6-2024

# Development and optimization of object detection technology in pavement engineering: A literature review

Hui Yao
*Beijing University of Technology*

Yaning Fan
*Beijing University of Technology*

Yanhao Liu
*Beijing University of Technology*

Dandan Cao
*Beijing University of Technology*

Ning Chen
*Beijing University of Technology*

*See next page for additional authors*



## Recommended Citation

Yao, H., Fan, Y., Liu, Y., Cao, D., Chen, N., Luo, T., Yang, J., Hu, X., Ji, J., & You, Z. (2024). Development and optimization of object detection technology in pavement engineering: A literature review. *Journal of Road Engineering, 4*(2), 163-188. http://doi.org/10.1016/j.jreng.2024.01.006
Retrieved from: https://digitalcommons.mtu.edu/michigantech-p2/842

## Authors

Hui Yao, Yaning Fan, Yanhao Liu, Dandan Cao, Ning Chen, Tiancheng Luo, Jingyu Yang, Xueyi Hu, Jie Ji, and Zhanping You

Journal of Road Engineering 4 (2024) 163–188

Contents lists available at ScienceDirect

Journal of Road Engineering

journal homepage: www.keaipublishing.com/en/journals/journal-of-road-engineering

Review Article

# Development and optimization of object detection technology in pavement engineering: A literature review

Hui Yao [a,*], Yaning Fan [a], Yanhao Liu [a], Dandan Cao [a], Ning Chen [a], Tiancheng Luo [a], Jingyu Yang [a], Xueyi Hu [a], Jie Ji [b], Zhanping You [c]

[a] *Beijing Key Laboratory of Transportation Engineering, Faculty of Architecture, Civil and Transportation Engineering, Beijing University of Technology, Beijing 100124, China*

[b] *School of Civil and Transportation Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China*

[c] *Department of Civil and Environmental Engineering, Michigan Technological University, Houghton, MI 49931, USA*

## HIGHLIGHTS

- A comprehensive discussion of the relevant models utilized in the field of pavement distress detection is provided.
- Two strategies, attention mechanisms and lightweight networks, are proposed to optimize the detection model.
- 3D object detection methods are categorized into four aspects: monocular/binocular vision, point cloud and multi-view.
- Several lightweight networks are trained and compared by utilizing pavement crack images and radar datasets.

## ARTICLE INFO

*Keywords:*
Pavement engineering
Object detection
Lightweight network
Attention mechanism
Convolutional neural network

## ABSTRACT

Due to the rapid advancement of the transportation industry and the continual increase in pavement infrastructure, it is difficult to keep up with the huge road maintenance task by relying only on the traditional manual detection method. Intelligent pavement detection technology with deep learning techniques is available for the research and industry areas by the gradual development of computer vision technology. Due to the different characteristics of pavement distress and the uncertainty of the external environment, this kind of object detection technology for distress classification and location still faces great challenges. This paper discusses the development of object detection technology and analyzes classical convolutional neural network (CNN) architecture. In addition to the one-stage and two-stage object detection frameworks, object detection without anchor frames is introduced, which is divided according to whether the anchor box is used or not. This paper also introduces attention mechanisms based on convolutional neural networks and emphasizes the performance of these mechanisms to further enhance the accuracy of object recognition. Lightweight network architecture is introduced for mobile and industrial deployment. Since stereo cameras and sensors are rapidly developed, a detailed summary of three-dimensional object detection algorithms is also provided. While reviewing the history of the development of object detection, the scope of this review is not only limited to the area of pavement crack detection but also guidance for researchers in related fields is shared.

## 1. Introduction

According to the latest statistics in 2022, the total length of highways in China has reached 5.3548 million kilometers, with a maintenance coverage of 5.3503 million kilometers, accounting for 99.9% of the total. Relying solely on manual inspections is insufficient to meet the demands of large-scale daily highway maintenance management. With the development of computer vision, there has been a gradual exploration of

* Corresponding author.
*E-mail addresses:* huiyao@mtu.edu (H. Yao), fanyaning@emails.bjut.edu.cn (Y. Fan), liuyh0929@emails.bjut.edu.cn (Y. Liu), dandan_cao@bjut.edu.cn (D. Cao), chenningbjut@bjut.edu.cn (N. Chen), stevenluo@emails.bjut.edu.cn (T. Luo), yangjingyu@emails.bjut.edu.cn (J. Yang), huxueyi@emails.bjut.edu.cn (X. Hu), jijie@bucea.edu.cn (J. Ji), zyou@mtu.edu (Z. You).

Peer review under responsibility of Chang'an University.

https://doi.org/10.1016/j.jreng.2024.01.006
Received 2 September 2023; Received in revised form 8 January 2024; Accepted 22 January 2024
Available online 1 June 2024

automated detection methods for pavement surface defects (Ali et al., 2023; JRE Editorial Office et al., 2023; Yao et al., 2023a,b; Yin et al., 2023). The acquisition of detection data and the selection of algorithms are two major challenges that urgently need to be addressed. A detection vehicle equipped with a linear array image acquisition system is used to obtain high-resolution pavement surface data. In addition, the vehicle-mounted three-dimensional (3D) ground penetrating radar is used to obtain subgrade data information. The application of detection vehicles allows for uninterrupted information acquisition without impeding traffic flow, effectively accomplishing pavement surface inspection tasks through non-destructive testing. Therefore, this paper focuses primarily on the selection of detection algorithms. It provides an overview of the development of object detection in the field of artificial intelligence and explores the optimization and application of detection models.

In 1957, the first artificial neural network, perceptron (Rosenblatt, 1958), was invented, which served as a foundational concept in the development of artificial neural networks. The perceptron consists of an input layer, weights, an activation function, and an output layer. While it is capable of adjusting the weights to approximate the optimal classification outcome based on error, it is unable to solve non-linear problems. Subsequent advancements led to the introduction of hidden layers and non-linear activation functions in the form of multilayer perceptrons, which effectively tackled non-linear problems. Therefore, the perceptron, serving as a precursor to neural networks, has played a fundamental and influential role in advancing the field of neural networks and laying the foundation for modern deep learning models.

Advances in artificial intelligence have also played a role in driving pavement detection technology from manual to automatic detection. Crack detection can be classified into two methods: destructive tests and non-destructive tests (Zhang et al., 2017a). Automatic crack detection without damage has become the mainstream to realize fast and reliable pavement detection and defect analysis. Compared with automated detection, manual inspection presents difficulties in the objective evaluation of the progression of pavement distress. Automated detection technology has been developed to the application level, and the accuracy of distress identification under ideal conditions even exceeds that of manual detection. Meanwhile, conventional methods for detecting pavement distress are characterized by a high demand for time and labor. The emergence of automatic detection systems has completely changed the situation of manual inspection on pavement surfaces, greatly reducing the impact of detection work on traffic flow. The early automatic detection system has not been well popularized due to a large amount of post-processing work and single detection content. Fig. 1 shows several events in the progression of a convolutional neural network (CNN). The improvement of automatic pavement detection technology is mainly due to the rapid rise of CNN.

The proposal of the backpropagation algorithm addressed the optimization and adjustment issue of weights and biases in deep neural networks (DNN), enabling the training of multi-layer feedforward neural networks (Rumelhart and Hinton, 1986). LeNet, one of the pioneering

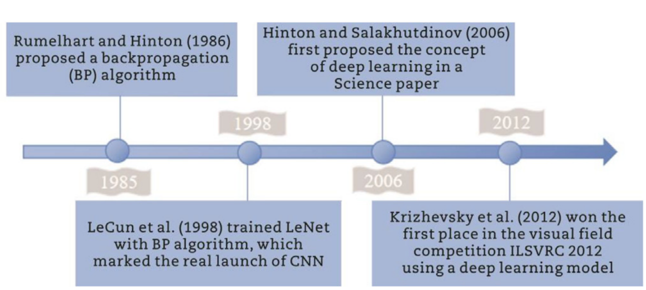


**Fig. 1.** Development of convolutional networks (Hinton and Salakhutdinov, 2006; Krizhevsky et al., 2012; LeCun et al., 1998; Rumelhart et al., 1986).

works in convolutional neural networks, combined with the BP algorithm, demonstrated the potential of deep learning and laid the foundation for complex convolutional neural networks (LeCun et al., 1998). Deep learning (Krizhevsky et al., 2012) (AlexNet) stood at a turning point after winning the image recognition competition, The imagenet large scale visual recognition challenge (ILSVRC) in 2012, and it occupied an unshakable position in the following competition. The present-day prevalent categorization of object detection architecture can be classified into two major categories: one-stage detection, for instance, you only look once (YOLO), single shot detection (SSD), and two-stage detection based on the region-based convolutional neural network (RCNN). The two-stage detection algorithm entails the generation of a sequence of regions from the input image, followed by utilizing CNN to extract features from the generated regions for constructing an object classifier. Subsequently, classification and regression are performed on the candidate regions. The prevalent approach for implementing the two-stage algorithm involves devising algorithms in the RCNN series that are based on the Region Proposal. The RCNN family encompasses Fast RCNN and Faster RCNN, among other extended processing algorithms. While exhibiting high accuracy, this method is computationally intensive and time-consuming. Zhang et al. (2022a) conducted a comparative analysis of the disease classification performance between convolutional neural network and support vector machine (SVM) using 500 images captured by a mobile phone camera at a resolution of 3264 pixels $\times$ 2448 pixels. Despite the limitation in training sample size, CNN and SVM achieved F1 scores of 0.89 and 0.73, respectively, with CNN demonstrating a significant advantage. In contrast, the one-stage approach possesses a stronger advantage in terms of rapid detection applications. Previous studies have compared various architectures, including YOLOv2-tiny, YOLOv2 based on Darknet19, YOLOv2 based on ResNet50, YOLOv3, and YOLOv4-tiny, to detect sidewalk cracks using unmanned aerial vehicles (Qiu and Lau, 2023). The attained accuracy of 94.54% serves as a compelling testament to the robust adaptability of the proposed method in arduous scenarios, encompassing challenging environmental factors such as the presence of shadows and moisture. In the context of bridge crack detection, research has been conducted to improve the YOLOv4 model by optimizing anchor boxes prior to model training and incorporating non-maximum suppression during testing (Zhang et al., 2023). This approach has yielded impressive results, with accuracy and recall rates reaching 93.96% and 90.12% respectively. Although the lightweight processing slightly compromises the precision of the model, it achieves a remarkable frame rate of 140.2 frames per second, making it significantly faster than the Faster RCNN model. From an industry application perspective, the YOLO series algorithms are gradually becoming the mainstream choice.

This paper introduces attention mechanisms similar to human visual features to better identify pavement distresses. The attention mechanism serves to alleviate redundancy and prioritize information deemed most pertinent to the current objective. The lightweight transformation of the network and the tradeoff between model size and reasoning speed represent potential directions for future development. The deep learning network model is easy to deploy on mobile or embedded devices, which is necessary for industrialization. The latter part of this paper covers this module. Considering the current two-dimensional (2D) object detection architecture based on deep learning is relatively mature, this paper introduces the research status of 3D object detection. In conjunction with the aforementioned content, prospects for object detection development are also contemplated. The content framework of this paper is shown in Fig. 2 below.

## 2. Pavement distress identification technology

This section focuses on pavement distress detection technology, which can be broadly categorized into image-processing algorithms and machine learning-based detection algorithms. Image-processing techniques are proposed to address the challenges posed by external factors such as lighting conditions and oil stains during pavement image

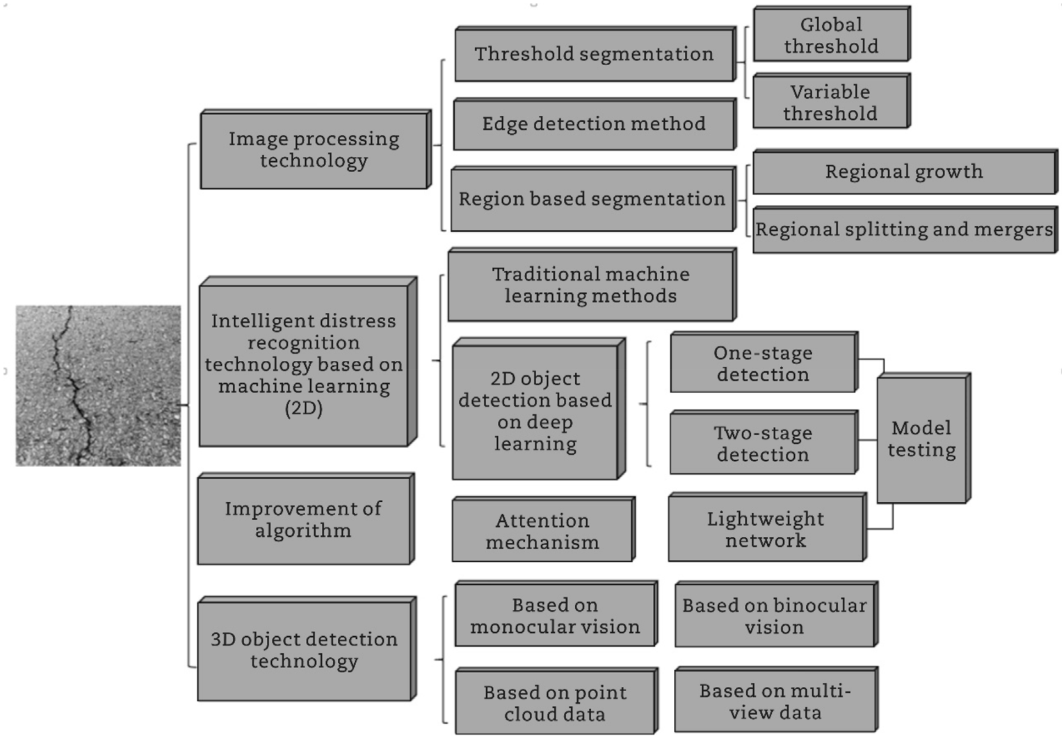


**Fig. 2.** Main structures of this project or manuscript.

acquisition. Therefore, this technology can be utilized for data preprocessing of the images before feeding them into the detection model. Image processing algorithms can be classified into threshold segmentation, edge detection, and region-based segmentation. Threshold segmentation is a technique that divides an image into two categories based on the grayscale values of pixels and a predefined threshold. Edge detection marks regions in an image where significant changes in grayscale or color values occur. region-based segmentation is based on pixel similarity and iteratively merges similar pixels. Deep learning, as a subfield of machine learning, is the main part of this introduction. Compared to traditional machine learning methods, deep learning learns higher-level feature representations directly from raw data through multi-layer neural networks. The detection model can be divided into one-stage and two-stage detection. These two types of models mainly complete the detection task based on anchor. In contrast, the anchor-free object detection algorithm is an emerging technology, so this paper classifies such models separately. The content integrated deep learning technology with the practical application of pavement detection.

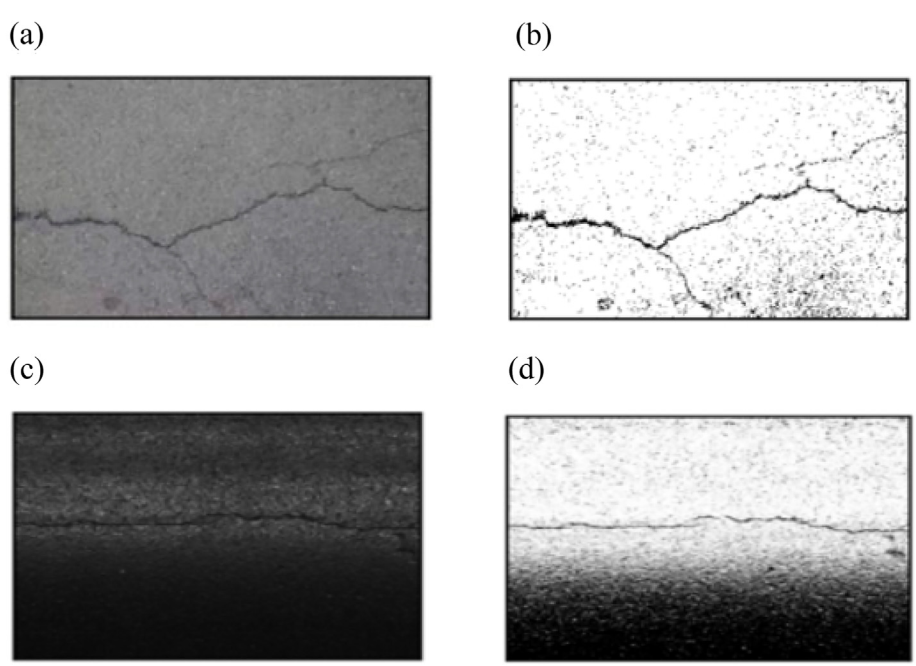


**Fig. 3.** Image segmentation algorithms based on threshold processing. (a) Original image 1. (b) Processed image 1. (c) Original images 2. (d) Processed images 2.

### 2.1. Image processing

#### 2.1.1. Threshold segmentation

Threshold segmentation represents the most frequently applied algorithm in image segmentation. By judging whether the feature attributes meet the threshold standard, the foreground can be discriminated from the background (Zhu et al., 2007). This requires transforming greyscale images into binary images. A binary image involves vital information as to the shape and position of objects. The threshold segmentation method contains global thresholding, variable thresholding, and multiple thresholding (Al-Amri and Kalyankar, 2010). Obtaining a credible threshold value is the key to this approach. In addition, the dynamic threshold is presented and it achieves good results in pavement distress detection (Oliveira and Correia, 2009). Presuming that crack intensity typically registers lower than the background, methods utilizing intensity thresholding are widely employed in crack detection (Fig. 3). A neighborhood difference histogram method (NDHM) was proposed (Li and Liu, 2008), which outperformed the classical thresholding method for crack detection. However, the application of a solitary threshold across an entire image is rendered unsuitable when the image contains sources of non-uniform illumination, such as shadows. In summary, under strong lighting conditions, a fixed threshold may struggle to accurately segment the target contours, while a lower threshold in dim lighting can lead to higher false detection rates. Therefore, dynamically adjusting the threshold based on the actual circumstances allows for better adaptation to different application scenarios.

Based on the multi-scale and local optimal thresholding method, segmentation through crack density distribution is more effective and robust compared to traditional global thresholding algorithms (Wang and Tang, 2012). In a recent study, an unsupervised crack detection method that leverages the grayscale histogram and the Otsu thresholding method is proposed (Akagic et al., 2018). Input images were separated into four autonomous sub-images. Cracks were detected by examining the ratio between the Otsu threshold and the highest histogram values for each sub-image. Subsequently, the sub-images were amalgamated into the final output images. By testing different pavement images and cracks of various types, results have revealed that the proposed method is capable of attaining satisfactory performance. Li et al. (2015) employed a dual-threshold segmentation technique, utilizing an enhanced Otsu thresholding algorithm to remove pavement markings. Subsequently, an adaptive iterative threshold segmentation algorithm was applied to further partition the image, and morphological denoising was utilized to obtain the final crack contours.

To sum up, the thresholding method may deliver superior segmentation outcomes when there exists a significant disparity in the gray levels between the target and background regions. When such a notable variation in gray levels is represented in the image histogram, it manifests as a characteristic of bimodal distribution in the local image's gray histogram. The utilization of this technique is notably impacted by image noise, uniformity of the light source, uniformity of reflection, as well as the relative sizes of the object and the background.

#### 2.1.2. Edge detection method

An edge could be defined as the boundary between two areas with definite attributes of grey level. Edge detection operators, such as the Roberts operator, Sobel operator (Sobel, 1990), and Prewitt operator, are commonly used. The Roberts operator is particularly effective on images that exhibit steep edges and minimal noise. As shown in Fig. 4, the Sobel operator is a classic first-order derivative-based edge detection operator that can effectively suppress noise and mitigate its impact on the tested image.

Edge detection techniques were frequently employed for pavement detection, prior to the emergence of deep learning (Liu et al., 2008; Maode et al., 2007). Three novel fast edge detection approaches for detecting rough edge mapping through fuzzy logic have been proposed. Liang and Looney (2003) proposed a competitive fuzzy edge detection (CFED) technique. Both methods leverage a fuzzy classifier to determine the pattern to which a given edge type pertains. However, in certain intricate regions of the image, CFED may fail to identify subtle textures, resulting in the emergence of spots.

Lim et al. (2013) proposed sketch labeling, to detect the representation of middle-level features based on local contours, which uses supervised middle-level information to learn what appears in images as hand-drawn outlines. Learning markers from images with ground truth outlines significantly enhances detection accuracy. To account for the high statistical dependence among pixels within an object, Isola et al. (2014) introduced a novel approach that achieves precise pixel-level boundaries. Kanga and Wang (2007) proposed a new edge detection method for grayscale images and color images that uses two sets of pixels in a 3 × 3 mask to define an objective function. The objective function of

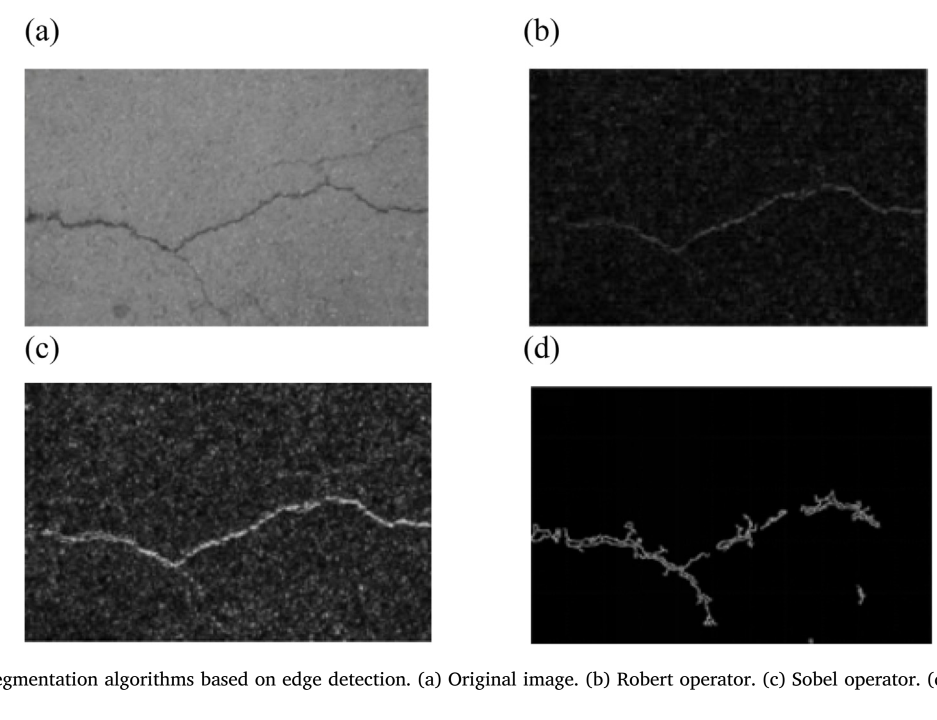


**Fig. 4.** Image segmentation algorithms based on edge detection. (a) Original image. (b) Robert operator. (c) Sobel operator. (d) Canny operator.

the four directions is utilized to determine the edge intensity and direction for each pixel within the mask. The effect of this method is thinner and more realistic than the edge lines detected with the Sobel method.

#### 2.1.3. Region based segmentation

This method can be used to divide the region straight. Specifically, regional growth is the process by which, starting from a set of initial seed points, similar domain pixels are constantly added to each seed point by pre-defined regional growth rules, and the final growth area forms when the termination conditions for regional growth are met (Sarathi et al., 2013). Therefore, the choice of seeds directly affects the accuracy of the image segmentation. A region growth segmentation algorithm based on wavelet features is applied. In the seed point selection process, the pre-processed images undergo convolution and wavelet feature extraction. Prominent wavelet feature points are selected as the candidate seeds and rotated to obtain the optimal region of interest (RoI). In addition, a hierarchical region-based solution is proposed to integrate object detection and image segmentation (Gould et al., 2009), explain each pixel in the image, and strengthen the global consistency between all random variables in the model. Therefore, this region-based image segmentation technique results in a significant enhancement in the accuracy of detection results. In the method proposed by Zhou et al. (2016), crack seeds were selected by grid cell analysis and they were connected using a Euclidean minimum spanning tree structure. The experiment indicated that the technique greatly improves detection efficiency.

In addition to regional growth, there are regional splitting and mergers. In short, regional splitting is viewing the entire graph as a region. This region is divided into a series of mutually consistent discrete regions. Regional consolidation is the opposite aspect of regional division. It starts with fine regions and incorporates regions with similar properties, such as variance, etc.

### 2.2. Detection algorithms based on traditional machine learning

Achieving a variety of tasks with outstanding performance, machine learning is used in almost every field. This section will classify machine learning algorithms as supervised and unsupervised learning methods. The distinction between supervised and unsupervised learning is whether labeled data is used for training. Supervised learning requires data with labels, whereas unsupervised learning does not. The learning algorithm predicts the output of unknown data by learning the mapping between input variables and output variables. Supervised learning tasks can be classified into classification and regression. Supervised learning is usually exploited to solve classification problems since the aim is generally to get the machine to learn a classification system that users have created. The classification task uses scattered labels, while the regression task uses continuous labels. Supervised learning algorithms contain logistic regression (LaValley, 2008), Naive Bayesian (Leung, 2007), SVM (Burges, 1998), artificial neural networks (Jain et al., 1996), and random forest (Cutler et al., 2012). Marques and Correia (2012) proposed a novel automated pavement detection method. In order to detect cracks, the crack image is initially preprocessed to accentuate existing cracks. Then, the image is partitioned into non-overlapping blocks, each of which yields a feature vector. The supervised learning technique SVM is utilized in this method for crack detection. Xu et al. (2008) converted the image recognition of each sub-block of the crack into a crack probability judgment by using the self-learning characteristics of the neural network. In this method, the parameters that capture the essential characteristics of cracks are first extracted from sub-images obtained through segmentation of the binary crack image. These representative images are then chosen to train a backpropagation neural network.

Unsupervised learning eliminates the need for labeled data, so the detection task reduces human interference. Unsupervised learning can be viewed as finding patterns in huge amounts of data. Akagic et al. (2018) presented an approach utilizing gray histogram analysis and Otsu's method. The input image is partitioned into four mutually exclusive sub-images of equal size in this technique. Cracks are subsequently detected by examining the ratio between the Otsu threshold and the maximum histogram value for each sub-image. The final image is generated by combining all of the sub-images. Amhaz et al. (2016) proposed an enhanced unsupervised learning algorithm that relies on minimum path selection. This method assesses the width of cracks and is effective in reducing loop and peak artifacts during detection. Before the emergence of deep learning, traditional machine learning necessitated a pre-defined feature extraction phase, which reduced data complexity and helped algorithms identify patterns more accurately. A drawback of traditional machine learning methods is that they are limited to shallow learning techniques, which are inadequate in handling the complex information present in images without learning higher-level features. For images that are greatly interfered by environmental factors such as lighting, traditional machine learning methods clearly cannot produce ideal results.

### 2.3. Typical structure of object detection based on deep learning

CCD cameras are mainly used to obtain pavement surface images for the detection of pavement crack images, such as the detection vehicle system developed by the Australian Highway Research Institute (Cafiso et al., 2006), which has been applied to the detection of pavement defects. Since cracks on highway pavement are often in various forms, subjective judgment has a great impact on distress identification. Therefore, deep learning technology is introduced into the field of pavement detection to realize the automatic identification of pavement defects. Convolutional neural networks are high-profile models in this technology. In contrast to machine learning, deep learning automatically extracts data features through convolution kernels during training. Specifically, deep learning doesn't require a predefined feature extraction stage, which uses substantial data to train a generic and robust model. As a classic object detection method, the sliding window method slides windows of different sizes on the image, similar to an exhaustive search of image subregion, and generates candidate boxes for feature extraction, including HOG, SIFT, etc. After conducting the convolution operation and training the classifier (SVM (Burges, 1998), random forest (Liaw and Wiener, 2002)), the probability of object existence is determined. In general, the target accounts for only a small part of the image, so the calculation of producing candidate regions is very costly and difficult to apply in practice. A selection search method is proposed to search the area in the image most likely to contain objects to improve computational efficiency. A subregion merging method is used to extract candidate boundary boxes.

Object detection can be classified into two types based on the technical approaches, namely anchor-based object detection and non-anchor-based object detection. On the other hand, based on the detection process prescribed by the model structure, object detection can be categorized as either one-stage or two-stage detection (Fig. 5).

The subsequent content introduces the two-stage and one-stage object detection in detail according to the division of deep learning technology. The above-mentioned techniques are primarily based on anchor boxes for performing object detection tasks, a separate introduction to anchor-free object detection techniques is provided.

#### 2.3.1. Two-stage object detection algorithm for RCNN structure

Two-stage object detection tasks can be decomposed into the generation of region proposals and subsequent classification. Convolutional neural networks are employed to extract features and perform classification, enabling precise object detection and localization. RCNN can be regarded as the cornerstone of two-stage object detection algorithms. RCNN received great attention when it was proposed in 2014. The innovation of this structure lies in the combination of AlexNet (Krizhevsky et al., 2012) and selective search algorithms (Uijlings et al., 2013). The detection task is divided into four main steps. First, candidate regions are extracted from the input image by a selective search algorithm, and then

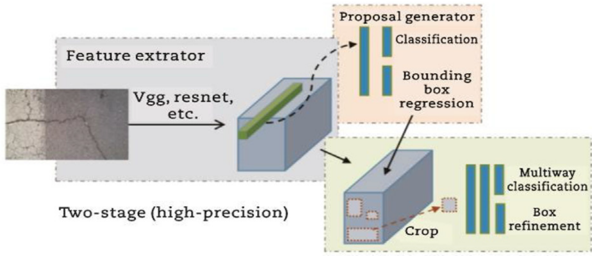


**Fig. 5.** One-stage detection and two-stage detection model scheme.

each candidate region is normalized and input to CNN one by one to extract features. Finally, SVM classification and regional regression are performed for these features. Although the detection accuracy of this method far exceeds that of traditional detection, it is time-consuming to use in selective research, and there are a lot of repeated calculations for extracting features for each region proposal. Feature extraction, image classification, and border regression are three independent steps that need to be trained separately, so the detection speed is low. Moreover, the fully connected layer in the RCNN structure requires a fixed-length input. Cropping or packing to obtain a fixed size may lead to unnecessary geometric distortions, and the cropped region may not contain the whole object. SPP-net (He et al., 2015) is a novel network architecture that addresses the issue of CNN networks requiring fixed-size inputs, which can decrease image recognition accuracy. This enhanced structure expands upon the CNN network and increases scale invariance while mitigating overfitting risk through multi-scale training. Inspired by SPP-net, the Fast RCNN model (Girshick, 2015) was proposed in 2015. By inputting the whole image into the model, the overall feature map of the picture is obtained. Through candidate region mapping onto the feature map, the RoI pooling algorithm is used to generate features for each candidate box. The RoI pooling layer meets fixed dimension input requirements in the fully connected layer and can be batch-processed simultaneously. The speed still needs to be improved due to the use of selective search. In 2016, Faster RCNN (Ren et al., 2015) replaced selective search with regional proposal network (RPN) in the extraction of candidate regions, which realized end-to-end training and quite improved detection speed. Fig. 6 illustrates the improvements of the RCNN series models in detecting pavement distress images. In practical applications, Yang et al. (2020) utilized the Fast RCNN model to effectively improve the accuracy and efficiency of recognition by integrating image recognition, GPS location information, and in-vehicle signal information (Tian et al., 2020). Currently, significant progress has been made in the detection of pavement surface defects, and the accuracy of pavement damage identification by combining automatic detection and segmentation tasks through two-step convolutional neural networks has reached more than 90% (Liu et al., 2020). This method greatly overcomes the technical bottleneck of intelligent pavement detection.

(1) Research on feature extraction network

While deep convolutional layers exhibit high recall in identifying objects of interest, their localization performance is often poor due to feature map roughness. On the other hand, lower network layers are better at localizing objects of interest, yet this is accompanied by a reduced recall rate.

Therefore, it is difficult to achieve ideal results by relying only on the features of the final convolutional layer to detect objects of different scales. For Fast RCNN, the size of the feature map is too coarse to classify some small-size instances. The full convolutional network (FCN) shows good performance in semantic segmentation, so Long et al. (2015) combined coarse layer and high layer information with fine layer and low layer information for semantic segmentation. Inspired by this, Kong et al. (2016) developed a new hyperfeature that integrates both deep and coarse information with shallow and fine information, producing more comprehensive features. In 2017, a top-down hierarchical structure (Lin et al., 2017a) with horizontal connections was proposed to extract multiple features of different scales. This approach involves fusing high-level and shallow features, yielding scale features that contain rich geometric information along with semantic information. Through experimentation with classifying small objects on ImageNet using different network architectures, it was discovered that deep convolutional neural networks are not robust to scaling variations.

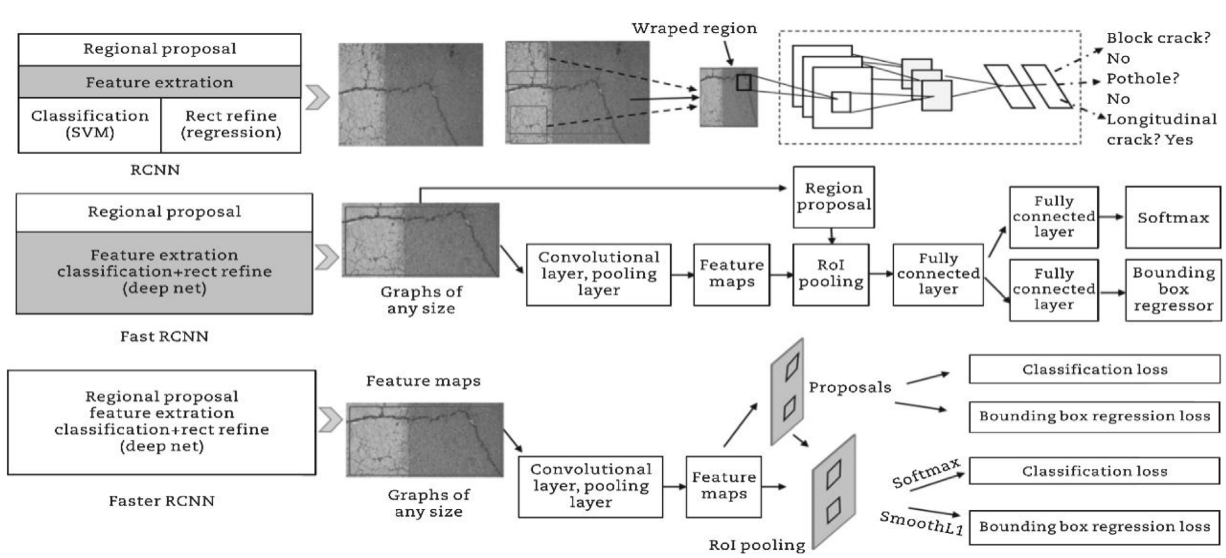


**Fig. 6.** RCNN series network architecture development process (Girshick et al., 2013, 2015; Ren et al., 2015).

A novel training approach referred to as scale normalization of image pyramids (SNIP) was proposed (Singh and Davis, 2018), which selectively backpropagates the gradients of object instances based on their size in relation to the image scale. This strategy aims to enhance the training process by considering object instances of different sizes differently. In this way, small target objects can be detected with high resolution, medium targets can be detected with medium resolution and large targets can be detected with low resolution. Like VGG (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al., 2015) and ResNet (He et al., 2016) have already achieved good results in image networks and are extensively used as feature extraction networks in object detection. However, these networks are memory intensive. For this reason, Fu et al. (2018) proposed a kind of high-precision head detection framework and investigated the feasibility of an embedded implementation. As shown in Fig. 7 below, the VGG network has a neat and simple structure that can reach up to 19 layers. GoogLeNet replaces fully connected layers with average pooling layers and achieves a similar deep network structure by stacking repeated units.

(2) Research on the RoI pooling layer

RoI pooling features fixed-size feature maps obtained from deep networks with multiple convolution kernels pooling, which significantly shortens the time of training and greatly improves the detection speed, thus helping to realize real-time detection. RoI pooling serves two primary purposes: it reduces the number of parameters, resulting in a reduction in computational costs, and it helps regulate overfitting in the network. Fundamentally, an ideal pooling technique should extract only relevant information while discarding irrelevant details. Average pooling is initially applied to deep neural networks. It performs downsampling by partitioning the input into rectangular regions and computing the average value for each region (LeCun et al., 1998). The max pooling obtains the largest value of rectangular pool regions through the filter. In addition, there are other methods such as mixed pools (Yu et al., 2014). In actual image detection applications, max pooling is generally better. Position-sensitive score maps were introduced to address the trade-off between translation invariance in image classification and translation variance in object detection (Li et al., 2016).

To comprehensively investigate both local and global characteristics, Zhu et al. (2017) proposed an innovative fully convolutional network named CoupleNet, which integrates global structure with local parts. Concretely, the object proposals generated by the RPN are fed into a coupled module composed of two branches. One branch tries to employ position-sensitive RoI (PSRoI) pooling to capture local partial object information, while the other branch uses RoI pooling to encode global and contextual features. To improve the transformation modeling ability of CNN, deformable convolutions and deformable RoI pooling were proposed (Dai et al., 2017a). Both methods involve augmenting the spatially sampled locations in the module by adding additional offsets and training these offsets from the target task in an unsupervised manner, without requiring any additional supervision. He et al. (2017) introduced a novel RoIAlign layer that replaces the coarse quantization of RoIPool and accurately aligns extracted features with the input. Additionally, Sun et al. (2019) proposed an innovative algorithm, termed RoI pool correlation filter (RPCF), for robust visual tracking. By means of mathematical derivation, they established that imposing supplementary constraints on learned weights of filters can result in an implementation equivalent to that of the RoI-based filter pool.

(3) Improvement of RPN and NMS

The concept of RPN was first proposed in a Faster RCNN model (Ren et al., 2015). The core work of Faster RCNN is the incorporation of candidate box extraction during feature extraction, which greatly accelerates the speed of object detection. Through the design of the anchor box, including aspect ratio and laying interval, the classification and regression of the anchor box are realized. To improve the quality and detection performance of regional proposals, The cascade region proposal network (cascaded RPN) is proposed (Vu et al., 2019). This architecture replaces the use of multiple anchors with predetermined scales and aspect ratios and instead adopts a single anchor at each location while executing multi-stage optimization for improved performance. Chen et al. (2018a) introduced a context refinement algorithm with the objective of refining each proposed region in a more effective manner. The proposed algorithm explores abundant contextual information and significantly improves the quality of regional proposals and detection outcomes. Optimization and research on RPN are still being proposed. Inspired by the RPN of Faster RCNN, a new method for proposal generation (Chen et al., 2018b), namely enhanced region proposal network (ERPN) was proposed. Generally speaking, the non-maximum suppression (NMS) algorithm operates by setting a pre-defined threshold and sorting all detection boxes based on their classification confidence scores. The algorithm then selects the detected results with the highest score and compares them using intersection over union (IoU) in order to remove detection bounding boxes that exceed the threshold.

However, the setting of a single threshold often leads to missed detection and a high miss rate. To solve this kind of problem, the Soft NMS algorithm was proposed (Bodla et al., 2017). Instead of directly removing neighboring results that exceed the threshold of IoU, it uses linear or Gaussian-weighted methods to reduce its confidence value. Then, choosing a suitable confidence threshold can effectively eliminate false detection boxes and enhance the model's sensitivity towards missed detections. The coordinates of all detection boxes with the highest confidence scores and a ratio above a certain threshold are weighted and averaged to obtain the final detection result. This approach significantly improves the localization accuracy of the object. He et al. (2018) used another way, which does not involve a direct selection classification confidence test box with the highest scores as the final test results. Instead, the coordinates of all detection boxes with the highest confidence scores and the ratio greater than a certain threshold are weighted and averaged as the final detection result, so as to locate the object more accurately. Dai et al. (2017b) proposed an object relation module, which deals with a group of objects concurrently through the interaction between their appearance features and geometric shapes. Instead of the traditional non-maximum suppression algorithm, the proposed object

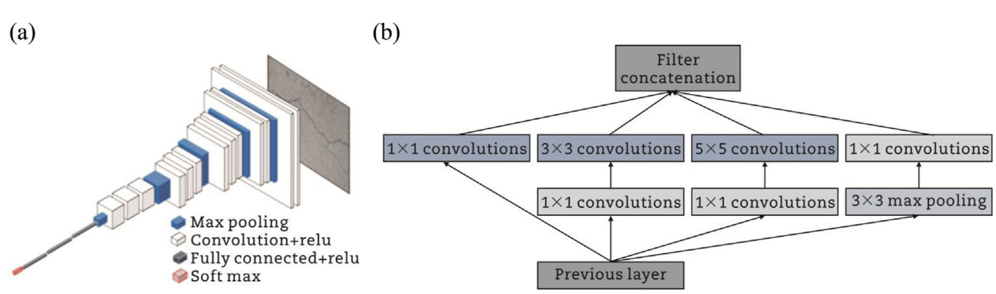


**Fig. 7.** Classical network models. (a) VGG model (Simonyan and Zisserman, 2014). (b) GoogLeNet inception module (Szegedy et al., 2015).

relation module is effective in enhancing object identification and removing repeat processing steps.

In the research of automated detection models for pavement defects, Alzraiee et al. (2021) utilized Faster RCNN to accomplish the detection task, but with only 1040 instances of distress data. Gou et al. (2019) improved the region proposal network and feature extraction module based on the Faster RCNN model, resulting in a model that exhibits significantly better performance than traditional feature extraction methods and possesses better generalization capabilities. For the detection of various defects such as transverse cracks and longitudinal cracks, the use of Faster RCNN in a full-size pavement image framework reduced the sliding window size (Ibragimov et al., 2022), enabling the model to detect larger images. Mask RCNN has also been widely applied in applications such as concrete crack (Attard et al., 2019) and asphalt crack detection (Liu et al., 2023).

### 2.3.2. One-stage object detection algorithm

(1) YOLO series network

The YOLO algorithm was proposed to integrate the classification, localization (Redmon et al., 2016), and detection functions in a network. The bounding box and category probability of the object in the input image can be obtained directly after only one network calculation. The YOLO algorithm completely abandons the candidate region generation step, which improves the detection rate and meets the speed requirements. Fig. 8 illustrates the development process of the entire YOLO algorithm. YOLOv1 divides the input picture into $7 \times 7$ grids, with each grid predicting two boundary boxes, so there is a $7 \times 7 \times 2$ boundary box to recognize a maximum of 49 targets. However, due to the rough design of the network, YOLOv1 is far from meeting the accuracy requirements of real-time object detection, and there are challenges such as inaccurate localization, frequently missed detections, and poor detection performance for small or multiple targets. The above problems leave plenty of room for further improvements. In 2017, the proposed YOLO9000 (Redmon and Farhadi, 2017) focused on improving recall and location rates while maintaining classification accuracy. This approach adopts the anchor mechanism from the Faster RCNN algorithm and eliminates the fully connected layer. Instead, it employs the convolutional layer to forecast the positional offset and category information of the detection box. In contrast to the manual design of the original anchor mechanism, this approach utilizes *K*-means clustering to acquire the optimal initial anchor template from the training set. Meanwhile, YOLO9000 introduces a pass-through layer that links the shallow feature map with the deep feature map to obtain fine-grained features. In 2018, Redmon et al. used the concept of jump connections in the residual network to construct a benchmark network called Darknet-53 (Redmon and Farhadi, 2018), consisting of 53 layers. This network achieves classification accuracy comparable to that of ResNet152 (He et al., 2016). This improvement mainly reduces the amount of calculation.

To deal with multi-scale objects, YOLOv3 uses feature maps of three different scales for object detection. Every feature map is obtained by fusing high-level and shallow feature maps. When predicting categories, independent logistic classifiers are used instead of the Softmax method. After two years, YOLOv4 (Bochkovskiy et al., 2020) adopted a fresh backbone and integrated spatial pyramid pooling with a path aggregation network (PAN) for feature fusion to achieve enhanced performance (Liu et al., 2018). Based on YOLOv3, CSP-Darknet53 was built to strengthen CNN learning ability while reducing memory usage. The spatial pyramid pool used by YOLOv3 is still used to solve the multi-scale problem, and YOLOv4 is better trained on a single video card. The structure of YOLOv5 is similar to that of YOLOv4. The biggest difference is that according to the scale of different channels, five models of YOLOv5-N/S/M/L/X are constructed from small to large. The architecture of the YOLOv5s is depicted in Fig. 9. The network also contains the Focus and space pyramid pool (SPP) modules. The input image is self-copied through the Focus module and then sliced to reduce the amount of computation. The sliced feature vectors were convolved, normalized, and activated by the CSP module.

YOLOv6 uses a more concise anchor-free detection method. The anchor-based detector needs to perform cluster analysis before training to determine the optimal anchor set (Li et al., 2022b), which increases the complexity of the detector to some extent. At the same time, in some edge applications, the need to move a large number of detection results between the hardware can introduce additional delays. To obtain more high-quality positive samples, YOLOv6 introduced the SimOTA algorithm to dynamically allocate positive samples for enhancing accuracy. SIoU loss was used in the YOLOv6s experiment. Compared with CIoU loss, the average detection accuracy was improved by 0.3% AP. YOLOv7 proposed a training method for the auxiliary head (Wang et al., 2023), the main purpose of which was to increase training cost and accuracy while not impacting the reasoning time. The auxiliary head would appear only during training.

For pavement distress detection, the YOLO series of algorithms has undergone extensive optimization and research. Huang et al. (2023) reduced model complexity and achieved a 91% average accuracy by employing a grouped spatial pyramid pooling module, resulting in a 29.3% reduction in the model's parameter count. For the detection of pavement potholes, BV-YOLOv5s, a lightweight target detection, was proposed. bidirectional feature pyramid network (BIFBN) was used to optimize the model structure and improve the detection accuracy of pavement defects through multi-scale fusion (Du and Jiao, 2022). Similarly, Park et al. (2021) applied YOLOv4, YOLOv4-tiny, and YOLOv5s models to effectively identify pothole, achieving 77.7%, 78.7%, and 74.8% mean average precision (mAP), respectively. Liu et al. (2022) conducted model architecture optimization based on YOLOv3 and combined it with data augmentation techniques to ensure the robust performance of the model in the presence of noisy interference for defect detection (Li et al., 2022a). Given the difficulty of detecting the pavement oil repair area, Liu et al. (2022) applied MobileNetV3 to the detection network of YOLOv5, combined with the SPP-net structure, and proposed that the M-YOLO model mAP reached 95.5%. At present, many improved algorithms based on the YOLO model have been proposed, most of which combine attention mechanisms to explore the optimization of the model.

(2) SSD series algorithm

YOLO and RCNN series algorithms have their advantages. On the whole, the RCNN algorithm has higher detection accuracy but a slower speed than YOLO. Meanwhile, YOLO has good generalization ability for large dimensional objects but a poor detection effect for small objects. Based on the understanding of the Faster RCNN and YOLO algorithms, Liu et al. (2016) proposed the SSD algorithm, combining the advantages of the above two. As illustrated in Fig. 10, the SSD network is designed with the idea of extracting features hierarchically (Liu et al., 2016). In addition, SSD draws on the anchor mechanism of the Faster RCNN

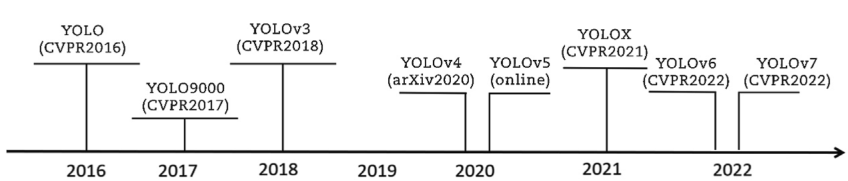


**Fig. 8.** YOLO model development process.

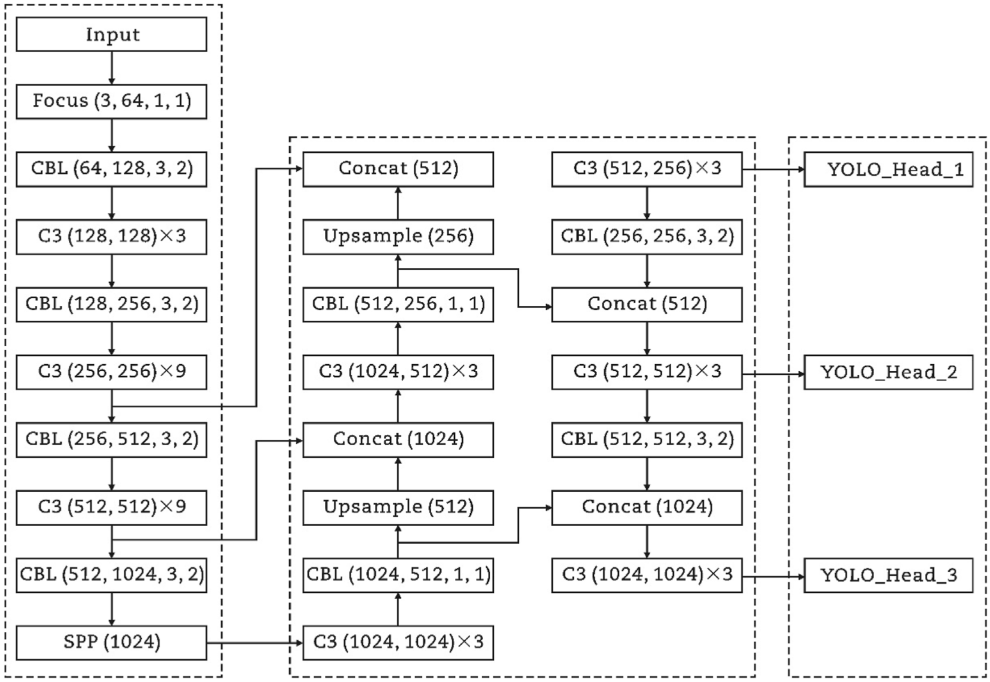


**Fig. 9.** YOLOv5s algorithms (Xu et al., 2022).

algorithm and presets a fixed number of default boxes of different scales and aspect ratios at each position of the extracted feature map. The network can directly perform intensive sampling on the feature map to extract candidate boxes for prediction.

To address the limitation of SSD in detecting small targets and further enhance its feature representation capability in low-level feature maps, the deconvolutional single shot detector (DSSD) algorithm was proposed (Fu et al., 2017). It replaced the VGG16 used by SSD with ResNet101 (He et al., 2016) to enhance the capacity of network feature extraction. DSSD introduced two novel modules, namely the prediction module and the deconvolution module. The prediction module, similar to the residual module, merges the features from different layers through skip connections to enhance the feature representation capability. By employing the deconvolution operation, the deconvolution module creates a top-down pathway to generate new feature maps with various scales. The feature maps integrate advanced and shallow features and introduce rich spatial background information, but the ResNet101 backbone network is too deep, resulting in slow training and detection speed. Similarly, in feature fusion single shot multi-box detector (FSSD), a single-shot multi-box detector with feature fusion was proposed (Li and Zhou, 2017), which integrates features from different scales and layers. The approach downsamples the features to produce a new feature pyramid, which is then fed into the multi-box detector to detect the final results.

To address the issue of object scale variations, multi-scale features with rich contextual and spatial information need to be extracted. Many improvements in SSD algorithms pay more attention to the application of feature pyramid networks (FPN). The NAS-FPN algorithm was proposed by Ghiasi et al. (2019). The model automatically searches and designs the optimal FPN structure to realize cross-scale feature fusion. It exceeded Mask RCNN on network performance, but the model of training needs a lot of GPUs. In addition, the content in the candidate box generated by the SSD algorithm has the problem of category imbalance. Lin et al. (2017b) presented the RetinaNet algorithm, which employs focus loss to adjust the contribution of positive and negative samples to the loss based on the contents in the candidate box. This method enhances the attention to samples with low confidence and mitigates the impact of class imbalance on the algorithm's accuracy.

In the crack detection of asphalt pavement, Han et al. (2021) selected the SSD model for target classification, making the detection accuracy reach 96% on small-scale data sets, but the model trained with only 480 data sets is difficult to apply in practice. Also for pavement cracks, Feng et al. (2020) combined SSD and U-shaped image segmentation network (U-Net) networks to achieve accurate crack detection and geometric parameter calculation, and the accuracy of the method reached 86.8%, 87.6% and 85.5% for transverse, longitudinal and crocodile cracks. In the context of pavement defect detection, the SSD series models have been more extensively involved in comparative

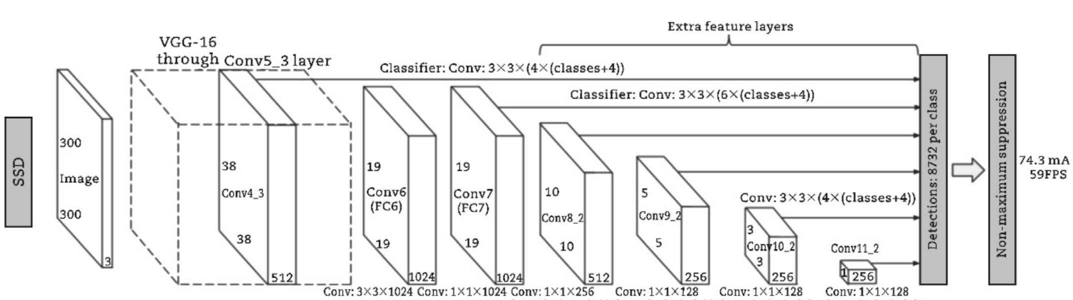


**Fig. 10.** SSD algorithm (Liu et al., 2016).

experiments. Haris and Glowacz (2021) conducted comparative experiments on SSD, YOLOv4, RetinaNet, Mask RCNN, and other models. The SSD model obtained the lowest recall rate and average precision in the detection process.

*2.3.3. Anchor-free detection algorithm*

The one-stage and two-stage object detection algorithms mainly complete the detection task through anchor. The proposed anchor-free algorithm aims primarily to address the limitations of anchor-based detection algorithms. In short, this method removes the process of presetting the anchor and predicts boxes directly. As mentioned before, models based on the CNN series as well as SSD are based on anchor box algorithms. However, an anchor-based object detection algorithm generally requires determining hyperparameters such as different sizes, aspect ratios, and the number of generated anchor boxes according to the training data. The selection of hyperparameters has a direct impact on the accuracy of the algorithm. The anchor-free algorithm has the advantage of being more robust because the parameters of the anchor box are learned from the data. As there is no need to calculate the IoU, this approach reduces training time. The anchor-free object detection can be categorized into two types: algorithms based on key points and those based on the central domain.

Object detection based on the central domain predicts the coordinates of the object's central region and the scale information of the bounding box. The DenseBox algorithm is an early algorithm based on the central domain (Huang et al., 2015). It takes each pixel as the center point and predicts the distance from the point to the upper, lower, left, and right boundaries of the object where it is located and the probability that the object belongs to different categories, that is, it predicts a boundary box for each pixel of the input image. Fully convolutional one-stage object detection (FCOS) algorithm adopts the idea of predicting the bounding box by pixel (Tian et al., 2020), which is similar to the DenseBox algorithm. Meanwhile, it gives a method of overlapping object detection based on multi-scale features. The negative factor of the FCOS algorithm is that when the predicted pixel is far away from the actual center point of the object, the detection result is easily affected by overlapping object semantics.

It is noteworthy that the YOLOv1 algorithm also belongs to an early object detection method based on the central domain. As mentioned above, this algorithm predicts the object's center position within a grid and detects only the object closest to the center point. However, this approach leads to lower accuracy compared with one-stage algorithms that rely on anchor boxes. Therefore, the anchor-free detection algorithm was not adopted in the subsequent YOLO series, but its idea provided a reference for the subsequent anchor-free detection. The CornerNet algorithm (Law and Deng, 2018) is a representative keypoint-based algorithm (Fig. 11). The detection box is obtained by directly predicting the upper left and lower right points of the object, and the detection problem is solved using keypoint detection. Because the features of the object to be detected are easily ignored, the algorithm is not accurate enough for object classification. On this basis, the ExtremeNet algorithm (Zhou et al., 2019b) was proposed, and it made innovations in the selection and combination of key points. To improve the recall rate, CenterNet (Zhou et al., 2019a) added the center point prediction branch on the basis of CornerNet, making it necessary to form an object detection frame not only to match the left and right corners but also to match the center point of the detection frame. CenterNet no longer uses non-maximum suppression to eliminate redundant boxes, resulting in increased algorithm speed to a certain extent.

In scenarios where human intervention or equipment deployment is challenging, such as the inspection of bridge undersides or urban underground culverts, CenWholeNet, an anchor-free detection network employing center points and holistic information, has been introduced (He et al., 2022). Comparative analysis with models like Faster RCNN reveals that CenWholeNet is better suited for detecting intricate defects characterized by substantial variations in length-to-width ratios and intricate environmental conditions. Similar anchor-free detection algorithms have been applied in many fields of traffic, such as pedestrian detection (Liu et al., 2019) and lane detection (Efrat et al., 2020). However, in general, in pavement detection, the performance of the anchor-free algorithm is often worse, although it consumes less memory than the anchor-based detection algorithm.

## 3. Related improvements to enhance performance

From LeNet (LeCun et al., 1998) to the Network with ResNet as the skeleton, the network improved at more levels and showed excellent detection results. VGGNet (Simonyan and Zisserman, 2014), through the construction of 16-layer (13Conv+3FC) and 19-layer (16Conv+3FC) neural networks, showed that the stacking of modules with the same structure can get good results. ResNet (He et al., 2016) proved that it was possible to build a very deep architecture while maintaining stable performance. The research of GoogLeNet (Szegedy et al., 2015) demonstrated that increasing the width of the model is another factor that can lead to improved performance. Xception (Chollet, 2017) and ResNeXt (Xie et al., 2017) proposed that increasing the cardinality of networks can lead to even stronger representations of features. Numerous improvement strategies have been proposed to optimize the model. Among them, the introduction of attention modules has been suggested to enhance the recognition capability of pavement defects and improve detection performance in complex environments. To meet the practical engineering requirements and facilitate the deployment of the model on edge platforms with limited computational capacity, a lightweight design strategy is employed. This

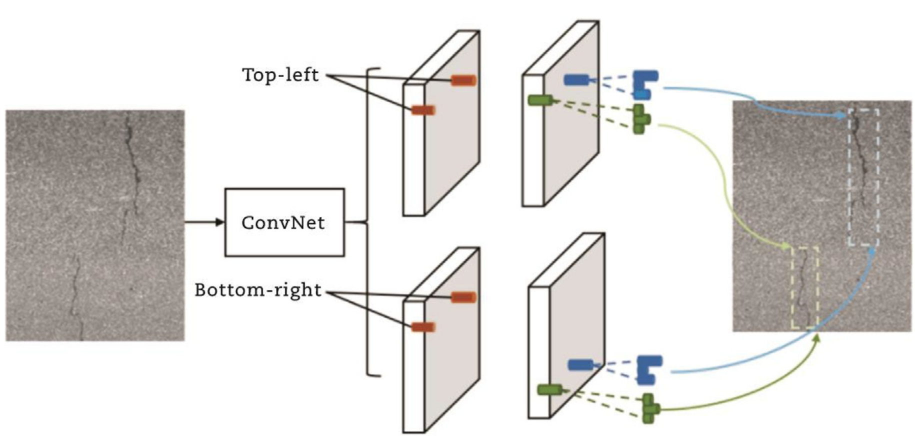


**Fig. 11.** CornerNet model (Law and Deng, 2018).

design approach aims to reduce the model's storage space occupation and accelerate the detection speed. These two aspects constitute the primary focus of this chapter.

### 3.1. Attention mechanism

Attention mechanisms were introduced to enhance the recognition performance of distress targets at a lower computational cost. The idea of the attention mechanism is to use all the information in the context, not just the information in the last state, during the decoding process. Attention mechanisms were first proposed in the field of visual images (Tsotsos et al., 1995). The attention mechanism was first utilized for machine translation in 2014 as a part of the recurrent neural network (RNN) encoder-decoder framework. Current studies mostly focus on natural language processing. This section will cover the attention mechanism in CNNs, which can be divided into two types: single-way and multi-way attention.

#### 3.1.1. Single-way attention mechanism

Squeeze and excitation network (SE-Net) was proposed in the papers included in the computer vision and pattern recognition conference (CVPR), and SE-Net won the first prize in the image classification task of the ImageNet 2017 competition (Hu et al., 2018). The module explicitly models the interdependence between feature graphs, adaptively assigns the weight of each feature map through the learning mode, and reasonably invests computing resources into different channels. The concept of SE-Net is to learn feature weight based on the loss function through the network to train the model by increasing the weight of the feature graph with a more obvious effect on the task and decreasing the weight of the feature graph with a less obvious effect, so as to achieve better results. SE-Net block is a plug-and-play lightweight module. By embedding this module into the network, it can greatly improve the network performance, albeit with a slight increase in the number of parameters.

In the papers included in CVPR in 2020, it was proposed that the efficient channel attention network (ECA-Net) can upgrade SE-Net by incorporating a local cross-channel interaction strategy (ECA block) without dimensionality reduction and a method for adaptively selecting the size of the one-dimensional convolution kernel (Fig. 12). One-dimensional convolution layer is used to summarize cross-channel information to get more accurate information (Wang et al., 2020). The premise of the ECA block is that it is necessary to exchange information across channels, while the SE-Net block focuses only on the integration of information inside channels without considering the importance of information from adjacent channels. Ren et al. (2022) incorporated multiple attention mechanism modules, such as SE and ECA, in the detection of cracks. Instead of replacing the subnetworks in the YOLOv5 model architecture, they trained the model by adding these modules. The original model achieved a precision of 90.91%. After adding the SE and ECA modules, the precision improved to 93.21% and 94.76%, respectively. Similarly, Yao et al. (2022) explored the insertion positions and quantities of various modules, including SCSE and CBAM, based on pavement distress data.

#### 3.1.2. Multi-way attention mechanism

In 2018, convolutional block attention module network (CBAM) was proposed, whose innovation lies in that the feature graph in the convolutional network contains not only rich attention information in the channels but also a large amount of attention information between the pixels of the feature graph (Woo et al., 2018). Previous attention mechanisms focused only on capturing information within channels, resulting in a waste of attentional information across spatial dimensions (Wang et al., 2021). CBAM constructs two sub-modules (Fig. 13), namely the spatial attention module (SAM) and channel attention module (CAM), to capture attentional information across both spatial and channel dimensions, thereby summarizing the relevant information and synthesizing it to some degree. Thus, more comprehensive attention information can be obtained, and more reasonable guidance can be realized for the allocation of computing resources. In the recognition of bridge cracks, CBAM has been employed to optimize the U-Net recognition algorithm (Su et al., 2022), resulting in more comprehensive and accurate identification of cracks. The accuracy achieved is 92.66%.

Pyramid feature attention network also uses the idea that a feature graph is rich in a large amount of attention information in pixels between and within channels (Zhao and Wu, 2019). The structure of the CA branch is the same as that of SE-Net, which is to extract the channel attention information through global average pooling and obtain the weight diagram of each channel through full connection. The spatial attention module (SA) uses the same convolution layer of alternating convolution kernels to extract the attention information between pixel positions within the channel and obtain the correlation and importance of different pixel positions within the channel, as shown in Fig. 14. The two modules extract information from the feature map or the direction of channel and space, respectively. They extract the weight information of different pixel positions in different channels and spaces and update the feature map adaptively.

The development of the mobile terminal network is relatively slow, and the limitation of computing resources prevents the addition of the

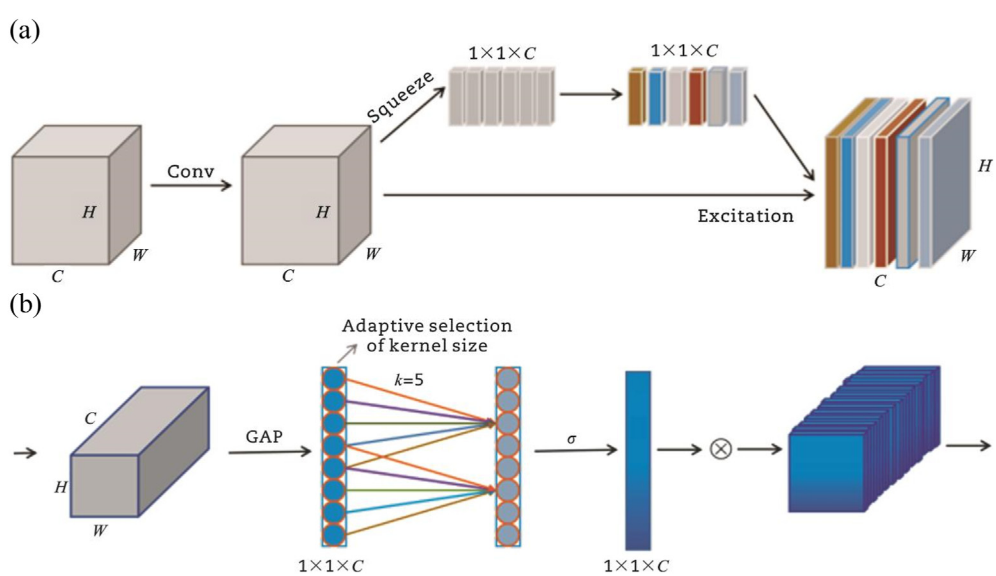


**Fig. 12.** Typical single-way attention mechanism. (a) Squeeze and excitation operation (Hu et al., 2018). (b) ECA-Net (Wang et al., 2020).

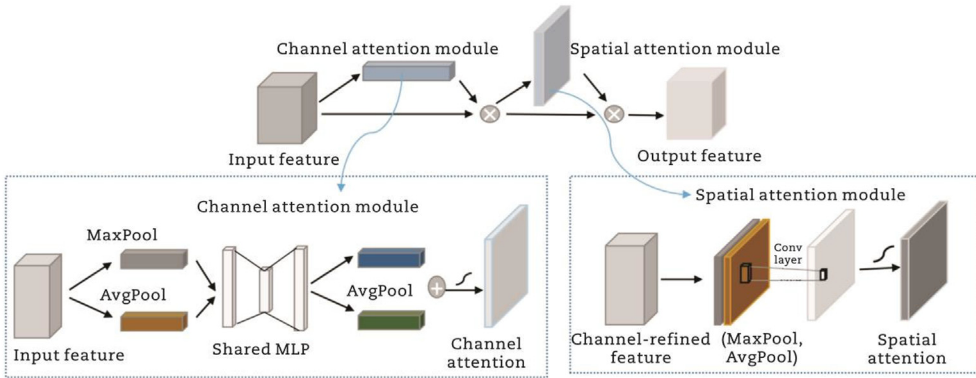


**Fig. 13.** CBAM structure (Woo et al., 2018).

attention mechanism module (Fig. 15). Hou et al. (2021) Introduced the coordinate attention mechanism suitable for the mobile terminal network. While reducing the amount of calculation, the location information is embedded in the channel attention, which increases the receptive field. The issue of location data loss in global pooling is solved by equating channel attention to two parallel one-dimensional features. Specifically, CA uses global pooling in $x$ and $y$ directions to aggregate features in vertical and horizontal directions respectively, so that the input feature map location information is embedded into the aggregation feature vector of channel attention. Conv and concat in Fig. 15 are abbreviations for convolution and concatenation, respectively.

### 3.1.3. Optimization based on attention mechanism

Compared to other common objects, cracks are relatively smaller in size. Therefore, incorporating an attention mechanism subnetwork into pavement crack detection models often enhances the capability of extracting crack features.

In the automated detection models for cracks, the integration of attention modules has been found to enhance the feature extraction capability (Ren et al., 2022). In terms of bridge crack detection, the model is improved by adding the SE module, ECA module, and CBAM module. Comparative experiments show mAP can be increased from 80.5% to 87.0% with the addition of modules. CBAM convolutional module is introduced into the YOLOv7 model to enrich the acquisition of backbone semantic information (Huang et al., 2023), and the $K$-means clustering algorithm is combined to significantly improve the detection accuracy of pavement pothole distress. Yao et al. (2022) conducted focused research on optimizing pavement detection models using attention modules. They extensively investigated the placement and integration methods of different attention modules, resulting in a performance improvement of 6.7% in the enhanced model. Yao et al. (2023a,b) developed the pyramid region attention module (PRAM) to achieve precise extraction of crack information, enabling efficient global multi-scale context integration and capturing long-range dependencies with reduced computational cost. Also based on pavement crack detection, the CrackResAttentionNet network effectively integrates positional attention and channel attention to capture long-range contextual information more comprehensively (Wan et al., 2021).

To assess the severity of pavement defects, Zhang et al. (2022b) manually annotated data with different levels of defect types. They employed an attention fusion module based on edge detection to suppress non-salient features, thereby reducing the false detection rate (Zhang et al., 2022b). The attention mechanism is also applied in the anchor-free detection network, which is better used in the detection of facility security conditions (He et al., 2022). Attention mechanisms have proven to be instrumental in enhancing the detection performance of infrastructure-related tasks, particularly in the identification of large-scale structural cracks (Su et al., 2022). When applied to bridge detection, attention mechanisms effectively enhance the representation of shallow-level information, thereby facilitating more comprehensive and precise crack detection. Attention mechanisms have also gained attention in the field of object segmentation models due to their ability to achieve significant improvements in accuracy with relatively fewer parameters (He et al., 2023). Meanwhile, its plug-and-play feature brings great convenience to the designer, making it an efficient and highly practical mechanism in deep learning.

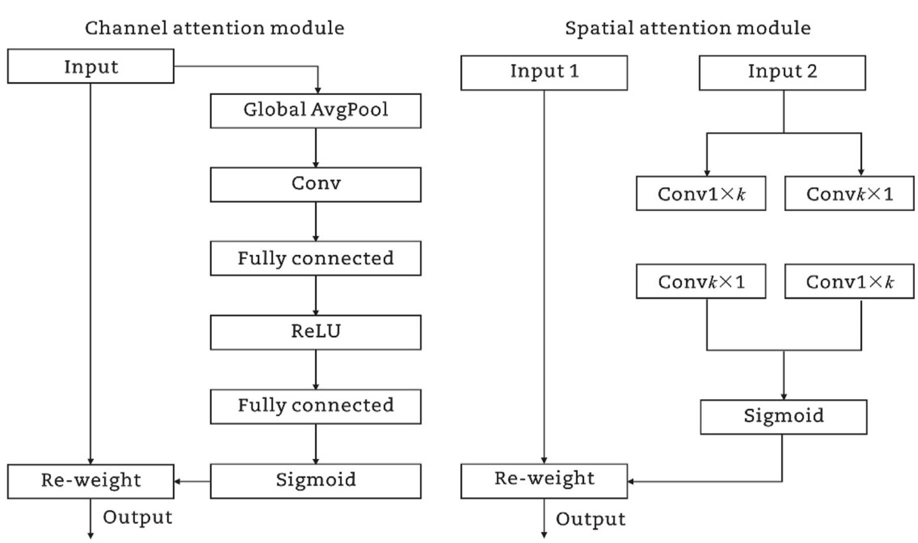


**Fig. 14.** CA module and SA module in pyramid feature attention network (Zhao and Wu, 2019).

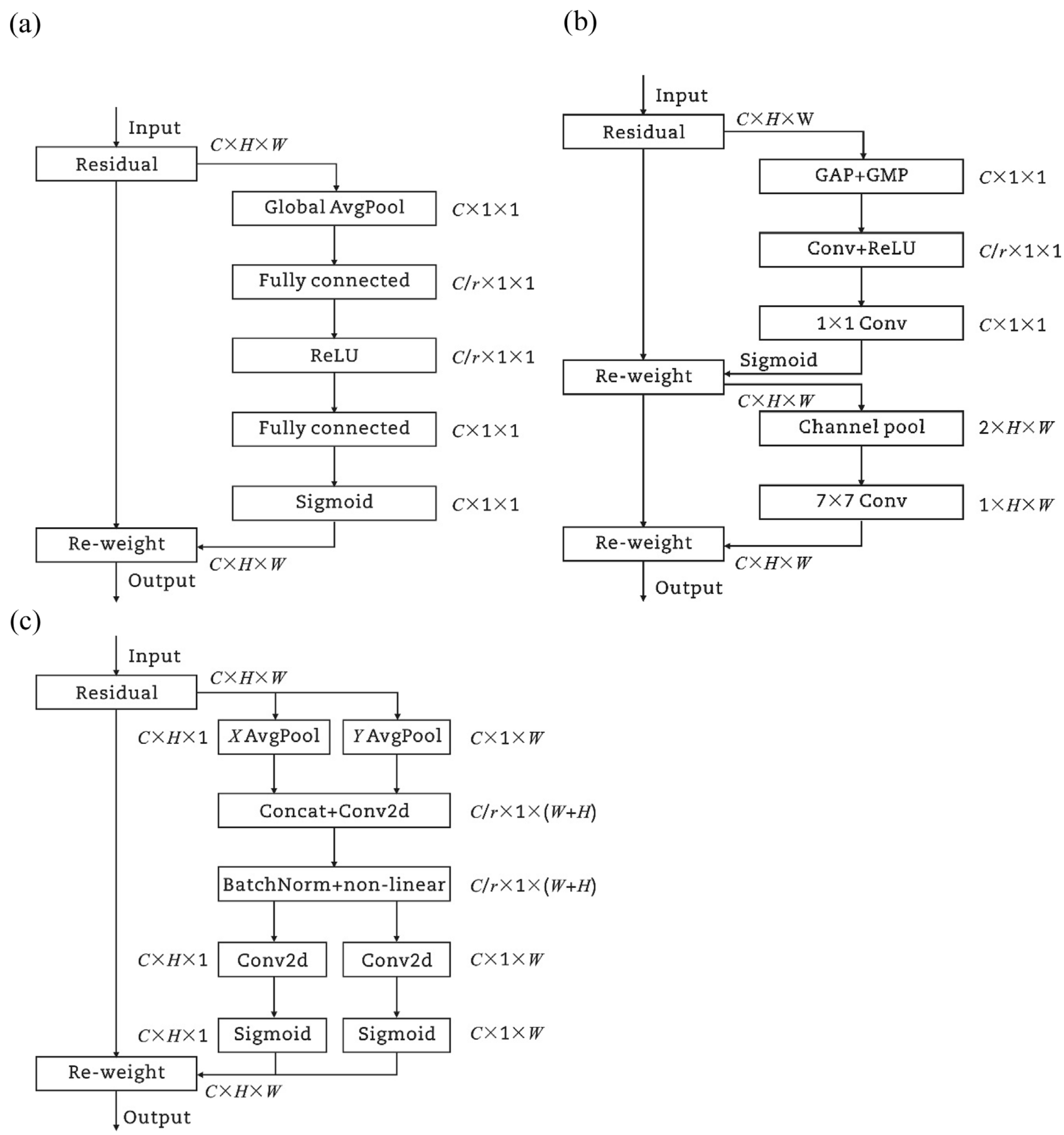


**Fig. 15.** Structural comparison of attention mechanisms. (a) SE attention mechanism (Hu et al., 2018). (b) CBAM attention mechanism (Woo et al., 2018). (c) Coordinate attention mechanism (Hou et al., 2021).

### 3.2. Lightweight network model

There is a demand for integrating object detection functionality directly onto devices. Due to limited hardware capabilities, the detection models need to be lightweight. This means optimizing the network structure while also improving computational efficiency to meet these requirements. A large number of lightweight target detection algorithms, including Tiny-Yolo (Redmon et al., 2016) and Tiny-SSD (Womg et al., 2018), have been developed successively, which effectively improves the performance of lightweight object detection algorithms. This section introduces several typical lightweight object detection architectures and summarizes their innovations.

#### 3.2.1. Lightweight backbone network

(1) **MobileNet**

MobileNetV1, as a lightweight network model based on a manual design method, has the advantages of a few model parameters and fast computing speed, which is suitable for mobile terminal deployment (Howard et al., 2017). Its core is to use deep separable convolution. Meanwhile, the original ReLU activation function is improved by using the ReLU6 activation function ReLU6 = min (max ($x$, 0), 6). The loss of ReLU to low-dimensional feature information is reduced. The model itself has some defects, such as the fact that the structure is similar to the convolution operation stack of VGGNet, each channel is independent of the other, and the features cannot be fused.

Fig. 16 shows the difference between the modules of MobileNetV2 and MobileNetV3. As an improved version of MobileNetV1, MobileNetV2 (Sandler et al., 2018) innovates in two ways. First, the inverted residual block structure was proposed on the basis of the inverted residual block idea. That is, the volume kernel with the size of $1 \times 1$ is first used for dimensional enhancement, then the depth separable convolution with $3 \times 3$ is used for feature extraction, and finally, the filter with the size of $1 \times 1$ is used for feature dimension reduction. Second, a linear bottleneck structure was proposed, that is, the ReLU6 is not used after the $1 \times 1$ convolution operation at the bottom of the bottleneck residual block structure, which avoids data loss caused by the extensive use of

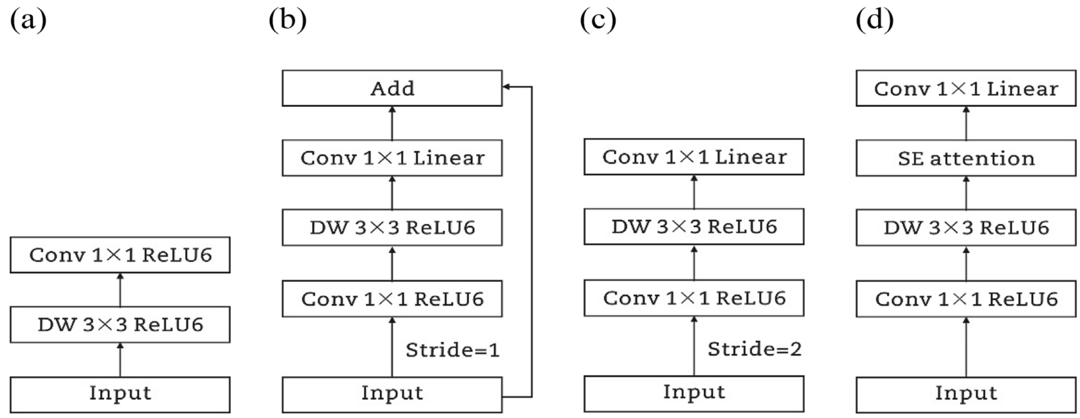


**Fig. 16.** Basic structures of MobileNetV1, MobileNetV2 (Sandler et al., 2018) and MobileNetV3 (Qian et al., 2021). (a) MobileNetV1. (b) MobileNetV2, stride = 1. (c) MobileNetV2, stride = 2. (d) MobileNetV3.

nonlinear mappings. Based on these two innovations, MobileNetV2 has designed a basic structure with steps 1 and 2.

MobileNetV3 (Qian et al., 2021) is the last improved algorithm in this series. Its innovations are the introduction of SE lightweight attention mechanism, changes in nonlinear transformation, modification of the tail structure, and reduction of the number of output channels. These tricks have a great impact on improving classification accuracy and reducing parameters and computations. In the process of lightweight model improvement, MobileNet is commonly chosen as a classical network. The deployment of the model on JETSON XAVIER NX is utilized to validate the proposed method's real-time execution on edge devices with limited computational capabilities. Zhang et al. (2023) attempted to replace the feature extraction network CSP-Darknet53 of YOLOv4 with MobileNetV1, MobileNetV2, MobileNetV3, and GhostNet in order to facilitate model deployment on edge devices with limited computational power. The optimized network occupies approximately five times less memory compared to the original model. Hou et al. (2022) proposed an engineering methodology that combines transfer learning with lightweight models to classify and detect stress on concrete bridges. They utilized a dataset comprising 2500 images of distressed asphalt pavement (asphalt distress dataset, DDAP) and 906 images of distressed concrete bridges (concrete bridge distress dataset, DDCB). In the classification task, they achieved the highest accuracy of 97.8%, presenting a novel approach for the detection and maintenance of intelligent transportation infrastructure.

(2) **ShuffleNetV1-2**

The core of ShuffleNetV1 (Zhang et al., 2017b) is that it uses the strategy combined with group convolution and channel shuffle to reduce the computation cost of 1 × 1 point-wise convolutions (Fig. 17). ShuffleNetV1 is a relatively efficient network model based on the manual design method. Conventional group convolution means that the channels of input maps are split into *n* groups, and every group is learned separately without any influence on each other, which leads to a marginal effect of an output channel learning only a part of the input channel. In order to break through this limitation, ShuffleNetV1 introduces the method of channel shuffle, that is, the features of each group are dispersed into different groups, and then the next group convolution is performed. ShuffleNetV1 is also stacked with some basic units (Shuffle Units).

The author of ShuffleNetV2 (Ma et al., 2018) proposed the guiding principles of efficient network design. (1) Ensure consistency between input and output feature channels. (2) Minimize the number of group convolutions used. (3) Reduce network fragmentation as much as possible. And (4) Reduce usage because although the per-element operation has low FLOPs, its MAC is too high. Accordingly, ShuffleNetV2 is obtained (Fig. 17), which uses a new channel segmentation operation, so that the input channels are split into two sections. One part carries out the real deeply separable convolution operation, and channels splice the results with the other part. Eventually, the channel shuffle operation is carried out to complete the interworking of information. The 1 × 1 group convolution operation is not used in the whole process, and the point-by-point addition operation is also avoided. ShuffleNetV2 removes the channel splitting operation when downsampling is required, so the channels are doubled when the final channel concatenation is performed.

(3) **GhostNet**

The concept of GhostNet (Han et al., 2020) is to utilize fewer parameters to generate more features to achieve a lightweight network structure (Fig. 18). The core innovation of GhostNet lies in the proposal of cheap operation, that is, to obtain a piece of a feature graph through small-scale convolution operation and then perform the linear transformation on each layer of this feature graph to produce a Ghost feature graph. Finally, the above feature graph and the ghost feature graph after the transformation of channel superposition achieve the goal of reducing the amount of computation. It does operations with a conventional convolution kernel of arbitrary size and combines residuary structures in ResNet to propose ghost bottleneck, which are structures with steps equal to one and two, respectively. DWConv still represents depthwise separable convolution in the structure with step two.

GhostNetV2 (Tang et al., 2023) was proposed in 2022 as an enhanced version of GhostNet (Fig. 19), introducing a new attention mechanism (DFC attention) to capture spatial information over long distances while maintaining a lightweight network. decoupled fully connected (DFC) attention mechanism achieves the aggregation of feature map pixels by decomposing the FC layer into horizontal FC and vertical FC. This mechanism is used to introduce the original GhostNet module to strengthen the attention to the information in the middle of the feature graph.

#### *3.2.2. Improvements based on lightweight networks*

Lightweight models enable better utilization of limited computational resources, facilitating fast inference and promoting the deployment and industrial adoption of models on mobile devices. In the process of model lightweight, it is common to utilize the aforementioned model backbone or modules to modify the existing model and achieve the objective of lightweight.

For pavement defect detection, Huang et al. (2023) improved YOLOv7 by adding the Ghost Conv module. This enhancement not only enhances feature extraction but also minimizes the number of parameters in the model, thus minimizing the consumption of computational resources. To address the challenges of false positives and false negatives commonly

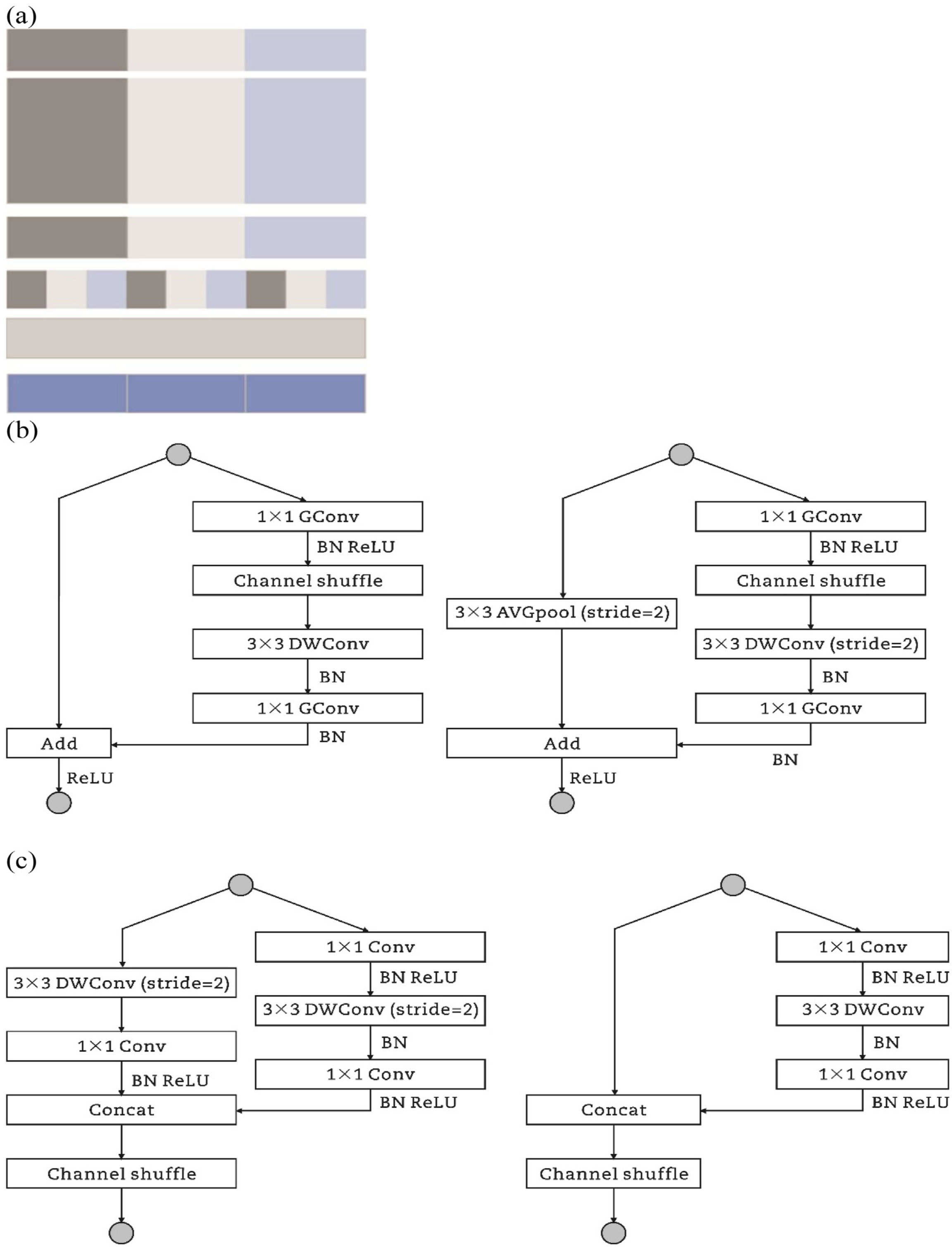


**Fig. 17.** ShuffleNet series modules. (a) Channel shuffle. (b) ShuffleNetV1 basic unit (Zhang et al., 2017b). (c) ShuffleNetV2 basic unit (Ma et al., 2018).

encountered in complex pavement environments, Wang et al. (2022a) employed the quality focal loss technique to enhance the precision of object localization. Additionally, they incorporated the GhostNet concept by substituting the bottleneck of the neck with GhostBottleneck. The algorithm's backbone was replaced with MobileNetV3. As a result, the algorithm achieved a 62% reduction in computational complexity while improving the detection speed by 6 frames per second (FPS) on CPUs. Also, to solve the problem of computing time, Wan et al. (2022) added ECA attention modules to the ShuffleNet lightweight network. Based on YOLOv5, the YOLO-LRDD model is proposed, which can improve the recognition speed of a single image by 22.3%. In practical pavement detection scenarios, MobileNet is commonly utilized as an optimization strategy and a standard for lightweight models to reduce model parameters or optimize model structure. Ahmad et al. (2023) conducted a study on pothole recognition using MobileNetV2 and the ResNet series models. Among them, MobileNetV2 achieved a pothole detection accuracy of 98% (Ahmad et al., 2023). In the improvement of the YOLO model, the application of MobileNetV3 improves the speed of model detection (Liu

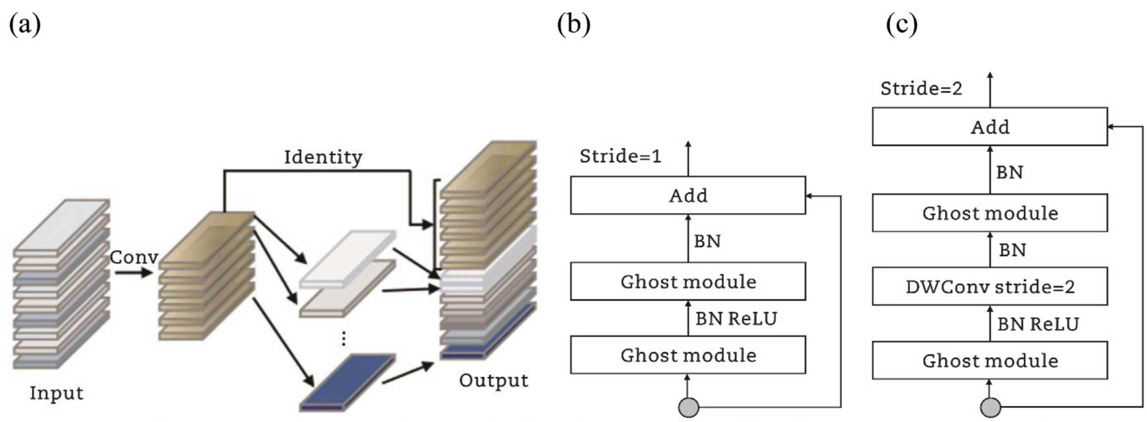


**Fig. 18.** GhostNetV1 (Han et al., 2020). (a) Ghost module. (b) GhostNetV1 bottleneck, stride = 1. (c) GhostNetV1 bottleneck, stride = 2.

et al., 2022). There have been studies specifically focusing on training MobileNet and RetinaNet models and comparing them with other models (Angulo et al., 2019). Qayyum et al. (2023) conducted a directional evaluation of convolutional neural networks used for crack detection, involving GoogLeNet, MobileNetV2, ResNet18, ResNet50, ResNet101, and ShuffleNet. It is worth mentioning that different networks may exhibit variations in performance on different datasets due to the distinct characteristics of various data objects. Subsequently, MobileNetV2, MobileNetV3, ShuffleNetV2, and GhostNetV2 are tested based on the pavement distress data set and radar roadbed data set.

This section describes several lightweight networks and the attention mechanisms commonly used in convolutional networks. In general, the lightweight of the convolutional network is achieved by the replacement of a large convolutional kernel with a small convolutional kernel, a single-size convolutional kernel with a multi-size convolutional kernel, a fixed shape convolutional kernel with a deformable convolutional kernel, etc. On the convolutional layer channel, standard convolution is replaced by group convolution, channel shuffle is performed before group convolution, and channel weighting calculation is performed. Current studies often combine the above two strategies, such as MobileNeXt (Zhou et al., 2020). The above strategies are not only for object detection but also for the dilemma of unstructured pavement with multiple features and complex pavement structures. The semantic segmentation model can also use lightweight feature extraction networks to improve the problem of excessive discrete computation in feature extraction networks. Meanwhile, an attention mechanism is introduced to optimize the

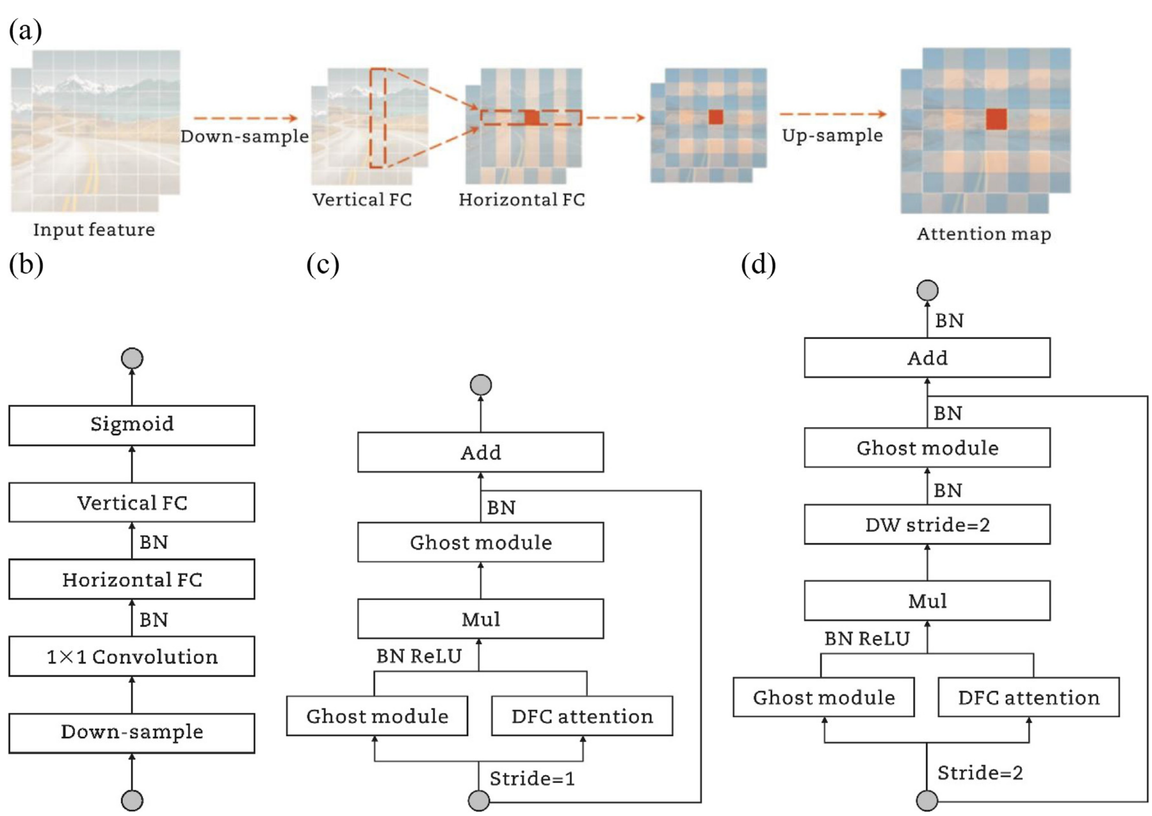


**Fig. 19.** GhostNetV2 (Tang et al., 2023). (a) The information flow of DFC attention. (b) Structural diagram of DFC attention mechanism. (c) GhostNetV2 bottleneck, stride = 1. (d) GhostNetV2 bottleneck, stride = 2.

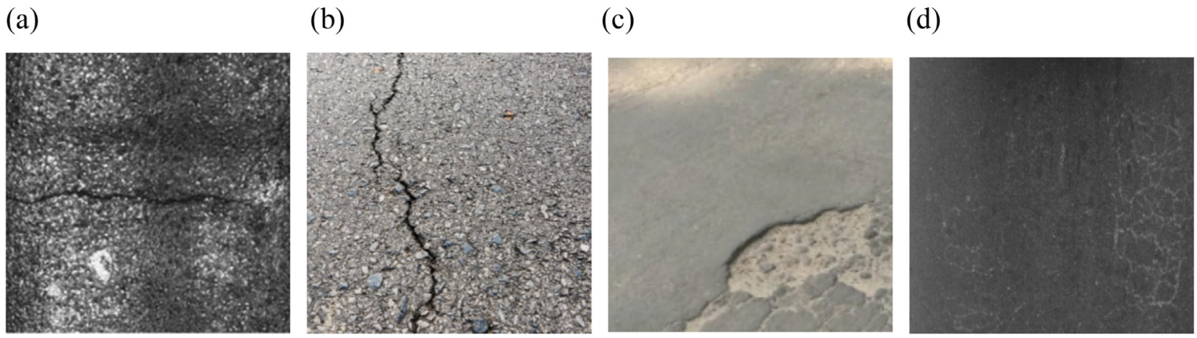


**Fig. 20.** Pavement defect data. (a) Transverse cracks. (b) Longitudinal cracks. (c) Potholes. (d) Alligator cracks.

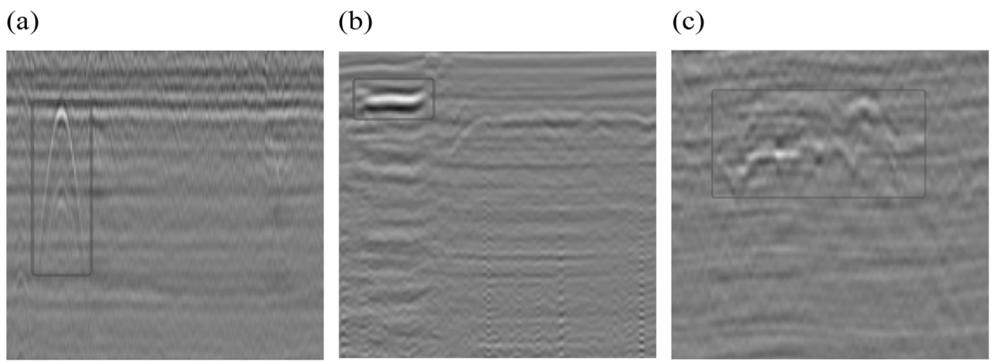


**Fig. 21.** Dataset contains target categories. (a) Pipeline facilities. (b) Loose diseases. (c) Void diseases.

processing of high-level and low-level feature maps to enhance the sensitivity and accuracy of feature extraction (Wang et al., 2022b).

### 3.3. Simple tests of the model

#### 3.3.1. Data set

The performance of mainstream models has been validated on large-scale publicly available datasets. Nevertheless, the applicability of these models for pavement defect detection still lacks clarity. Training and testing were carried out using small-scale datasets to gain deep insights into the generalization capabilities of the model. The pavement distress dataset encompasses 7270 images, featuring four distinct categories: transverse crack, longitudinal crack, alligator crack, and pothole (Fig. 20).

In the subsequent testing phase of lightweight backbone networks, the aforementioned dataset was also employed. Additionally, to assess the generalization of lightweight backbone networks, a comparative evaluation was conducted using a separate dataset consisting of radar-based roadbed data. A total of 2980 radar data samples were collected from actual road scenarios. These samples were processed using the 3DR-Examiner software to visualize cross-sectional profiles, including longitudinal sections, transverse sections, and horizontal slices. The dataset is split into training, validation, and test sets in a 6:2:2 ratio. The data collected in this study include facilities in underground pipes, voids, and anomalous bodies that may be loose. Fig. 21 shows the radar data of the three categories of targets below.

#### 3.3.2. Evaluation metrics

The assessment of the aforementioned model's performance incorporates the concepts of IoU, precision, recall, and mAP. IoU is introduced to evaluate the accuracy of predicted results compared to annotated ground truth. It represents the ratio of the intersection between the real bounding box and the predicted bounding box to their union. A higher IoU value indicates a more accurate prediction.

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

During the training, the IoU can be adjusted to determine the classification of predicted bounding boxes. If the IoU between the predicted anchor and the ground truth exceeds the threshold, it is classified as a true positive (TP). Otherwise, it is classified as a false positive (FP).

As depicted in Fig. 22, the non-overlapping region between the predicted bounding box and the ground truth annotation is considered a FP. The actual overlapping region between the predicted bounding box and

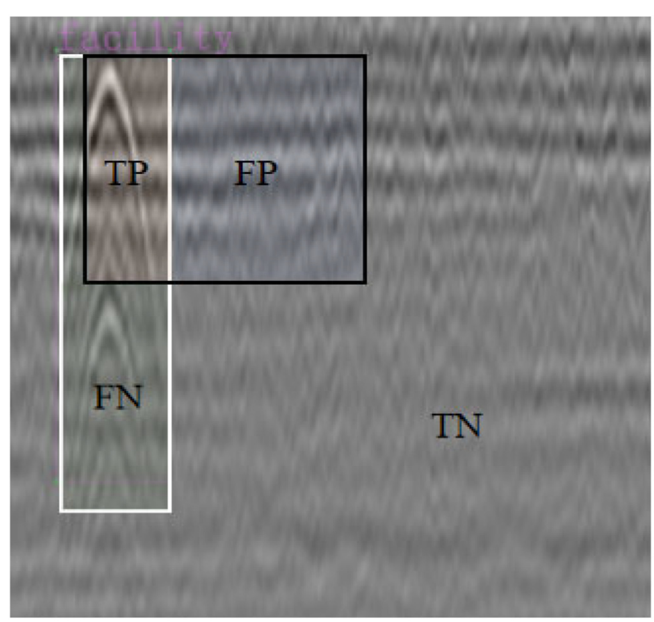


**Fig. 22.** Explanation of IoU diagram.

**Table 1**
Detection results of pavement distress with different algorithms.

| Model | AP (%) | | | | mAP (%) |
|---|---|---|---|---|---|
| | Transverse crack | Longitudinal crack | Alligator crack | Pothole | |
| Faster-RCNN-ResNet50 | 88.38 | 88.57 | 98.44 | 99.87 | 93.81 |
| Faster-RCNN-VGG | 87.17 | 86.27 | 97.99 | 100.00 | 92.86 |
| YOLOv5s | 86.80 | 89.40 | 89.00 | 98.90 | 91.00 |
| YOLOv7 | 88.80 | 86.60 | 99.40 | 98.10 | 93.20 |

the ground truth box is classified as TP. The parts of the ground truth box that are not detected are referred to as false negatives (FN), while the regions that are neither part of the target nor detected are defined as true negatives (TN).

Precision (*P*) is used to measure the proportion of correctly predicted instances in the actual annotation.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Recall (*R*) is a metric that measures the proportion of correctly predicted instances among the predicted results.

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

By varying the IoU thresholds, multiple sets of *P* and *R* values can be obtained. The precision values are plotted on the *y*-axis, while the recall values are plotted on the *x*-axis, thereby forming the precision-recall curve (*P-R* curve). The area under this curve, which is enclosed by the curve and the coordinate axes, represents the average precision (AP). By averaging the AP values obtained for all detected targets, we derive the mAP.

#### 3.3.3. Model performance test

To explore the effect of the model on pavement disease data detection, the current mainstream target detection algorithms were tested. The results are shown in Table 1.

From the performance comparison, it is evident that the two-stage model Faster RCNN with ResNet as the backbone network can achieve higher average precision. However, this model requires more training time. The YOLOv5s model, on the other hand, has a smaller weight file size of 13.7 MB compared to 108 MB for Faster-RCNN-ResNet50, 521 MB for Faster-RCNN-VGG, and 284 MB for YOLOv7. In cases where the difference in average precision is not significant, the YOLOv5s model seems to be more suitable due to its smaller parameters.

Subsequently, the detection effect of the lightweight model was tested based on the YOLOv5 model. The radar data set is relatively small, so the YOLOv5n model is selected for training. The results of the model based on pavement distress data are shown in Table 2. The results of the model based on radar data are shown in Table 3.

In terms of the overall effect, the lightweight of the model causes the reduced performance of the model. Among the four network architectures, ShuffleNetV2 performs poorly on the two data sets, while GhostNetV2 performs well. It is not absolute that the lightweight model may reduce the model detection effect. Compared with the original YOLOv5s model, the number of parameters of GhostNetV2 in pavement distress detection is reduced, but the performance is improved to a certain extent.

The shortcoming of this experiment is that only the mainstream model is applied to radar and pavement distress data to compare the detection effect and generalization of the model. The optimization strategy of the model was not explored in depth.

## 4. Object detection based on 3D data

Reviewing the history of artificial intelligence, it is not difficult to find that the breakthrough of hardware equipment is of great help to the research of algorithms. As hardware continues to change, the cost of acquiring and processing 3D data continues to decrease.

### 4.1. 2D to 3D detection

At present, mainstream computer vision is based on 2D data, while 3D object detection still needs to be improved. The range of information extracted from 2D space is limited, and any type of distressed 2D image analysis is a challenging task. The cracks within the image are characterized as elongated, uneven, and dark lines of strength, which are enclosed by robust texture noise. 2D object detection only relies on regressing the pixels of an object from a flat surface, which results in a lack of depth and other deformation-related information for pavement distress, such as rutting. Moreover, the 2D image detection methods struggle to differentiate between pavement defects and the influence of shadows in dark areas (Zhang et al., 2017a).

3D laser scanning data, which acquires depth information unaffected by lighting conditions, is gradually becoming the mainstream method for pavement data collection. In recent studies, it is proposed a sparse processing algorithm for extracting the 3D pavement contour (Tsai and Chatterjee, 2018). By combining the extracted candidate points with an improved minimum-cost spanning tree algorithm, crack information extraction was achieved. This approach achieved a crack detection accuracy of over 98% due to the utilization of 3D information data. Although 3D imaging technology can provide more information, it involves expensive equipment and high computational costs. Image data, point cloud, and voxel data are the primary data types used in current 3D object detection. From the current development level, visual methods mainly use CNN based on deep learning, and the recognition of images has reached a high accuracy. This kind of network is widely used due to the advantages of low cost and rich features. Laser point cloud data can obtain obvious 3D features, so it is widely utilized in 3D object detection.

Unlike the review, this chapter analyzes the imaging principles of different devices and compares the pros and cons of various methods (Mathavan et al., 2015). In this chapter, according to the different data detection methods, 3D object detection is classified into image-based and

**Table 2**
Detection results based on the pavement distress dataset.

| Model | Precision (%) | Recall (%) | mAP_0.5 (%) | Parameter |
|---|---|---|---|---|
| GhostNetV2 | 95.08 | 94.47 | 97.31 | 6,765,137 |
| MobileNetV3 | 88.20 | 92.42 | 94.74 | 1,381,247 |
| MobileNetV2 | 92.86 | 91.73 | 95.73 | 1,381,247 |
| ShuffleNetV2 | 67.60 | 70.00 | 70.40 | 846,417 |
| YOLOv5s | 93.91 | 93.46 | 96.98 | 7,071,633 |

**Table 3**
Detection results based on the radar data.

| Model | Precision (%) | Recall (%) | mAP_0.5 (%) | Parameter |
|---|---|---|---|---|
| GhostNetV2 | 92.50 | 88.12 | 90.31 | 1,708,608 |
| MobileNetV3 | 88.91 | 86.63 | 89.70 | 801,942 |
| MobileNetV2 | 84.00 | 80.81 | 85.32 | 221,672 |
| ShuffleNetV2 | 76.14 | 77.60 | 80.50 | 217,256 |
| YOLOv5n | 94.42 | 87.71 | 91.90 | 1,767,976 |

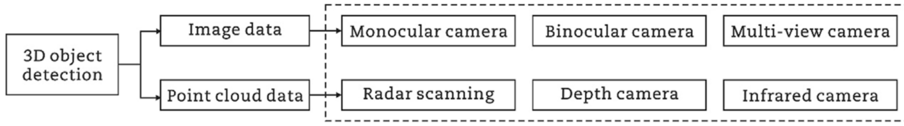


**Fig. 23.** 3D object detection classification based on data.

point cloud-based data. The boxes with the dashed lines are the part not covered in this article (Fig. 23).

### 4.2. Methods based on monocular images

Vision sensors (e.g., monocular cameras, binocular cameras) cost much less than LiDAR. Especially in recent years, the research on depth estimation of image data has also made relatively outstanding achievements, which also shows that effective depth features can be obtained through image data. Monocular vision provides a stable, rich, accurate appearance and other semantic information for object detection networks, and it is not easily interfered with by external factors. The data processing process is simple, so it has attracted much attention in practical applications. However, due to the lack of depth information for 3D object detection, it is essential to improve positioning accuracy. Although binocular vision can obtain more accurate depth through image processing, the sensor is poor, and it is susceptible to interference from external factors.

The Mono3D object detection method was first proposed by Chen et al. (2016). The Mono3D algorithm consists of two main steps for 3D object detection. RPN network is used to perform dense sampling of 3D space and generate several 3D candidate box regions. Then, the 3D detection box is used to generate the corresponding 2D image region by projection transformation, and the confidence degree of the 3D candidate box is calculated in the image plane. After non-maximum suppression and other post-processing methods, the final 3D object detection result is obtained. As an earlier proposed 3D object detection algorithm based on monocular vision, its performance in object positioning accuracy is far below the requirements of practical applications. Its algorithm uses the method of dense sampling and multi-data prior information fusion, which not only makes the calculation of the network very large but also makes the stability of the model poor. Mousavian et al. (2017) proposed the Deep3Dbbox algorithm, which utilizes the well-established 2D object detector and introduces a multi-bins hybrid discrete-continuous regression approach. This method simplifies the network structure by using the 2D object detection method and uses the regression method to get the depth information of the object. However, due to the imprecision of in-depth information, this method may not lead to a significant improvement in detection accuracy. Researchers have recommended various solutions for depth estimation due to the dearth of depth information in monocular vision, including the fusion visual depth estimation feature method and computer-aided design and drafting (CAD) template matching constraint method. For instance, the MF3D method proposed by Xu and Chen (2018) introduced the sub-NET module of the depth estimation network of monocular vision. The location information estimation is achieved by fusing the RoI region features produced by the Deep3Dbbox algorithm with the depth map features. As depicted in Fig. 24, the use of both view synthesis and stereo-matching networks enhances the geometric reasoning capabilities. In Fig. 24, the combination of view synthesis and stereo-matching networks effectively enhances the geometric reasoning capabilities.

In the application of monocular vision, Liebold and Maas (2020) employed image sequence analysis techniques to observe crack samples through a sequence of monocular images. They utilized a least squares method to achieve fitting of narrower crack widths. Zhang et al. (2020) addressed the issue of low efficiency in concrete crack measurement by combining monocular vision and image processing techniques to develop the efficient and automated Mono-Harris method. To address the challenge of distinguishing 2D image defects from the surrounding materials, an approach was adopted to utilize a depth estimator for prediction (Duy and Natori, 2020). The training was conducted in an unsupervised manner using three consecutive monocular image sequences to identify surface defects. This method aimed to overcome the difficulty of distinguishing defects from the surrounding materials by incorporating depth estimation into the prediction process.

### 4.3. Methods based on binocular images

By calculating the relative positions of the cameras, binocular stereo vision can obtain stronger spatial constraint relationships than monocular vision. A comprehensive summary of the typical methods for monocular and binocular vision is presented based on algorithmic mechanisms, as depicted in Table 4. Building upon fast RCNN, the 3D object proposals (3DOP) network (Chen et al., 2017) is structured to generate 3D candidate regions through encoding object size priority, ground plane estimation, and depth information. Additionally, a loss function based on multi-feature prior information is proposed by the 3DOP network. Then, the 3D candidate box is regressed and classified by the RCNN network to complete the detection task. Li and Qin (2018) extended the dynamic structure method to dynamic object detection. By fusing spatiotemporal information, 3D objects, and camera poses were continuously tracked. A 3D object detection algorithm based on Stereo RCNN (Li et al., 2019) has been proposed by building upon the Mask RCNN architecture. This algorithm utilizes an alignment algorithm based on dense regions to ensure the accurate positioning of 3D objects during detection. Despite utilizing depth information, the localization effect of this method cannot be matched by detectors based on laser point clouds.

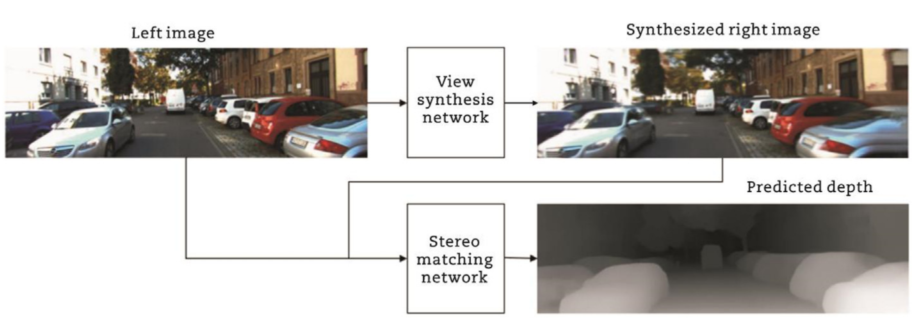


**Fig. 24.** Monocular depth estimation method (Luo et al., 2018).

**Table 4**
Algorithm comparison of monocular vision and binocular vision.

| Type | Algorithm mechanism | Correlation method | Characteristic analysis |
|---|---|---|---|
| Monocular vision | Depth estimation | MF3D | Combining a 2D detection network and depth estimation network, the realization method is simple, but the depth estimation error is large under monocular vision, and the multi-network structures easily cause error accumulation. |
| | Fusion of prior information | Mono3D | A substantial amount of prior information needs to be fused, and the prior information is complex, so the end-to-end detection network cannot be realized. |
| Binocular vision | Depth map features | 3DOP | Combined with depth map features, the 2.5D region proposal network is more accurate. The single processing of image features and depth map features ignores the sterility of space. |
| | 3D convolution | Stereo RCNN | 3D spatial mapping is established by binoculars, and a 3D convolutional network is realized. The large amount of computation leads to the waste of computing resources. |

In the recognition of cracks, Liu (2021) conducted research on the effectiveness of binocular vision technology for long-distance crack measurement using a comparative and analytical approach. Through experimental measurements, they were able to control the error within 9%. The study aimed to assess the accuracy of crack measurement using binocular vision techniques and demonstrated promising results in effectively measuring cracks from a distance (Liu, 2021). Tang et al. (2018) proposed a least squares-based circle fitting algorithm and utilized binocular technology to measure concrete deformation. Their study concluded that this method can serve as a viable alternative to traditional contact-based deformation measurement techniques. In order to conduct pavement smoothness testing, utilized binocular disparity to obtain the 3D coordinates of points on the pavement surface (Yi et al., 2021). By leveraging binocular vision and extracting depth information through disparity, the proposed method enabled precise assessment of pavement flatness. The integration of Gabor filtering and multi-frame fusion further enhanced the accuracy of the detection process.

### *4.4. Methods based on point cloud*

Point cloud expresses surface feature information and spatial distribution information of objects through a vast dataset of data points. Compared with image information, point clouds can better restore the 3D features of the target.

The efficient utilization of point cloud data has become a key focus for researchers because of the disorder, lack of structure, and sparsity that characterizes this type of data. The PointNet proposed by Qi et al. (2017a) solves the disorder problem by max pooling with symmetry. It extracts features through the convolution layer and then selects the most influential features as the global features of the point cloud via maximum pooling. These selected global features are then inputted into a subsequent network for classification and segmentation tasks. Li et al. (2018a) constructed a convolution operator to tackle the issue of point cloud disorder. Liang et al. (2019) processed the disordered point cloud data through graphic convolution, which exhibits a strong capacity to extract local point cloud features. The multi-view convolutional neural network (MVCNN) method was the first to obtain images of 3D point cloud objects from different perspectives, project the 3D information into multiple 2D image information, and process the data using convolution and other techniques (Su et al., 2015). The primary objective of this method is to structure the data and leverage mature 2D image processing techniques for analysis. On this basis, Feng et al. (2018) proposed the group-view convolutional neural network (GVCNN) framework, which added a grouping mechanism to strengthen feature association between multi-view data. The spatial distance difference between sensors and different objects leads to different sparsity of 3D point clouds. One issue is that the weight of updating network parameters is not always reasonable when upsampling or undersampling in network training. For this reason, different density modules are used to alleviate the impact of point cloud sparsity to a certain extent (Li et al., 2020; Wen et al., 2021). When point cloud data is converted to 3D grids, voxel data are generated. Voxels (Meng et al., 2019; Xu et al., 2017) are similar to images that expand from 2D pixels to 3D cube units to address the issue of disorder and disorganization in point cloud data.

Additionally, research that directly analyzes the raw point cloud data can fully utilize the information contained in the point cloud, avoiding potential information loss or distortion that may occur when converting it into image data. PointNet++ (Qi et al., 2017b) innovatively solves the problem of how to extract features directly from 3D point clouds and applies it to classification and segmentation tasks. PointRCNN (Shi et al., 2019) uses the foreground point as the target center to represent the object in the 3D scene, transforming 3D target detection into the localization and classification task of sampled target points in the point cloud.

Overall, due to hardware cost considerations, spatial scanning points exhibit sparsity. Applying projection methods to downsample high-resolution maps can lead to the disappearance of distant or small-sized targets. Consequently, there has been increased attention on geometric inference using local point clouds (Liu et al., 2020). Point clouds are essentially located in non-Euclidean metric spaces, where they can be regarded as a topological graph with corresponding relationships between vertices and edges, enabling the learning of spatial characteristics on this graph. In recent research, the utilization of graph convolutional networks (GCNs) to process raw point clouds has been mentioned, introducing learnable convolutional parameters (Shi and Rajkumar, 2020). This approach optimizes the topological graph and facilitates the learning of point cloud features. In the research of pavement crack width acquisition technology, a method to detect pavement cracks from moving laser scanning data is proposed (Zhong et al., 2020). The 2D refractive index of each laser point is provided by laser scanning angle or acquisition time, which solves the problem of topological correctness well and makes the average prediction accuracy of crack width exceed 0.812. Feng et al. (2022) conducted a comparative study of multiple detection models for laser scanner point clouds and calculated the differences between each algorithm's estimation of crack area and reference data.

### *4.5. Methods based on multi-views*

Laser scanners are advantageous in providing precise depth information, whereas cameras are capable of capturing more detailed semantic information. The main purpose of this method is to take images from different perspectives, such as photos obtained from cameras and projection images of point clouds, as the input of the deep model and use CNN to learn features and output 3D rectangular boxes. There are two main methods for multimodal data fusion: pre-fusion and post-fusion. Pre-fusion first maps the data of different modalities into the same feature space and then processes them uniformly, while post-fusion processes the data of different modalities separately and then fuses the results. SPLATNet proposed by Su et al. (2018) utilizes the idea of pre-fusion. Related studies, such as Qi et al. (2018), Shi et al. (2019), refer to a typical approach of post-fusion.

A combination of LiDAR and images has been used for 2D detection, but for 3D object detection, a model that takes full advantage of the strengths of multiple modes is needed (González et al., 2017). An MV 3D model was proposed by Chen et al. (2017b). The RGB image, radar aerial view, and radar front view are used as the input of the network to achieve accurate vehicle identification and 3D rectangular box regression. The model uses a region-based fusion network, using RoI pools to obtain the same feature vectors for each pattern. The deep integration approach enhances the interaction between the middle layers from different perspectives. The AVOD model (Ku et al., 2018) was proposed by improving the MV3D method. An FPN network is used to convert RGB and aerial view images into full-resolution images. Through feature fusion, a 3D region proposal is selected to realize 3D object detection.

To achieve accurate crack detection, Chen et al. (2021) proposed an automated solution for crack detection by combining 3D point clouds and 2D images using an improved OTSU algorithm. The proposed method was tested on various types of concrete and cracks of different scales. The new approach achieved an average precision of 89.0% and an F1 score of 86.7% (Chen et al., 2021). In contrast, the methods that relied on a single technique for detection achieved F1 scores of 67.6% and 76.0% respectively. In order to detect pothole distress, Wu et al. (2019) used convolution technology to extract 2D potholes in images, point cloud technology to obtain 3D potholes, and in-depth analysis to finally determine the pavement distress. A simulation case and a real case of a 26.4 km expressway were used for verification, and the average accuracy of potholes measured was 1.5–2.8 cm.

## 5. Discussions

This paper provides a summary and overview of object detection techniques in both the 2D and 3D domains. Deep learning-based algorithms for 2D object detection, regardless of the framework used, have shown significant improvements compared to traditional machine learning techniques. To overcome the limitations of applications due to the lack of depth and other parameter information, 3D object detection techniques have been introduced. On the whole, the YOLO series algorithm in the field of 2D object detection has stronger generalization and is still the mainstream application. In 3D object detection, the accuracy of monocular vision is poor for multi-object detection. The utilization of binocular vision for establishing 3D spatial mapping often leads to wastage of computational resources due to excessive computational requirements. Although the point cloud makes the 3D features more prominent, the acquisition cost is high and the network structure is too complex. Multi-view enhances the accuracy of the algorithm through multi-angle fusion, but it requires additional processing operations such as multi-camera and data fusion, and the computational cost is also high. Therefore, a careful balance needs to be struck when choosing 3D detection techniques, considering various factors.

Nowadays, the surge of deep learning has led to many articles being reviewed and utilized from different views. Researchers classify the various object detection mechanisms (Sharma and Thakur, 2017) based on the types of search, classification, and others. The researchers are more inclined to introduce detection equipment for pavement distress classification and pavement condition evaluation research (Coenen and Golroo, 2017). Furthermore, the field of target detection has been classified into two main lines: one-stage and two-stage detections (Zou et al., 2019). After sorting and summarizing the literature, the following research and hot topics can be summarized as follows.

- Object detection accuracy in specific scenes still needs to be further improved, especially for pavement distress characteristics with diverse shapes. For instance, to address the issue of shadows that have a similar intensity to cracks, a geodesic shadow removal algorithm is employed, followed by the establishment of a crack probability map using tensor voting (Zou et al., 2012). For crack detection on pavement characterized by low contrast and complex textured backgrounds, an unsupervised multi-scale fusion crack detection approach is employed (Li et al., 2018b), eliminating the need for training data. Due to the limitations of accurately detecting cracks in concrete caused by bubbles, stains, and shadows, a denoising method based on permeation modeling is proposed (Yamaguchi et al., 2008). Further investigation into the technical effectiveness of specific application scenarios is warranted.
- It is necessary to research the backbone network to obtain high performance. Improving the specificity of the backbone network for different application tasks can build a high-resolution lightweight network and optimize feature fusion to achieve performance improvement.
- For complex realistic scenes with small objects and easily blocked objects, how to effectively combine contextual information for detection remains to be broken through. Part of the improvement of the YOLOv1 subsequent algorithm is to solve this problem.
- From the perspective of the macro system of 3D object detection, there are few 3D object detection methods based on binocular vision, and there is still much room for breakthroughs in related research.
- The detection model of 3D detection also needs lightweight processing. Generally, in deep learning tasks, the model performance is often improved by adding components, which makes the network more complex. Network pruning can be used to remove network redundancy (Liu et al., 2017). In addition, an effective network structure can be designed by reducing the size of the convolutional kernel and the number of channels.
- The advantages of different data modalities can compensate for the limitations of individual sensors, thereby enhancing the accuracy of 3D detection. By leveraging multiple sensors, such as LiDAR point clouds and RGB images captured by cameras, the complementary nature of these data sources can improve the overall detection performance. It is necessary to further explore the selective combination of multiple sensors as inputs at different structural levels of neural networks. Furthermore, LiDAR is susceptible to the interference of rain particles, resulting in the generation of noisy points. The development of robust algorithms to handle noise is a promising direction for further advancement.
- In 2D object detection, the convolutional neural network improves its performance through the introduction of the attention mechanism. At present, obtaining spatial attention and channel attention information is the mainstream method. Exploring new directions to obtain richer attention information from feature maps, or obtaining more accurate attention information in new ways or means, is also a key point that needs attention in the future.
- The original target detection framework is conducive to obtaining more target information by combining other learning tasks and establishing a parameter-sharing mechanism through multi-task learning can significantly improve the performance of target detection. For example, semantic segmentation is used to obtain the edge information and context information of the target image, which is helpful for the subsequent target detection task to identify and locate the target better.
- LiDAR is expensive, and point cloud data is difficult to obtain and label. 2D image data is easily obtainable and cost-effective, but the depth information of the target is lost. Using 2D image data to drive 3D object detection to reduce the dependence on point cloud data may also be a feasible direction.

It is worth mentioning that the transformer structure has demonstrated significant success in the domain of natural language processing, and in 2020, it was successfully employed in image classification tasks, surpassing the performance of CNN. Transformer multi-head self-attention mechanism and efficient computing greatly promote the development of image-related tasks, and networks such as ViT transformer (Dosovitskiy et al., 2020) and Swin transformer (Liu et al., 2021) are bound to get more attention from researchers.

## 6. Conclusions

Based on the background of object detection technology in intelligent pavement maintenance, this paper summarizes the algorithms of object detection for automatic application. The two-stage detection algorithm can compose the detection target into candidate regions, perform the corresponding regression and classification processing, and then realize the detection candidate processing. Subsequent relevant studies have been carried out on improving the feature extraction network, the region pooling layer of interest, the region extraction network, the NMS, and other modules. Considering the comparison with engineering applications, it becomes evident that one-stage models have emerged as the prevailing choice. Notably, subsequent improvements have led to the one-stage model's widespread superiority over two-stage models, both in terms of final testing performance and model training cost. This paper also introduces structural models and relevant techniques for 3D object detection based on pavement distress detection.

The optimization of pavement detection models can prioritize enhancing the data aspect by accounting for the environmental complexity and the diverse nature of pavement defects. In practical application, precision and speed need to be balanced in combination with the actual scene. It is noteworthy that lightweight detection algorithms are being proposed and improved to enable target detection network architectures to be deployed on industrial platforms. In the follow-up work, lightweight networks have been applied in YOLO series or RCNN series algorithms, and attention mechanisms have been added to continuously break through the current network performance. From the perspective of whether an anchor frame is applied, the algorithm based on the anchor frame has developed into a relatively complete system rapidly. The frameless object detection algorithm performs better than the frame-based target detection algorithm in small target detection in specific scenes. The frameless target detection is proposed to make up for the defects of the frame-based target detection algorithm in small target detection. This kind of detection algorithm appeared relatively late and still has much room for research.

Overall, this paper provides a comprehensive review of object detection algorithms. It combines practical detection scenarios to elaborate on the mainstream models and relevant optimization strategies for the 2D and 3D detection of pavement. The exposition adheres to the evolving trends in object detection and endeavors to offer novel insights for the utilization of artificial intelligence in the realm of road engineering.

## Declaration of competing interest

Zhanping You and Hui Yao are the editorial boad member and young academic editor of Journal of Road Engineering respectively. They were not involved in the editorial review or the decision to publish this article. All authors declare that there are no competing interests.

## Acknowledgments

The first author appreciates the financial support from Hunan Provincial Expressway Group Co., Ltd. and the Hunan Department of Transportation (No. 202152) in China. The first author also appreciates the funding support from the National Natural Science Foundation of China (No. 51778038) and the Beijing high-level overseas talents in China. Any opinion, finding, and conclusion expressed in this paper are those of the authors and do not necessarily represent the view of any organization.

## References

Ahmad, C.F., Cheema, A., Qayyum, W., et al., 2023. Classification of Potholes Based on Surface Area Using Pre-trained Models of Convolutional Neural Network. Available at: https://arxiv.org/abs/2309.17426 (Accessed 7 January 2024).

Akagic, A., Buza, E., Omanovic, S., et al., 2018. Pavement crack detection using OTSU thresholding for image segmentation. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, Opatija, 2018.

Al-Amri, S.S., Kalyankar, N.V., Khamitkar, S.D., 2010. Image Segmentation by Using Threshold Techniques. Available at: https://arxivorg/abs/1005.4020 (Accessed 7 January 2024).

Ali, A.A., Milad, A., Hussein, A., 2023. Predicting pavement condition index based on the utilization of machine learning techniques: a case study, 2023. Journal of Road Engineering 3 (3), 266–278.

Alzraiee, H., Leal Ruiz, A., Sprotte, R., 2021. Detecting of pavement marking defects using faster R-CNN. Journal of Performance of Constructed Facilities 35 (4), 04021035.

Amhaz, R., Chambon, S., Idier, J., et al., 2016. Automatic crack detection on two-dimensional pavement images: an algorithm based on minimal path selection. IEEE Transactions on Intelligent Transportation Systems 17 (10), 2718–2729.

Angulo, A., Vega-Fernández, J.A., Aguilar-Lobo, L.M., et al., 2019. Road damage detection acquisition system based on deep neural networks for physical asset management. In: 18th Mexican International Conference on Artificial Intelligence, Xalapa, 2019.

Attard, L., Debono, C.J., Valentino, G., et al., 2019. Automatic crack detection using mask R-CNN. In: 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA), Dubrovnik, 2019.

Bochkovskiy, A., Wang, C., Liao, H., et al., 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. Available at: https://arxir.org/abs/2004.10934 (Accessed 7 January 2024).

Bodla, N., Singh, B., Chellappa, R., et al., 2017. Soft-NMS–improving object detection with one line of code. In: 2017 IEEE International Conference on Computer Vision, Venice, 2017.

Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2 (2), 121–167.

Cafiso, S., Di Graziano, A., Battiato, S., 2006. Evaluation of pavement surface distress using digital image collection and analysis. In: Seventh International Congress on Advances in Civil Engineering, Istanbul, 2006.

Chen, Z., Huang, S., Tao, D., 2018a. Context refinement for object detection. In: 15th European Conference on Computer Vision (ECCV), Munich, 2018.

Chen, X., Kundu, K., Zhang, Z., et al., 2016. Monocular 3D object detection for autonomous driving. In: IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016.

Chen, X., Kundu, K., Zhu, Y., et al., 2017a. 3D object proposals using stereo imagery for accurate object class detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (5), 1259–1272.

Chen, X., Li, J., Huang, S., et al., 2021. An automatic concrete crack-detection method fusing point clouds and images based on improved OTSU's algorithm. Sensors 21 (5), 1581.

Chen, Y., Li, Y., Wang, G., 2018b. An enhanced region proposal network for object detection using deep learning method. PLoS One 13 (9), e0203897.

Chen, X., Ma, H., Wan, J., et al., 2017b. Multi-view 3D object detection network for autonomous driving. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 2017.

Chollet, F., 2017. Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017.

Coenen, T., Golroo, A., 2017. A review on automated pavement distress detection methods. Cogent Engineering 4 (1), 1374822.

Cutler, A., Cutler, D.R., Stevens, J.R., 2012. Random forests. Ensemble Machine Learning: Methods and Applications 157–175.

Dai, J., Qi, H., Xiong, Y., et al., 2017a. Deformable convolutional networks. In: 2017 IEEE International Conference on Computer Vision, Venice, 2017.

Dai, B., Zhang, Y., Lin, D., 2017b. Detecting visual relationships with deep relational networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Available at: https://arxiv.org/abs/2010.11929 (Accessed 7 January 2024).

Du, F., Jiao, S., 2022. Improvement of lightweight convolutional neural network model based on YOLO algorithm and its research in pavement defect detection. Sensors 22 (9), 3537.

Duy, T.V.T., Natori, N., 2020. Efficient defect detection from consecutive monocular images by deep learning. In: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, 2020.

Efrat, N., Bluvstein, M., Oron, S., et al., 2020. 3d-LaneNet+: Anchor Free Lane Detection Using a Semi-local Representation. Available at: https://arxiv.org/abs/2011.01535 (Accessed 7 January 2024).

Feng, Z., El Issaoui, A., Lehtomäki, M., et al., 2022. Pavement distress detection using terrestrial laser scanning point clouds–accuracy evaluation and algorithm comparison. ISPRS Open Journal of Photogrammetry and Remote Sensing 3, 100010.

Feng, X., Xiao, L., Li, W., et al., 2020. Pavement crack detection and segmentation method based on improved deep learning fusion model. Mathematical Problems in Engineering 2020, 8515213.

Feng, Y., Zhang, Z., Zhao, X., et al., 2018. GvCNN: group-view convolutional neural networks for 3d shape recognition. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018.

Fu, C., Liu, W., Ranga, A., et al., 2017. DSSD: Deconvolutional Single Shot Detector. Available at: https://arxiv.org/abs/1701.06659 (Accessed 7 January 2024).

Fu, C., Yuan, Y., Zeng, Q., et al., 2018. An effective head detection framework via convolutional neural networks. In: 18th Pacific-Rim Conference on Multimedia, Harbin, 2017.

Ghiasi, G., Lin, T., Le, Q.V., 2019. NAS-FPN: learning scalable feature pyramid architecture for object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beoch, 2019.

Girshick, R., 2015. Fast R-CNN. Available at: https://arxiv.org/abs/1054.08083 (Accessed 7 January 2024).

Girshick, R., Donahue, J., Darrell, T., et al., 2013. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. Available at: https://arxiv.org/abs/1311.2524 (Accessed 7 January 2024).

González, A., Vazquez, D., López, A.M., et al., 2017. On-board object detection: multicue, multimodal, and multiview random forest of local experts. IEEE Transactions on Cybernetics 47 (11), 3980–3990.
Gou, C., Peng, B., Li, T., et al., 2019. Pavement crack detection based on the improved faster-RCNN. In: 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Dalian, 2019.
Gould, S., Gao, T., Koller, D., 2009. Region-based segmentation and object detection. In: 22nd International Conference on Neural Information Processing Systems, Red Hook, 2009.
Han, Z., Chen, H., Liu, Y., et al., 2021. Vision-based crack detection of asphalt pavement using deep convolutional neural network. Iranian Journal of Science and Technology, Transactions of Civil Engineering 45, 2047–2055.
Han, K., Wang, Y., Tian, Q., et al., 2020. GhostNet: more features from cheap operations. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, 2020.
Haris, M., Glowacz, A., 2021. Road object detection: a comparative study of deep learning-based algorithms. Electronics 10 (16), 1932.
He, A., Dong, Z., Zhang, H., et al., 2023. Automated pixel-level detection of expansion joints on asphalt pavement using a deep-learning-based approach. Structural Control and Health Monitoring 2023, 7552337.
He, K., Gkioxari, G., Dollár, P., et al., 2017. Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision, Venice, 2017.
He, Z., Jiang, S., Zhang, J., et al., 2022. Automatic damage detection using anchor-free method and unmanned surface vessel. Automation in Construction 133, 104017.
He, K., Zhang, X., Ren, S., et al., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (9), 1904–1916.
He, K., Zhang, X., Ren, S., et al., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016.
He, Y., Zhang, X., Savvides, M., et al., 2018. Softer-NMS: Rethinking Bounding Box Regression for Accurate Object Detection. Available at: https://arxiv.org/abs/1809.08545 (Accessed 7 January 2024).
Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507.
Hou, Y., Shi, H., Chen, N., et al., 2022. Vision image monitoring on transportation infrastructures: a lightweight transfer learning approach. IEEE Transactions on Intelligent Transportation Systems 24 (11), 12888–12899.
Hou, Q., Zhou, D., Feng, J., 2021. Coordinate attention for efficient mobile network design. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Washville, 2021.
Howard, A., Zhu, M., Chen, B., et al., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. Available at: https://arxiv.org/abs/1704.04861 (Accessed 7 January 2024).
Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt, Lake City, 2018.
Huang, P., Wang, S., Chen, J., et al., 2023. Lightweight model for pavement defect detection based on improved YOLOv7. Sensors 23 (16), 7112.
Huang, L., Yang, Y., Deng, Y., et al., 2015. Densebox: Unifying Landmark Localization With End to End Object Detection. Available at: https://arxiv.org/abs/1509.04874 (Accessed 7 January 2024).
Ibragimov, E., Lee, H., Lee, J., et al., 2022. Automated pavement distress detection using region based convolutional neural networks. International Journal of Pavement Engineering 23 (6), 1981–1992.
Isola, P., Zoran, D., Krishnan, D., et al., 2014. Crisp boundary detection using pointwise mutual information. In: 13th European Conference, Zurich, 2014.
Jain, A.K., Mao, J., Mohiuddin, K.M., 1996. Artificial neural networks: a tutorial. Computer 29 (3), 31–44.
JRE Editorial Office, Cavalli, M.C., Chen, D., et al., 2023. Review of advanced road materials, structures, equipment, and detection technologies. Journal of Road Engineering 3 (4), 3700, 468.
Kang, C., Wang, W., 2007. A novel edge detection method based on the maximizing objective function. Pattern Recognition 40 (2), 609–618.
Kong, T., Yao, A., Chen, Y., et al., 2016. HyperNet: towards accurate region proposal generation and joint object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with deep convolutional neural networks. Communications of the ACM 60 (6), 84–90.
Ku, J., Mozifian, M., Lee, J., et al., 2018. Joint 3D proposal generation and object detection from view aggregation. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, 2018.
LaValley, M.P., 2008. Logistic regression. Circulation 117 (18), 2395–2399.
Law, H., Deng, J., 2018. Cornernet: detecting objects as paired keypoints. Proceedings of the International Journal of Computer Vision 128, 642–656.
LeCun, Y., Bottou, L., Bengio, Y., et al., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86 (11), 2278–2324.
Leung, K.M., 2007. Naive bayesian classifier, 2007. Polytechnic University Department of Computer Science/Finance and Risk Engineering 2007, 123–156.
Li, Y., Bu, R., Sun, M., et al., 2018a. PointCNN: convolution on x-transformed points. In: 32nd Conference on Neural Information Processing Systems, Montréal, 2018.
Li, P., Chao, W., Li, S., et al., 2015. Research on crack detection method of airport runway based on twice-threshold segmentation. In: 2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC), Qinhuangdao, 2015.
Li, P., Chen, X., Shen, S., 2019. Stereo R-CNN based 3d object detection for autonomous driving. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2019.
Li, L., Fang, B., Zhu, J., 2022a. Performance analysis of the YOLOv4 algorithm for pavement damage image detection with different embedding positions of CBAM modules. Applied Sciences 12 (19), 10180.
Li, Y., He, K., Sun, J., 2016. R-FCN: object detection via region-based fully convolutional networks. In: 30th International Conference on Neural Information Processing Systems, Red Hook, 2016.
Li, C., Li, L., Jiang, H., et al., 2022b. YOLOv6: A Single-stage Object Detection Framework for Industrial Applications. Available at: https://arxiv.org/abs/2209.02976 (Accessed 7 January 2024).
Li, Q., Liu, X., 2008. Novel approach to pavement image segmentation based on neighboring difference histogram method. In: 2008 Congress on Image and Signal Processing, Sanya, 2008.
Li, H., Song, D., Liu, Y., et al., 2018b. Automatic pavement crack detection by multi-scale image fusion. IEEE Transactions on Intelligent Transportation Systems 20 (6), 2025–2036.
Li, P., Qin, T., 2018. Stereo vision-based semantic 3D object and ego-motion tracking for autonomous driving. In: European Conference on Computer Vision (ECCV), Munich, 2018.
Li, X., Wang, L., Wang, M., et al., 2020. DANCE-NET: density-aware convolution networks with context encoding for airborne LiDAR point cloud classification. ISPRS Journal of Photogrammetry and Remote Sensing 166, 128–139.
Li, Z., Zhou, F., 2017. FSSD: Feature Fusion Single Shot Multibox Detector. Available at: https://arxiv.org/abs/1712.00960 (Accessed 7 January 2024).
Liang, L., Looney, C.G., 2003. Competitive fuzzy edge detection. Applied Soft Computing 3 (2), 123–137.
Liang, Z., Yang, M., Deng, L., et al., 2019. Hierarchical depthwise graph convolutional neural network for 3D semantic segmentation of point clouds. In: 2019 International Conference on Robotics and Automation (ICRA), Montreal, 2019.
Liaw, A., Wiener, M., 2002. Classification and regression by random Forest. R News 2 (3), 18–22.
Liebold, F., Maas, H.-G., 2020. Strategy for crack width measurement of multiple crack patterns in civil engineering material testing using a monocular image sequence analysis. PFG–Journal of Photogrammetry. Remote Sensing and Geoinformation Science 88, 219–238.
Lim, J.J., Zitnick, C.L., Dollár, P., 2013. Sketch tokens: a learned mid-level representation for contour and object detection. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, 2013.
Lin, T.-Y., Dollár, P., Girshick, R., et al., 2017a. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017.
Lin, T.-Y., Goyal, P., Girshick, R., et al., 2017b. Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision, Venice, 2017.
Liu, B., 2021. Long-distance recognition of crack width in building wall based on binocular vision. In: 2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture, Manchester, 2021.
Liu, W., Anguelov, D., Erhan, D., et al., 2016. SSD: single shot multibox detector. In: 14th European Conference, Amsterdam, 2016.
Liu, W., Hasan, I., Liao, S., 2019. Center and scale prediction: anchor-free approach for pedestrian and face detection. Pattern Recognition 135, 109071.
Liu, Z., Li, J., Shen, Z., et al., 2017. Learning efficient convolutional networks through network slimming. In: 2017 IEEE International Conference on Computer Vision, Venice, 2017.
Liu, Z., Lin, Y., Cao, Y., et al., 2021. Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision, Montreal, 2021.
Liu, S., Qi, L., Qin, H., et al., 2018. Path aggregation network for instance segmentation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018.
Liu, M., Sheng, L., Yang, S., et al., 2020. Morphing and sampling network for dense point cloud completion. In: 2020 AAAI Conference on Artificial Intelligence, New York, 2020.
Liu, Y., Shi, G., Li, Y., et al., 2022. M-YOLO based detection and recognition of highway surface oil filling with unmanned aerial vehicle. In: 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, 2022.
Liu, F., Xu, G., Yang, Y., et al., 2008. Novel approach to pavement cracking automatic detection based on segment extending. In: 2008 International Symposium on Knowledge Acquisition and Modeling, Washington DC, 2008.
Liu, Z., Yeoh, J.K., Gu, X., et al., 2023. Automatic pixel-level detection of vertical cracks in asphalt pavement based on GPR investigation and improved mask R-CNN. Automation in Construction 146, 104689.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition, Washington DC, 2015.
Luo, Y., Ren, J., Lin, M., et al., 2018. Single view stereo matching. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018.
Ma, N., Zhang, X., Zheng, H., et al., 2018. ShuffleNetV2: practical guidelines for efficient CNN architecture design. In: European Conference on Computer Vision, Munich, 2018.
Maode, Y., Shaobo, B., Kun, X., et al., 2007. Pavement crack detection and analysis for high-grade highway. In: 8th International Conference on Electronic Measurement and Instruments, Xi'an, 2007.
Marques, A., Correia, P.L., 2012. Automatic Road Pavement Crack Detection Using SVM. Instituto Superior Técnico, Lisbon.
Mathavan, S., Kamal, K., Rahman, M., 2015. A review of three-dimensional imaging technologies for pavement distress detection and measurements. IEEE Transactions on Intelligent Transportation Systems 16 (5), 2353–2362.

Meng, H., Gao, L., Lai, Y., et al., 2019. VV-Net: voxel VAE net with group convolutions for point cloud segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision, Seoul, 2019.
Oliveira, H., Correia, P.L., 2009. Automatic road crack segmentation using entropy and image dynamic thresholding. In: 17th European Signal Processing Conference, Glasgow, 2009.
Park, S.S., Tran, V.T., Lee, D.E., 2021. Application of various YOLO models for computer vision-based real-time pothole detection. Applied Sciences 11 (23), 11229.
Qayyum, W., Ehtisham, R., Bahrami, A., et al., 2023. Assessment of convolutional neural network pre-trained models for detection and orientation of cracks. Materials 16 (2), 826.
Qi, C., Liu, W., Wu, C., et al., 2018. Frustum PointNets for 3D object detection from RGB-D data. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2018.
Qi, C., Su, H., Mo, K., et al., 2017a. PointNet: deep learning on point sets for 3d classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017.
Qi, C., Yi, L., Su, H., et al., 2017b. PointNet++: deep hierarchical feature learning on point sets in a metric space. In: 31st International Conference on Neural Information Processing Systems, Red Hook, 2017.
Qian, S., Ning, C., Hu, Y., 2021. MobileNetV3 for image classification. In: 2021 IEEE 2nd International Conference on Big Date, Artificial Intelligence and Internet of Things Engineering, Nanchang, 2021.
Qiu, Q., Lau, D., 2023. Real-time detection of cracks in tiled sidewalks using YOLO-based method applied to unmanned aerial vehicle (UAV) images. Automation in Construction 147, 104745.
Redmon, J., Divvala, S., Girshick, R., et al., 2016. You Only Look Once: Unified, Real-time Object Detection. Available at: https://arxiv.Org/abs/1506.021040 (Accessed 7 January 2024).
Redmon, J., Farhadi, A., 2017. YOLO9000: Better, Faster, Stronger. Available at: https://arxiv.org/abs/1612.08242 (Accessed 7 January 2024).
Redmon, J., Farhadi, A., 2018. YOLOv3: an Incremental Improvement. Available at: https://arxiv.orglabs/1804.02767 (Accessed 7 January 2024).
Ren, S., He, K., Girshick, R., et al., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In: 28th International Conference on Neural Information Processing Systems, Cambridge, 2015.
Ren, J., Zhao, G., Ma, Y., et al., 2022. Automatic pavement crack detection fusing attention mechanism. Electronics 11 (21), 3622.
Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review 65 (6), 386–408.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536.
Sandler, M., Howard, A., Zhu, M., et al., 2018. MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018.
Sarathi, M.P., Ansari, M.A., Uher, V., et al., 2013. Automated Brain Tumor segmentation using novel feature point detector and seeded region growing. In: 36th International Conference on Telecommunications and Signal Processing (TSP), Rome, 2013.
Sharma, K.U., Thakur, N.V., 2017. A review and an approach for object detection in images. International Journal of Computational Vision and Robotics 7 (1–2), 196–237.
Shi, W., Rajkumar, R., 2020. Point-GNN: graph neural network for 3d object detection in a point cloud. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, 2020.
Shi, S., Wang, X., Li, H., 2019. PointrCNN: 3d object proposal generation and detection from point cloud. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2019.
Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-scale Image Recognition. Availalbe at: https://arxiv.org/abs/1409.1556 (Accessed 7 January 2024).
Singh, B., Davis, L.S., 2018. An analysis of scale invariance in object detection SNIP. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018.
Sobel, I., 1990. An Isotropic 3×3 Gradient Operator, Machine Vision for Three-Dimensional Scenes. Academic Press, New York.
Su, H., Jampani, V., Sun, D., et al., 2018. SPLATNet: sparse lattice networks for point cloud processing. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018.
Su, H., Maji, S., Kalogerakis, E., et al., 2015. Multi-view convolutional neural networks for 3D shape recognition. In: 2015 IEEE International Conference on Computer Vision, Santiago, 2015.
Su, H., Wang, X., Han, T., et al., 2022. Research on a U-Net bridge crack identification and feature-calculation methods based on a CBAM attention mechanism. Buildings 12 (10), 1561.
Sun, Y., Sun, C., Wang, D., et al., 2019. RoI pooled correlation filters for visual tracking. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2019.
Szegedy, C., Liu, W., Jia, Y., et al., 2015. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015.
Tang, Y., Han, K., Guo, J., et al., 2023. GhostNetV2: Enhance Cheap Operation with Long-range Attention. Available at: http://arxiv.org/abs/2211.12905 (Accessed 7 January 2024).
Tang, Y., Li, L., Feng, W., et al., 2018. Binocular vision measurement and its application in full-field convex deformation of concrete-filled steel tubular columns. Measurement 130, 372–383.
Tian, Z., Shen, C., Chen, H., et al., 2020. FCOS: a simple and strong anchor-free object detector. IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (4), 1922–1933.
Tsai, Y., Chatterjee, A., 2018. Pothole detection and classification using 3D technology and watershed method. Journal of Computing in Civil Engineering 32 (2), 04017078.
Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., et al., 1995. Modeling visual attention via selective tuning. Artificial Intelligence 78 (1–2), 507–545.
Uijlings, J.R.R., Van De Sande, K.E.A., Gevers, T., et al., 2013. Selective search for object recognition. International Journal of Computer Vision 104, 154–171.
Vu, T., Jang, H., Pham, T.X., et al., 2019. Cascade RPN: delving into high-quality region proposal network with adaptive convolution. In: 33rd International Conferenceon Neural Information Processing Systems, Red Hook, 2019.
Wan, H., Gao, L., Su, M., et al., 2021. Attention-based convolutional neural network for pavement crack detection. Advances in Materials Science and Engineering 2021, 5520515.
Wan, F., Sun, C., He, H., et al., 2022. YOLO-LRDD: a lightweight method for road damage detection based on improved YOLOv5s. EURASIP Journal on Applied Signal Processing 2022 (1), 1–18.
Wang, C.-Y., Bochkovskiy, A., Liao, H., 2023. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, 2023.
Wang, P., Fu, S., Cao, X., 2022a. Improved lightweight target detection algorithm for complex roads with YOLOv5. In: 2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), Guangzhou, 2022.
Wang, S., Tang, W., 2012. Pavement crack segmentation algorithm based on local optimal threshold of cracks density distribution. In: 7th International Conference on Intelligent Computing, Zhengzhou, 2011.
Wang, Y., Wang, H., Peng, Z., 2021. Rice diseases detection and classification using attention based neural network and bayesian optimization. Expert Systems with Applications 178, 114770.
Wang, Q., Wu, B., Zhu, P., et al., 2020. ECA-Net: efficient channel attention for deep convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, 2020.
Wang, Z., Zhang, H., Huang, Z., et al., 2022b. Multi-scale dense and attention mechanism for image semantic segmentation based on improved DeepLabv3+. Journal of Electronic Imaging 31 (5), 053006.
Wen, C., Li, X., Yao, X., et al., 2021. Airborne LiDAR point cloud classification with global-local graph attention convolution neural network. ISPRS Journal of Photogrammetry and Remote Sensing 173, 181–194.
Womg, A., Shafiee, M.J., Li, F., et al., 2018. Tiny SSD: a tiny single-shot detection deep convolutional neural network for real-time embedded object detection. In: 15th Conference on Computer and Robot Vision (CRV), Toronto, 2018.
Woo, S., Park, J., Lee, J.-Y., et al., 2018. CBAM: convolutional block attention module. In: 15th European Conference on Computer Vision (ECCV), Munich, 2018.
Wu, H., Yao, L., Xu, Z., et al., 2019. Road pothole extraction and safety evaluation by integration of point cloud and images derived from mobile mapping sensors. Advanced Engineering Informatics 42, 100936.
Xie, S., Girshick, R., Dollár, P., et al., 2017. Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017.
Xu, B., Chen, Z., 2018. Multi-level fusion based 3d object detection from monocular images. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018.
Xu, Y., Hoegner, L., Tuttas, S., et al., 2017. Voxel-and graph-based point cloud segmentation of 3D scenes using perceptual grouping laws. ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences 4, 43–50.
Xu, G., Ma, J., Liu, F., et al., 2008. Automatic recognition of pavement surface crack based on BP neural network. In: 2008 International Conference on Computer and Electrical Engineering, Phuket, 2008.
Xu, X., Zhang, X., Zhang, T., 2022. Lite-yolov5: a lightweight deep learning detector for on-board ship detection in large-scene sentinel-1 sar images. Remote Sensing 14 (4), 1018.
Yamaguchi, T., Nakamura, S., Saegusa, R., et al., 2008. Image-based crack detection for real concrete surfaces. IEEJ Transactions on Electrical and Electronic Engineering 3 (1), 128–135.
Yao, H., Liu, Y., Li, X., et al., 2022. A detection method for pavement cracks combining object detection and attention mechanism. IEEE Transactions on Intelligent Transportation Systems 23 (11), 22179–22189.
Yao, H., Liu, Y., Lyu, H., et al., 2023a. Encoder-decoder with pyramid region attention for pixel-level pavement crack recognition. Computer-Aided Civil and Infrastructure Engineering 34 (8), 713–727.
Yao, H., Xu, Z., Hou, Y., et al., 2023b. Advanced industrial informatics towards smart, safe and sustainable roads: a state of the art. Journal of Traffic and Transportation Engineering (English Edition) 10 (2), 143–158.
Yi, Y., Peng, J., Pei, S., et al., 2021. Pavement flatness detection based on binocular vision. In: 2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI), Beijing, 2021.
Yin, G., Gao, J., Gao, J., et al., 2023. Crack identification method of highway tunnel based on image processing. Journal of Traffic and Transportation Engineering (English Edition) 10 (3), 469–484.
Yu, D., Wang, H., Chen, P., et al., 2014. Mixed pooling for convolutional neural networks. In: 9th International Conference on Rough Sets and Knowledge Technology, Shanghai, 2014.
Zhang, C., Nateghinia, E., Miranda-Moreno, L.F., et al., 2022a. Pavement distress detection using convolutional neural network (CNN): a case study in Montreal, Canada. International Journal of Transportation Science and Technology 11 (2), 298–309.
Zhang, J., Qian, S., Tan, C., 2023. Automated bridge crack detection method based on lightweight vision models. Complex & Intelligent Systems 9 (2), 1639–1652.
Zhang, A., Wang, K., Li, B., et al., 2017a. Automated pixel-level pavement crack detection on 3D asphalt surfaces using a deep-learning network. Computer-Aided Civil and Infrastructure Engineering 32 (10), 805–819.

Zhang, L., Wang, C., Yap, J.B.H., et al., 2020. Developing novel monocular-vision-based standard operational procedures for nondestructive inspection on constructed concrete cracks. Journal of Performance of Constructed Facilities 34 (2), 04020012.
Zhang, H., Wu, Z., Qiu, Y., et al., 2022b. A new road damage detection baseline with attention learning. Applied Sciences 12 (15), 7594.
Zhang, X., Zhou, X., Lin, M., et al., 2017b. ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: 2018 IEEE/CVF Conferenceon Computer Vision and Pattern Recognition, Salt Lake City, 2018.
Zhao, T., Wu, X., 2019. Pyramid feature attention network for saliency detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2019.
Zhong, M., Sui, L., Wang, Z., et al., 2020. Pavement crack detection from mobile laser scanning point clouds using a time grid. Sensors 20 (15), 4198.
Zhou, D., Hou, Q., Chen, Y., et al., 2020. Rethinking bottleneck structure for efficient mobile network design. In: 2020 European Conference on Computer Vision, Glasgow, 2020.
Zhou, X., Wang, D., Krähenbühl, P., 2019a. Objects as Points. Available at: https://arxiv.org/abs/1904.07850 (Accessed 7 January 2024).
Zhou, Y., Wang, F., Meghanathan, N., et al., 2016. Seed-based approach for automated crack detection from pavement images. Transportation Research Record 2589, 162–171.
Zhou, X., Zhuo, J., Krahenbuhl, P., 2019b. Bottom-up object detection by grouping extreme and center points. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2019.
Zhu, S., Xia, X., Zhang, Q., et al., 2007. An image segmentation algorithm in image processing based on threshold segmentation. In: Third International IEEE Conference on Signal-Image Technology and Internet-Based Systems, Shanghai, 2007.
Zhu, Y., Zhao, C., Wang, J., et al., 2017. CoupleNet: coupling global structure with local parts for object detection. In: 2017 IEEE International Conference on Computer Vision, Venice, 2017.
Zou, Q., Cao, Y., Li, Q., et al., 2012. CrackTree: automatic crack detection from pavement images. Pattern Recognition Letters 33 (3), 227–238.
Zou, Z., Shi, Z., Guo, Y., et al., 2019. Object detection in 20 years: a survey. Proceedings of the IEEE 11 (3), 257–276.

**Dr. Hui Yao** received a PhD degree in the field of civil engineering and transportation engineering from Central South University in China and Michigan Technological University in the United States. He is serving as a professor position at Faculty of Architecture, Civil and Transportation Engineering, Beijing University of Technology in China. His research interests include materials for civil engineering, smart maintenance, and numerical simulations.

**Mr. Yaning Fan** is a graduate student at Faculty of Architecture, Civil and Transportation Engineering, Beijing University of Technology in China. His research direction focuses on the intelligent identification of pavement distresses based on radar technology.

**Mr. Yanhao Liu** is a PhD student at Faculty of Architecture, Civil and Transportation Engineering, Beijing University of Technology, China. His research interest focuses on the application of machine learning to transportation engineering and smart maintenance.

**Dr. Dandan Cao** received a PhD degree in municipal engineering from the Dalian University of Technology in 2018. He is currently at Faculty of Architecture, Civil and Transportation Engineering, Beijing University of Technology in China. He presided the project of the National Natural Science Foundation of China and published 10 SCI and EI papers. His research is mainly focused on transportation infrastructure simulation and pavement analysis and maintenance.

**Dr. Ning Chen** received the BE degree in traffic engineering from the South China University of Technology in 2007, the M.E. degree in transportation of planning and management from Southwest Jiaotong University in 2011, and the PhD degree in transportation of planning and management from the Beijing University of Technology in 2015. He is currently at Faculty of Architecture, Civil and Transportation Engineering, Beijing University of Technology in China. His research mainly focuses on transportation infrastructure monitoring, universal on-board units of intelligent vehicles, and proactive safety systems at the network level.

**Mr. Tiancheng Luo** is an undergraduate student at Faculty of Architecture, Civil and Transportation Engineering, Beijing University of Technology, China. His research direction is transportation planning and intelligent management.

**Ms. Jingyu Yang** is an undergraduate student at Faculty of Architecture, Civil and Transportation Engineering, Beijing University of Technology, China. Her research direction is transportation planning and intelligent management.

**Ms. Xueyi Hu** is an undergraduate student at Faculty of Architecture, Civil and Transportation Engineering, Beijing University of Technology, China. Her research direction is road and railway engineering.

**Dr. Jie Ji** is a professor and doctoral supervisor of Beijing University of Civil Engineering and Architecture. Currently, she is the academic leader of transportation engineering in the School of Civil and Transportation Engineering, Beijing University of Civil Engineering and Architecture in China. Her research interest includes structure and material of subgrade and pavement and application for solid waste.

**Dr. Zhanping You** is a distinguished professor of transportation engineering and materials at Michigan Technological University, United States. Dr. You has published over 400 papers in peer reviewed journals and conference proceedings. He earned the prestigious Michigan Tech Research Award and University Distinguished Professorship. He was elected as a Fellow of ASCE and a Fellow of ASCE Engineering Mechanics Institute.