ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0016

A Lightweight Road Damage Detection Model based on Structural Improvement and Knowledge Distillation

Xiaohui Shi

College of Metropolitan Transportation, Beijing University of Technology, Beijing 100124, China

Abstract

To address the problem of large model parameters in existing high-precision road damage detection models, this paper proposes a lightweight detector called YOLO-LWD based on structural optimization and knowledge distillation. First, standard convolutions in the backbone are replaced with GhostConv, an efficient channel attention (ECA) module is introduced into the neck, and the detection head is replaced with DyHead, with its channel number reduced from 512 to 256. Then, a hybrid knowledge distillation method is adopted to transfer soft labels from the output layer and feature knowledge from intermediate layers of a teacher model to the student model. Experiments on the Aug-RDD dataset show that YOLO-LWD achieves 78.4% mAP@0.5 with only 8.1 MB parameters and 15.8 GFLOPs, which outperforms current mainstream models in both model complexity and detection accuracy.

Keywords

Road Damage Detection; Lightweight Model; Knowledge Distillation; YOLO.

1. Introduction

China has made significant progress in developing its transportation infrastructure, with its highway network continuously expanding. By the end of 2024, the total mileage of highways in China had reached 5.4904 million kilometers. However, during service, roads are subject to various factors such as traffic volume, vehicle load, temperature variation, humidity fluctuation, and natural weathering, gradually leading to defects such as cracks and potholes [1]. These defects not only accelerate road aging but also significantly reduce driving comfort, increase vehicle wear, and pose potential threats to road safety. Consequently, timely detection of road damage is essential for reducing maintenance costs and ensuring road safety [2]. Traditional road damage detection methods mainly rely on manual visual inspection, which suffers from long detection cycles, high costs, and strong subjectivity, making it difficult to meet the demands of rapid and large-scale automated detection in modern road management [3]. In recent years, the rapid development of deep learning has enabled image recognition methods that no longer require manual feature design, achieving excellent recognition accuracy and efficiency, thus providing a new approach for road damage detection [4]. Traditional object detection models are often parameter-heavy and computationally complex. Therefore, lightweight improvement of these models has become a key research direction for deployment on resource-constrained mobile devices. Current lightweight design of object detection algorithms primarily relies on network structure optimization, which reduces the number of parameters and computational complexity by redesigning neural network architectures. This approach typically involves introducing efficient components such as depthwise separable convolutions, bottleneck structures, or attention mechanisms. Typical representatives include MobileNet proposed by Howard et al. [5] and EfficientNet proposed by Tan et al. [6]. The former significantly reduces parameters and computation through depthwise separable convolutions, laying a foundation for subsequent efficient

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0016

architectures, while the latter further improves detection efficiency using compound scaling and neural architecture search. Chen et al. [7] provided a systematic review of lightweight deep convolutional neural networks, summarizing recent research progress in this field.

Recently emerged Transformer-based architectures, such as the Detection Transformer (DETR), offer an end-to-end detection scheme without non-maximum suppression. However, they often underperform on small-sized defects and suffer from long training cycles and slow convergence, which limits their applicability in scenarios requiring rapid iteration and deployment [8].

2. Model Lightweight Strategy

2.1 Model Structure Improvement

2.1.1 Ghost Convolution

In the backbone network, this study replaces all standard convolutional layers with Ghost Convolution (GhostConv) to maintain the model's feature representation capability while significantly reducing model complexity.

The core idea of GhostConv is to generate "ghost" feature maps through depthwise separable convolution and cheap linear operations, thereby reducing redundant feature generation in conventional convolution. The convolution structure is shown in Fig. 1.

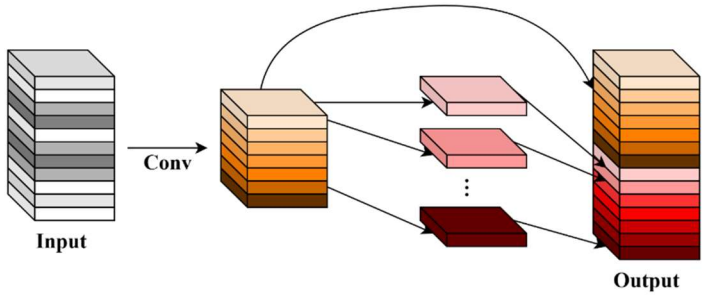

Fig. 1 Schematic diagram of GhostConv

The operation of GhostConv can be divided into two stages: first, a partial set of feature maps is generated using conventional convolution; then, the remaining feature maps are generated via lightweight linear transformations (such as depthwise convolution). This process can be formally expressed as:

$$Y = \text{Concat}\left(Y', \Phi(Y')\right) \tag{1}$$

where $Y' = \text{Conv}_m(X) \in \mathrm{R}^{H \times W \times m}$ is the output of conventional convolution, and $\Phi(\cdot)$ denotes the linear transformation operation. The final number of output channels is $s \times m$, where s is the expansion coefficient, which is set to 2 in this study.

2.1.2 Efficient Channel Attention Mechanism

To enhance feature representation while maintaining computational efficiency, this paper introduces the Efficient Channel Attention (ECA) mechanism into the neck network. ECA captures cross-channel interactions through lightweight one-dimensional convolution without dimensionality reduction, preserving channel information while keeping parameters to a minimum. This mechanism enables the model to adaptively recalibrate channel features, improving road damage detection performance without significantly increasing model complexity.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0016

The ECA mechanism eliminates the dimensionality reduction operation commonly used in traditional channel attention mechanisms. It directly captures cross-channel interactions through one-dimensional convolution, avoiding the negative impact of dimensionality reduction on channel attention. The schematic diagram is shown in Fig. 2.

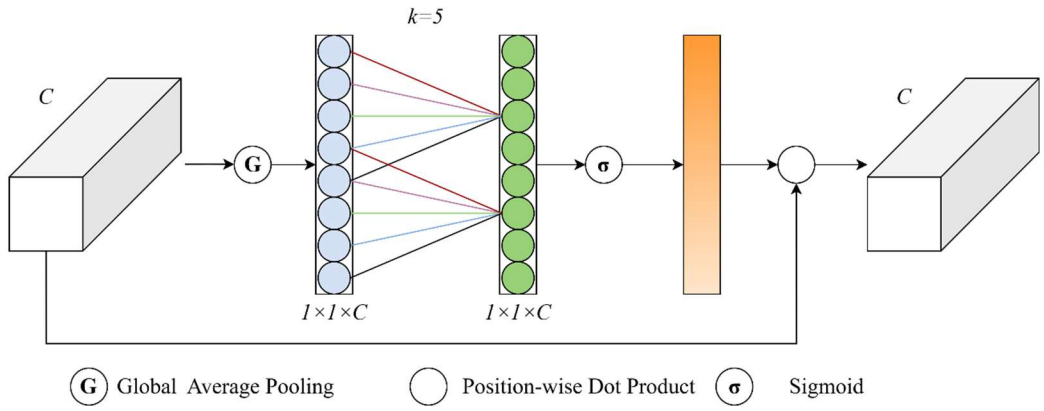

Fig. 2 Schematic diagram of the ECA mechanism (with k=5 as an example)

First, ECA performs global average pooling on the input features, then generates channel weights via 1D convolution, and finally normalizes them using the Sigmoid function. The process can be expressed as:

$$\omega = \sigma\left(\mathrm{Conv1D}_k\left(g(X)\right)\right) \tag{2}$$

where $g(X)$ denotes the global average pooling operation, $\mathrm{Conv1D}_k$ denotes 1D convolution with kernel size k, and σ denotes the Sigmoid activation function.

2.1.3 Lightweight Design of Detection Head

The original detection head of YOLOv8s adopts a decoupled structure, with separate branches for classification and localization to perform object detection tasks. However, in practical applications, this detection head has certain limitations. First, YOLOv8s uses PANet for multi-scale feature fusion, but the fusion process is relatively fixed and lacks adaptive weight adjustment, making it difficult to meet the feature requirements of different types of targets. Additionally, although the classification and localization branches are independent, they share the same convolutional structure and lack dynamic optimization mechanisms tailored to their respective tasks.

To address these issues, this study introduces Dynamic Head (DyHead) to replace the original detection head, as shown in Fig. 3. DyHead integrates scale-aware, space-aware, and task-aware attention mechanisms, demonstrating excellent performance in the classification and localization of complex road damage. However, its multi-dimensional attention computation and large number of convolutional parameters introduce significant computational overhead. Therefore, after adopting DyHead, this study further performs lightweight optimization on it.

The core idea of lightweight detection head design is to reduce channel redundancy in feature maps. In object detection tasks, the parameter count of the detection head is proportional to the square of the input channel dimension. Since GhostConv in the backbone and ECA in the neck have already effectively compressed and enhanced the features, the feature maps fed into the detection head have high semantic compactness and low channel redundancy. Based on this characteristic, this study reduces the internal channel number of DyHead from 512 to 256 by applying a channel reduction

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0016

factor. This optimization significantly reduces model complexity while preserving the original decoupled structure, meeting the real-time requirements for practical deployment.

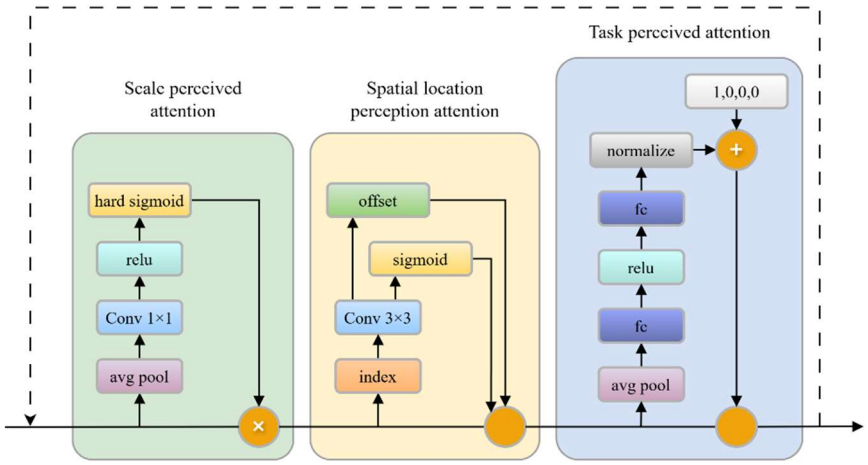

Fig. 3 Schematic diagram of DyHead structure

2.2 Hybrid Knowledge Distillation Combining Output and Features

To compensate for the accuracy fluctuation caused by structural improvements, this study adopts a hybrid knowledge distillation strategy that combines feature-based and output-based distillation. The teacher model used is YOLO-DMD, a high-accuracy detector proposed in our previous work, which achieves 79.8% mAP@0.5 on the Aug-RDD dataset. It is pretrained on the same training set and remains frozen during the distillation process.

To effectively transfer knowledge from a high-precision teacher model to student model, a well-designed knowledge distillation framework is essential. Knowledge distillation techniques can be broadly categorized into three types based on the knowledge dimension transferred: output distillation (transferring softened probability distributions), feature distillation (aligning intermediate feature maps), and relation distillation (transferring relationships between samples or features).

Output distillation softens the predicted probability distribution of the teacher model to convey inter-class similarity information, which enhances the model's generalization ability. This is particularly useful for distinguishing visually similar defect categories, such as cracks and pavement repairs. Feature distillation aligns the intermediate feature representations between teacher and student, directly transferring the teacher's capability to model geometric shapes, texture details, and multi-scale contextual information. This is critical for handling the irregular shapes and varying scales of road damage. Relation distillation focuses on transferring structured relationships, but despite its effectiveness in certain tasks, it introduces high computational complexity and offers limited benefits for instance detection tasks like the one in this study.

Considering the need for robust detection of road damage with diverse morphologies and scales, along with the potential feature representation limitations of the simplified student model, this study proposes a hybrid distillation strategy that combines output distillation and feature distillation.

Soft-label distillation transfers the teacher's smoothed category knowledge, helping the student better distinguish visually similar defects. Feature distillation enforces alignment of intermediate and high-level feature maps between the student and teacher networks, enabling the student to inherit the teacher's ability to model irregular defect structures and multi-scale context. This hybrid approach addresses the shortcomings of single distillation methods in complex detection tasks and establishes a solid foundation for developing an efficient lightweight model.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0016

Fig. 4 illustrates the knowledge distillation framework. The teacher and student models perform forward propagation in parallel. The softened output probabilities and selected intermediate feature maps from the teacher are extracted to compute the distillation loss, which guides the student model's parameter updates.

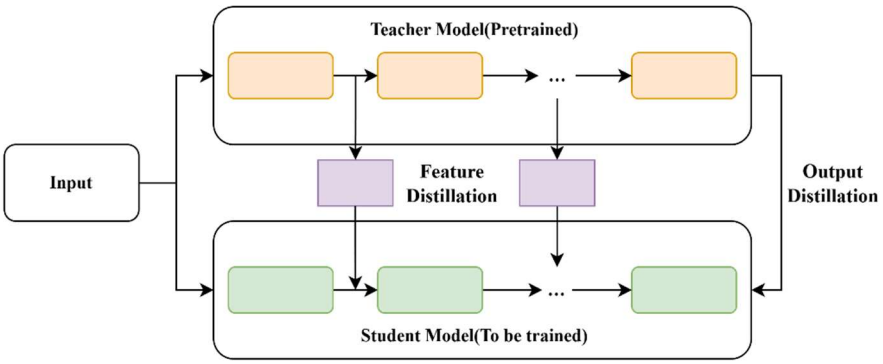

Fig. 4 Hybrid Knowledge distillation combining Feature and Output

3. Experiment

3.1 Experimental Setup

3.1.1 Dataset and Experimental Environment

To comprehensively evaluate the effectiveness of the proposed lightweight improvement strategy, a systematic experimental scheme is designed in this section. The dataset used in this study is SVRDD, a public dataset released by the Institute of Remote Sensing, Peking University. It contains six common types of road damage: longitudinal cracks, transverse cracks, alligator cracks, potholes, longitudinal repairs, and transverse repairs. In previous experiments, manhole covers were often misidentified as potholes, so they were separately annotated in SVRDD. Due to the imbalanced distribution of samples in the original data, targeted data augmentation was applied to several defect types. The resulting augmented dataset, named Aug-RDD, contains 22,204 annotated bounding boxes across 8,000 images. The experimental environment is shown in Table 1.

Table 1. Experimental Environment

Name	Configuration
CPU	13th Gen Intel(R) Core(TM) i9 13900K 3.0GHz
GPU	NVIDIA GeForce RTX 4090 24G
Operating System	Windows 11 Professional
Framework	Pytorch 2.0.1
Development Environment	Python 3.9.20，CUDA 11.7
GPU Acceleration	cuDNN 8.5.0

3.1.2 Experimental Training Parameters

During training, the input image size is set to 640×640. The dataset is divided into training, validation, and test sets with a ratio of 8:1:1, and the batch size is set to 8. All models are trained using the SGD optimizer with a momentum of 0.937 and weight decay of 5×10−4.

The initial learning rate is set to 0.01, with linear warm-up for the first 3 epochs, followed by a learning rate decay strategy. The model is trained for a total of 300 epochs, and an early stopping mechanism is enabled if the mAP does not improve for 40 consecutive epochs.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0016

3.1.3 Evaluation Metrics

To comprehensively evaluate the object detection model, the following commonly used evaluation metrics are taken as references: mean Average Precision (mAP), Parameters (Mb), Giga Floating-point Operations Per Second (GFLOPs), and FPS.

The mAP metric is calculated based on Precision (*P*) and Recall (*R*). *P* refers to the proportion of correctly predicted samples among those predicted as positive by the model. *R* refers to the proportion of actual positive samples that are correctly predicted as positive. The Average Precision (*AP*) is equal to the area under the P-R curve, representing the average precision of the model for a certain category under different confidence thresholds.

They can be expressed by the formulas:

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

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

$$AP = \sum_{k=0}^{k=n-1} [R(k) - R(k-1)] * P(k) \tag{5}$$

The mAP is used to evaluate the overall detection accuracy of the model, where (n) denotes the number of object categories. When only a single category is detected (n=1), mAP is equal to the *AP* of that category.

Thus, mAP can be expressed as:

$$mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k \tag{6}$$

3.2 Experimental Results and Analysis

Table 2. Comparative Experiment Results

Model	Paras/Mb	GFLOPs	mAP@0.5(%)	FPS
YOLOv8s	11.1	28.6	76.4	63.9
YOLO-LWD (before distillation)	8.1	15.8	76.1	113.7
SSD	23.8	56.4	58.9	44.7
RT-DETR-R18	19.8	60	75.6	58
DAMO-YOLO-M	28.2	61.8	77.3	55.3
YOLO-LWD (after distillation)	8.1	15.8	78.4	115.3

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0016

To verify the effectiveness of the collaborative lightweight strategy of "structure improvement + knowledge distillation" proposed in this study, we conducted comparative experiments between the final distilled YOLO-LWD model and several mainstream lightweight models. The experimental results are shown in Table 2.

Compared with other models, the proposed YOLO-LWD shows significant advantages in model complexity. With only 8.1 MB parameters and 15.8 GFLOPs, YOLO-LWD is considerably smaller than the other models, demonstrating the effectiveness of the proposed network structure in lightweight design.

In terms of detection performance, after hybrid knowledge distillation, YOLO-LWD achieves 78.4% mAP@0.5, which is a 2.3% improvement over the version without distillation. This indicates that the proposed lightweight strategy not only keeps the model compact and efficient but also improves detection accuracy.

Overall, the experimental results demonstrate that YOLO-LWD achieves a good balance between efficiency and accuracy in road damage detection tasks. It offers strong practical applicability and promotional value, providing effective support for real-world road damage detection.

4. Conclusion

This study addresses the challenges of deploying high-precision models on mobile devices and the significant accuracy loss of lightweight models in road damage detection tasks. We propose YOLO-LWD, a lightweight detection model that combines network structure optimization with hybrid knowledge distillation. By introducing GhostConv modules to replace standard convolutions in the backbone network, redundant computation in the feature extraction stage is significantly reduced. The ECA attention mechanism is adopted in the neck network to enhance feature representation while maintaining computational efficiency, further compressing model complexity without sacrificing detection performance. Additionally, the number of channels in the detection head is reduced, achieving an overall lightweight model structure. To compensate for the accuracy loss caused by structural simplification, a hybrid knowledge distillation strategy is designed, which integrates soft labels from the output layer and feature maps from intermediate layers to effectively transfer the teacher model's discriminative knowledge and structural perception ability to the student model.

Experimental results on the Aug-RDD dataset show that YOLO-LWD achieves 78.4% mAP@0.5 with only 8.1 MB parameters and 15.8 GFLOPs. Compared with mainstream lightweight models, YOLO-LWD demonstrates significant comprehensive advantages in detection accuracy, model size, and inference speed, validating the effectiveness of the proposed collaborative lightweight strategy.

In conclusion, YOLO-LWD achieves a good balance between accuracy and efficiency in road damage detection tasks, showing strong practical applicability and promotional value. Future work can explore even lighter model structures and conduct deployment and validation on edge computing devices in real-world scenarios to promote the practical application of automated road inspection technology.

References

[1] Yang, X., Zhang, J., Liu, W., Jing, J., Zheng, H., & Xu, W. (2024). Automation in road distress detection, diagnosis and treatment. Journal of Road Engineering, 4, 1–26.

[2] Manjusha, M., & Sunitha, V. (2025). Optimizing YOLO models for high-accuracy automated detection and classification of road surface distresses. Innovative Infrastructure Solutions, 10, 381.

[3] Zhang, Y., & Wang, Y. (2020). Machine learning for pavement condition assessment: A review. Journal of Transportation Engineering.

[4] Botezatu, A.-P., Burlacu, A., & Orhei, C. (2024). A review of deep learning advancements in road analysis for autonomous driving. Applied Sciences, 14(11), 4705.

[5] Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[EB/OL]. arXiv, 2017. https://arxiv.org/abs/1704.04861.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0016

[6] Tan M, Le Q V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks[C]//International Conference on Machine Learning. PMLR, 2019: 6105-6114.

[7] Chen F H, Li S L, Han J L, et al. Review of lightweight deep convolutional neural networks[J]. Archives of Computational Methods in Engineering, 2024, 31(4): 1915-1937.

[8] Golizadeh, M., et al. (2025). Architectural insights into knowledge distillation for object detection: A comprehensive review. arXiv preprint arXiv:2508.03317.