Robust Pothole Detection through dual-pass RGB and Depth Fusion

Percy Lam[1]
phl25@cam.ac.uk
Weiwei Chen[1,2]
weiwei.chen@ucl.ac.uk
Lavindra de Silva[1]
lpd25@cam.ac.uk
Ioannis Brilakis[1]
ib340@cam.ac.uk

[1] Department of Engineering
Civil Engineering Building
University of Cambridge
7a JJ Thomson Avenue
Cambridge, UK

[2] Bartlett School of Sustainable Construction
University College London
1-19 Torrington Place
London, UK

Abstract

Having potholes detected and repaired is essential for road maintenance and uninterrupted transport. Generations of researchers and engineers have been innovating on using computer vision and sensors to detect potholes, however, gaps in knowledge remain in the lack of robustness of detectors trained with RGB images in varied environments and the inability to harness depth in detecting potholes with monocular images. This research proposes a solution by fusing predictions made by RGB-trained detectors and monocular depth estimation. The solution first orthorectifies the pavement from perspective RGB images. It then receives predictions from RGB-trained detectors and a depth detector enabled by DINOv2 independently. These predictions are subsequently fused by weighted bounding box fusion and have masks predicted by Segment Anything. Whereas RGB-trained detectors perform well in test sets within the training context, they show a drastic loss of performance in out-of-context situations, such as in images taken on different roads in more challenging environmental conditions. Fusing predictions with depth enhances F1 scores by 16% to 81% in out-of-context situations, reinforcing detection robustness. This solution paves the way for further research on expanding detection to motorways and overcoming shadows in images.

1 Introduction

Well-maintained roads are crucial for providing reliable connectivity. Road conditions have traditionally been assessed by engineering judgement, while since the 1970s automation began with computerised analytical methods on measurements and priority setting [29]. The current practice commonly measures road defects with automatic scanning and vision processing, such as SCANNER and TRACS in the UK. They produce features such as transverse profile (rut depth), longitudinal profile (bumpiness), cracks and surface textures [6] for calculating the Road Condition Index and assigning maintenance priorities. The current survey techniques however cannot identify potholes as a specific defect or record their deterioration

[5]. These vehicles are currently deployed only on local and trunk roads and do not cover unclassified roads.

The strengths and the room for improvement of the current practice highlight the need for robust pothole detection. When using computer vision technique on RGB images, potholes are **detected** when the solution locates and produces bounding boxes (bbox) and segmentation masks within the bbox to indicate where a pothole is in an image, and this process needs to be sufficiently **robust** to be reliable in images taken in variable scenes and conditions. This research explores the possibility of **fusion** of **dual-pass** results, where each image is read twice at inference by two different detectors and their results are combined to produce the final predictions. Each image is read by a deep learning object detector trained on annotations made on RGB images (RGB-trained models), and once by a detector utilising monocular depth estimation designed in this research, to be further discussed in later sections. This research concretely makes the following contributions:

- Improve the generalisation of traditional RGB-trained models across different road scenes and environmental conditions by incorporating monocular depth estimation. This is achieved with monocular camera images, without special data modalities such as stereo images or LiDAR.
- Provides extra interpretability of the pothole detection process by estimating depths from an image and extracting regions of greater depths in an image.
- Performs experiments to evaluate the vulnerability of out-of-context detection by using images of varied geolocation and environmental conditions

2 Related Work

Pothole detection has been thoroughly investigated by researchers using different techniques and data modalities. Fusion of different modalities and detection models has been widely investigated in more general domains. Relevant previous work is synthesised as follows.

2.1 Pothole Detection

Deep learning was the most prevalent technique in recent research. A common setup involved collecting images with cameras or smartphones, annotating the images, and training object detection models [2, 21]. Smartphone footage may also be used to locate or validate potholes measured with sensor data [23, 35]. Datasets were available publicly on Kaggle and Roboflow and offered for benchmarking in Road Defect Detection Dataset [18] for experiments and prototyping [11, 14, 33]. Deep learning could alternatively be used to segment point clouds [32] or images to help reconstruct and make measurements on point clouds [25, 37].

Another common stream of research utilised depth differences in detecting potholes. Some research commenced with self-collected LiDAR [8, 22], while others began on stereo image pairs [2, 25] or drone images [4] to reconstruct point clouds. The 3D point clouds would then search for the pavement by line or plane fitting [4, 7] and segment points below the surface by thresholding. The 3D point clouds also enable contour plotting, pothole size estimation and mesh reconstructions [37].

Other research relied on sensor data. They typically measured vibrations and movements with accelerometers and gyroscopes and detected potholes with machine learning techniques such as Artificial Neural Network [24] or Long Short-Term Memory [35]. The moments where the vehicle experienced potholes can be correlated with the GPS trajectory and/or video footage to obtain the pothole locations.

2.2 Fusion

Fusion is commonly employed to maximise positive detection of a dataset and minimise false detections. This is motivated by having limited labelled data, needing to detect objects with difficult physical presence (such as occlusion, small in size, relationship between objects) in complex backgrounds and detecting objects among highly varied images. Sometimes objects require different detection models or even detection methods to be captured. Out-of-context degradation [9, 26] and constraints in computing resources [9, 27] also contribute to the need to employ multiple detection and fusion for the final prediction results.

Fusion techniques can be divided by the stage ensembling occurs. The earliest stage ensembling occurs is at the input data, by creating multiple candidate regions [9] or augmented images [34] for the downstream classifier. Some ensembling takes place at the feature level, where the solution combines features obtained from different models (or modalities) to influence downstream detection. Earlier research combined more primitive features such as contrast and histogram [26], while more recent research may aggregate or cluster features [27, 31] and saliency maps [17], or employ self and cross attention to tokens [19].

Ensembling can alternatively happen towards the final prediction. Some researchers opted for fusion at the detection head before finalising the predictions, such as combining multiple prediction verdicts on shared region proposals [1], region proposals in a teacher-student network [10] and features at the instance and pixel level [3]. The majority of the literature, however, ensembled predicted instances at the output. This removed the need to interfere with ithe internal structures of detection models and provided flexibility to the models used. Typical ensembling techniques included voting [13, 30], non-maximum suppression [27, 36] and weighted bbox fusion [1, 28].

The current landscape of pothole detection reveals several gaps in knowledge. When deep learning is used, the current practice overwhelmingly relies on training some images and testing other images in the same dataset. The tolerance of road scene variation relies on the variability of images in the dataset, which is commonly limited to hundreds or a few thousand images. In road scenes at different locations or real-world environmental conditions such as fallen leaves and standing water after rain, RGB-trained detectors have not been shown to maintain consistent performance, and may in fact perform significantly worse. This poor generalisation inhibits more widespread adoption in the industry.

On the patterns adopted to detect potholes from the surrounding pavement, the current research landscape requires point clouds collected on-site or reconstructed from stereo cameras in order to utilise depth for pothole detection. The otherwise indiscriminate use of supervised training with RGB image annotation negates the fact that potholes are fundamentally potholes because they form depressions in a continuous surface of road pavement, not because they possess a rough texture that an object detector can pick up. The demands for generalisation and understanding depths inspire this research.

3 Methodology

This research project assumes the use of monocular RGB images as they are vastly prevalent in road vehicles. The wide availability of RGB images helps data collection and facilitates wider adoption in infrastructure maintenance. The proposed solution uses RGB images to detect potholes with RGB patterns and depth.

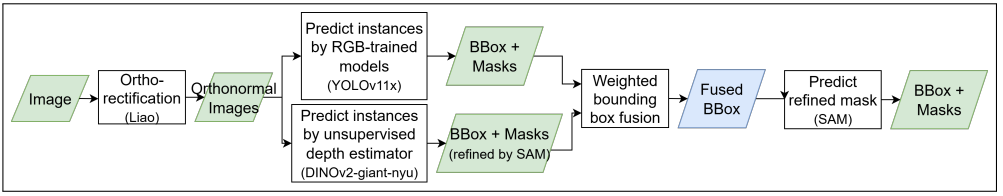

Figure 1: The overall solution to detect potholes from RGB images with two passes

3.1 The Overall Solution

The overall solution is depicted in Figure 1. In preparation, the solution first rotates images to an orthonormal projection to correct for perspective distortion by inverse perspective mapping [16]. Object detectors will be trained with annotated RGB orthorectified images at the instance segmentation level by YOLOv11x-seg [12]. The pothole detector via depth (depth detector) utilises monocular depth estimation powered by DINOv2 with a ViT-g/14 backbone fine-tuned on the NYU-depth dataset [20]. The training setup of the RGB-trained models and the design of the whole depth detector will be explained in detail.

In inference, the first pass on the orthorectified testing images predicts instances from the RGB-trained model. The second pass through the pre-trained depth estimator returns depth maps with pixel-wise estimated depth. The depth detector then corrects the depth maps with a fitted plane and extracts features of varying size with a feature pyramid in 3 scales. The bboxes that surround the features are fed into Segment Anything [15] to refine masks and bboxes that represent potholes detected by estimated depths. The predicted bboxes from the RGB-trained model and the depth detector are fused by weighted bounding box fusion (WBF) [28] and are processed by Segment Anything to generate the final masks and bboxes.

3.2 The RGB-trained Models

Deep learning object detectors are trained with self-collected and prepared monocular RGB images to make inferences. The experiment first trains weights with the two training datasets, West Road (WR) and Tennis Court Road (TCR) in Cambridge, collected with a smartphone or an action camera mounted on a bicycle. The recorded videos were sliced into individual image frames and rotated into orthonormal projection for annotation, training and testing. The exact data split and extra out-of-context testing sets are further illustrated in Section 3.5.

The two training datasets were then trained on YOLOv11x-seg. This was the best RGB-image-based model that provided results at an instance segmentation level and could be trained on a desktop computer with Nvidia 3080Ti GPU with 12GB VRAM. Each dataset would be trained twice, with the best checkpoints of the two models chosen to represent the road section. Each model (referred to as RGB-trained model) was trained for a maximum of 200 epochs.

3.3 The Depth Detector

The depth detector comprises a monocular depth estimator, depth correction, feature extraction and prediction refinement as shown in Figure 2. Depths of the monocular RGB images are first estimated by DINOv2 with a ViT-g/14 backbone, finetuned on NYU-depth dataset. Unlike other monocular depth estimation models that are distilled for perspective views of common scenes, this original model is trained in an unsupervised manner. The model is thus learned to extract depth relationships between pixels even in unfamiliar views, such as orthonormal projections. Orthorectification is necessary because the depth detector needs

to differentiate potholes from the surrounding pavement surface, which has a much smaller depth difference than the near and far views in images taken from a perspective.

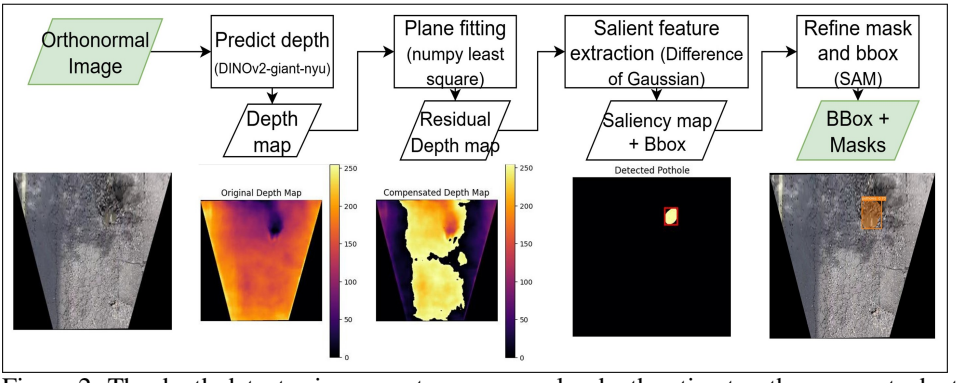

Figure 2: The depth detector incorporates a monocular depth estimator, then corrects depth, extracts features and refines depth regions from the orthonormal image

Bearing the need to eliminate perspective effects in the image, the detector proceeds to remove lingering global perspective effects or inherent tilts on the pavement surface. The algorithm normalises the depth matrix outputted from DINOv2 to the interval of -1 and 1, and fits the matrix to a plane by minimum square distance with a numpy package. The residual depth is calculated by the orthogonal distance to the plane. Only depth values below the plane (indicating a depression) are retained in the residual depth map.

The residual depth map is then fed into a feature pyramid network with 3 scales, achieved by downsizing the image by half in each iteration. Points of interest at each scale are extracted by a difference of Gaussians (of $\sigma = 2$ *and* $2\sqrt{2}$). The points exceeding a preset threshold intensity (empirically tested to be 0.6), indicating significant depressions from the surface, will be accepted as regions of interest and be drawn with preliminary bboxes and masks. The preliminary bboxes are offset by 20 pixels and used to prompt Segment Anything, creating refined masks that include the pothole's periphery.

3.4 Fusion of predicted instances

Fusion takes place at the prediction outputs. The predicted bboxes from the RGB-trained model and depth detector are fed into the weighted bbox fusion tool [28] to receive fused bboxes. The weights are the ratios of the F1 score of potholes achieved by the RGB-trained model and the depth estimator on the particular testing image dataset. This design enables a more flexible choice of detectors when better models emerge in the future and a more modular design for easier maintenance. The fused bboxes are fed into Segment Anything to generate instance masks and refined bboxes as the final combined outcomes.

3.5 Experimental Setup

As described in Section 3.2, the authors collected RGB videos and prepared them into orthonormal images in two roads in Cambridge, namely West Road (WR) and Tennis Court Road (TCR). Images from each street formed a dataset, which was split into training and testing sets and contained three classes: potholes, patches and alligator cracks. Detailed numbers of images and instances are in Table 1.

In addition to the two testing sets, two more testing sets of still camera RGB images were created to evaluate models in out-of-context and challenging environmental conditions. They

Dataset	WR		TCR		Camb-still	Pothole-special
	Train	Test	Train	Test	Test	Test
Nos. Images	873	190	765	175	112	89
Positive Images	693	148	528	135	109	89
Nos. Instances	894	356	871	280	290	225
Instance distribution						
potholes	301	201	76	27	149	150
patch	453	138	238	93	134	75
alligator cracks	140	17	557	160	7	0

Table 1: Data distribution of the two training sets and the four testing sets

were collected in neighbourhoods with an asphalt surface beyond the two roads, as shown in the map in Figure 3. These two image sets were not trained by any models and were prepared for testing only. The first test-only dataset (Camb-still) was taken on days with fair weather and clear conditions. The second (Pothole-special) was taken in more challenging conditions, either taken after rain with ponding water in potholes, covered by fallen leaves or with significant shade.

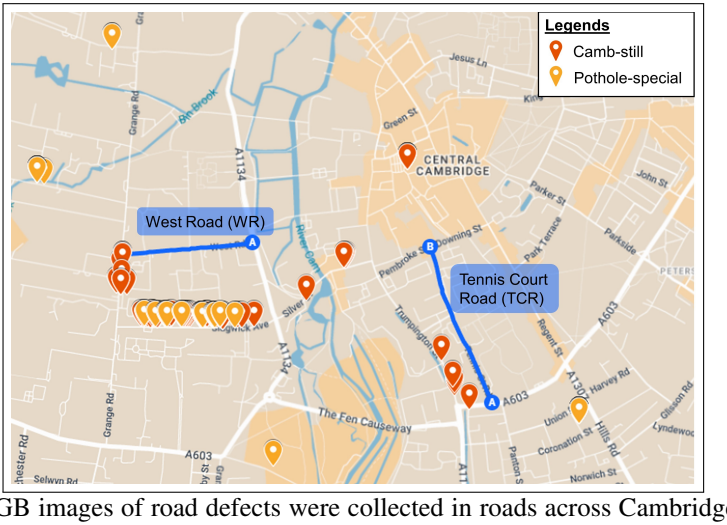

Figure 3: RGB images of road defects were collected in roads across Cambridge. Pothole-special specifically collects images in challenging conditions.

The RGB-trained models were trained with all three classes to improve the models' inter-class discrimination and enhance the results for the target class. In testing, the RGB-trained models of the two streets and the depth detector inferred instances on each of the testing sets: WR, TCR, Camb-still and Pothole-special. The testing adopted metrics of precision, recall and F1 scores for masks at the Intersection-over-Union (IoU) threshold of 0.5. This paper only presented metrics evaluated on the potholes category, as the depth detector only impacted the detection of potholes.

Note the confidence score thresholds at the output of RGB-trained models. When evaluating the models against the four testing sets, the confidence score threshold was set at 0.001 for bboxes and masks to capture predictions of the full range of confidence scores when evaluating the average precisions and recalls. To prevent garbage bboxes from being fused, RGB-trained models produced another set of predictions with a confidence score threshold

of 0.25 and fused with those from the depth detector with WBF.

4 Results and Discussions

Table 2 and Table 3 show the F1 scores, precision and recall of the two RGB-trained models, the depth detector alone and the fusion results with both RGB-trained models. The cell colour shows the alignment with the trained context. Cells in dark blue have testing images taken in the same perspective, road scene and geolocation as the trained models. Cells in magenta show combinations in different geolocations. Cells in beige show combinations in different geolocations and road scenes.

Trained \ Test	Pothole only (F1_50, M)			
	WR	TCR	Camb-still	Pothole-special
WR	0.335	0.090	0.376	0.149
TCR	0.048	0.765	0.189	0.031
Depth only	0.056	0.004	0.139	0.161
Fused (using $conf = 0.25$):				
WR+Depth	0.388	0.064	0.281	0.204
TCR+Depth	0.096	0.603	0.306	0.165

Legends			
Perspective	✓	✓	✓
Scene	✓	✓	×
Geolocation	✓	×	×

Table 2: F1 score of RGB-trained, Depth Detector and Fused. Fusion improved the F1 scores when detecting out of context.

RGB-trained models performed the best on the testing set that had the same perspective, scene and geolocation, in line with expectation. What caught a surprise was the significant drop of performance even by just changing the geolocation (WR and TCR exchanged, or both on Camb-still and pothole-special) and further in more challenging scenes and conditions (pothole-special). Metrics by the depth estimator alone were consistently mediocre in all datasets regardless of scenes and geolocations, performing at about the level of RGB-trained models in different geolocations.

When fused, the F1 score improved predominantly when RGB-trained models performed poorly out of context. This was mainly achieved by having the depth estimator to improve the precision. When the RGB-trained models detected well (TCR on TCR-val, WR on Camb-still), predictions from the depth estimator caused confusion and dragged the results.

4.1 Discussions

The sharp decline in performance on images taken in a different street, even at the same perspective and similar street scene in the same city, highlights the poor generalisation of

Trained \ Test	AP_50 (M)				AR_50 (M)			
	WR	TCR	Camb-still	Pothole-special	WR	TCR	Camb-still	Pothole-special
WR	0.243	0.049	0.278	0.094	0.542	0.519	0.584	0.360
TCR	0.032	0.721	0.124	0.016	0.100	0.815	0.396	0.247
Depth only	0.034	0.002	0.093	0.132	0.159	0.111	0.275	0.207
Fused (using $conf = 0.25$):								
WR + Depth	0.351	0.045	0.254	0.171	0.433	0.111	0.315	0.253
TCR + Depth	0.065	0.578	0.255	0.138	0.179	0.630	0.383	0.207

Table 3: Average precision and recall @ $IoU = 0.5$ of RGB-trained, Depth Detector and Fused. Fusion mainly improved the precision.

RGB-trained models. The performance worsens when the images are taken in more challenging environmental conditions. When a model performs so bad that it shows little ability to extrapolate and detect instances of the same category beyond the learned context, results highlight the need for some safeguard to supplement detections, especially when deployed for critical applications. In real practice, developers will collect all annotated images they have and train a holistic model to harness the benefits of the scaling law. The experiments postulate "what-if" situations when road scenes go out of context, even as benign as RGB patterns of potholes change with fallen leaves and ponding water.

The depth estimator supplements predictions by the RGB-trained models, especially out of their training context. It also provides a relatively consistent performance in clear and challenging conditions, without the acute plunge in performance by RGB-trained models in more complex conditions. This is likely caused by the fact that DINOv2 was trained in an unsupervised manner that enables it to draw relationships relative to other pixels in the same image, instead of relying explicitly on trained patterns in RGB-trained models. A deeper review of predictions by the depth estimator suggests room for improvement in shady environments. This may potentially be achieved by removing shadows in pre-processing, or fine-tuning a depth estimator in future development.

While local roads and inspections with simple apparatus may improve local residents' satisfaction, maintaining motorways is vital to sustain uninterrupted logistics and connectivity between key transport hubs and cities. Preliminary studies on images with more radical differences of road scenes were performed, such as motorway images inferred on models trained with local street images. The results tended to zero and were not presented. Further studies on fusing motorway and local road images are encouraged.

5 Conclusions

This research illustrates the problem of poor generalisation of object detectors trained with limited annotations of RGB images. In the use case of pothole detection, monocular RGB images were collected to evaluate and propose improvements to the problem. Experiments found a remarkable performance drop of 73% (WR weight) and 94% (TCR weight) in F1 score when transiting to an unfamiliar context, even when the transition was merely to another road in the same city.

The proposed solution targeted the out-of-context issue and addressed it by fusing depth detection to supplement RGB-trained model detections. As opposed to previous work that required stereo images or point clouds to reconstruct in 3D and threshold by depth, this research utilises a pre-trained DINOv2 depth estimator to predict depths on RGB images. The solution integrates orthorectification, depth correction, feature extraction and mask refinement to detect potholes with monocular RGB images. The predictions by RGB-trained models and depth are fused with weighted bounding box fusion. Experimental results reveal that while the depth detector standalone performed at about the level of out-of-context RGB-trained models, the fused results improved F1 scores by 0.05 (16%) to 0.13 (81%) when the RGB-trained models operate out-of-context. Further research may include addressing detections on motorways and tackling shadows in images.

Acknowledgement

The author (P Lam) is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Future Infrastructure and Built Environment:

Resilience in a Changing World (FIBE2) [grant number EP/S02302X/1] and sponsored by the National Highways, Costain and Trimble Solutions. This work is supported by the Digital Roads, UK EPSRC [grant number EP/V056441/1].

References

[1] Denis Mbey Akola and Gianni Franchi. How To Effectively Train An Ensemble Of Faster R-CNN Object Detectors To Quantify Uncertainty. *arXiv preprint*, 12 2023.

[2] Anas Al-Shaghouri, Rami Alkhatib, and Samir Berjaoui. Real-Time Pothole Detection Using Deep Learning, 2021.

[3] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8570–8578, Seattle, WA, USA, 2020. IEEE.

[4] Siyuan Chen, Debra F. Laefer, Xiangding Zeng, Linh Truong-Hong, and Eleni Mangina. Volumetric Pothole Detection from UAV-Based Imagery. *Journal of Surveying Engineering*, 150(2), 5 2024. ISSN 0733-9453. doi: 10.1061/jsued2.sueng-1458.

[5] Department for Transport. Road condition statistics - a basic guide and quality assessment, 12 2024. URL https://www.gov.uk/government/publications/road-network-size-and-condition-statistics-guidance/road-condition-statistics-a-basic-guide-and-quality-assessme

[6] Department for Transport. Section 2: measuring surface condition using automated visual methods, 12 2024. URL https://www.gov.uk/government/publications/road-condition-statistics-technical-note/section-2-measuring-surface-condition-using-automated-visual

[7] Amita Dhiman, Hsiang-Jen Chien, and Reinhard Klette. Road Surface Distress Detection in Disparity Space. In *2017 International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE, 12 2017. ISBN 9781538642764. doi: 10.1109/IVCNZ.2017.8402459.

[8] Ali Faisal and Suliman Gargoum. Cost-effective LiDAR for pothole detection and quantification using a low-point-density approach. *Automation in Construction*, 172, 4 2025. ISSN 09265805. doi: 10.1016/j.autcon.2025.106006.

[9] Feng Gao, Caimei Wang, and Caihong Li. A combined object detection method with application to pedestrian detection. *IEEE Access*, 8:194457–194465, 2020. ISSN 21693536. doi: 10.1109/ACCESS.2020.3031005.

[10] Jiyang Gao, Jiang Wang, Shengyang Dai, Li-Jia Li, and Ram Nevatia. NOTE-RCNN: NOise Tolerant Ensemble RCNN for Semi-Supervised Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9507–9516, Seoul, Korea, 2019. IEEE.

[11] Zineb Haimer, Khalid Mateur, Youssef Farhan, and Abdessalam Ait Madi. Pothole Detection: A Performance Comparison Between YOLOv7 and YOLOv8. In *2023 9th*

International Conference on Optimization and Applications, ICOA 2023 - Proceedings. Institute of Electrical and Electronics Engineers Inc., 2023. ISBN 9798350312546. doi: 10.1109/ICOA58279.2023.10308849.

[12] Glenn Jocher and Ultralytics. Ultralytics YOLO, 2024. URL https://github.com/ultralytics/ultralytics?tab=readme-ov-file.

[13] Gargi Joshi, Amey Joshi, Mranmay Shetty, Rahee Walambe, Ketan Kotecha, Fabio Scotti, and Vincenzo Piuri. Ensemble learning and EigenCAM-based feature analysis for improving the performance and explainability of object detection in drone imagery. *Discover Applied Sciences*, 7(5), 5 2025. ISSN 30049261. doi: 10.1007/s42452-025-06879-5.

[14] Malhar Khan, Muhammad Amir Raza, Ghulam Abbas, Salwa Othmen, Amr Yousef, and Touqeer Ahmed Jumani. Pothole detection for autonomous vehicles using deep learning: a robust and efficient solution. *Frontiers in Built Environment*, 9, 2023. ISSN 22973362. doi: 10.3389/fbuil.2023.1323792.

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. 4 2023. URL https://segment-anything.com.

[16] James Liao. Inverse Perspective Mapping, 2019. URL https://github.com/JamesLiao714/IPM-master.

[17] Kangjie Liu, Borui Zhang, Jiwen Lu, and Haibin Yan. Toward Integrity and Detail with Ensemble Learning for Salient Object Detection in Optical Remote-Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. ISSN 15580644. doi: 10.1109/TGRS.2024.3400032.

[18] Hiroya Maeda, Yoshihide Sekimoto, Toshikazu Seto, Takehiro Kashiyama, and Hiroshi Omata. Road Damage Detection Using Deep Neural Networks with Images Captured Through a Smartphone. *Computer Aided Civil and Infrastructure Engineering*, 1 2018. doi: 10.1111/mice.12387.

[19] Zhixiong Nan, Xianghong Li, Tao Xiang, and Jifeng Dai. DI-MaskDINO: A Joint Object Detection and Instance Segmentation Model. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024).*, Vancouver, Canada, 12 2024.

[20] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, 2 2024.

[21] Arjun Paramarthalingam, Jegan Sivaraman, Prasannavenkatesan Theerthagiri, Balaji Vijayakumar, and Vignesh Baskaran. A deep learning model to assist visually impaired in pothole detection using computer vision. *Decision Analytics Journal*, 12, 9 2024. ISSN 27726622. doi: 10.1016/j.dajour.2024.100507.

[22] J. S. Park, K. S. Lee, and S. Kim. Assessment for a condition using terrestrial lidar data. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, volume 42, pages 311–314. International Society for Photogrammetry and Remote Sensing, 8 2019. doi: 10.5194/isprs-archives-XLII-3-W8-311-2019.

[23] Atharv Patawar, Mohammed Mehdi, Bhaumik Kore, and Pradnya Saval. A pothole can be seen with two eyes: an ensemble approach to pothole detection. *Machine Vision and Applications*, 36(3), 5 2025. ISSN 14321769. doi: 10.1007/s00138-025-01679-8.

[24] Yaning Qiao, Jia Wang, Ximeng Zhang, Jiandong Huang, Liang He, and Runhua Zhang. Enhancing automatic pothole detection in non-motorized corridors using smartphones: an integrated algorithm. *Measurement Science and Technology*, 36(3), 3 2025. ISSN 13616501. doi: 10.1088/1361-6501/adb641.

[25] Mario Roman-Garay, Hector Rodriguez-Rangel, Carlos Beltran Hernandez-Beltran, Peter Lepej, José Eleazar Arreygue-Rocha, and Luis Alberto Morales-Rosales. Architecture for pavement pothole evaluation using deep learning, machine vision, and fuzzy logic. *Case Studies in Construction Materials*, 22, 7 2025. ISSN 22145095. doi: 10.1016/j.cscm.2025.e04440.

[26] Navjot Singh, Rinki Arya, and R. K. Agrawal. A novel approach to combine features for salient object detection using constrained particle swarm optimization. *Pattern Recognition*, 47(4):1731–1739, 4 2014. ISSN 00313203. doi: 10.1016/j.patcog.2013.11.012.

[27] Priya Singh and Rajalakshmi Krishnamurthi. Object detection using deep ensemble model for enhancing security towards sustainable agriculture. *International Journal of Information Technology (Singapore)*, 15(6):3113–3126, 8 2023. ISSN 25112112. doi: 10.1007/s41870-023-01341-4.

[28] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107, 3 2021. ISSN 02628856. doi: 10.1016/j.imavis.2021.104117.

[29] The Chartered Institution of Highways & Transportation. Sixty Years of Highways Maintenance. Technical report, CIHT North Eastern Branch, 2012.

[30] Rahee Walambe, Aboli Marathe, Ketan Kotecha, and George Ghinea. Lightweight Object Detection Ensemble Framework for Autonomous Vehicles in Challenging Weather Conditions. *Computational Intelligence and Neuroscience*, 2021, 2021. ISSN 16875273. doi: 10.1155/2021/5278820.

[31] Juan Wang, Zetao Zhang, Minghu Wu, Yonggang Ye, Sheng Wang, Ye Cao, and Hao Yang. Improved BlendMask: Nuclei instance segmentation for medical microscopy images. *IET Image Processing*, 17(7):2284–2296, 5 2023. ISSN 17519667. doi: 10.1049/ipr2.12792.

[32] Niannian Wang, Jiaxiu Dong, Hongyuan Fang, Bin Li, Kejie Zhai, Duo Ma, Yibo Shen, and Haobang Hu. 3D reconstruction and segmentation system for pavement potholes based on improved structure-from-motion (SFM) and deep learning. *Construction and*

Building Materials, 398, 9 2023. ISSN 09500618. doi: 10.1016/j.conbuildmat.2023.132499.

[33] Wenzhe Wang, Bin Wu, Sixiong Yang, and Zhixiang Wang. Road Damage Detection and Classification with Faster R-CNN. In *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, 12 2018. IEEE. ISBN 9781538650356.

[34] Pan Wei, John E. Ball, and Derek T. Anderson. Fusion of an ensemble of augmented image detectors for robust object detection. *Sensors (Switzerland)*, 18(3), 3 2018. ISSN 14248220. doi: 10.3390/s18030894.

[35] Hanyu Xin, Yin Ye, Xiaoxiang Na, Huan Hu, Gaoang Wang, Chao Wu, and Simon Hu. Sustainable Road Pothole Detection: A Crowdsourcing Based Multi-Sensors Fusion Approach. *Sustainability (Switzerland)*, 15(8), 4 2023. ISSN 20711050. doi: 10.3390/su15086610.

[36] Qihang Yang, Yang Zhao, and Hong Cheng. MMLF: Multi-modal Multi-class Late Fusion for Object Detection with Uncertainty Estimation. *arXiv preprint*, 10 2024.

[37] Junkui Zhong, Deyi Kong, Yuliang Wei, and Bin Pan. YOLOv8 and point cloud fusion for enhanced road pothole detection and quantification. *Scientific Reports*, 15(1), 12 2025. ISSN 20452322. doi: 10.1038/s41598-025-94993-0.