Traffic & Transportation

Automatic Road Damage Detection Based on Improved YOLO11

Siwei WEI[1], Yujian PENG[2], Hongfang LUO[3], Chunzhi WANG[4]

Original Scientific Paper
Submitted: 4 Mar 2025
Accepted: 3 Aug 2025
Published: 28 Apr 2026

[1] waosfengw@whut.edu.cn, School of Computer Science, Hubei University of Technology, Wuhan, China; CCCC Second Highway Consultants Company Ltd., Wuhan, China
[2] 1807752343@qq.com, School of Computer Science, Hubei University of Technology, Wuhan, China
[3] Corresponding author, luohongfang@hbut.edu.cn, Engineering and Technology College, Hubei University of Technology, Wuhan, China
[4] chunzhiwang@hbut.edu.cn, School of Computer Science, Hubei University of Technology, Wuhan, China

Publisher:
Faculty of Transport and Traffic Sciences, University of Zagreb

ABSTRACT
Road damage detection is vital for effective road maintenance and ensuring traffic safety. However, existing object detection models struggle with small objects, interference from complex backgrounds and difficulty handling multi-scale object features. To tackle these challenges, this study proposes an improved road damage detection model based on YOLO11. A novel RoadRep-C3 module is introduced to improve feature extraction, while an efficient multi-scale attention (EMA) mechanism captures multi-scale damage features more effectively. Additionally, a hypergraph structure is incorporated into the neck network to enable cross-stage information fusion, improving the detection of small objects. The proposed model also utilises a slide loss function to optimise performance on challenging samples. Experimental results on the RDD2022 dataset show a 2% increase in mean average precision (mAP@0.5) over the original YOLO11, with a reduced model size. These findings demonstrate the model's high accuracy and efficiency, offering a practical solution for detecting road damage and enhancing traffic safety.

KEYWORDS
intelligent transportation system; road damage detection; object detection; deep learning.

1. INTRODUCTION

With the continuous development of urban transportation networks, the global road mileage is steadily increasing. However, as traffic volume grows and the service life of roads extends, road ageing and damage have become increasingly severe. Road damage, including potholes and cracks, not only affects the aesthetic appearance of cities but also poses significant risks to vehicle and pedestrian safety. If damaged roads are not detected in time, they may continue to deteriorate, resulting in greater losses. Timely, precise and efficient identification of road damage is essential for initiating early repairs and mitigating potential risks [1].

In recent years, the swift advancement of deep learning, especially in object detection algorithms, has opened up new opportunities for road damage detection [2-3]. However, several challenges remain in practical applications. First, road damage varies significantly in size and shape; for example, cracks are often narrow and elongated, while potholes are larger and irregular, making multi-scale object detection difficult. Second, some road damage areas are small, occupying a low proportion of the overall image, which makes them prone to missed detections. Additionally, the complexity of road environments often leads to false detections. Finally, road damage can appear differently across various road types, weather conditions and times of day, imposing higher requirements on model generalisation.

To tackle these challenges, this paper applies the state-of-the-art YOLO11 (You Only Look Once11) object detection model for the detection of road damage and further enhances its accuracy and generalisation capabilities. The main contributions of this paper are as follows:

1) A novel feature extraction module, RoadRep-C3, is proposed, which leverages the RepVIt module (a lightweight vision transformer-based module) to redesign the bottleneck structure within the C3K2 module (a custom-designed block within YOLO11), significantly enhancing the accuracy and efficiency of feature extraction in the model.
2) By integrating the EMA (efficient multi-scale attention module) attention mechanism with the RoadRep-C3 module, the model's ability to capture multi-scale damage features is enhanced.
3) The YOLO11 neck network is replaced with the HyperC2Net (hypergraph-based cross-level and cross-position representation network), which utilises a hypergraph structure to capture high-order semantic relationships in the feature space. This allows comprehensive fusion of information from the backbone network, improving the model's robustness to complex road environments and enhancing its small-object detection capabilities.
4) To mitigate the problem of sample classification imbalance in road damage detection, the slide loss function is employed. It places greater emphasis on difficult samples, such as small cracks, thereby improving detection accuracy.

In summary, this paper presents a high-precision road damage detection model based on YOLO11, which introduces four improvements over the original YOLO11. These enhancements are specifically designed to optimise detection in complex road environments and for small cracks. The model significantly improves detection capabilities for challenging road damage, contributing to safer urban road traffic.

2. RELATED WORK

2.1 Road damage detection

Road damage detection was traditionally done through manual inspections or dedicated sensor-equipped vehicles [4]. However, manual detection presents several drawbacks, including being time-consuming, labour-intensive and inefficient. Moreover, its accuracy is heavily reliant on the experience of the personnel, making it challenging to ensure consistently reliable results [5]. Additionally, manual inspection may disrupt traffic flow, potentially causing safety hazards or even accidents, thereby endangering the safety of the inspectors. Dedicated detection vehicles improve efficiency, but their sensors may struggle under complex conditions like water or mud [6-8]. Moreover, these systems are prohibitively expensive, with high acquisition and maintenance costs, placing a heavy financial burden on some regions [9].

With the swift advancement of object detection algorithms driven by deep learning, their application in road damage detection has become increasingly widespread [10-16]. These algorithms can autonomously and accurately detect the location, type and boundaries of damage within images, facilitating quick, efficient and precise damage identification. At present, mainstream object detection algorithms are classified into two categories: two-stage and single-stage algorithms [17-18]. Two-stage algorithms perform object detection through candidate region generation followed by classification and regression, offering high detection accuracy but incurring significant computational costs, which makes them unsuitable for real-time detection [19]. Notable examples of two-stage algorithms include R-CNN, faster R-CNN and mask R-CNN [20-22]. For instance, Li et al. proposed an improved faster R-CNN model for crack recognition, incorporating attention mechanisms. This model employs ResNet50 as the backbone for feature extraction and integrates the squeeze-and-excitation network to strengthen attention mechanisms, optimising the model's ability to recognise complex patterns in images [23]. Furthermore, Li et al. constructed a pothole dataset encompassing diverse road conditions and environments and proposed an enhanced mask R-CNN model capable of accurately extracting the geometric characteristics and area data of identified potholes [24]. In contrast, single-stage algorithms directly perform object classification and localisation tasks on input images. These algorithms are faster, simpler and more efficient [25], with representative examples including the YOLO series and SSD object detection algorithms [26-28]. For example, Andika et al. applied image enhancement techniques such as smoothing and sharpening to a road damage dataset and then used SSD Mobilenet for damage detection, achieving a 2% accuracy improvement compared to the model without image enhancement. Ning et al., based on YOLOv7, replaced standard convolution with a combination of distributed shift convolution (DSConv) and efficient lightweight aggregation network (EDC), and improved the spatial pyramid feature network. This approach achieved a nearly 54% reduction in runtime while enhancing accuracy [29]. Zeng et al. introduced an optimised lightweight road damage detection algorithm built on YOLOv8. By incorporating Ghost convolution and a lightweight shared convolutional detection head, the model achieved a 1.4% increase in

accuracy with a parameter size of only 2.3 M [30]. To overcome the challenge of low detection accuracy in complex background settings, Zhang et al. designed a street-view-based pavement defect detection and statistical system. This system, built on YOLOv8 combined with the DeepSORT tracking algorithm, enables a single camera to perform road damage detection and statistical analysis [31].

2.2 YOLO11

YOLO11, the newest generation of object detection algorithms in the Ultralytics YOLO series, was officially released by the Ultralytics team in September 2024 [32]. YOLO11 introduces an entirely new framework, demonstrating outstanding performance in detection accuracy, speed and efficiency. In comparison to previous versions, YOLO11 integrates lightweight design principles with enhanced feature extraction capabilities, enabling it to effectively capture intricate details from complex scenes. In addition, YOLO11 has achieved significant improvements in processing speed. Testing on the COCO dataset revealed that YOLO11m not only delivers higher mean average precision (mAP) and faster processing speeds compared to YOLOv8m but also reduces the parameter count by 22%. These lightweight and efficient characteristics make YOLO11 an ideal choice for road damage detection tasks [33].

Figure 1 shows the architecture of YOLO11. Similar to its predecessor, YOLO11 employs a layered design, comprising a backbone network, a neck network and a prediction head. The backbone network comprises Conv modules, C3K2 modules, SPPF modules and C2PSA modules, which are responsible for multi-scale feature extraction from the input image. The neck network, incorporating Conv modules and C3K2 modules, integrates multi-scale features from the backbone-generated feature maps, enhancing the model's ability to detect objects across different scales. The head is tasked with generating bounding boxes and performing class predictions. Moreover, YOLO11 is versatile and can accomplish various computer vision tasks like instance segmentation, image classification and pose estimation, simply by replacing the head module [34].

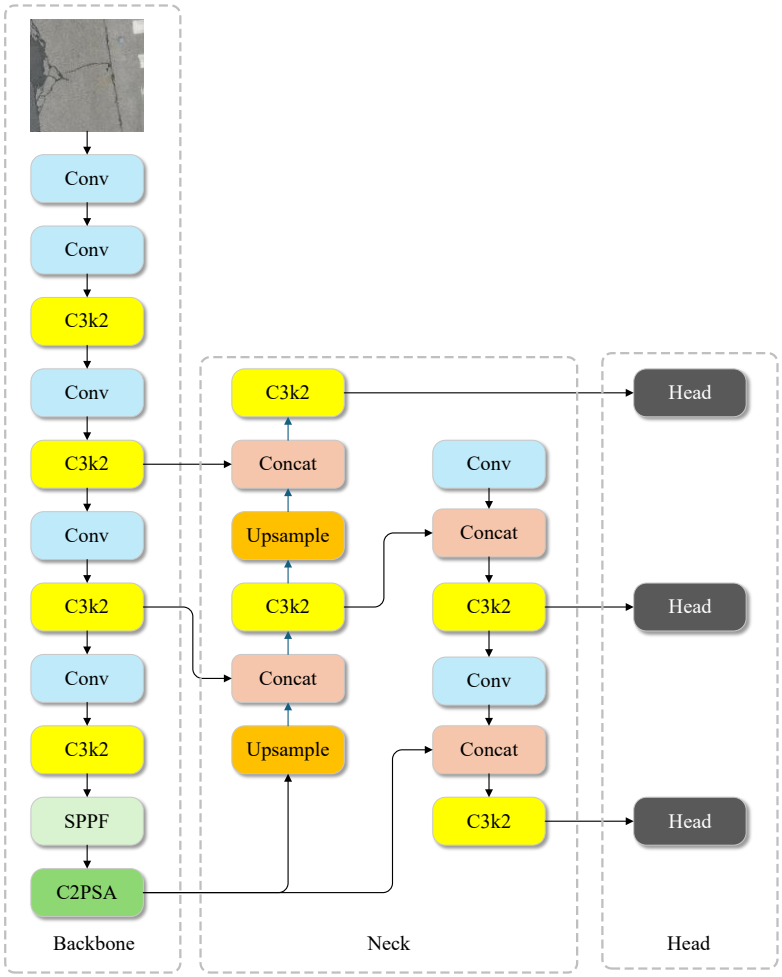


Figure 1 – The structure of YOLO11

Compared to its predecessor, YOLO11 introduces two core innovations. First, the conventional C2f module is substituted with the enhanced C3K2 module. Second, a C2PSA module is added after the SPPF (spatial pyramid pooling – fast) module, significantly improving the model's capacity to extract and amalgamate features. The C3K2 module is an optimised variant of the cross-stage partial network (CSPNet). It employs two smaller convolutional kernels to replace the traditional large kernel, reducing computational cost while enhancing feature representation. Additionally, the C3K2 module allows the kernel size to be adjusted through parameters, enabling flexible adaptation to complex features under different receptive fields and improving the model's proficiency in managing objects of varying sizes [35]. The C2PSA module (cross-stage partial with spatial attention) combines the advantages of CSPNet and spatial attention mechanisms. By introducing spatial attention, the C2PSA module effectively highlights significant features in the target region, strengthening the model's attention on target regions while minimising reliance on extraneous background elements. This mechanism significantly lowers the false positive rate in complex scenarios, improving both detection accuracy and model robustness [36]. These innovations position YOLO11 as a more efficient and adaptable detection model for handling complex and multi-scale visual tasks.

3. METHODOLOGY

Road damage detection faces numerous challenges, such as difficulties in detecting small objects, false positives and missed detections in complex scenarios, and significant differences in the scales of various types of damage. Moreover, the detection accuracy of the original YOLO model remains insufficient for practical applications, limiting its widespread adoption in real-world road scenarios. To overcome these limitations, this study introduces a set of improvements tailored to the YOLO11s. First, a new feature extraction module, RoadRep-C3, is designed by combining the strengths of the RepViT and C3K2 modules. This module improves both feature extraction accuracy and efficiency, allowing the model to more effectively capture key features of road damage. Next, the EMA attention mechanism is integrated into the 7th and 9th layers of the backbone network, boosting the model's ability to detect multi-scale damage features and improving its adaptability to complex conditions. Then, the original neck network is substituted with HyperC2Net, which leverages hypergraph structures to capture high-order semantic relationships in the feature space. This approach empowers the model to cohesively unify information propagated through the backbone network, strengthening its ability to process complex scenes and detect small objects. Finally, slide loss is utilised to address sample distribution imbalance while focusing more on difficult samples, thereby improving the model's performance in detecting small objects. The structure of the improved network is illustrated in *Figure 2*. By integrating these advancements, the model achieves heightened detection fidelity, stability and generalisability, positioning it as a more effective solution for authentic road defect recognition workflows.

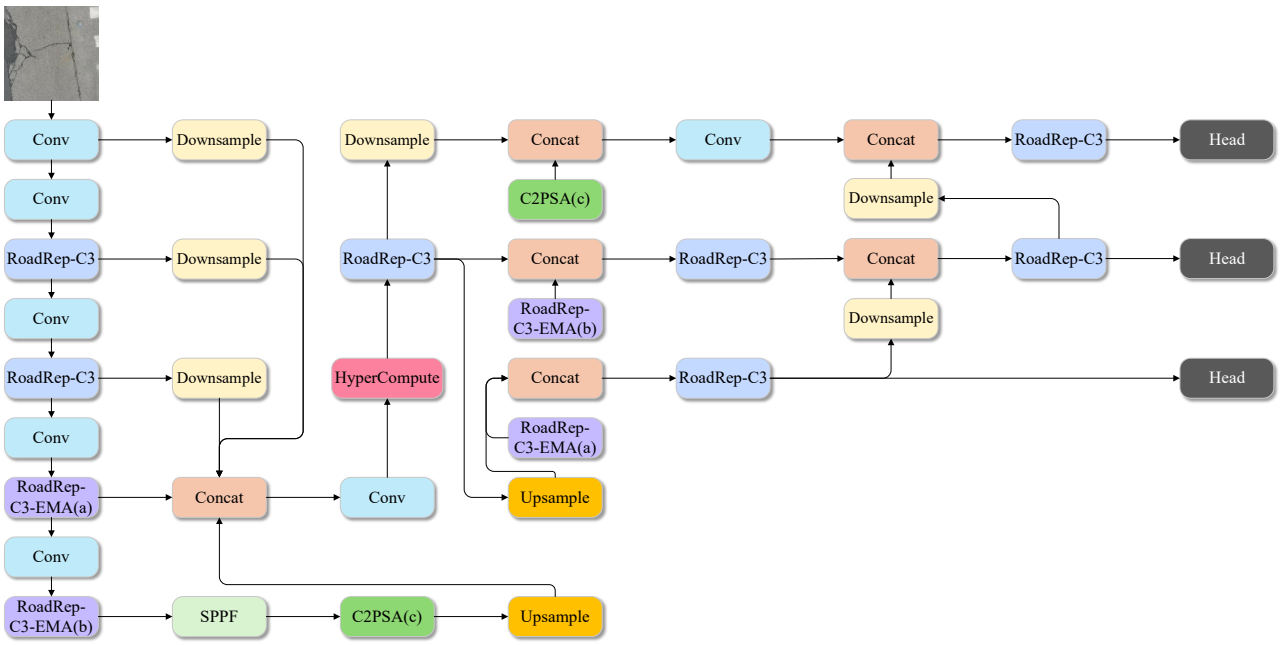


Figure 2 – The structure of improved YOLO11

3.1 RoadRep-C3

To strengthen the model's robustness in identifying targets under diverse and challenging road conditions, this paper proposes a novel feature extraction module called RoadRep-C3. The original C3K2 module, one of the key innovations in YOLO11, replaces a single large convolutional kernel with two smaller ones, thereby improving feature extraction efficiency while reducing computational costs. However, its design primarily focuses on computational efficiency. This limitation may hinder the model's ability to accurately capture detailed spatial features in complex road environments. To address this, the RoadRep-C3 enhances the model's capacity to detect subtle and small features, such as cracks, which are often difficult to capture due to their size and irregular shape. Cracks, unlike larger and prominent objects, require specialised attention from the model. The RoadRep-C3 aids in this by providing better feature extraction capabilities, specifically for fine-grained details.

Wang et al. combined lightweight vision transformer (ViT) structures with traditional CNNs to propose a novel lightweight network called RepViT [37]. The core of the RepViT network lies in its RepViTBlock, which is inspired by the inverted residuals structure in MobileNetV3. The RepViTBlock separates the token mixer, responsible for processing spatial features, from the channel mixer, which manages information exchange along the channel dimension. Unlike traditional CNNs, where convolutional layers simultaneously handle both spatial and channel information, RepViTBlock adopts a ViT-like design by splitting these two operations. A 3x3 depthwise separable convolution is used to extract spatial features, followed by a 1x1 convolution for channel expansion and mapping [38]. This design not only reduces computational overhead but also enhances the network's representational capacity, allowing it to capture key features in complex scenarios more effectively. Additionally, during inference, the multi-branch topology of the token mixer is merged into a single depthwise convolution through a structural re-parameterisation process. This approach effectively reduces computational and storage overhead, thereby improving computational efficiency [39-40]. The network structure of RoadRep-C3 is depicted in *Figure 3*. By replacing the original bottleneck structure with the RepViTBlock and maintaining the adjustable architecture of C3K2, RoadRep-C3 can adapt its internal structure through changes to the C3K parameters. This new feature extraction module achieves stronger feature extraction capability and higher computational efficiency.

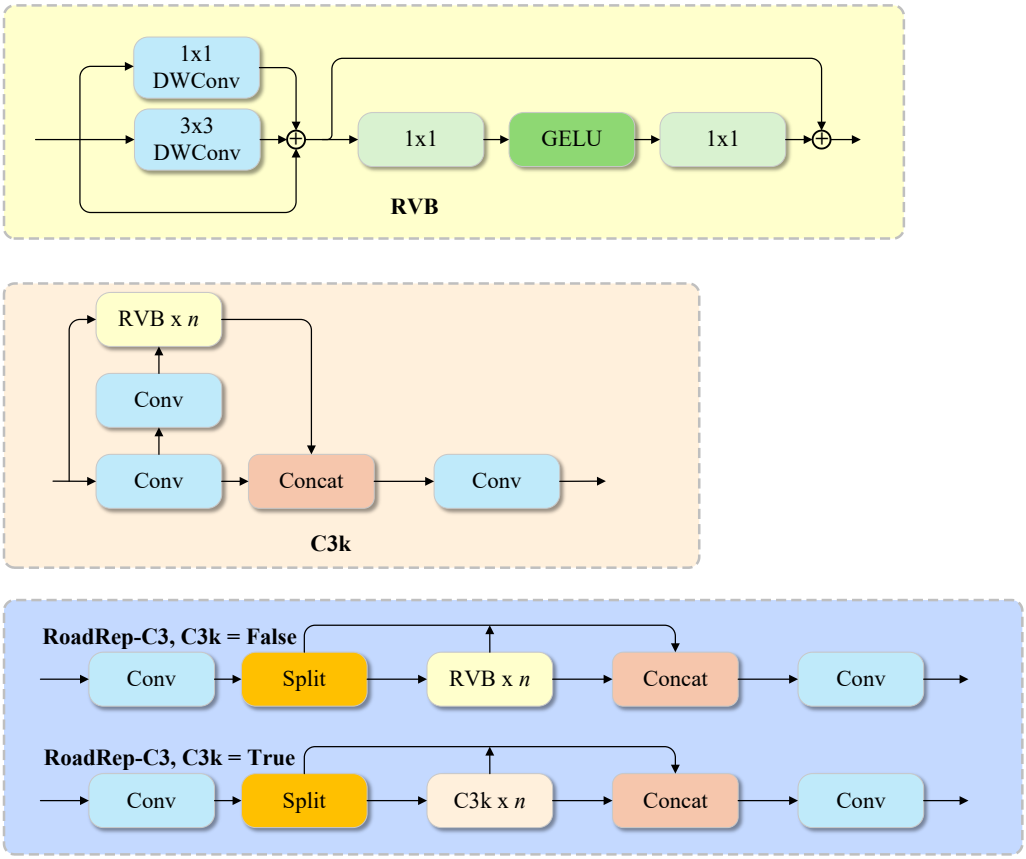


Figure 3 – The structure of RoadRep-C3

3.2 Efficient multi-scale attention mechanism

To tackle the difficulties posed by the significant scale differences in various types of road damage and the complexity of certain road environments, this paper introduces the efficient multi-scale attention (EMA) mechanism [41]. Attention mechanisms dynamically assign different weights to various feature information, strategically amplifying the most discriminative attributes. This refinement optimises feature representation, leading to enhanced detection precision in the model. Unlike traditional channel attention and spatial attention mechanisms, EMA combines the advantages of both. Although channel attention offers high computational efficiency, it fails to capture spatial information, whereas spatial attention can capture spatial relationships but significantly increases computational complexity. EMA achieves an optimal balance between multi-scale feature representation and computational efficiency, preserving hierarchical spatial information while minimising resource-intensive operations. The design of EMA is inspired by the coordinate attention (CA) mechanism, particularly the integration of spatial positional information with channel attention [42]. The structure of EMA is shown in *Figure 4*. EMA employs a multi-scale design: a 1x1 convolution branch captures local inter-channel feature information, while a 3x3 convolution branch captures global feature information. These two branches are further integrated using a cross-space learning approach. The multi-scale feature extraction capability of EMA plays a crucial role in detecting small, subtle features like cracks. Cracks, which often vary in size and shape, can be easily overlooked unless the model effectively captures both fine-grained local details and broader spatial patterns. Cross-space learning fuses data from various spatial dimensions, constructing pixel-level correspondences between the branches. This operation not only enables the integration of global spatial information and local details but also improves the model's ability to effectively capture multi-scale features. Consequently, EMA markedly enhances the model's efficacy in discerning essential features across multiple scales and intricate road environments.

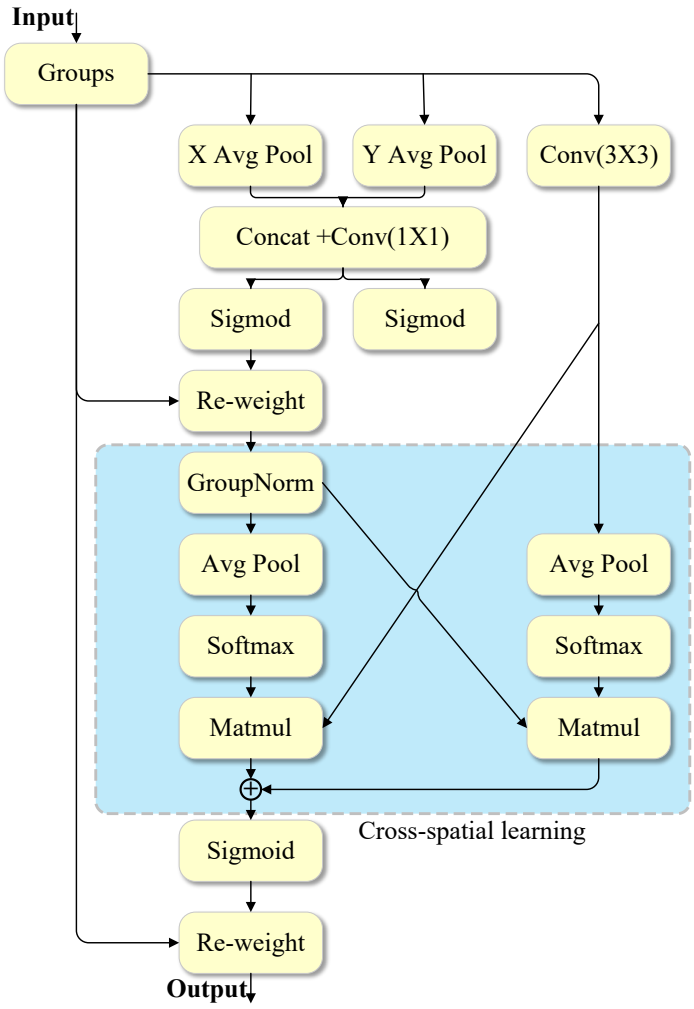


Figure 4 – The structure of EMA

3.3 Improved neck network

HyperC2Net is a cross-layer and cross-position feature representation network based on hypergraph computation, proposed by Feng et al. [43]. By introducing hypergraph computation and high-order information propagation, HyperC2Net substantially improves the model's ability to interpret and analyse road environments, while also boosting its ability to detect small objects. The structure of HyperC2Net is illustrated in *Figure 5*. The network consists of four stages: semantic collecting, hypercompute, semantic scattering and bottom-up.

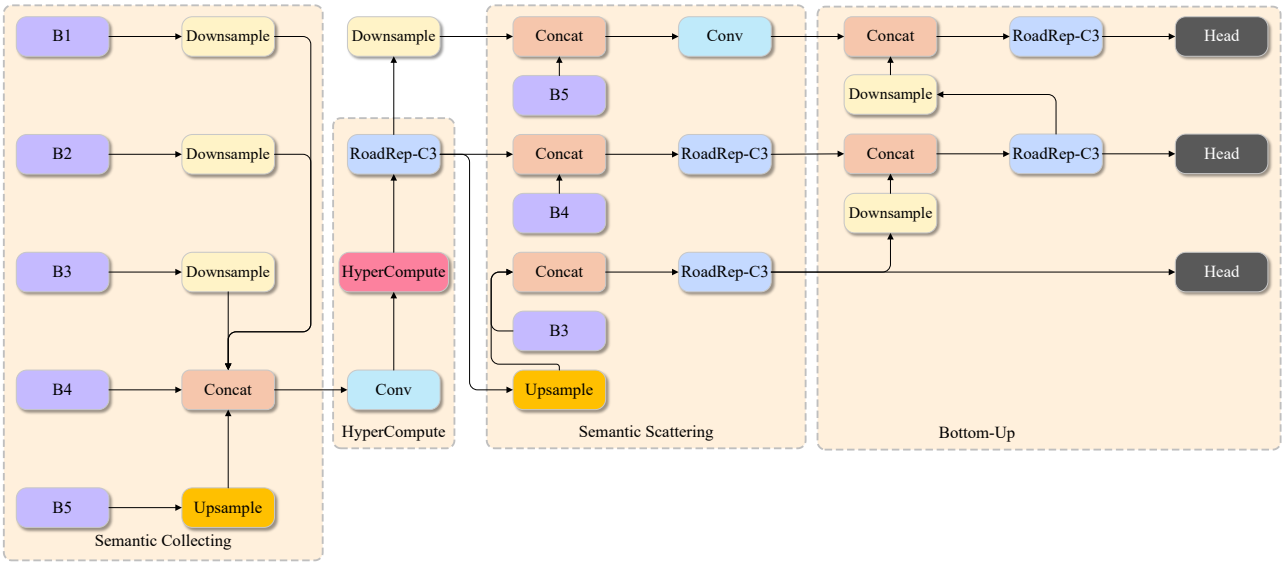


Figure 5 – The structure of HyperC2Net

In the semantic collecting stage, B1 to B5 represent the five feature extraction layers of the backbone network, ranging from shallow to deep layers. The purpose of this stage is to fuse multi-scale feature information from the backbone into a hybrid feature map, which combines features across multiple levels. The core innovation of HyperC2Net lies in its hypergraph computation, which utilises high-order information from the hypergraph to uncover complex feature relationships in the semantic space [44]. A hypergraph G={V,E} is defined by a set of nodes V and a set of hyperedges E. In HyperC2Net, each point in the hybrid feature map is considered a vertex of the hypergraph. Hyperedges are constructed based on the Euclidean distance between feature points: if the distance between two nodes falls below a predefined threshold, they are connected to the same hyperedge. Once the hyperedges are constructed, the next step is to perform hypergraph convolution. The purpose of hypergraph convolution is to propagate information through the hyperedges and capture complex high-order relationships among nodes in the hypergraph. This process enriches the feature representations, particularly in capturing subtle and complex dependencies in the data. The mathematical formula for hypergraph convolution is as follows [45]:

$$\begin{cases} \boldsymbol{x}_e = \dfrac{1}{|\mathcal{N}_v(e)|} \displaystyle\sum_{v \in \mathcal{N}_v(e)} \boldsymbol{x}_v \boldsymbol{\Theta} \\ \boldsymbol{x}'_v = \boldsymbol{x}_v + \dfrac{1}{|\mathcal{N}_e(v)|} \displaystyle\sum_{e \in \mathcal{N}_e(v)} \boldsymbol{x}_e \end{cases} \tag{1}$$

In hypergraph convolution, $\boldsymbol{x}_v$represents the feature vector of node v, indicating its original feature in the hypergraph. $\mathcal{N}_v(e)$ denotes the set of nodes connected by hyperedge e, and $\mathcal{N}_e(v)$ represents the set of hyperedges associated with node v. $\boldsymbol{x}_e$ is the feature vector of hyperedge e, and since a hyperedge can connect more than two nodes, the feature of a hyperedge is an aggregation of the features of all its connected nodes.$|\mathcal{N}_v(e)|$ and $|\mathcal{N}_e(v)|$ denote the number of nodes within hyperedge e and the number of hyperedges connected to node v, respectively. $\boldsymbol{\Theta}$ is a parameter matrix used to learn and adjust the relationship between node features and hyperedge features. The first formula aggregates the features of nodes associated with hyperedge e, resulting in a global feature $\boldsymbol{x}_e$ for the hyperedge. The second formula describes how to update the feature of node v, where the updated feature is determined by taking a weighted average of the original

feature and the features of all hyperedges connected to the node. Thus, hypergraph convolution aggregates the features of multiple nodes via hyperedges, constructing high-order relationships between nodes and enhancing the model's understanding of the semantic space.

After completing hypergraph computation, the semantic scattering stage distributes the generated high-order features across different semantic layers, further achieving cross-level information fusion. The high-order features are fused with the original feature layers from the backbone network, enabling each feature layer to simultaneously incorporate both high-order features and low-level details, enriching the semantic richness of the feature maps. Finally, the bottom-up stage further integrates the multi-level features to produce the final multi-scale feature map, which is then forwarded to the detection head for processing.

3.4 Slide loss

In object detection, the scarcity of hard samples compared to their simpler counterparts often induces a training bias, where models disproportionately emphasise easily learnable features while underrepresenting complex or rare patterns. Consequently, in road damage detection, small, hard-to-identify targets are often overlooked, resulting in frequent missed detections. To mitigate this challenge, this paper introduces the slide loss function. Slide loss is a simple yet efficient weighted loss function that combines traditional loss functions with a weighted mixing mechanism, employing dynamic weight assignment to focus more on learning from hard samples [46]. Its core concept lies in the slide weight function, as defined by the following formula:

$$f(x) = \begin{cases} 1 & x \leq \mu - 0.1 \\ e^{1-\mu} & \mu < x < \mu - 0.1 \\ e^{1-x} & x \geq \mu \end{cases} \tag{2}$$

In the slide weight function, x denotes the IoU (Intersection over Union) between the predicted bounding box and the ground truth box. μ is the average IoU of the samples, acting as a dynamic threshold to categorise samples into three types. When $x \leq \mu - 0.1$, the sample has low overlap with the ground truth and is treated as a negative sample, assigned the lowest weight. When $x \geq \mu$, the sample is considered a positive sample, assigned a higher weight. When $\mu < x < \mu - 0.1$, the sample lies near the decision boundary between positive and negative categories. These samples are the focus of the slide weight function and are assigned the highest weights. Slide loss demonstrates significant improvements in scenarios with imbalanced sample distributions. In the challenging task of road damage detection, slide loss substantially enhances the model's ability to detect small-scale targets and capture intricate road surface patterns. Additionally, its dynamic thresholding characteristic reduces the need for manually setting hyperparameters, enabling automatic optimisation for learning hard samples.

4. EXPERIMENTS

4.1 Dataset

The road damage dataset used in this experiment, RDD2022, forms part of the Crowdsensing-based Road Damage Detection Challenge (CRDDC 2022) [47-48]. It includes road damage images from six countries: the United States, China, India, Japan, Norway and the Czech Republic. For this study, 4,378 images of road damage from China were selected, captured by motorcycle-mounted cameras and low-altitude drones. The collection environments included normal lighting, shadowed conditions and post-rain scenarios.

The dataset comprises five types of damage: vertical cracks (D00), horizontal cracks (D10), alligator cracks (D20), potholes (D40) and patches. The first four categories are road damages that require timely repair, while patches are maintenance objects requiring regular inspection. *Figure 6* shows typical images of the five damage types from the RDD2022 dataset. All images have a resolution of 512x512. *Figure 7* illustrates the distribution of the five damage types, revealing that vertical cracks are the most common road damage, and potholes are the least frequent. The labels in the original dataset follow the Pascal VOC format. To align with the YOLO model requirements, the labels were converted into the YOLO format using a Python script. For the experiment, the dataset was randomly split into training, validation and testing sets with an 8:1:1 ratio, resulting in 3,502 images for the training set, and 438 images each for the validation and testing sets. This setup ensures a balanced evaluation across different stages of model training and testing. Since the YOLO framework incorporates various built-in data augmentation techniques, such as spatial transformations, colour adjustments and mosaic augmentation, no additional data augmentation was applied to the training set.

(a) *(b)* *(c)* *(d)* *(e)*

Figure 6 – The five types of road damage: a) D00; b) D10; c) D20; d) D40; e) Repair

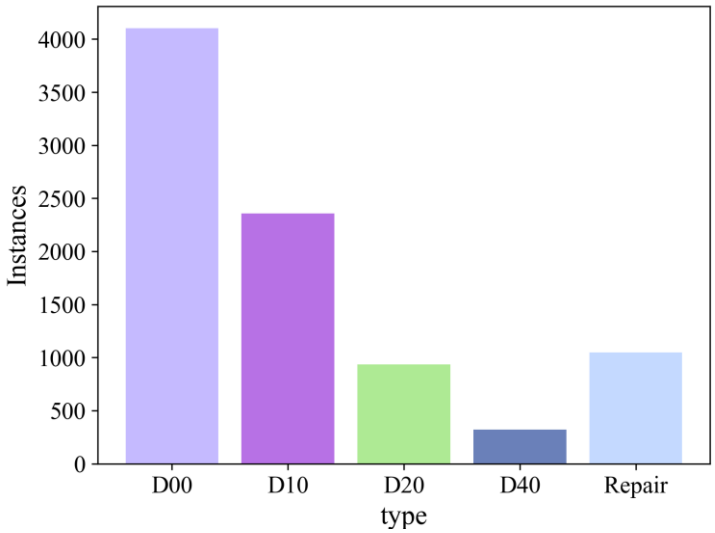


Figure 7 – Distribution of the five types of road damage

4.2 Experimental environment and parameter settings

The computer configuration used for this experiment includes an AMD Ryzen 7 7700 8-Core Processor, 32GB RAM and an NVIDIA GeForce RTX 4080 SUPER GPU with 16GB VRAM. The operating system is Windows 11, and the Python interpreter version is 3.10.15. The deep learning framework is PyTorch 2.5.0, with CUDA 12.4 for GPU acceleration. The hyperparameters used during training are shown in *Table 1*.

Table 1 – Hyperparameter settings

Hyperparameters	Value
Image size	640x640
Learning rate	0.01
Optimiser	SGD
Batch size	32
Workers	8
Epochs	300

4.3 Evaluation metrics

The evaluation metrics used in this experiment include mean average precision (mAP), F1 score, params and GFLOPs. These metrics collectively enable a holistic evaluation of the model's efficacy, robustness and computational efficiency in real-world road damage identification scenarios. mAP and F1 score are the core indicators for measuring the detection accuracy and overall performance of the model, directly reflecting its

detection capability. Parameter count and GFLOPs measure the model's lightweight design and computational complexity, ensuring that the model improves performance without significantly increasing its size, rendering it practical for real-world applications with strict hardware or energy constraints.

The computation of mAP and F1 score requires two intermediate metrics: precision and recall, which are defined as follows:

$$Precision=\frac{TP}{TP+FP} \tag{3}$$

$$Recall=\frac{TP}{TP+FN} \tag{4}$$

TP refers to samples that are correctly identified as positive by the model. FP indicates samples that the model predicts as positive but are actually negative, while FN represents samples that the model incorrectly classifies as negative when they are actually positive (missed detections). Therefore, precision calculates the ratio of true positive samples to all those predicted as positive by the model, whereas recall assesses the proportion of actual positives accurately detected by the model.

To offer a thorough assessment of the model's performance, mAP and F1 score are used, and their formulas are as follows:

$$mAP=\frac{1}{n}\sum_{i=1}^{n}\int_0^1 Precison(Recall)d(Recall) \tag{5}$$

$$F1=2\frac{Precision\cdot Recall}{Precision+Recall} \tag{6}$$

In object detection tasks, mAP is the most widely used performance metric, providing an overall measure of detection accuracy across multiple categories. The F1 score, defined as the harmonic mean of precision and recall, reflects the model's balance between missed detections (FN) and false detections (FP). A higher F1 score signifies better performance. Parms represents the number of learnable parameters in the model, while GFLOPs refers to the number of billions of floating-point operations the model performs per second.

4.4 Ablation experiments

To investigate whether all proposed improvement modules are effective for road damage detection, ablation experiments were conducted on the RDD2022 dataset. The results are summarised in *Table 2*, where A represents the RoadRep-C3 module, B represents RoadRep-C3 with EMA attention, C represents the improved neck network and D represents the slide loss. The training process of the original YOLO11 model and the improved YOLO11 model is illustrated in *Figure 8*.

The RoadRep-C3 module was designed to enhance feature extraction by focusing on critical road damage characteristics, such as cracks and textures, which are typically challenging to detect under complex road conditions. Experimental results show that integrating the RoadRep-C3 module led to a 1.2% increase in mAP@0.5, while the model's parameter count and computational cost decreased by 1.22 M and 2.3 GFLOPs, respectively. These results demonstrate the module's effectiveness in improving detection accuracy while simultaneously reducing model complexity. Theoretically, the EMA mechanism enhances multi-scale feature representation by dynamically adjusting attention across various scales, thus improving the model's capability to handle road damages of varying sizes. Experimental results highlight that the EMA mechanism led to the most significant improvement in the F1 score, validating its effectiveness in enhancing feature representation across different damage scales. The improved neck network, which incorporates hypergraph-based computation, enables more efficient fusion of features from different stages of the network. Theoretically, this approach improves the model's ability to handle complex road damage scenarios by combining fine-grained features with higher-level contextual information. However, experimental results show that while the mAP increased to 86.4%, the parameter count and computational cost also increased significantly. This trade-off suggests that while hypergraph-based computation enhances performance in complex scenarios, it also demands higher computational resources. Experimental results show that the addition of slide loss led to improvements in detection accuracy. The dynamic weight adjustment mechanism allowed the model to focus on harder samples, which would have otherwise been overlooked due to their proximity to the decision

boundary. Moreover, slide loss helped alleviate issues related to sample imbalance, ensuring that rare and small road damage instances were more effectively learned by the model.

Combining all the proposed modules resulted in a complete model with an mAP@0.5 of 87.0%, a 2% improvement compared to the baseline, while reducing the parameter count by 0.18 M. The ablation experiment results demonstrate that all the proposed modules effectively improve and enhance the model's detection capability in road damage detection tasks.

Table 2 – Ablation experiment results

Model	A	B	C	D	mAP@0.5	F1/%	Params/M	GFLOPs
YOLO11s					85.0	81.2	9.41	21.3
YOLO11s	√				86.2	81.6	8.19	19.0
YOLO11s	√	√			86.4	83.0	8.20	19.1
YOLO11s	√	√	√		86.8	82.1	9.23	21.5
YOLO11s	√	√	√	√	87.0	83.2	9.23	21.5

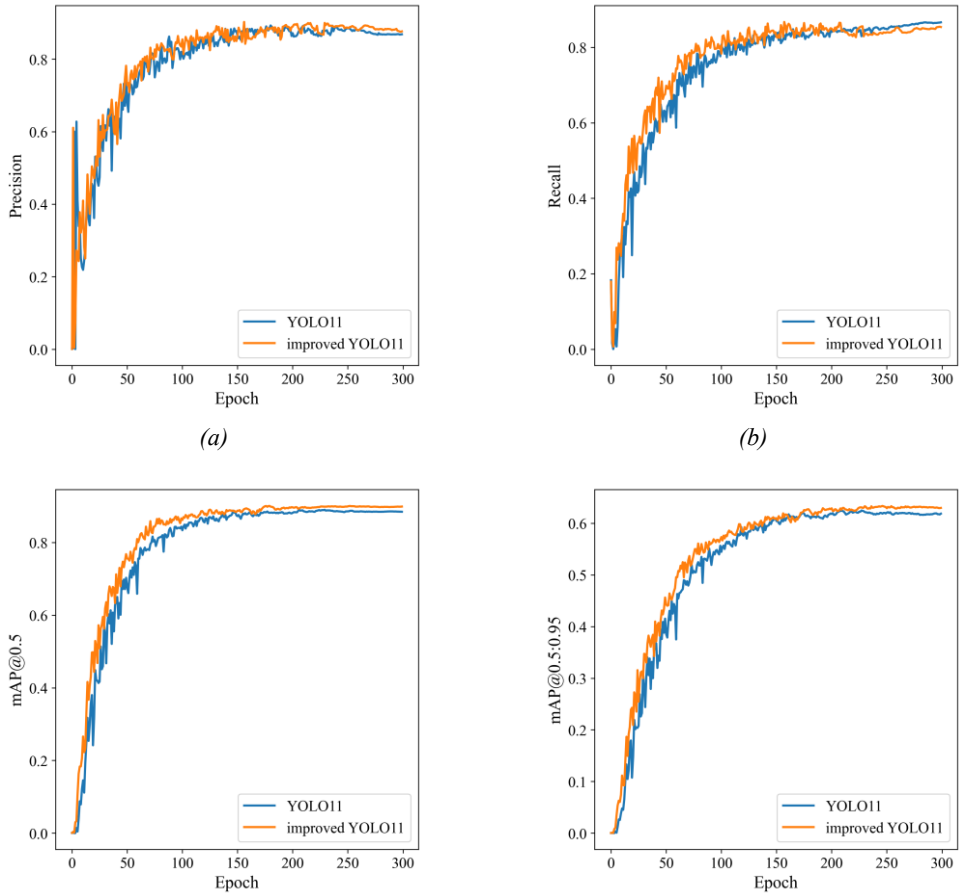


Figure 8 – Training process of YOLO11 and the improved YOLO11: a) Precision curve; b) Recall curve; c) mAP@0.5 curve; d) mAP@0.5:0.95 curve

4.5 Comparison experiments

To further assess the practical effectiveness of the proposed improvements, comparative experiments were performed on the RDD2022 dataset with mainstream algorithms. The improved YOLO11 proposed in this paper was compared with other state-of-the-art object detection models. The YOLO11 was used as the baseline, and the comparison included YOLO series models such as YOLOv8s, YOLOv9s and YOLOv10s, as

well as the Faster-RCNN based on a two-stage detection framework and RT-DETR-L, which is transformer-based. As seen in *Table 2*, the improved YOLO11 significantly outperforms other models in mAP@0.5, F1 score and parameter count, while the GFLOPs only increased by 0.2. Considering both accuracy and computational complexity, the proposed model delivers excellent detection capabilities while preserving a streamlined and efficient structure, making it a balanced and efficient solution for road damage detection tasks.

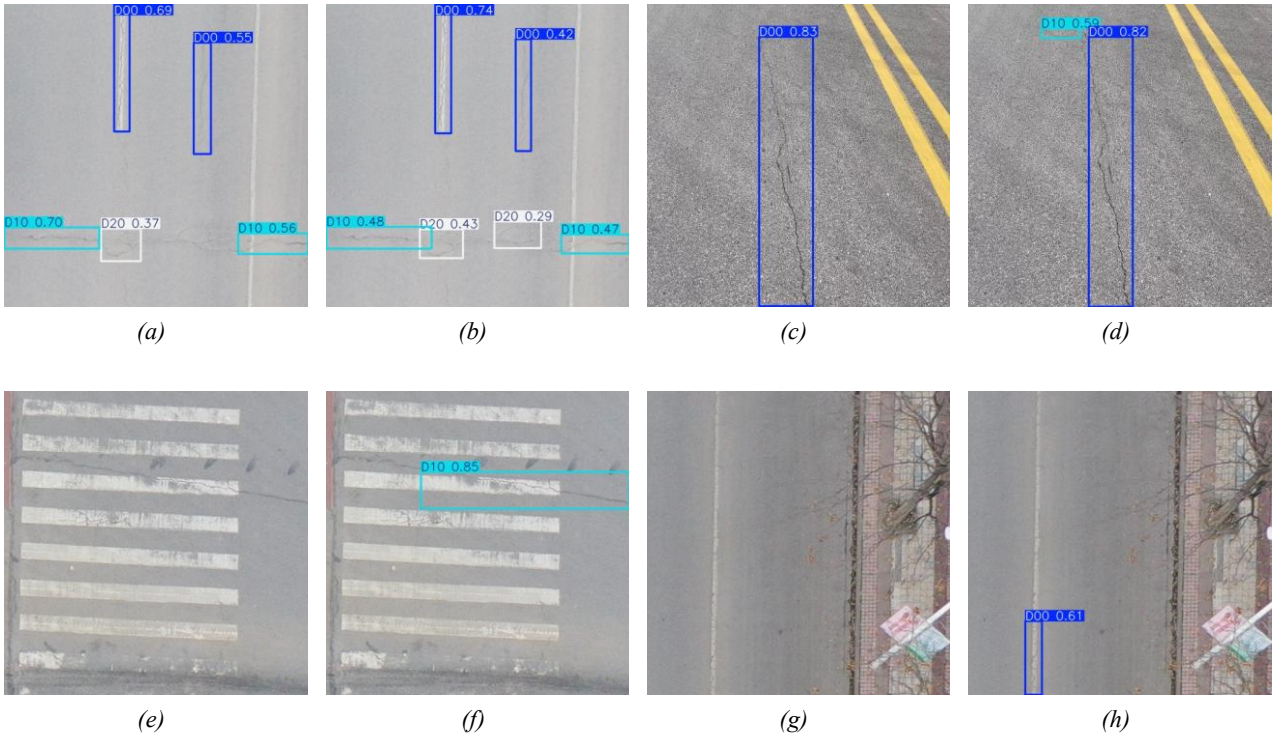
Table 3 – Comparison of experiment results

Model	mAP@0.5/%	F1/%	Params/M	GFLOPs
Faster-RCNN	*77.6*	*63.0*	*41.37*	-
RT-DETR-L	*81.2*	*77.2*	*28.45*	*100.6*
YOLOv8s	*83.9*	*81.6*	*11.12*	*28.4*
YOLOv9s	*84.5*	*80.7*	*7.17*	*26.7*
YOLOv10s	*81.9*	*78.5*	*8.04*	*24.5*
YOLO11s	*85.0*	*81.2*	*9.41*	*21.3*
Ours	*87.0*	*83.2*	*9.23*	*21.5*

4.6 Visualisation

To visually assess detection performance, a comparative visualisation experiment was performed between the improved YOLO11 and the original YOLO11, focusing on four typical scenarios. The detection results are presented in *Figure 9*. In *Figure 9a-9b,* in addition to the prominent road cracks, minor fissures are also visible. In *Figure 9e-9h*, the cracks are concealed within the zebra crossings or lane markings. *Figure 9i-9l* depicts a complex road environment, where lane markings, manhole covers and speed bumps overlap, with cracks or repair patches interspersed. *Figure 9m-9p* displays scenarios involving occlusions caused by vegetation, vehicles and shadows.

In these complex scenarios, the original YOLO11 failed to detect some targets, resulting in missed detections, while the improved YOLO11 successfully identified these targets. The experimental results indicate that the improved YOLO11 model performs better in complex scenarios, offering higher accuracy for small targets and significantly reducing missed detection rates.



(a) *(b)* *(c)* *(d)*

(e) *(f)* *(g)* *(h)*

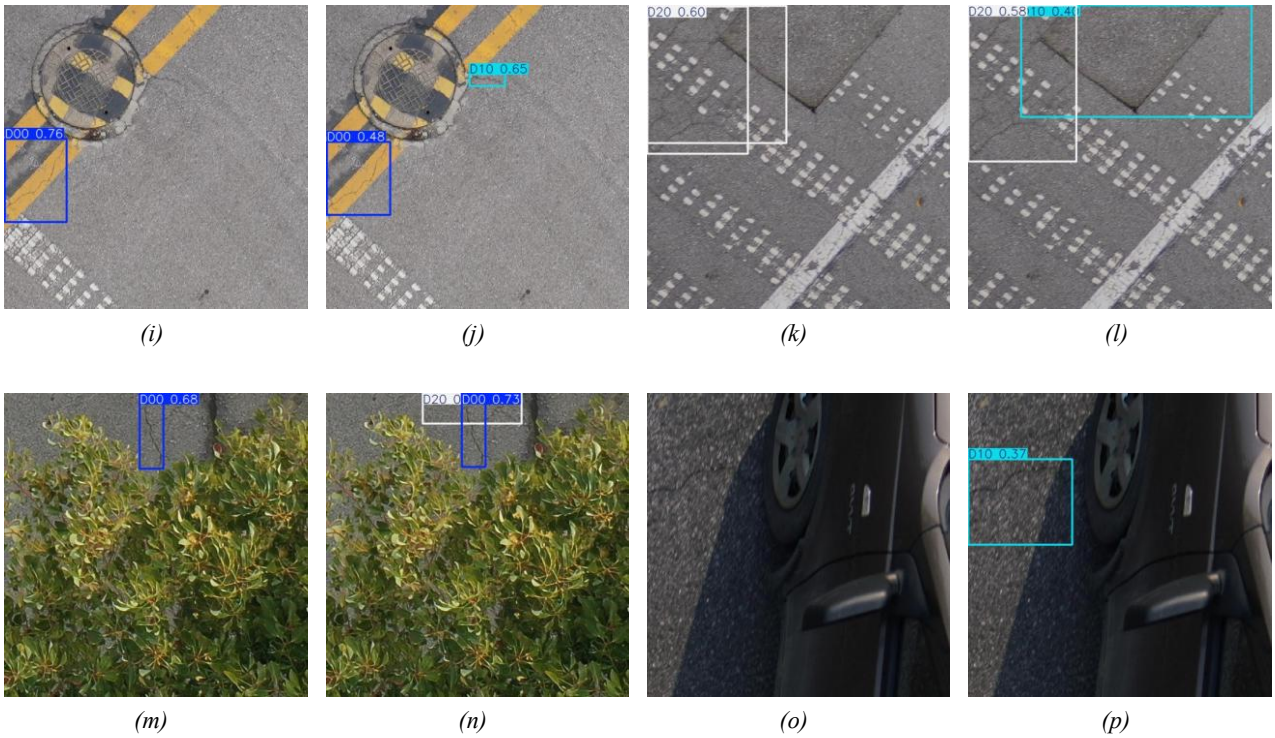


Figure 9 – Comparison of detection results between YOLO11 and the improved YOLO11: a) YOLO11; b) Improved YOLO11; c) YOLO11; d) Improved YOLO11; e) YOLO11; f) Improved YOLO11; g) YOLO11; h) Improved YOLO11; i) YOLO11; j) Improved YOLO11; k) YOLO11; l) Improved YOLO11; m) YOLO11; n) Improved YOLO11; o) YOLO11; p) Improved YOLO11

5. DISCUSSION

This study proposes an improved YOLO11 model for road damage detection, demonstrating significant performance enhancements in detecting small objects, handling complex road environments and addressing multi-scale road damage features. The experimental results show a 2% improvement in mAP@0.5 compared to the original YOLO11 model, with particularly noticeable gains in challenging scenarios, such as detecting small cracks and damage hidden in complex backgrounds like road markings or shadow-covered damage.

In terms of dataset selection, this study uses the Chinese subset of the RDD2022 dataset, which includes two collection methods: motorcycle-mounted cameras and low-altitude drones. This multi-perspective data collection approach provides a more comprehensive set of damage information for the model to learn from. Road conditions, climate and driving habits in different regions can significantly impact the formation of road damage, leading to various types and characteristics of road damage. By using the Chinese dataset, this study eliminates the interference from diverse regional road materials, allowing for an evaluation of the model's performance improvement in a relatively consistent and controlled environment, ensuring the reliability of the results. Nevertheless, relying on data from a single region may limit the diversity of damage types and styles. Future studies can expand the dataset to test the model's robustness across different regions.

In practical applications, the improved YOLO11 model shows considerable potential. The improved model has a lightweight design with only 21.5 GFLOPs, making it suitable for deployment on edge devices or drones for real-time road damage detection. The inclusion of drone-collected images in training allows for the potential direct deployment of this model on drones for road inspection. Drone-based periodic and scheduled patrols will enable all-weather, real-time road damage monitoring, ensuring timely identification of damage and the implementation of preventive measures. Such advancements aim to reduce accidents and provide more reliable technological support for intelligent transportation systems and infrastructure maintenance.

6. CONCLUSIONS

This paper introduces a road damage detection model based on YOLO11, overcoming challenges such as small-object detection difficulties, interference from intricate backgrounds and significant multi-scale target differences in road damage detection tasks. Several optimisation modules were introduced to improve the

original YOLO11. First, the RoadRep-C3 module was incorporated to significantly enhance the precision and efficiency of the feature extraction process. Second, the integration of the EMA mechanism strengthened the model's capacity to identify and process multi-scale damage characteristics. Third, the hypergraph structure was utilised in the neck network to capture high-order information in the semantic space, achieving cross-stage information fusion and feature enhancement, which effectively improved the detection capability for small targets. Finally, the slide loss function was introduced to address the disparity between simple and complex samples, thereby boosting the model's efficacy in extracting insights from challenging data instances.

Experimental results demonstrate that the improved YOLO11 achieves a 2% increase in mAP on the RDD2022 dataset while reducing parameter count by 0.18 M. Moreover, in various complex scenarios, such as small objects, occlusions and complex backgrounds, the improved model outperformed the original YOLO11, with a significant reduction in missed detections.

REFERENCES

[1] Ji A, et al. Scientometric analysis of pavement maintenance: A twenty-year review. *Journal of Civil Engineering and Management*. 2023;29(5):439–462. DOI: 10.3846/jcem.2023.19031.

[2] Rathee M, Bačić B, Doborjeh, M. Automated road defect and anomaly detection for traffic safety: A systematic review. *Sensors*. 2023;23(12):5656. DOI: 10.3390/s23125656.

[3] Zhou Y, et al. Review of intelligent road defects detection technology. *Sustainability*. 2022;14(10):6306. DOI: 10.3390/su14106306.

[4] Ragnoli A, De Blasiis MR, Di Benedetto A. Pavement distress detection methods: A Review. *Infrastructures*. 2018;3(4):58. DOI: 10.3390/infrastructures3040058.

[5] Yang X, et al. Automation in road distress detection, diagnosis and treatment. *Journal of Road Engineering*. 2024;4(1):1–26. DOI: 10.1016/j.jreng.2024.01.005.

[6] Rizelioğlu, M. An extensive bibliometric analysis of pavement deterioration detection using sensors and machine learning: Trends, innovations, and future directions. *Alexandria Engineering Journal*. 2025;112:349–366. DOI: 10.1016/j.aej.2024.09.097.

[7] Zhang AA, et al. Intelligent pavement condition survey: Overview of current researches and practices. *Journal of Road Engineering*. 2024;4(3):257–281. DOI: 10.1016/j.jreng.2024.04.003.

[8] Yu J, et al. Road surface defect detection—From image-based to non-image-based: A survey. *IEEE Transactions on Intelligent Transportation Systems*. 2024;25(9):10581–10603. DOI: 10.1109/tits.2024.3382837.

[9] Bonilla M, et al. A review of practices for municipal road maintenance construction funding allocation in the United States. *Proceedings of the Construction Research Congress* 2020, 8-11 Mar. 2020, Tempe, AZ, USA. 2020. p. 832-840. DOI: 10.1061/9780784482889.088.

[10] Zheng L, et al. Deep learning-based intelligent detection of pavement distress. *Automation in Construction*. 2024;168:105772. DOI: 10.1016/j.autcon.2024.105772.

[11] Safyari Y, Mahdianpari M, Shiri, H. A review of vision-based pothole detection methods using computer vision and machine learning. *Sensors*. 2024;24(17):5652. DOI: 10.3390/s24175652.

[12] Kothai R, et al. Pavement distress detection, classification, and analysis using machine learning algorithms: A survey. *IEEE Access*. 2024;12: 126943–126960. DOI: 10.1109/access.2024.3455093.

[13] Wang Q, et al. Fusing visual quantified features for heterogeneous traffic flow prediction. *Promet - Traffic& Transportation*. 2024;36(6):1068–1077. DOI: 10.7307/ptt.v36i6.667.

[14] Wu X, et al. Vessel trajectory prediction method based on the time series data fusion model. *Promet - Traffic&Transportation*. 2024;36(6):1160–1175. DOI: 10.7307/ptt.v36i6.772.

[15] Du Z, et al. Application of image technology on pavement distress detection: A review. *Measurement*. 2021;184:109900. DOI: 10.1016/j.measurement.2021.109900.

[16] Li L, et al. Automatic pavement crack recognition based on BP Neural Network. *PROMET - Traffic&Transportation*. 2014;26(1):11–22. DOI: 10.7307/ptt.v26i1.1477.

[17] Zou Z, et al. Object detection in 20 years: A survey. Proceedings of the IEEE. 2023;111(3):257–276. DOI: 10.1109/jproc.2023.3238524.

[18] Zhao ZQ, et al. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*. 2019;30(11):3212–3232. DOI: 10.1109/tnnls.2018.2876865.

[19] Du L, Zhang R, Wang X. Overview of two-stage object detection algorithms. *Journal of Physics: Conference Series*. 2020;1544(1):012033. DOI: 10.1088/1742-6596/1544/1/012033.

[20] Girshick R, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 23-28 June 2014, Columbus, OH, USA*. 2014. p. 580-587. DOI: 10.1109/CVPR.2014.81.

[21] Ren S, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017;39(6):1137–1149. DOI: 10.1109/tpami.2016.2577031.

[22] He K, et al. Mask R-CNN,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2018;42(2):386–397. DOI: 10.1109/tpami.2018.2844175.

[23] Li Q, et al. The improvement of faster-RCNN crack recognition model and parameters based on attention mechanism. *Symmetry*. 2024;16(8):1027. DOI: 10.3390/sym16081027.

[24] Li L, et al. Road pothole detection based on crowdsourced data and extended mask R-CNN. *IEEE Transactions on Intelligent Transportation Systems*. 2024;25(9):12504–12516. DOI: 10.1109/tits.2024.3360725.

[25] Deng J, et al. A review of research on object detection based on deep learning. *Journal of Physics: Conference Series*. 2020;1684(1):012028. DOI: 10.1088/1742-6596/1684/1/012028.

[26] Redmon J, et al. You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 27-30 June 2016, Las Vegas, NV, USA*. 2016. p. 779-788. DOI: 10.1109/CVPR.2016.91.

[27] Terven J, Córdova-Esparza DM, Romero-González JA. A comprehensive review of YOLO architectures in Computer Vision: From YOLOV1 to Yolov8 and Yolo-Nas. *Machine Learning and Knowledge Extraction*. 2023;5(4):1680–1716. DOI: 10.3390/make5040083.

[28] Liu W, et al. SSD: Single shot multibox detector. *Proceedings of the European Conference on Computer Vision, 11-14 Oct. 2016, Amsterdam, The Netherlands*. 2016. p. 21-37. DOI: 10.1007/978-3-319-46448-0.

[29] Ning Z, et al. Yolov7-RDD: A lightweight efficient pavement distress detection model. *IEEE Transactions on Intelligent Transportation Systems*. 2024;25(7):6994–7003. DOI: 10.1109/tits.2023.3347034.

[30] Zeng J, Zhong H. Yolov8-PD: An improved road damage detection algorithm based on yolov8n model. *Scientific Reports*. 2024;14(1):12052. DOI: 10.1038/s41598-024-62933-z.

[31] Zhang Z, et al. Detection and statistics system of pavement distresses based on street view videos. *IEEE Transactions on Intelligent Transportation Systems*. 2024;25(10):15106–15115. DOI: 10.1109/tits.2024.3401150.

[32] Ultralytics. YOLO11 NEW - Ultralytics YOLO Docs. 2025. Available at: https://docs.ultralytics.com/models/yolo11/ [Accessed 25th Feb. 2025].

[33] Khanam R, Hussain M. YOLOv11: An overview of the key architectural enhancements. *arXiv*. 2024. Available at: https://arxiv.org/abs/2410.17725 [Accessed 25th Feb. 2025].

[34] Jegham N, et al. YOLO Evolution: A comprehensive benchmark and architectural review of YOLOv12, YOLOv11, and their previous versions. *arXiv*. 2025. Available at: https://arxiv.org/abs/2411.00201 [Accessed 25th Feb. 2025].

[35] Alif MAR. YOLOv11 for vehicle detection: Advancements, performance, and applications in Intelligent Transportation Systems. *arXiv*. 2024. Available at: https://arxiv.org/abs/2410.22898 [Accessed 25th Feb. 2025].

[36] Wei J, et al. GFS-yolo11: A maturity detection model for Multi-Variety Tomato. *Agronomy*. 2024;14(11): 2644. DOI: 10.3390/agronomy14112644.

[37] Ding X, et al. RepVGG: Making VGG-style convnets great again. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 19-25 June 2021, Nashville, TN, USA*. 2021. p. 13728-13737. DOI: 10.1109/CVPR46437.2021.01352.

[38] Chollet F. Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 21-26 July 2017, Honolulu, HI, USA*. 2017. p. 1800-1807. DOI: 10.1109/CVPR.2017.195.

[39] Ding X, et al. ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks. *Proceedings of the IEEE/CVF International Conference on Computer Vision, 27 Oct - 2 Nov 2019, Venice, Italy*. 2019. p. 1911-1920. DOI: 10.1109/ICCV.2019.00200.

[40] Wang A, et al. RepVIT: Revisiting mobile CNN from ViT perspective. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16-22 June 2024, Seattle, WA, USA*. 2024. p. 15909-15920. DOI: 10.1109/CVPR52733.2024.01506.

[41] Ouyang D, et al. Efficient multi-scale attention module with cross-spatial learning. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 15-19 May 2023, Rhodes, Greece*. 2023. p. 1-5. DOI: 10.1109/ICASSP49357.2023.10096516.

[42] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 19-25 June 2021, Nashville, TN, USA*. 2021. p. 13713-13722. DOI: 10.1109/CVPR46437.2021.01350.

[43] Feng Y, et al. Hyper-YOLO: When visual object detection meets hypergraph computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2024:1–14. DOI: 10.1109/tpami.2024.3524377.

[44] Feng Y, et al. Hypergraph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019;33(01):3558–3565. DOI: 10.1609/aaai.v33i01.33013558.

[45] Bai S, Zhang F, Torr HS. Hypergraph convolution and hypergraph attention. *Pattern Recognition*. 2021;110: 107637. DOI: 10.1016/j.patcog.2020.107637.

[46] Yu Z, et al. Yolo-FaceV2: A scale and occlusion aware face detector. *Pattern Recognition*. 2024;155:110714. DOI: 10.1016/j.patcog.2024.110714.

[47] Arya D, et al. Crowdsensing-based road damage detection challenge (CRDDC'2022). *Proceedings of the IEEE International Conference on Big Data, 9-12 Dec. 2022, Osaka, Japan*. 2022. p. 6378-6386. DOI: 10.1109/BIGDATA55660.2022.10021040.

[48] Arya D, et al. RDD2022: A multi-national image dataset for automatic road damage detection. *Geoscience Data Journal*. 2024;11(4):846-862. DOI: 10.1002/gdj3.260.