Predicting Damage of Dutch Road Markings

Amanda Jansen[1,2], Vivienne Jansen[1,2], Jurjen Helmus[1], and Arnoud Visser[2]

[1] *Maintenance Lab, Applied University of Amsterdam, Amsterdam, The Netherlands*
[2] *Intelligent Robotics Lab, University of Amsterdam, Amsterdam, The Netherlands*

Abstract. Road markings play a crucial role in road safety by guiding traffic and ensuring visibility. As markings deteriorate over time, their effectiveness diminishes, necessitating timely maintenance. This paper studies two methods to classify the road-marking damage according to the Dutch CROW guidelines. The first is a model based approach, which first uses a regression model to estimate the marking damage, and then applies the thresholds in the CROW guidelines to classify the damage class. In contrast, a data driven approach is used, classifying directly the damage class with a YOLOv8 classifier. This results in state-of-the-art accuracy, demonstrating strong potential for real-time deployment.

Keywords: computer vision · segmentation · classification

1 Introduction

As cities grow and mobility increases, the pressure on public infrastructure and the need for efficient maintenance strategies intensifies [1]. Road markings, which include painted lines, symbols, and patterns on the road surface, play a key role in managing road safety. These markings help warn road users and ensure smooth traffic flow [16]. However, they degrade over time and currently rely on manual inspections that are time-consuming, costly, and often inconsistent [17].

Therefore, Velotech[3], in collaboration with Amsterdam University of Applied Sciences, has developed a Smart Bikes project that uses artificial intelligence (AI) to automate the inspection of urban infrastructure. By equipping bikes with edge AI, a technology that processes data locally on the device, the system captures and analyzes road condition data in real-time. This reduces latency, minimizes dependence on external servers and aligns with municipal privacy standards, as sensitive visual data never leaves the bike [9]. This integrated approach allows municipalities to efficiently assess the condition of road markings and prioritize repairs.

In the Netherlands, such maintenance decisions are guided by CROW guidelines, which serve as the national standard for evaluating road infrastructure. These guidelines categorize road markings into four classes, from A to D, based

[3] `https://velotech.ai/`

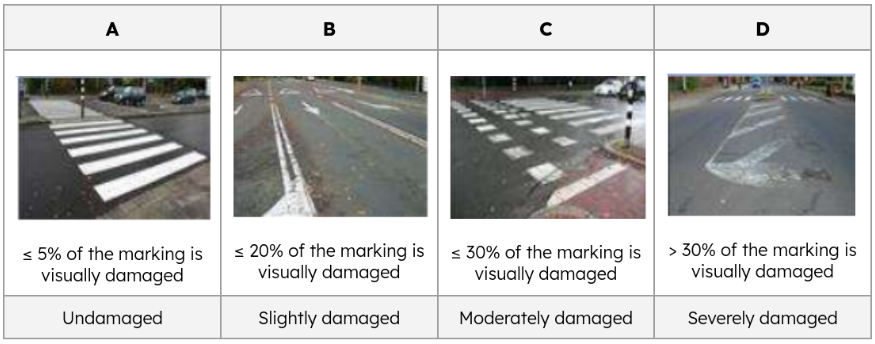


Fig. 1: CROW guidelines per severity category (classes A, B, C and D) [6].

on the damage severity [6]. Class A represents road markings in excellent condition, while class D indicates markings in poor condition that require repair (see Fig. 1). To automate this classification process, this research uses both a model-based and a data-driven approach.

2 Method

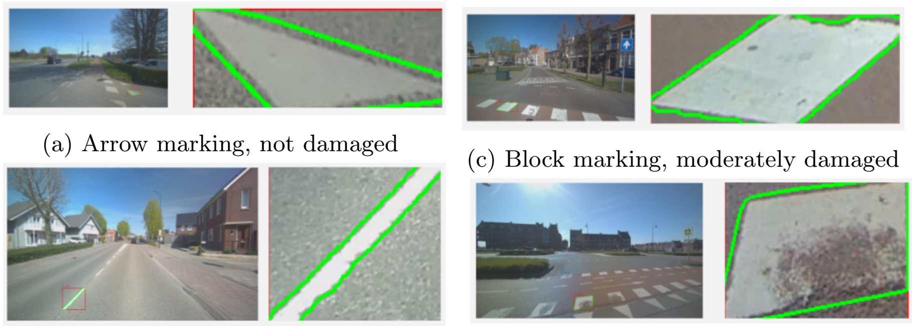
Estimation of the road-marking damage is performed in this study in two different ways; first by a regression method which estimates the amount of damaged paint followed by a decision-model based on the thresholds in the CROW guidelines. Because this model has a tendency to overpredict damage severity, this method is compared with a fully data-driven approach.

2.1 Dataset

The data that was used in this study was collected by Velotech with a ZED-X stereo camera that was developed by Stereolabs[4]. The camera was mounted on both bicycles and cars to simulate real-world mobile inspection scenarios [7]. Data acquisition took place in two distinct urban environments in the Netherlands: Geertruidenberg, a small municipality characterized by relatively calm residential streets, and Amsterdam Oud-Zuid, a densely populated urban district with a high volume of traffic, varied infrastructure, and complex road markings. This geographical variation introduces diversity in lighting conditions, road surface materials, marking styles, and levels of wear, and ensures that the dataset reflects a broad range of real-world conditions that are relevant to road marking assessment (see Fig. 2).

The combined recordings from Geertruidenberg and Amsterdam Oud Zuid resulted in a dataset consisting of 15.745 high-resolution images of road markings. The images were each accompanied by annotations that were stored in

[4] https://www.stereolabs.com/



(a) Arrow marking, not damaged

(b) Line marking, slightly damaged

(c) Block marking, moderately damaged

(d) Block marking, severely damaged

Fig. 2: Examples Velotech recordings and the individual marking selected for classification.

a structured JSON file. Each entry described an individual road marking and included a polygon delineating its shape, a bounding box, a manually assigned CROW severity class (A, B, C, or D), and a 'superclass' indicating the broader category of the road marking (e.g., line, arrow, or zebrawalking). Polygon and bounding box annotations were automatically generated using a YOLOv8 segmentation model developed by Velotech. After removing images with missing severity or polygon annotations, the final dataset contained 15.723 annotated road markings.

The majority of markings fall into severity classes A (#5.002) and B (#6.572), while fewer instances are observed in classes C (#1.907) and D (#2.242). The dataset is strongly dominated by the line category (#13.720), with considerably fewer examples of other types such as giveawayrow (#807), block (#368), zebrawalking (#314), and smaller categories such as arrow (#63) and stopping (#48) marking.

Although the distribution is imbalanced, care has been taken not to influence our methods. For the data-driven approach the dataset was divided into training (70%), test (15%), and validation (15%) subsets using stratified splitting.

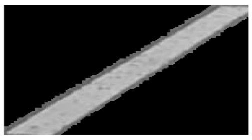
2.2 Regression Method

The regression method starts with a high-resolution image recorded by Velotech. Velotech has an accurate detection model (98% pixel accuracy) [11] which localizes road markings and generates a polygon and bounding-box around the road marking. The polygon is used here to get a region of interest (ROI). The polygon is not intended to be tight around the road marking, so part of the road surface is visible at the edges. To prevent that these edges contribute to estimation of the amount of damaged paint, an erosion algorithm with a kernel of 7×7 is applied to be able to concentrate on the core of the road marking (see Fig. 3).

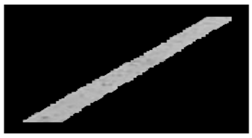
An estimate of the color of the road surface (in grayscale) surrounding the road marker is also important, because when the road marking is damaged the

(a) Velotech recording

(b) Mask from detection



(c) Zoom in on mask



(d) Erosion on mask

(e) Outlier mask

Fig. 3: Preprocessing method

road surface shines through. Yet, the color of the road surface is not always the same, nor the lighting conditions, so an outlier mask (see Fig. 3.e) can be used to estimate the color of the road surface near the road marking.

To distinguish between intact and damaged areas within the marking a dynamic threshold is applied, based on Otsu's method [10]. Otsu's method used the histogram of the image to define two clusters of bright and dark pixels as classes and maximizes the between-class variance to find the optimal threshold. By calculating the proportion of dark/damaged pixels inside the eroded mask the damage ratio can be calculated, which can directly be mapped to the CROW guidelines.

2.3 Data-Driven Method

The road markings are already detected and localized in the images with a YOLOv8-based detector [11], so it is logical to see how well a YOLOv8-based classification would work on this problem. You Only Look Once (YOLO) is a real-time object detection algorithm known for its speed and accuracy in identifying and classifying visual elements within images [13]. YOLOv8 is an algorithm that is slightly easier to fine-tune on new types of objects than YOLOv9 [15]. An alternative would be the most recent YOLOv12 [14], although the attention model relies on FlashAttention for optimal speed. FlashAttention is only supported on relatively modern GPU architectures and is less suitable for edge-computing.

Two models were trained with YOLOv8; one for a binary (damaged/undamaged) classification task and one for a multiclass (A/B/C/D) classification task. The models were initialized with pre-trained weights and trained on road-marking images. These images were obtained by cropping the original images to the bounding boxes of the road markings. Training was conducted for 100 epochs with an input image resolution of 640×640 pixels, a batch size of 32, and 8 data loader workers. Early stopping was applied with a patience of 10 epochs to prevent overfitting.

To evaluate the learning behavior and generalization capability of both models during training, the progression of the training and validation loss was monitored. The trained YOLOv8 models were both evaluated on the 15% validation set, consisting of 2,359 previously unseen road marking instances, each labeled with one of the four CROW-defined severity classes (A-D). Fig. 4 shows the loss curves for the binary classification model, and Fig. 5 presents those for the multiclass model.

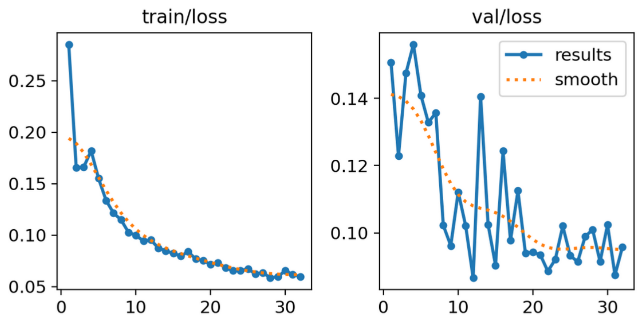


Fig. 4: Training and validation loss functions for the YOLOv8 binary (damaged and undamaged) model classification model. Training loss (left) and validation loss (right) curves over 32 training epochs.

For both models, the training loss steadily decreases, demonstrating that the model is effectively learning to minimize the error on the training dataset. Simultaneously, the validation loss shows a similar downward trend and closely tracks the training loss, indicating that the model generalizes well to unseen data. The absence of any increase or divergence in validation loss gives confidence that overfitting did not occur (yet).

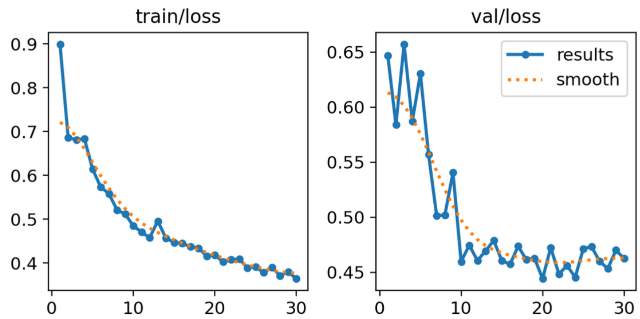


Fig. 5: Training and validation loss functions for the YOLOv8 multiclass (classes A, B, C and D) model. Training loss (left) and validation loss (right) curves over 30 training epochs

3 Results

Both the regression as data-driven methods were studied extensively, including multiclass classification with the regression method and analysis of the model performance per road marking type. For more details, see the theses [4,5].

3.1 Regression Results

When the regression method described in Sec. 2.2 is applied to a binary classification task, distinguishing between undamaged (class A) and damaged (class B,C,D) markings, the precision for the undamaged class (0.92) and the recall of the damaged class (0.99) are quite good. So, from the 15.723 instances in the dataset, 99% of the damaged markings were correctly identified and when a road marking was predicted to be undamaged, it is highly likely to be correct.

Table 1: Binary Classification with Dynamic Thresholding

Class	Precision	Recall	F1-score	Instances
Undamaged (class A)	**0.92**	0.26	0.41	5002
Damaged (class B/C/D)	0.74	**0.99**	0.85	10721
Accuracy	—	—	0.76	15723

Yet, as can be seen from Table 1, the model tends to over-detect damage and frequently misclassifies undamaged markings as damaged. The undamaged recall is only 0.26 and the precision on damaged markings is only 0.74, leading to an overall F1-score of 0.76. This makes this model only useful as pre-filtering tool to reduce the volume of markings requiring manual inspection within the maintenance workflow of Velotech.

3.2 Data-Driven Results

The data-driven approach described in Sec. 2.3 improves these results both for the undamaged recall and the precision on recognizing damaged markings (Table 2). Although there is still a slight bias towards flagging damage, the false negatives are so low that it approaches the operational goal of Velotech.

Table 2: Binary Classification with YOLOv8

Class	Precision	Recall	F1-score	Instances
Undamaged (class A)	0.98	**0.95**	0.96	749
Damaged (class B/C/D)	**0.97**	0.99	0.98	1608
Accuracy	—	—	**0.97**	2357

Based on these preliminary numbers, it becomes interesting to look at the multi-class results, if the data-driven approach makes the distinction between slightly damaged, moderately damaged, severely damaged (CROW classes B/C/D).

Table 3: Multi-class Classification with YOLO v8

Class	Precision	Recall	F1-score	Instances
A	**0.95**	**0.95**	**0.95**	749
B	**0.85**	**0.89**	**0.87**	986
C	0.54	0.40	0.46	288
D	0.70	0.76	0.72	334
Macro-average	0.76	0.75	0.75	2357
Accuracy	—	—	**0.83**	2357

The results for road markings in good condition or only slightly deteriorated (class A/B) are good, as can be seen in Table 3. In contrast, the performance is notably lower for moderate or severely damaged road markings (class C/D).

Most errors originate from class C. As can be seen from the confusion matrix in Fig. 6, where road markings annotated as 'moderated damaged' are predicted in 32% of the cases as 'slightly damaged', 40% of the cases as 'moderated damaged' and in 27% of the cases as 'severely damaged'. The boundaries between class B/C (20% damaged surface) and between class C/D (30% damaged surface) seem to be hard to estimate. Also human annotators have difficulty making this distinction, as can be seen in Section 4.1.

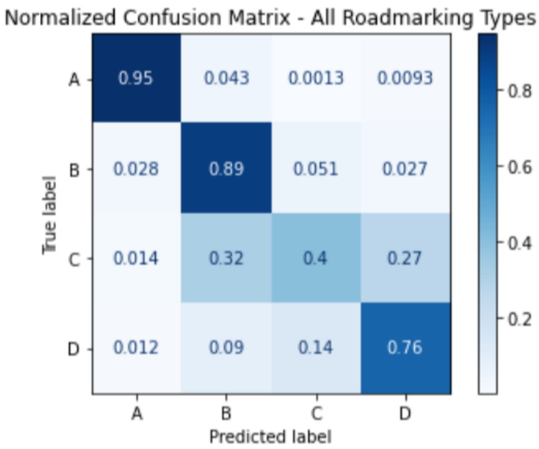


Fig. 6: Confusion matrix of overall severity classification on the test set. The confusion matrix displays the number of predicted labels versus ground truth labels across the four CROW severity classes. The diagonal cells represent correct predictions.

4 Discussion

4.1 Damage severity

The manual annotations of the severity classes were outsourced by Velotech and carried out by multiple individuals. Velotech has implemented quality control measures to get consistent annotations between the annotators. Still, differentiating between severity levels B ($\leq$ 20% damage), C ($\leq$ 30% damage), and D ($>$ 30% damage) requires annotators to estimate the proportion of damaged surface area by eye, which is a task prone to individual interpretation and variability.

Consequently, even experienced annotators may produce inconsistent or inaccurate labels, particularly when the extent of damage is near the boundary between two severity classes. In cases where the model correctly detects damage to a road marking, the damage ratio given by the regression model appears to provide a more objective and consistent assessment of damage severity compared to manual evaluations of the severity class (see Fig. 7).

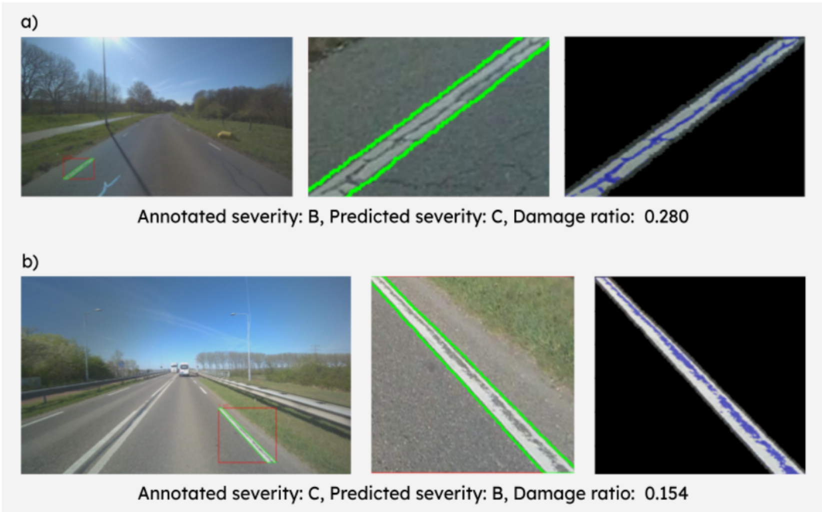


Fig. 7: Examples in which the model provides a reasonable damage ratio other than the severity classification from the manual annotations.

Unfortunately, the regression model fails to always detect damage correctly. Two common failure cases can be easily demonstrated with two examples. The first failure case is due to partial shadows. Because the algorithm depends on brightness, a sharp shadow can be easily judged to be a damaged area (see Fig. 8).

Fig. 8: Example of a road marking partially covered by shadow. The marking is labeled as class A (undamaged) in the ground truth, but the model predicts it as Class C due to the presence of a partial shadow cast by a streetlight. The reduced brightness in the shadowed region is incorrectly interpreted by the model as surface damage.

Another common failure case originates from complex road marking shapes. For some road marking symbols it is difficult to clearly define the inside and outside regions. When the mask is just a square bounding box, a lot of the road surface is still visible, which could result in unrealistic high damage ratio estimates (see Fig. 9).

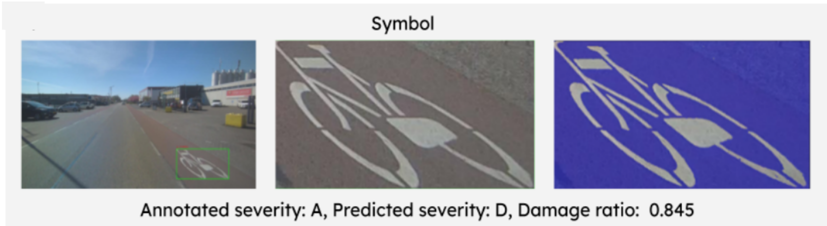


Fig. 9: Examples of highly imprecise polygon annotations that resemble bounding boxes rather than accurate segmentations, shown for the bicycle symbol.

Yet, with many borderline cases where class boundaries are open for interpretation, it is hard to train a data-driven approach like YOLO on learning the decision boundaries more precisely.

4.2 Related research

Datasets focused on road marking damage are scarce. A study of Iparraguirre *et al* [3] combined two datasets from Japan & Spain. They added 971 new labeled images for Spanish roads. They performed binary classification with three different convolutional neural networks. Their best result used EfficientDet v1 D0 [12], and achieved an F1-score of 0.93, which was a large improvement compared to the previous result on the Japanese dataset (F1-score of 0.72) [8]. The regression results from this study are comparable with the results reported by Maeda *et al*, and the data-driven approach is comparable with the results of Iparraguirre *et al*. Yet, unfortunately the combined dataset from Japan & Spain does not contain a labeled level of the severity of the damage.

This information was available in a dataset from the USA. Recent work of Antariska *et al* also used a data-driven approach based on YOLOv8 [2]. This method was trained on 865 images collected along New Jersey State routes. The dataset concentrated on a subset of road markings used in this study, namely the center line. Instead of four damage classes, three damage classes were used (good, moderate, poor). This system achieved a macro-averaged precision of 0.51, considerably lower than the results in this study. Partly this can be contributed to the smaller dataset (they only annotated 865 images from the 15,536 available images). Their study was also limited by the resolution of the images collected along the New Jersey routes. The road markings cover only part of the image, so you have to zoom in at the road marking and make (implicit or explicit) an estimate of the damage ratio. In that case high-resolution images, as provided by the ZED-X stereo camera, can make the difference.

5 Conclusion

The model-based approach showed mixed results for different road-marking types. Simple road markings such as lines and blocks with clear boundaries could be classified quite well, but the results for complex road markings such as the bicycle symbol pixel-wise segmentation was required. This method showed also to be sensitive to the lighting conditions, such as the occurrence of partial shadows over the road markings.

In contrast, the data-driven approach worked well under all circumstances, although it suffered from class-imbalance in the training dataset. Yet, the thresholds of the CROW guidelines are quite strict, while the YOLOv8 classification confused the 'moderately damaged' class easily with 'slightly damaged' and 'severely damaged' class. Also the human annotators had difficulty with this distinction.

So, manual verification of the damage is still required. Automatic classification could still benefit maintenance operations, by excluding clearly undamaged road markings and allowing them to give priority to severely damaged road markings.

Acknowledgement

We would like to express our sincere gratitude to Velotech for their valuable collaboration, for providing us with access to their data, and for their guidance throughout the research process. Their support and the availability of their dataset have been fundamental to this research.

References

1. Ai, D., Jiang, G., Lam, S.K., He, P., Li, C.: Computer vision framework for crack detection of civil infrastructure—a review. Engineering Applications of Artificial Intelligence **117**, 105478 (Jan 2023)
2. Antariksa, G., Chakraborty, R., Somvanshi, S., Das, S., Jalayer, M., Patel, D.R., Mills, D.: Comparative analysis of advanced ai-based object detection models for pavement marking quality assessment during daytime. preprint arXiv:2503.11008 (Mar 2025)
3. Iparraguirre, O., Iturbe-Olleta, N., Brazalez, A., Borro, D.: Road marking damage detection based on deep learning for infrastructure evaluation in emerging autonomous driving. IEEE Transactions on Intelligent Transportation Systems **23**(11), 22378–22385 (Jul 2022)
4. Jansen, A.: Classifying the damage severity of road markings using computer vision. Bachelor thesis, University of Amsterdam (Jul 2025)
5. Jansen, V.: Automated assessment of the damage severity of road markings using yolo. Bachelor thesis, University of Amsterdam (Jul 2025)
6. Kennisplatform CROW: Kwaliteitscatalogus openbare ruimte 2023. published online (Aug 2023)
7. Koomen, D.: Determining the geolocation of road markings based on stereo camera vision and a Bird's Eye View. Bachelor thesis, University of Amsterdam (Jun 2024)

8. Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T., Omata, H.: Road damage detection and classification using deep neural networks with smartphone images. Computer-Aided Civil and Infrastructure Engineering **33**(12), 1127–1141 (Jun 2018)
9. Meuser, T., Lovén, L., Bhuyan, M., Patil, S.G., Dustdar, S., Aral, A., Bayhan, S., Becker, C., Lara, E.d., Ding, A.Y., Edinger, J., Gross, J., Mohan, N., Pimentel, A.D., Rivière, E., Schulzrinne, H., Simoens, P., Solmaz, G., Welzl, M.: Revisiting edge ai: Opportunities and challenges. IEEE Internet Computing **28**(4), 49–59 (Aug 2024)
10. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (Jan 1979)
11. Schimmelpennink, M.: Segmentation of road markings in the Netherlands for automatic inspections, using YOLOv8-seg on an edge system. Bachelor thesis, University of Amsterdam (Jul 2024)
12. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10778–10787 (Aug 2020)
13. Terven, J., Córdova-Esparza, D.M., Romero-González, J.A.: A comprehensive review of yolo architectures in computer vision. Machine Learning and Knowledge Extraction **5**(4), 1680–1716 (Nov 2023)
14. Tian, Y., Ye, Q., Doermann, D.: Yolov12: Attention-centric real-time object detectors. preprint arXiv 2502.12524 (Feb 2025)
15. Wang, C.Y., Yeh, I.H., Mark Liao, H.Y.: Yolov9: Learning what you want to learn using programmable gradient information. In: European conference on computer vision. pp. 1–21. Springer (Oct 2024)
16. Wu, J., Liu, W., Maruyama, Y.: Street view image-based road marking inspection system using computer vision and deep learning techniques. Sensors **24**(23), 7724 (Nov 2024)
17. Younesi Heravi, M., Dola, I.S., Jang, Y., Jeong, I.: Edge ai-enabled road fixture monitoring system. Buildings **14**(5) (Apr 2024)