Inf. time (ms): Inference time is evaluated on a GeForce GTX 1080Ti GPU.
†: The models are trained from scratch.
Red color: Best competitor.

TABLE IV
THE EVALUATION RESULTS ON DIFFERENT VALIDATION SETS.

Dataset	mAP@0.5 (%)	mAP@0.5:0.95(%)	Inf. time
Evening	37.7	13.9	18.4
Night	16.3	5.54	20.4
Rain	37.4	13.1	18.9
Sunset	32.5	11.6	21.0

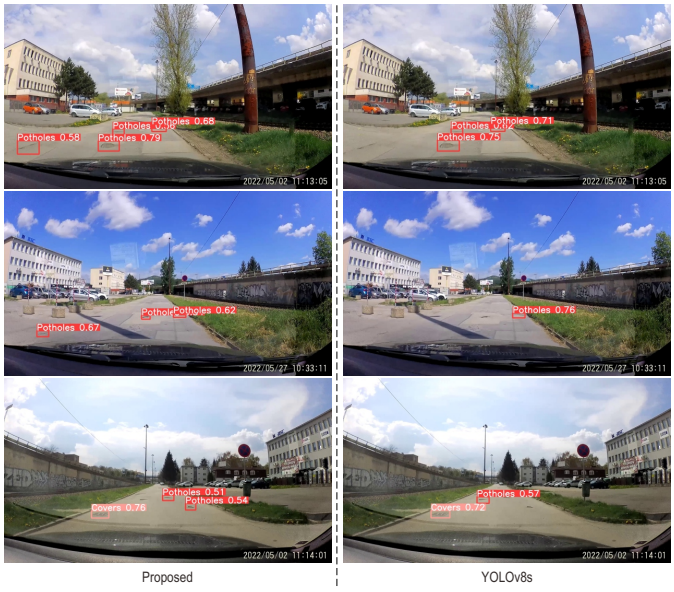

Fig. 4. The comparison result between proposed method and YOLOv8s on validation set.

devices. In the future, the detector will be developed with other novel attention techniques to enhance small-size object detection and compare the performance to the latest YOLOv9.

ACKNOWLEDGEMENT

This result was supported by the "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE)(2021RIS-003).

TABLE V
ABLATION STUDIES WITH DIFFERENT PROPOSED NETWORKS ON THE VALIDATION SET WITH INPUT IMAGE SIZE SET AT 640 PIXELS.

Blocks	Proposed backbones			
First Conv			✓	✓
RFAConv	✓	✓	✓	✓
SPPF	✓			✓
SPP		✓	✓	
Parameter	11,234,901	11,234,901	11,234,502	11,234,502
GFLOPs	29.9	29.9	29.1	29.1
Weight (MB)	29.9	29.9	22.8	22.8
mAP@0.5	67.2	71.0	72.5	**74.2**
mAP@0.5:0.95	33.4	34.1	33.0	**34.1**
Inf. time (ms)	1.1	5.6	5.1	**2.2**

REFERENCES

[1] Y.-M. Kim, Y.-G. Kim, S.-Y. Son, S.-Y. Lim, B.-Y. Choi, and D.-H. Choi, "Review of recent automated pothole-detection methods," *Applied Sciences*, vol. 12, no. 11, 2022.
[2] A. Dhiman and R. Klette, "Pothole detection using computer vision and learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3536–3550, 2019.
[3] D. Wang, Z. Liu, X. Gu, W. Wu, Y. Chen, and L. Wang, "Automatic detection of pothole distress in asphalt pavement using improved convolutional neural networks," *Remote Sensing*, vol. 14, no. 16, 2022.
[4] H. Tahir and E.-S. Jung, "Comparative study on distributed lightweight deep learning models for road pothole detection," *Sensors*, vol. 23, no. 9, 2023.
[5] H.-W. Wang, C.-H. Chen, D.-Y. Cheng, C.-H. Lin, and C.-C. Lo, "A real-time pothole detection approach for intelligent transportation system," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–7, 08 2015.
[6] H. P M and V. Gopi, "Vehicle vibration signal processing for road surface monitoring," *IEEE Sensors Journal*, vol. PP, pp. 1–1, 06 2017.
[7] A. Talha, M. Karasneh, D. Manasreh, A. Oide, and M. Nazzal, "A lidar-camera fusion approach for automated detection and assessment of potholes using an autonomous vehicle platform," *Innovative Infrastructure Solutions*, vol. 8, 09 2023.
[8] Z. Zhang, X. Ai, C. Chan, and N. Dahnoun, "An efficient algorithm for pothole detection using stereo vision," pp. 564–568, 05 2014.
[9] C. Pena-Caballero, D. Kim, A. Gonzalez, O. Castellanos, A. Cantu, and J. Ho, "Real-time road hazard information system," *Infrastructures*, vol. 5, no. 9, 2020.

[10] B. Bucko, E. Lieskovska, K. Zábovská, and M. Zábovský, "Computer vision based pothole detection under challenging conditions," *Sensors*, vol. 22, p. 8878, 11 2022.
[11] S.-S. Park, V.-T. Tran, and D.-E. Lee, "Application of various yolo models for computer vision-based real-time pothole detection," *Applied Sciences*, vol. 11, no. 23, 2021.
[12] K. R. Ahmed, "Smart pothole detection using deep learning based on dilated convolution," *Sensors*, vol. 21, no. 24, 2021.
[13] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023.
[14] X. Zhang, C. Liu, D. Yang, T. Song, Y. Ye, K. Li, and Y. Song, "Rfaconv: Innovating spatial attention and standard convolutional operation," 2023.
[15] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, and W. Zuo, "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," *CoRR*, vol. abs/2005.03572, 2020.
[16] M. S. Hossain, J. M. Betts, and A. P. Paplinski, "Dual focal loss to address class imbalance in semantic segmentation," *Neurocomputing*, vol. 462, pp. 69–87, 2021.
[17] R. Rubinstein and D. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Information Science and Statistics, Springer New York, 2011.
[18] G. Jocher and et al., "ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements," Oct. 2020.