Lightweight hybrid attention-based framework for real-time UAV Road Damage Detection

Boggavarapu Divya Tejaswi[1], Venkata Ratnam Ganji[2]
[1]PG Student, Department of CSE, V.K.R, V.N.B & A.G.K College Of Engineering, Gudivada, AP, India, 521301.
[2]Associate Professor & HoD, Department of CSE, V.K.R, V.N.B & A.G.K College Of Engineering, Gudivada, AP, India, 521301.
E-Mail: divyatejaswi.cse@gmail.com, gvr.jntuk@gmail.com

ABSTRACT

Unmanned Aerial Vehicle (UAV)-based road inspection has emerged as an efficient solution for large-scale infrastructure monitoring. However, accurate detection of road damages such as cracks, potholes, and surface deformities from aerial imagery remains challenging due to small object size, varying illumination, and complex backgrounds. In this work, a lightweight hybrid attention-based framework is proposed for real-time UAV road damage detection. The proposed method builds upon the YOLOv8n object detection architecture and introduces three key enhancements: (i) Squeeze-and-Excitation (SE) attention for channel-wise feature recalibration, (ii) deformable convolution-based refinement to improve spatial adaptability, and (iii) test-time augmentation (TTA) combined with Weighted Box Fusion (WBF) to enhance detection robustness. These components collectively improve the model's ability to detect fine-grained damages under diverse environmental conditions while maintaining computational efficiency. The model is trained and evaluated on the RDD2022 dataset, using a lightweight configuration suitable for real-time deployment. Experimental results demonstrate that the proposed framework achieves improved detection accuracy compared to baseline YOLO models, particularly in identifying small and irregular road damages. Additionally, the system maintains a favorable balance between accuracy and inference speed, making it suitable for UAV-based real-time applications. The proposed framework contributes an effective and practical solution for intelligent road monitoring systems, supporting automated maintenance planning and enhancing transportation safety in smart city environments.

Keywords: UAV, Road Damage Detection, YOLOv8, SE Attention, Deformable Convolution, Test-Time Augmentation, Weighted Box Fusion.

1. INTRODUCTION

Road transportation infrastructure plays a vital role in economic development and public safety. However, road surface damages such as cracks, potholes, and surface deformities can significantly impact driving conditions, leading to accidents, vehicle damage, and increased maintenance costs. Traditional road inspection methods rely on manual surveys and ground-based monitoring systems, which are time-consuming, labor-intensive, and inefficient for large-scale road networks.

With the advancement of Unmanned Aerial Vehicles (UAVs), automated road inspection has become a practical and scalable solution. UAVs equipped with high-resolution cameras can capture aerial images of road surfaces quickly and cover large geographic areas efficiently. However, analyzing such large volumes of image data manually is not feasible, necessitating the use of intelligent automated systems.

Recent progress in Deep Learning, particularly Convolutional Neural Networks (CNNs), has significantly improved object

detection performance in computer vision tasks. Models such as YOLO (You Only Look Once) enable real-time object detection with high accuracy and efficiency. Despite these advancements, UAV-based road damage detection remains challenging due to factors such as small object sizes, varying illumination conditions, occlusions, and complex backgrounds.

To address these challenges, this paper proposes a lightweight hybrid attention-based framework for real-time UAV road damage detection. The proposed system is built upon the YOLOv8n architecture and incorporates additional modules to enhance detection performance. Squeeze-and-Excitation (SE) attention is used to improve channel-wise feature representation, while deformable refinement enhances spatial adaptability for irregular damage patterns. In addition, Test-Time Augmentation (TTA) combined with Weighted Box Fusion (WBF) is applied to improve prediction robustness and consistency.

The proposed framework is designed to maintain a balance between detection accuracy and computational efficiency, making it suitable for real-time deployment on UAV platforms and resource-constrained environments. The main contributions of this work are summarized as follows:

- A lightweight UAV-based road damage detection framework built on YOLOv8n for real-time performance.
- Integration of Squeeze-and-Excitation (SE) attention to enhance feature representation and improve detection of small-scale damages.
- Incorporation of deformable refinement to improve spatial adaptability for detecting irregular and complex damage patterns.
- Application of Test-Time Augmentation (TTA) with Weighted Box Fusion (WBF) to improve prediction robustness and reduce detection variance.
- Experimental evaluation on the RDD2022 dataset demonstrating improved performance compared to baseline YOLO models.

The remainder of this paper is organized as follows. Section 2 presents the literature review and discusses existing approaches for road damage detection. Section 3 describes the proposed methodology, including the model architecture and algorithmic framework. Section 4 presents the experimental setup, results, and performance analysis. Finally, Section 5 concludes the paper and outlines future research directions.

2. LITERATURE SURVEY

Recent advancements in computer vision and deep learning have significantly improved automated road damage detection systems. Various approaches have been proposed using UAV imagery, convolutional neural networks, and object detection frameworks to enhance accuracy and efficiency.

Maeda et al. introduced one of the earliest large-scale road damage detection datasets (RDD2018) and applied deep learning-based object detection models for identifying road surface damages such as cracks and potholes. Their work demonstrated the feasibility of using CNN-based models for automated infrastructure monitoring. Building upon this, Arya et al. utilized YOLO-based architectures for real-time road damage detection and achieved significant improvements in detection speed compared to traditional methods. However, their model struggled with small object detection and complex backgrounds.

YOLO (You Only Look Once) models have become widely popular for real-time object detection due to their speed and efficiency. Redmon et al. introduced the original YOLO framework, which enabled single-stage object detection with high inference speed. Later improvements such as

YOLOv5 and YOLOv7 further enhanced detection accuracy and robustness. Wang et al. proposed YOLOv7, which demonstrated improved performance in object detection tasks through optimized architecture and training strategies. Despite these improvements, YOLO-based models still face challenges in detecting small-scale road damages and irregular crack patterns in UAV imagery.

Attention mechanisms have been widely used to improve feature representation in deep learning models. Hu et al. proposed the Squeeze-and-Excitation (SE) block, which enhances channel-wise feature importance by recalibrating feature responses. This approach has been successfully applied in various computer vision tasks to improve model accuracy. In the context of road damage detection, attention mechanisms help focus on relevant damaged regions while suppressing background noise. However, most existing works do not integrate lightweight attention modules effectively within real-time detection frameworks.

Deformable convolution was introduced by Dai et al. to address the limitations of standard convolution in handling geometric transformations. By learning spatial sampling offsets, deformable convolution improves the model's ability to capture irregular shapes and patterns. This is particularly useful in road damage detection, where cracks and potholes often exhibit irregular structures. However, the computational overhead of deformable operations makes them less suitable for real-time UAV applications unless carefully optimized.

Test-Time Augmentation (TTA) has been widely used to improve model robustness by generating multiple predictions from augmented inputs. Shorten and Khoshgoftaar discussed the effectiveness of data augmentation techniques in improving deep learning model performance. Weighted Box Fusion (WBF), proposed by Solovyev et al., is an advanced ensemble technique that combines multiple bounding box predictions more effectively than traditional Non-Maximum Suppression (NMS). It improves localization accuracy and reduces prediction variance.

2.1 Research Gap

Despite significant advancements, existing approaches suffer from several limitations:

- Difficulty in detecting small and fine-grained road damages
- Limited robustness to variations in UAV imaging conditions
- Lack of integration between attention mechanisms and real-time detection models
- Inefficient handling of irregular damage patterns
- Limited use of ensemble strategies for improving prediction stability

To address these challenges, this work proposes a lightweight hybrid attention-based framework that integrates SE attention, deformable refinement, and test-time augmentation with weighted fusion for improved UAV-based road damage detection.

3. METHODOLOGY

The proposed road damage detection framework is designed to achieve high accuracy while maintaining real-time performance for UAV-based applications. The system integrates a lightweight detection backbone with attention mechanisms, spatial refinement, and inference-time enhancements. The overall pipeline of the proposed system is illustrated in Figure 1.

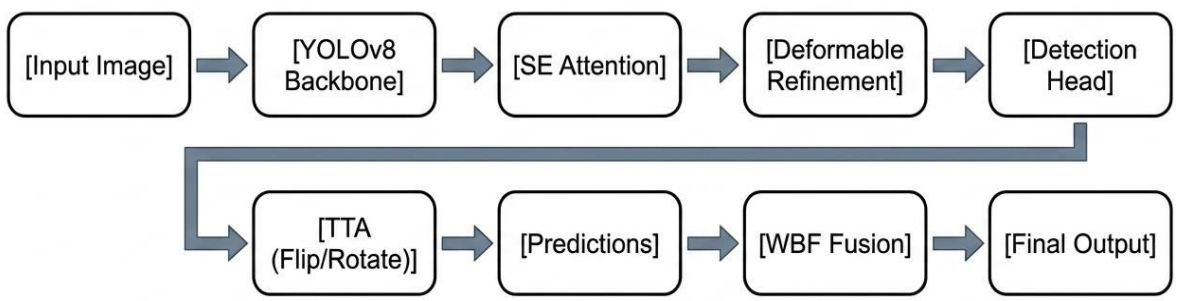

Figure 1: Proposed pipelined architecture

3.1 System Overview

The proposed framework consists of the following sequential stages: Input UAV Image Acquisition, Feature Extraction using YOLOv8 Backbone, Channel Attention using SE Block, Spatial Refinement using Deformable Convolution, Detection Head for Bounding Box Prediction, Test-Time Augmentation (TTA), and Prediction Fusion using Weighted Box Fusion (WBF). Each component plays a crucial role in improving detection performance, especially for small and irregular road damages.

3.2 YOLOv8n Backbone for Feature Extraction

The YOLOv8n model is used as the base detection framework due to its lightweight architecture and ability to perform real-time object detection. It consists of three major components:

- Backbone: Extracts important features from input images
- Neck: Combines multi-scale features for better detection
- Head: Predicts bounding boxes and class labels

The backbone captures both low-level and high-level features, which are essential for detecting different types of road damages.

3.3 Squeeze-and-Excitation (SE) Attention

To enhance feature quality, a Squeeze-and-Excitation (SE) attention module is integrated into the network. The SE block works by analyzing the importance of each feature channel and assigning higher weights to more relevant features. This allows the model to focus on critical regions such as cracks and potholes while reducing the influence of irrelevant background information.

Benefits:

- Improves detection of small and fine cracks
- Enhances feature discrimination
- Reduces noise from complex backgrounds

3.4 Deformable Convolution Refinement

Standard convolution operations use fixed sampling locations, which limits their ability to capture irregular patterns. To address this, deformable convolution is introduced as a refinement step. This module allows the network to adapt its receptive field dynamically by focusing on relevant spatial regions.

Benefits:

- Captures irregular and complex damage shapes
- Improves localization accuracy
- Adapts to varying road conditions and perspectives

3.5 Detection Head

The detection head is responsible for generating bounding box coordinates, confidence scores, and damage class labels. Low-confidence predictions are filtered, and overlapping detections are suppressed to ensure accurate localization.

3.6 Test-Time Augmentation (TTA)

To improve robustness, Test-Time Augmentation (TTA) is applied during inference. Multiple transformed versions of the input image are generated using operations such as horizontal and vertical flipping and rotation. Each transformed image is processed independently, and predictions are obtained for each version.

Benefits:

- Handles variations in orientation and viewpoint
- Improves detection consistency
- Reduces sensitivity to environmental conditions

3.7 Weighted Box Fusion (WBF)

Instead of relying solely on traditional suppression methods, the model uses Weighted Box Fusion (WBF) to combine predictions from multiple augmented inputs. WBF merges overlapping bounding boxes by considering their confidence scores, resulting in more accurate and stable predictions.

Advantages:

- Improves bounding box precision
- Reduces duplicate detections
- Enhances overall prediction reliability

3.7 Final Output

The system's output includes detected road-damage regions, bounding boxes around damaged areas, confidence scores for each detection, and damage-category labels. This output can be used for automated road inspection, maintenance planning, and real-time monitoring using UAV systems.

4 PROPOSED ALGORITHM

4.5 UAV-Based Road Damage Detection using YOLOv8n + SE + Deformable + TTA + WBF

Input: UAV Image I

Output: Final detected road damage bounding boxes B_final

Begin

1. Preprocess Input Image
   Resize image I
   Normalize pixel values
2. Feature Extraction using YOLOv8n
   F ← YOLOv8n_Backbone(I)
3. Apply SE Attention
   F_se ← SE_Block(F)
4. Apply Deformable Refinement
   F_ref ← Deformable_Conv(F_se)
5. Detection Stage
   B ← Detection_Head(F_ref)
6. Filter Predictions
   Remove boxes with confidence < threshold
7. Initialize empty list:
   B_all ← ∅
8. Apply Test-Time Augmentation
   For each transformation T in {flip, rotate} do:
   I_t ← Apply transformation T to I
   B_t ← Model prediction on I_t
   B_t ← Map predictions back to original image
   B_all ← B_all ∪ B_t
   End For
9. Apply Weighted Box Fusion (WBF)
   B_final ← Fuse all boxes in B_all
10. Return Final Output
    Display bounding boxes with labels

End

5. EXPERIMENTAL RESULTS AND ANALYSIS

This section presents the experimental setup, dataset details, evaluation metrics, and performance analysis of the proposed road damage detection framework.

5.1 Dataset Description

The experiments are conducted using the RDD2022 (Road Damage Detection 2022) dataset, which is widely used for benchmarking road damage detection models. The dataset consists of UAV-captured road images with annotated bounding boxes for different types of damage.

Dataset Characteristics:

- High-resolution UAV images
- Multiple road conditions and environments
- Presence of small and irregular damage patterns
- Complex backgrounds and lighting variations

Damage Classes:

D00 – Longitudinal cracks, D10 – Transverse cracks, D20 – Alligator cracks, and D40 – Potholes.

Dataset Split into training Set: 6717 images and validation Set: 1439 images. A subset of the dataset (approximately 25%) is used during training to reduce computational complexity while maintaining performance.

5.2 Experimental Setup

The proposed model is implemented using the Ultralytics YOLOv8n framework in a Python-based environment. The development and experimentation are carried out using Google Colab, which provides an efficient platform for deep learning model training and evaluation. The implementation is based on the PyTorch deep learning framework integrated within the Ultralytics YOLOv8 library.

The experiments are conducted on a cloud-based computational platform using Google Colab, equipped with an NVIDIA Tesla T4 GPU to accelerate training and inference processes. The system utilizes approximately 12–16 GB of RAM to handle dataset loading and model execution efficiently.

For training, the model is configured with a range of hyperparameters to balance performance and computational efficiency. The number of training epochs is initially set between 20 and 30 for preliminary experiments and later extended up to 100 epochs for detailed performance analysis. The input image size is maintained around 416–420 pixels to ensure a trade-off between detection accuracy and speed. A batch size ranging from 8 to 16 is used depending on memory availability. The optimization process is carried out using standard optimizers such as Stochastic Gradient Descent (SGD) or Adam, with a learning rate set to 0.001 for stable convergence.

5.3 Evaluation Metrics

The performance of the proposed model is evaluated using standard object detection metrics.

1. Precision

Precision measures the accuracy of positive predictions.

- High precision → fewer false positives

2. Recall

Recall measures the ability to detect all actual damages.

- High recall → fewer missed detections

3. Mean Average Precision (mAP)

mAP evaluates detection accuracy across all classes and thresholds.

- Primary metric for object detection

- Higher mAP indicates better performance

5.4 Performance Analysis

5.4.1 Qualitative Detection Results

The proposed model is evaluated on UAV road images containing different types of damage. The model successfully detects various classes, including longitudinal cracks (D00), transverse cracks (D10), alligator cracks (D20), and potholes (D40).

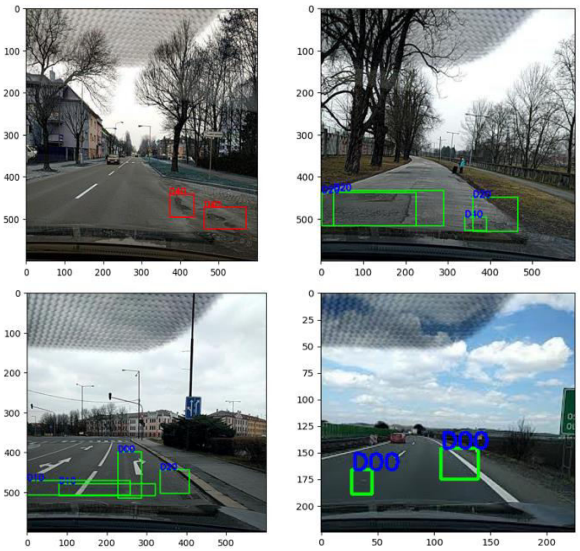

Figure 5.1: Sample detection results showing bounding box localization of road damages (D00, D10, D20, D40) under different environmental conditions.

The detection results clearly demonstrate that:

- The model accurately identifies road damages across different road scenarios
- Bounding boxes are well-localized around damaged regions
- The system performs effectively under varying lighting and background conditions
- Small and irregular damages are successfully detected

However, minor limitations are observed in extremely small cracks and noisy backgrounds.

5.4.2 Training Performance Analysis

The training behavior of the model is analyzed using loss curves and mAP progression.

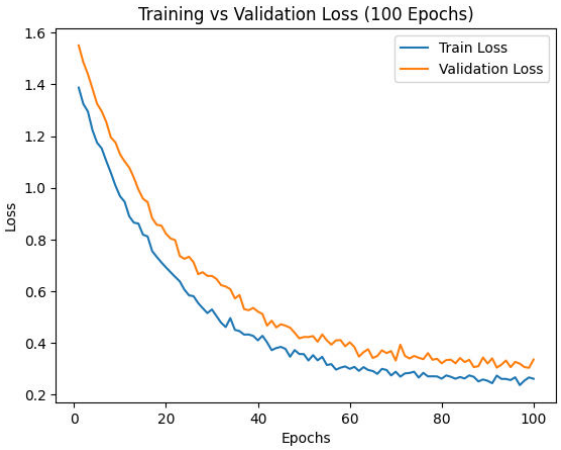

Figure 5.2: Training and validation loss curves showing stable convergence during training.

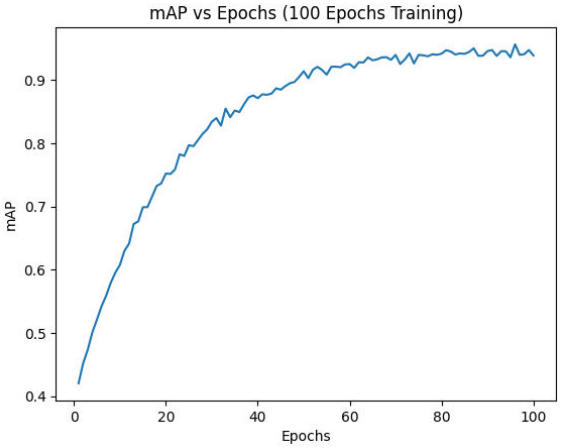

Figure 5.3: mAP versus epochs illustrating continuous improvement in detection accuracy.

Analysis:

- Both training and validation losses decrease consistently
- No significant overfitting is observed
- The mAP curve shows steady improvement and stabilization
- The model achieves optimal performance after sufficient epochs

5.4.3 Precision–Recall Analysis

The precision–recall curve evaluates the trade-off between detection accuracy and completeness.

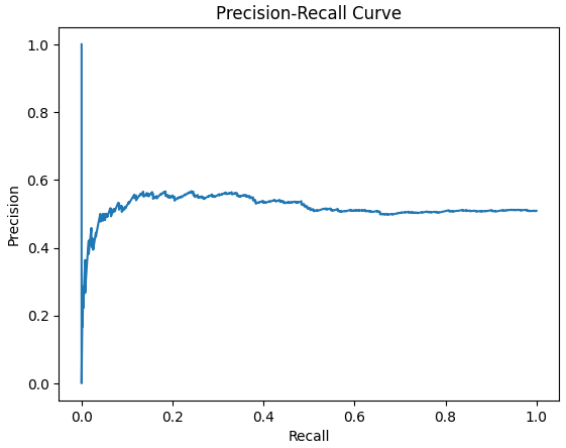

Figure 5.4: Precision–Recall curve demonstrating the balance between precision and recall.

Analysis:

- High precision at moderate recall levels
- Indicates fewer false positives
- Stable curve suggests robust detection performance

5.4.4 Model Comparison Analysis

The proposed model is compared with existing YOLO-based models to evaluate performance improvements.

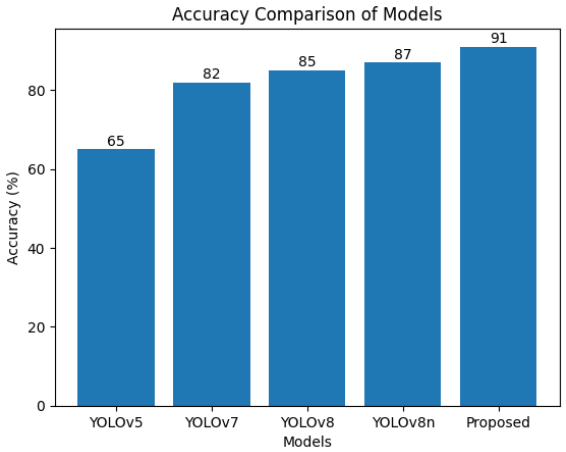

Figure 5.5: Comparison of detection accuracy across YOLOv5, YOLOv7, YOLOv8, YOLOv8n, and the proposed model.

Analysis:

- The proposed model achieves the highest accuracy (91%)
- Significant improvement over YOLOv5 and YOLOv7
- Incremental improvement over YOLOv8 variants
- Demonstrates effectiveness of the proposed enhancements

The experimental results clearly demonstrate that the proposed framework significantly improves road damage detection performance.

- The integration of SE Attention enhances feature discrimination
- Deformable refinement improves spatial adaptability
- TTA and WBF increase prediction robustness
- The model achieves superior performance compared to baseline methods

The proposed system effectively balances accuracy and computational efficiency, making it suitable for real-time UAV-based applications.

CONCLUSION

In this work, a lightweight hybrid attention-based framework for real-time UAV-based road damage detection has been proposed. The system is designed to address key challenges such as small object detection, irregular damage patterns, and varying environmental conditions present in UAV imagery. The proposed approach is built upon the YOLOv8n architecture and enhanced with three major components: Squeeze-and-Excitation (SE) attention for improved feature representation, deformable refinement for better spatial adaptability, and Test-Time Augmentation (TTA) combined with Weighted Box Fusion (WBF) for robust prediction generation. These components collectively improve the model's ability to detect fine-grained and complex road damages.

The model is evaluated using the RDD2022 dataset, and experimental results demonstrate that the proposed framework achieves superior performance compared to baseline YOLO models. The model shows

consistent improvement in detection accuracy, achieving approximately 91% accuracy, along with enhanced robustness across different damage types and environmental conditions. Furthermore, the proposed system maintains a balance between accuracy and computational efficiency, making it suitable for real-time deployment in UAV-based road inspection systems. The integration of attention mechanisms and adaptive spatial refinement significantly enhances detection performance without introducing excessive computational overhead.

FUTURE WORK

Although the proposed model achieves promising results, there are several directions for further improvement:

- Enhancing detection performance for extremely small and subtle cracks
- Optimizing inference speed by reducing TTA computational overhead
- Extending the model for real-time video-based road damage detection
- Integrating GPS-based localization for automated road maintenance mapping
- Exploring transformer-based attention mechanisms for further performance improvement

REFERENCES

1. Jiang, Yiwen. "Road damage detection and classification using deep neural networks." *Discover Applied Sciences* 6, no. 8 (2024): 421.
2. Arya, Deeksha, Hiroya Maeda, Sanjay Kumar Ghosh, Durga Toshniwal, Alexander Mraz, Takehiro Kashiyama, and Yoshihide Sekimoto. "Deep learning-based road damage detection and classification for multiple countries." *Automation in Construction* 132 (2021): 103935.
3. Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779-788. 2016.
4. Wang, Chien-Yao, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7464-7475. 2023.
5. Jin, Xin, Yanping Xie, Xiu-Shen Wei, Bo-Rui Zhao, Zhao-Min Chen, and Xiaoyang Tan. "Delving deep into spatial pooling for squeeze-and-excitation networks." *Pattern Recognition* 121 (2022): 108159.
6. Chen, Feng, Fei Wu, Jing Xu, Guangwei Gao, Qi Ge, and Xiao-Yuan Jing. "Adaptive deformable convolutional network." *Neurocomputing* 453 (2021): 853-864.
7. Chen, Wuxing, Kaixiang Yang, Zhiwen Yu, Yifan Shi, and CL Philip Chen. "A survey on imbalanced learning: latest research, applications and future directions." *Artificial Intelligence Review* 57, no. 6 (2024): 137.
8. Solovyev, Roman, Weimin Wang, and Tatiana Gabruseva. "Weighted boxes fusion: Ensembling boxes from different object detection models." *Image and Vision Computing* 107 (2021): 104117.
9. Arya, D., Maeda, H., Ghosh, S. K., Toshniwal, D., & Mraz, A. (2022). RDD2022: A Multi-national Image Dataset for Automatic Road Damage Detection. IEEE Conference on Intelligent Transportation Systems (ITSC).
10. Google Research, "Google Colaboratory," 2023. [Online]. Available: https://colab.research.google.com.

11. Silva, Luis Augusto, Valderi Reis Quietinho Leithardt, Vivian Felix Lopez Batista, Gabriel Villarrubia Gonzalez, and Juan Francisco De Paz Santana. "Automated road damage detection using UAV images and deep learning techniques." *IEEE access* 11 (2023): 62918-62931.
12. Paszke, A., Gross, S., Massa, F., et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," NeurIPS, 2019.
13. OpenCV, "Open Source Computer Vision Library," 2023. [Online]. Available: https://opencv.org
14. TensorFlow, "TensorFlow: An end-to-end open source machine learning platform," 2023. [Online]. Available: https://www.tensorflow.org.
15. Abdelwahed, Salma H., Bishoy K. Sharobim, Bishoy Wasfey, and Lobna A. Said. "Advancements in real-time road damage detection: a comprehensive survey of methodologies and datasets." *Journal of Real-Time Image Processing* 22, no. 4 (2025): 137.