2026 European Conference on Computing in Construction
Corfu, Greece
July 12-15, 2026

COMPARISON OF SEGMENTATION MODELS FOR POTHOLE DETECTION USING UAV IMAGERY

Varshney Muskan[1], Singh Aryan Raj[1], Garg Sahil[2]
[1,2]Department of Civil Engineering, Indian Institute of Technology Delhi (IIT Delhi), New Delhi, India

Abstract

Potholes are a significant cause of vehicle damage and road safety accidents. Existing inspection practices rely on manual surveys using inspection vehicles or vehicle-mounted cameras, which are time-consuming, non-scalable, and often provide incomplete visibility. This paper presents a comparative study of instance-segmentation-based deep learning models for pothole detection using UAV imagery, enabling wider and more consistent coverage. A custom dataset was developed and augmented using geometric transformations. YOLOv8 and transformer-based RF-DETR models were evaluated using precision, recall, and mAP@50. The best model achieved mAP@50 of 66.3%, highlighting challenges in detecting shallow potholes under limited training data.

Introduction

Potholes are a common form of pavement distress that develops due to repeated traffic loading, environmental exposure, and material fatigue, and they pose long-term risks to road safety, vehicle integrity, and transportation efficiency (Ibragimov et al. 2022). Even shallow potholes can progressively worsen over time, contributing to increased vehicle operating costs and a higher likelihood of traffic accidents when maintenance interventions are delayed (Ali et al. 2023).

Effective monitoring of road surface conditions is therefore essential for sustainable transportation infrastructure management. Conventional inspection practices typically rely on manual surveys conducted using inspection vehicles or visual assessments performed by trained personnel (Anand et al. 2025). Although widely used, such approaches are labor-intensive, subjective, and difficult to scale across extensive road networks. In addition, vehicle-mounted inspection systems often provide limited visibility of the road surface due to restricted viewing angles, occlusions, and perspective distortion, which can hinder reliable detection of subtle or spatially distributed defects (Yu et al. 2024).

Advances in computer vision and deep learning have enabled automated approaches for pavement distress detection using image and video data, offering more objective and repeatable alternatives to manual inspection (Safyari et al. 2024). Convolutional neural networks have demonstrated strong capability in learning discriminative visual features for pothole detection, significantly outperforming traditional image processing methods under diverse imaging conditions (Jakubec et al. 2023). Object detection frameworks such as YOLO have been widely explored due to their computational efficiency and suitability for large-scale deployment, while instance segmentation techniques provide pixel-level delineation of defect boundaries that supports more detailed geometric characterization (Li et al. 2024).

Despite methodological progress, most pothole detection studies continue to rely on ground-based or vehicle-mounted imaging platforms, where consistent surface coverage is difficult to achieve (Chen et al. 2024). Unmanned aerial vehicles (UAVs) offer a complementary sensing perspective by enabling wide-area, top-down observation of road surfaces with reduced occlusion and more uniform imaging geometry. Recent work has demonstrated the effectiveness of UAV imagery for pavement crack detection and surface distress assessment, highlighting its growing relevance for scalable road condition monitoring applications (Ellenberg et al. 2015).

In parallel with advances in sensing platforms, transformer-based vision models have emerged as an alternative to conventional convolutional architectures. By leveraging self-attention mechanisms, these models can capture long-range contextual relationships and improve generalization in visually complex scenes (Carion et al. 2020). Recent developments in lightweight and deformable transformer variants have further improved convergence behavior and performance under limited training data conditions (Zhu et al. 2020). However, the application of transformer-based instance segmentation models to pothole detection using UAV imagery remains insufficiently explored in existing literature. Recent studies have further shown that improved YOLOv8-based architectures are particularly effective for small object detection in UAV aerial imagery, making them suitable for detecting subtle pavement defects from aerial perspectives (Ni et al. 2024).

Addressing this gap requires systematic evaluation of segmentation-based approaches that can accurately localize potholes while accounting for the challenges associated with aerial imagery, subtle surface defects,

and limited annotated data. Such analysis is necessary to better understand model behavior and to support the development of reliable, scalable road surface monitoring systems.

Dataset and experimental setup

Data collection and annotation

A total of 196 road surface images were collected using a quad-copter unmanned aerial vehicle (UAV) equipped with an RGB camera. The UAV was operated to capture views of pavement surfaces, enabling wide-area coverage and consistent imaging geometry. The data acquisition strategy was designed to capture realistic road surface conditions under varying illumination levels, surface textures, and environmental influences. As a result, the dataset includes visual challenges such as shadows, debris, worn asphalt regions, uneven lighting, and surface irregularities that commonly affect road condition assessment.

All images were manually annotated using polygon-level masks to precisely delineate pothole regions. The annotation process was performed using the Roboflow Annotate tool, enabling accurate boundary-level labeling required for instance segmentation tasks. Shallow potholes with weak visual contrast posed particular challenges during annotation due to gradual depth transitions, shadows, and surface stains, leading to unavoidable annotation uncertainty. Despite these challenges, polygon-based annotations were preferred over bounding boxes, as they provide a more precise spatial representation of pothole boundaries and better support pixel-level segmentation, geometric analysis, and potential severity estimation in future extensions of this work.

Data augmentation

To address dataset scarcity and improve model generalization, the original dataset was expanded from 196 to 475 images using targeted geometric data augmentation techniques. The augmentation strategy focused on preserving the structural characteristics of potholes while introducing controlled variability in spatial orientation. The applied augmentations included horizontal flipping, small-angle rotations within a range of ±10°, and shear transformations limited to ±10°.

To prevent unrealistic geometric distortions and maintain annotation fidelity, no more than two augmentation operations were applied to any single image during training. This constraint ensured that the augmented samples remained representative of real-world aerial road scenes while avoiding excessive deformation of pothole boundaries.

The selected augmentation techniques were intended to enhance robustness to variations in UAV viewpoint, flight orientation, and minor perspective changes commonly encountered during aerial image acquisition. By increasing the effective diversity of the training data, the augmentation process contributed to improved detection and segmentation performance under diverse visual and geometric conditions.

Data augmentation was performed exclusively on the training set after dataset splitting, ensuring that the validation and test sets remained unaltered. This approach prevents data leakage and supports an unbiased evaluation of model performance

Models and training strategy

This study evaluates seven instance segmentation model configurations, comprising convolutional neural network–based and transformer-based architectures, to examine the influence of data augmentation, weight initialization, and model design on pothole detection from aerial imagery. All experiments were conducted under a consistent training environment to ensure comparability across configurations.

Four convolutional neural network–based configurations were developed using the YOLOv8 instance segmentation framework. The first configuration corresponds to a baseline YOLOv8 model trained without any data augmentation. A second YOLOv8 configuration incorporated additional training samples while maintaining the same augmentation-free setting. The third YOLOv8 configuration applied horizontal flipping as a single geometric augmentation to increase viewpoint invariance. The fourth YOLOv8 configuration employed a combination of horizontal flipping with mild shear and rotation transformations, enabling greater robustness to camera orientation and road alignment variations.

Three transformer-based configurations were implemented using the RF-DETR instance segmentation framework. All RF-DETR models were initialized using large-scale pretrained weights. The first RF-DETR configuration applied horizontal flipping combined with small-angle rotation. The second RF-DETR configuration used horizontal flipping together with shear and rotation transformations. The third RF-DETR configuration employed horizontal flipping with normal rotation only. These augmentation strategies were selected to balance increased data diversity with preservation of realistic road geometry.

Across all configurations, no more than two geometric augmentations were applied to any individual image during training in order to avoid unrealistic distortions of pothole boundaries. This constraint ensured that augmented samples remained representative of real-world aerial road scenes.

All models were trained within the Roboflow cloud-based training environment using available optimization settings and GPU acceleration. Maintaining a unified training platform and consistent preprocessing pipeline ensured that observed performance differences could be attributed to model architecture and augmentation strategy rather than implementation variability.

This structured experimental setup enables a controlled comparison between convolutional and transformer-based instance segmentation models, as well as an

assessment of the role of specific geometric augmentations in aerial pothole detection. Quantitative performance outcomes are presented separately in the Results section.

Training setup

The final dataset comprised 475 images, including both original and augmented samples. The original dataset was partitioned into training, validation, and test subsets using an approximate 70:20:10 split to ensure reliable performance evaluation and to mitigate overfitting. The training subset was used for model learning, while the validation and test sets were reserved for hyperparameter tuning and final performance assessment, respectively.

Two input resolutions, 640 × 640 and 432 × 432, were systematically explored during the training process. Both YOLOv8 and RF-DETR models were trained using each of these resolutions. It was observed that YOLOv8 converged most effectively at an input resolution of 640 × 640, whereas RF-DETR achieved the most stable and optimal convergence at 432 × 432. Consequently, these resolutions were adopted for the final experiments (640×640 for YOLOv8, 432×432 for RF-DETR). Model weights were initialized using pretrained checkpoints instead of random weights, enabling effective transfer learning and faster convergence, particularly under limited data availability.

Table 1: Summary of training configurations and hyperparameters

Model Name	Input Resolution	Augmentation Strategy	Weight Initialization
Road Hazards 2	640 × 640	None	Public Checkpoints (MS COCO)
Road Hazards 3	640 × 640	Horizontal Flip	Road Hazards 2
Road Hazards 6	640 × 640	Horizontal Flip	Public Checkpoints (MS COCO)
Road Hazards 7	640 × 640	Horizontal Flip + Shear Rotation ±10° horizontal, vertical	Road Hazards 6
Road Hazards 10	432 × 432	Horizontal Flip + Rotation ±10°	Objects365 Pretrained
Road Hazards 12	432 × 432	Horizontal Flip+ Shear Rotation ±10° horizontal, vertical	Objects365 Pretrained
Road Hazards 13	432 × 432	Horizontal Flip+ Rotation ±10°	Objects365 Pretrained

In addition to the configurations summarized in Table 1, RF-DETR models were trained using a batch size of 16, 100 epochs, a learning rate of 1×10^{-4}, and the AdamW optimizer. For YOLOv8, training was performed using the available hyperparameter settings provided by the Roboflow platform. These settings were not manually modified and were kept consistent across experiments to ensure fair comparison.

Evaluation metrics and loss functions

Model performance was evaluated using standard instance segmentation metrics commonly adopted in object detection and segmentation tasks. Mean Average Precision at an Intersection over Union threshold of 0.50 (mAP@50) was used as the primary evaluation metric, as it provides a balanced measure of detection accuracy under moderate localization constraints. In addition, precision and recall were reported to analyze false-positive and false-negative behavior, which is particularly important in safety-critical applications such as road condition monitoring.

During training, multiple loss components were continuously monitored to assess learning dynamics and segmentation quality. This included a Dice loss to evaluate the quality and overlap of predicted segmentation masks, classification loss to measure label prediction accuracy, and bounding box localization loss to quantify spatial regression performance. The evolution of both loss values and evaluation metrics across training epochs was analyzed to assess convergence behavior, detect potential overfitting, and guide model selection.

Results and discussion

The quantitative performance of all trained models is summarized in Table 1.

Table 2: Performance comparison of pothole detection models

Model Name	Type	mAP@50	Precision	Recall
Road Hazards 2	YOLOv8 (Baseline)	45.9%	74.1%	40.0%
Road Hazards 3	YOLOv8 (Extra data)	48.0%	56.5%	45.5%
Road Hazards 6	YOLOv8 (Augmented)	59.6%	74.8%	54.5%
Road Hazards 7	YOLOv8 (Double Aug)	56.0%	59.9%	60.5%
Road Hazards 10	RF-DETR (Transformer)	66.3%	87.5%	55.0%
Road Hazards 12	RF-DETR (Transformer)	65.9%	75.0%	59.0%
Road Hazards 13	RF-DETR (Transformer)	65.6%	78.0%	60.0%

The experimental results demonstrate the progressive improvement in pothole detection and segmentation performance achieved through model architecture

selection and data augmentation strategies. Comparative evaluation across convolutional and transformer-based models highlights both the benefits and limitations of current instance segmentation approaches when applied to aerial road surface imagery.

The initial YOLOv8 baseline models exhibited limited generalization capability. Early configurations achieved mAP@50 values in the range of approximately 46–48%, indicating difficulty in distinguishing shallow potholes from visually similar background artifacts such as shadows, stains, and surface irregularities. Precision was moderate, while recall remained comparatively low, reflecting a tendency to miss subtle, low-contrast pothole instances. These observations suggest that purely convolutional architectures struggle to capture sufficient contextual information when pothole boundaries are weak or ambiguous.

With the introduction of targeted data augmentation and incremental dataset expansion, YOLOv8 performance improved noticeably. Augmented variants achieved mAP@50 values in the range of approximately 56–60%, accompanied by an increase in recall to around 60%. This improvement highlights the importance of data diversity in enhancing robustness to variations in viewpoint, surface texture, and local appearance. Nevertheless, despite these gains, the convolutional models continued to face challenges in accurately delineating pothole boundaries under low-contrast conditions, indicating an architectural limitation rather than a data-only constraint.

The transformer-based RF-DETR models delivered the strongest overall performance across all evaluated metrics. The best-performing configuration achieved a mAP@50 of 66.3%, with precision reaching up to 87.5% and recall stabilizing in the range of 55–60%. Other RF-DETR variants produced comparable results, with mAP@50 values consistently above 65%. The transition from YOLOv8 to RF-DETR resulted in the largest single performance improvement, underscoring the effectiveness of global self-attention mechanisms and strong pretrained representations for capturing long-range contextual relationships in aerial imagery.

Figure 1 illustrates the training dynamics and loss behavior of the best-performing RF-DETR model. The mAP@50 curve stabilizes within the range of approximately 0.63–0.67 during the first 8–10 epochs, indicating rapid convergence driven by effective transfer learning. Minor oscillations observed after convergence suggest sensitivity to limited training data rather than training instability or overfitting. Dice loss remains relatively high and fluctuating, reflecting the inherent difficulty of producing consistent segmentation masks for shallow potholes with weak or gradual visual boundaries. Classification loss and bounding box localization loss remain generally stable, with occasional spikes attributable to small object sizes and annotation uncertainty.

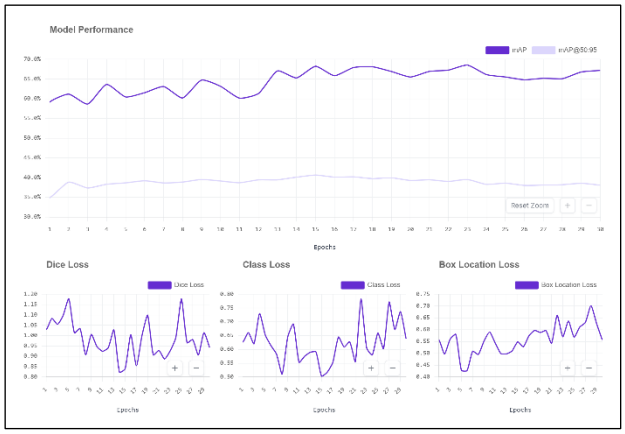

Figure 1: Performance and loss metrics of model Road Hazards 13

Qualitative evaluation further supports the quantitative findings. As illustrated in Figure 2, both YOLOv8 and RF-DETR models successfully detect potholes that are well illuminated and exhibit clear texture contrast relative to the surrounding road surface. Correct detections with high confidence scores demonstrate the capability of the models under favorable visual conditions. However, false positives are commonly triggered by dark stains, shadows, and worn surface patches that visually resemble pothole depressions. Missed detections primarily occur for extremely shallow potholes characterized by gradual depth transitions and weak contrast, particularly near road edges where surface wear blends into surrounding textures.

Figure 2: Single image showing all 3 cases: 1. Correct predictions (88%, 74%), 2. False positives (48%), and 3. False Negatives (below 88% mask)

Compared to YOLOv8, RF-DETR models exhibit a noticeable reduction in shadow-induced false positives and generate more spatially coherent segmentation masks, even in visually cluttered scenes. This behavior confirms the advantage of transformer-based architectures in leveraging broader contextual information, enabling more effective differentiation between true pothole regions and visually similar non-pothole artifacts.

Overall, the results indicate that while transformer-based instance segmentation models provide clear performance advantages over convolutional baselines, detection accuracy remains constrained by dataset scale, annotation ambiguity, and the subtle visual nature of shallow potholes. These findings emphasize that further improvements will likely require larger and more diverse

datasets, enhanced annotation strategies, and the incorporation of complementary geometric or contextual cues beyond RGB imagery alone.

Limitations

Despite the encouraging results obtained in this study, several limitations must be acknowledged. First, the dataset size remains relatively small, consisting of 196 original images, which restricts the diversity of visual scenarios available during training. Although geometric augmentation was employed to partially alleviate this constraint, the limited number of unique real-world samples may still impact generalization to unseen road conditions.

Second, all annotations were produced by a single annotator, introducing unavoidable subjectivity in boundary delineation, particularly for shallow potholes with weak visual contrast. Ambiguous transitions between damaged and undamaged road surfaces make precise annotation challenging and can lead to inconsistencies that affect pixel-level evaluation metrics.

Third, the applied data augmentation strategy was limited to basic geometric transformations. More complex augmentations that simulate real-world imaging conditions, such as motion blur, glare, weather effects, and illumination variation, were not incorporated. The absence of such augmentations may reduce robustness under highly variable environmental conditions commonly encountered in long-term field deployments.

Conclusion

This work presented the development of a custom instance-segmentation dataset for pothole detection and a systematic evaluation of deep learning models trained under realistic data constraints using aerial imagery. A total of 196 original road surface images were manually annotated and expanded through targeted geometric augmentation. Multiple model configurations were trained and evaluated within a unified cloud-based training environment to ensure consistent experimentation and reproducibility.

Baseline convolutional models based on YOLOv8 achieved mAP@50 values in the range of approximately 45–60% as dataset size and augmentation strategies were progressively improved. While these results demonstrate the feasibility of CNN-based approaches for pothole detection, they also revealed limitations in generalization, particularly when detecting shallow potholes with weak visual contrast and ambiguous boundaries. Transitioning to the transformer-based RF-DETR architecture resulted in the strongest overall performance, achieving mAP@50 values of approximately 65–66% and peak precision of up to 87.5%. These improvements confirm the advantage of strong pretraining and global attention mechanisms when operating under limited data availability and visually complex backgrounds.

The primary performance bottlenecks identified in this study were the limited scale of the annotated dataset and unavoidable annotation uncertainty associated with shallow, low-contrast pavement defects. Despite these constraints, the trained models exhibited stable convergence behavior and reliable qualitative performance, particularly in reducing shadow-induced false positives that commonly affect appearance-based detection methods. Such robustness is critical for practical deployment, where minimizing false alarms directly impacts the efficiency of inspection and maintenance workflows.

Overall, this study establishes a viable prototype for aerial pothole detection using instance segmentation and demonstrates the effectiveness of transformer-based architectures for challenging real-world road surface monitoring scenarios. Beyond detection, the generated segmentation masks provide a strong technical foundation for future extensions such as area-based pothole severity estimation, temporal damage progression analysis, and maintenance prioritization. With further dataset expansion, improved annotation strategies, and the integration of complementary geometric or sensing cues, the proposed approach has strong potential to support scalable and data-driven road infrastructure monitoring systems.

Future work

Future work will focus on improving robustness and scalability of the proposed pothole detection system. Expanding the dataset to include more diverse imaging conditions and adopting multi-annotator consensus labeling are expected to enhance segmentation reliability for shallow and ambiguous defects.

Beyond detection, integrating segmentation masks with geometric mapping and depth estimation techniques will enable physically meaningful severity metrics, supporting maintenance prioritization. Model optimization for efficient edge deployment and further improvements through transformer scaling and semi-supervised learning remain promising directions.

Acknowledgement

This project is funded by SPARC.

References.

Ali, F., Khan, Z.H., Khattak, K.S. and Gulliver, T.A., 2023. Evaluating the effect of road surface potholes using a microscopic traffic model. Applied Sciences, 13(15), p.8677.

Anand, R., Priyanka, S. and Goud, P.G., 2025. Assessment of Road Surface Conditions Using Machine Learning. Journal of Transportation Engineering and Its Applications ISSN: 3107-4499 (Online), 10(1).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S., 2020, August. End-to-end object detection with transformers. In European conference on computer vision (pp. 213–229). Cham: Springer International Publishing.

Chen, S., Laefer, D.F., Zeng, X., Truong-Hong, L. and Mangina, E., 2024. Volumetric pothole detection from UAV-based imagery. Journal of Surveying Engineering, 150(2), p.05024001.

Ellenberg, A., Branco, L., Krick, A., Bartoli, I. and Kontsos, A., 2015. Use of unmanned aerial vehicle for quantitative infrastructure evaluation. Journal of Infrastructure Systems, 21(3), p.04014054.

Ibragimov, E., Lee, H.J., Lee, J.J. and Kim, N., 2022. Automated pavement distress detection using region based convolutional neural networks. International Journal of Pavement Engineering, 23(6), pp.1981–1992.

Jakubec, M., Lieskovská, E., Bučko, B. and Zábovská, K., 2023. Comparison of CNN-based models for pothole detection in real-world adverse conditions: Overview and evaluation. Applied Sciences, 13(9), p.5810.

Li, Y., Yin, C., Lei, Y., Zhang, J. and Yan, Y., 2024. Rdd-yolo: Road damage detection algorithm based on improved you only look once version 8. Applied Sciences, 14(8), p.3360.

Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T. and Omata, H., 2018. Road damage detection and classification using deep neural networks with smartphone images. Computer-Aided Civil and Infrastructure Engineering, 33(12), pp.1127–1141.

Ni, J., Zhu, S., Tang, G., Ke, C. and Wang, T., 2024. A small-object detection model based on improved YOLOv8s for UAV image scenarios. Remote Sensing, 16(13), p.2465.

Safyari, Y., Mahdianpari, M. and Shiri, H., 2024. A review of vision-based pothole detection methods using computer vision and machine learning. Sensors, 24(17), p.5652.

Yu, J., Jiang, J., Fichera, S., Paoletti, P., Layzell, L., Mehta, D. and Luo, S., 2024. Road surface defect detection—from image-based to non-image-based: a survey. IEEE Transactions on Intelligent Transportation Systems, 25(9), pp.10581–10603.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X. and Dai, J., 2020. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. International Conference on Learning Representations (ICLR)