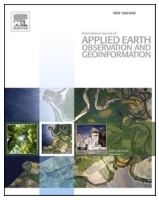
International Journal of Applied Earth Observation and Geoinformation 120 (2023) 103335

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



YOLOv5s-M: A deep learning network model for road pavement damage detection from urban street-view imagery

Miao Ren, Xianfeng Zhang *, Xiao Chen, Bo Zhou, Ziyuan Feng

Institute of Remote Sensing and Geographic Information System, Peking University, 5 Summer Palace Road, Beijing 100871, China

ARTICLE INFO

Keywords:
Pavement damage
YOLOv5
Street-view images
Deep learning
Feature fusion

ABSTRACT

Road pavement damage affects driving comfort markedly, threatens driving safety, and may even cause traffic accidents. The traffic management department conventionally captures pavement damage information mainly using manual and vehicle-mounted equipment, which is not conducive to the detection of large-scale road pavement distress. Street-view images can provide full-view images of urban roads where the data is updated regularly by navigation map service companies, making it possible to rapidly detect pavement damage in urban areas. This paper presents a new pavement damage detection approach that is built upon an improved YOLOv5 network and street-view images. The proposed model can deal with a large-scale detection layer to improve the detection precision of large distress targets, achieving thus both cross-layer and cross-scale feature fusion by using the Generalized Feature Pyramid Network (Generalized-FPN) structure. The improved network also involves a diagonal Intersection over Union loss for regression calculation of the boundary box and builds the decoupled Head structure to achieve the decoupling detection of prediction and regression. As a result, the fusion of the weak feature information in feature layers is enhanced at different spatial scales, a more suitable method is achieved for pavement damage detection in the complex context of multi-scale street-view images, and the accuracy of the modified network is much improved in the detection of pavement distress from street-view imagery. Furthermore, We created a large image sample set for model training and testing, and a total of 156,304 street-view images, obtained from Fengtai District, Beijing, China was used for demonstrating the usefulness of the proposed network. The findings indicated that the proposed approach could effectively achieve pavement damage detection of urban roads from street-view images, with a precision average of 79.8% on the test samples. Moreover, the developed model was applied for pavement damage detection for all the roads in Fengtai District, Beijing, indicating that our method can offer viable damage data for road maintenance planning.

1. Introduction

As one of the most important infrastructures, roads play a key role in regional and urban social & economic development due to transportation and communication establishment. However, road pavement structure is gradually affected by repeated vehicle loads and severe environmental factors, which eventually lead to pavement damages (Hou et al., 2021). Potholes and cracks are the two most common categories of pavement damages that are often encountered on the roads and have a significant impact on the vehicle running quality (Tedeschi and Benedetto, 2017). According to previous studies, for every single US dollar spent on road maintenance in developing countries, three US dollars are saved on-road use (Asian Development Bank, 2003). Therefore, with the rapid increase in the length of roads and the high percentage of road maintenance, detecting pavement damage quickly and effectively as well as understanding its distribution is of great practical significance.

Conventional methods for pavement damage detection rely much on the experience of road maintenance staff who uses ground measurements and vehicle-mounted multi-sensor inspection systems to instigate pavement conditions. These approaches are often time-consuming, inefficient, and hindering traffic; and consequently, are not suitable for monitoring a wide range of road pavement (Pan et al., 2017). With the development of 3D sensors (*e.g.*, laser profilers, LiDAR, and InSAR), many pavement damage detection studies have been conducted using 3D point cloud data (Coenen and Golroo, 2017; Fei et al., 2020; Jiang et al., 2018; Ragnoli et al., 2018). However, large-scale and complex point cloud scenarios demand increasing data collection and processing

* Corresponding author.
E-mail address: xfzhang@pku.edu.cn (X. Zhang).

https://doi.org/10.1016/j.jag.2023.103335
Received 22 January 2023; Received in revised form 23 April 2023; Accepted 26 April 2023
Available online 7 May 2023

requirements, and thus detecting fine and low-connectivity road damages remains challenging with these methods (Feng et al., 2022). On the other hand, when detecting pavement damage from multi-platform optical remote sensing imagery, machine learning-based methods have demonstrated proficiency in identifying pavement damage. Therefore, automated detection of pavement damage can be achieved by selecting and extracting pavement damage features and defining an appropriate classification model (Hoang and Nguyen, 2019; Mokhtari et al., 2016; Ng et al., 2019). However, since the feature set is usually selected interactively from specific pavement images, the conventional shallow learning has relatively poor robustness and weak generalization capability (Pan et al., 2021).

The rapid development of artificial intelligence enables deep learning techniques to be applied to pavement damage detection from various images. In the deep learning, the labeled sample dataset significantly affects the model performance as well as the training and testing model accuracy (Nguyen et al., 2022), whereas the time, difficulty, and cost of data collection dramatically affect the practical application of pavement damage detection models. Therefore, some researchers chose to equip a vehicle platform with a normal camera to acquire pavement images. For instance, Arya et al. (2021) used a smartphone, mounted on the windshield of the vehicle, to capture road images from its front view. Mei and Gül (2020) used a *GoPro Hero 7 Black*, installed on the rear license plate of the vehicle, to collect road images from its rear view. In contrast, when a model is built using pavement images that are captured from the front view of the vehicle, it is easier to train the model to fit practical situations (Maeda et al., 2018). For the dataset reported by Arya et al. (2021), two-stage target detection algorithms (*e.g.*, Faster R-CNN (Hascoet et al., 2020; Kortmann et al., 2020; Liu et al., 2020), Mask R-CNN (Singh and Shekhar, 2018), and Cascade R-CNN (Pei et al., 2020)) and single-stage target detection algorithms (*e.g.*, SSD (Camilleri and Gatt, 2020), EfficientDet (Naddaf-Sh et al., 2020), YOLOv2 (Mandal et al., 2018), YOLOv3 (Camilleri and Gatt, 2020), YOLOv4 (Liu et al., 2020), Scaled-YOLOv4 (Fassmeyer et al., 2021), and YOLOv5x (Jeong, 2020)) in deep learning were applied in pavement damage detection using road images captured from the front view of the vehicle and they all achieved high accuracy. In summary, these studies demonstrated the effectiveness of using vehicle-mounted platforms with cameras to acquire road images and the application of deep learning approaches.

Street-view images are usually captured along a road using a vehicle-mounted platform with a street-view photographic system equipped with multiple cameras (Anguelov et al., 2010), and the captured images include rear-view, side-view and front-view images of the vehicle. Street-view images are an abundant data source as they cover most roads in urban areas, and are collected by professional vehicles. They are in a large data volume and updated regularly. These low-cost byproducts of navigation services provide high-quality images with high spatial resolution, making them suitable for detecting a broad range of pavement damages of different sizes. In comparison with 3D point cloud data and UAV remote sensing data, street-view images are more accessible and appropriate for complex urban environments. Hence, street-view images offer a new data source for pavement damage detection and fulfill the data collection requirements for the application of deep learning algorithms in pavement damage detection. Compared to the front-view images frequently used in pavement damage detection, street-view images are not captured specifically for pavement damage detection. Therefore, the main factors affecting the precision of pavement damage detection from the street-view images include, *various types of urban roads, multi-scale features of pavement damage, and complex background and large interference.* Multiple grades of roads are observed in a city, and the pavement damage features differ with the different grades of roads. Due to the perspective effect, pavement damages show the characteristics of near large and far small in the front-view images. Moreover, the partition of a pavement damage target on multiple images is often observed when the target is large. In addition, the size of the cracks is usually diverse. The background of street-view images includes pedestrians, vehicles, buildings, viaducts, and trees, as well as their shadows. As for the shooting conditions, they include a variety of seasons, weather, and lighting conditions. These factors may have a certain impact on the pavement damage features and incur noise, resulting in a limited precision and a generalization capability of pavement damage extraction algorithms when using such images.

Some studies have currently attempted to detect pavement damage using street-view images. For example, Chacra and Zelek (2017) labeled the road damage in 250 street-view images where they extracted the textural features from the street-view images using Scale-Invariant Feature Transform (SIFT) features and Fisher vectors, and detected the pavement damages using the Support Vector Machine (SVM) algorithm. However, the feature extraction process in this method is complicated, and it may fail to detect diverse pavement damages under other generalized circumstances due to limited dataset availability. Moreover, Lei et al. (2020) used the YOLOv3 algorithm to detect the pavement damage from street-view images and they selected a typical road segment in Shanghai to verify the effectiveness of the model. Nevertheless, the algorithm was only optimized for the boundary box of the used dataset, and no corresponding improvements were made to the algorithm regarding the task of detecting pavement damage in street-view images. Furthermore, Majidifard et al. (2020) used a dataset composed of images extracted from Google street-view images, applied the integrated model of YOLO and U-Net to detect pavement damage and created a pavement condition index to evaluate the pavement damage conditions. Whereas, the street-view images were mainly a composite of wide-view images and bird's-eye view images, and the images had large distortion and misalignment, which limited the use of these approaches. As for Li et al. (2021), they used a dataset of multiple scenes, including street-view images to synthesize the data in new scenes by generating adversarial networks. They further implemented the YOLOv4 model to migrate the model, thus achieving a pavement damage detection in complex multi-scenes. This was a novel attempt to apply street-view images to detect pavement damage without the exclusive use of front-view street-view images. In summary, there is a lack of target detection algorithms considering the specific features of a variety of pavement damage in the street-view images. Furthermore, the effectiveness of the abovementioned methods for pavement damage detection may be limited by the constraints of the datasets under generalized circumstances.

Therefore, a new approach for pavement damage detection is proposed based on the modification of the YOLOv5 model and street-view images in our study. The contributions of our work are summarized as: 1) a diagonal IoU loss function is proposed to improve the capability of detecting boundary boxes by characterizing the difference of diagonal lengths of the boundary boxes in the regression calculation; 2) an improved YOLOv5 network model is developed for pavement damage detection from street-view images based on the Generalized-FPN and the Decoupled Head modules. This proposed network allows for sufficient information exchange between high and low layers, enhancing thus the fusion of weak feature information from the feature layers of various scales, and improving the detection precision of the pavement damage; 3) the manhole cover training samples are added in the construction of the training sample set, which effectively improves the detection capability for potholes; and 4) the street-view images of Fengtai District in Beijing within an area of 305.53 km^2 are used to demonstrated the usefulness of the proposed network, showing thus a good practicability of the proposed method.

The rest of this paper is organized as follows: in Section 2, the study area, data acquisition as well as the architecture of the YOLOv5 model and the improvements made in our work are presented. The training results and some related analysis of the network model are provided in Section 3. Discussions and analyses are conducted in Section 4, and the conclusions of this study as well as the proposition of some future works are drawn in Section 5.

2. Materials and methods

2.1. Study area and dataset

Fengtai District is located in the south of Beijing City, with a total area of 305.53 km^2 and a total roads length of 102.3 km, including the urban roads of various grades. All the major roads, including the 2nd, 3rd, 4th, 5th, and 6th Ring Roads, go through the Fengtai District, the highways such as Beijing–Hong Kong and Macau Expressway, Jingkai Expressway, and China National Highway No. 104 and No. 107 also pass through. Therefore, the Fengtai District is a suburban administrative unit with a high density of roads in Beijing, and the roads in this region are considered for further study as they can cover most of the typical urban roads in Beijing City. The road network data of the Fengtai District is retrieved from the Open Street Map (OSM) database. Accordingly, the roads can be divided into seven different types in the study, namely motorways, trunk roads, primary roads, secondary roads, tertiary roads, residential roads, and unclassified roads. Since OSM uses the WGS84 coordinate system, the OSM road network data was first converted to the BD09 coordinate system which is used in the *Baidu Maps*. Then, the sampling points were generated every five meters along the road network to extract the location information of the target from the street-view images.

The street-view images in our study were captured mainly in 2019 and 2020. The Web Service API of the Baidu Maps is available to any users of the Baidu Maps, and it is quite easy to apply for the access to the street-view images. The required URL parameters to retrieve the images include image width, height, location, heading, pitch, and field of view. In this study, two pitch images, with 0° and 45° angles at each location, were obtained respectively. Afterward, these images were vertically stitched together to obtain a complete front-view street-view image at a given location, and each image has a size of 1024 × 1024 pixels. The study area is illustrated in Fig. 1.

A total of 156,304 street-view images covering the entire Fengtai District of Beijing were collected among which 2,900 images with varying road grades were manually selected and labeled. Moreover, to balance the data categories, the number of labeled samples in all the categories was kept consistent. In this study, six common categories of pavement distress were identified in street-view images, including transverse crack, longitudinal crack, alligator crack, pothole, transverse patch and longitudinal patch (Arya et al., 2021). Since manhole covers are easily misclassified as potholes (Kortmann et al., 2020; Maeda et al., 2018), they were also labeled separately to clarify the difference between potholes and manhole covers. The sample dataset was randomly grouped into the training set (1740 images), validation set (580 images), and testing set (580 images), following the ratio of 6:2:2. The number of instances per pavement damage category is shown in Table 1.

2.2. Modification of the YOLOv5 model

2.2.1. The YOLOv5 model

YOLOv5 consists of a series of target detection architectures and models which are pre-trained using the COCO dataset with a network structure consisting of four main components (*i.e.*, the Input module, Backbone module, Neck module, and Head module). By combining different network depths and widths, the YOLOv5 family is categorized into four versions from small to large, that is, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Thus, YOLOv5s, being the smallest model, contains the least number of parameters and is easy to be deployed in practical application scenarios. Therefore, to reduce the consumption of computational resources, YOLOv5s is adopted in our study as a baseline for improvements and further experiments. Each of the four components of the YOLOv5 model is described below in detail.

Table 1
Number of instances per pavement damage category used in the study.

Class of pavement damages	Total	Training	Validation	Testing
Longitudinal Crack	1160	671	244	245
Transverse Crack	1148	672	233	243
Alligator Crack	1126	686	228	212
Pothole	1050	632	220	198
Manhole Cover	889	509	191	189
Longitudinal Patch	1149	682	231	236
Transverse Patch	1149	696	234	219

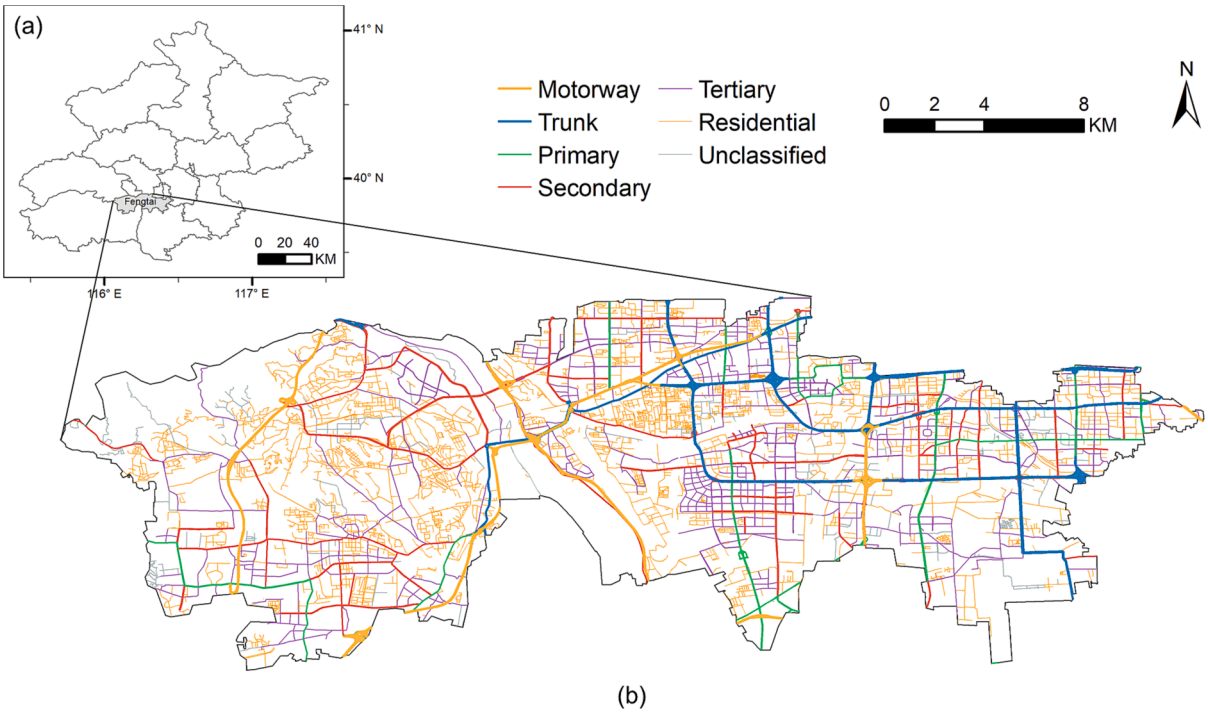


Fig. 1. Geographic location of the study area: (a) Beijing City, (b) Fengtai District.

YOLOv5 loads images through the ***Input module***, and several data augmentation methods are usually adopted at this stage, including HSV (Hue Saturation Value) alteration, panning, scaling, shear mapping, perspective, flipping, mosaic data augmentation, mix-up data augmentation, and copy-paste data augmentation. The data augmentation methods of the YOLOv5 network increase the background complexity, enhance the generalization ability of the models, and improve robustness and detection precision. Moreover, the YOLOv5 network uses an adaptive anchor frame detection approach, which applies the K-means clustering and the genetic learning algorithms to analyze the dataset and generates pre-defined anchor frames that fit the object boundary boxes in the dataset.

The ***Backbone module*** of the YOLOv5 model aggregates the input image information by the granularity of different types of imagery to extract the image features. Moreover, the final output feature map is used to characterize the object location, category, and other information. The module consists of the *Conv*, the *C3* and the *SPPF* modules. In more detail, the Conv module is responsible for the convolution operation for processing features, including a two-dimensional convolution (Conv2d), a Batch Normalization (BN), and a sigmoid-weighted linear unit (SiLU) activation function. As for the C3 module, it is a simplified CSP bottleneck block, which divides the input feature map into two branches where the first one performs the convolution operation and then passes through multiple bottlenecks to obtain the output, and the second one performs the convolution operation directly to obtain the output; the final outputs consists of fusing the outputs of both branches using the tensor concatenation (Concat) operation. This reduces the calculation intensity of the model, and furthermore enables a richer combination of features and better performance. Finally, the SPPF module in the YOLOv5 network is an improvement of the Spatial Pyramid Pooling (SPP) module in the YOLOv3 network, requiring less computational power and performing in a faster than the SPP module. By performing multiple maximum pooling operations on the feature map, the features at high layers are extracted and fused; therefore, the information on features at different scales can be obtained.

Concerning the ***Neck module***, it further enhances the features extracted by the Backbone module, making it possible for the model to recognize the required targets at different scales. In more detail, YOLOv5 adopts Path Aggregation Network (PANet) (Liu et al., 2018), which first transmits semantic information and fuses features through top-down paths, and then it transmits location information and fuses features through bottom-up paths; therefore, the feature information is being enhanced at different scales.

Finally, the ***Head module*** predicts the category probability and the boundary boxes of the target at multiple levels by using the multi-scale feature maps from the Neck module. Here again, YOLOv5 uses the Non-Maximum Suppression (NMS) method in the Head module to filter the detected boundary boxes.

Although the original YOLOv5s model has been widely used in image target detection, the model should be characterized when using street-view images, by multi-scale, complex background, and image distortion because the street-view imagery is not specifically captured for pavement damage. In our study, YOLOv5s was used as the baseline model for a smaller number of participants, and the large target detection layer, the GFPN module, the Decoupled Head module, and the LIoU loss method were integrated into the baseline network to form the improved model. With the following four improvements accorded to the original YOLOv5s network, an improved version of the YOLOv5s network model, YOLOv5s-M, is built for detecting pavement damage from street-view images (Fig. 2).

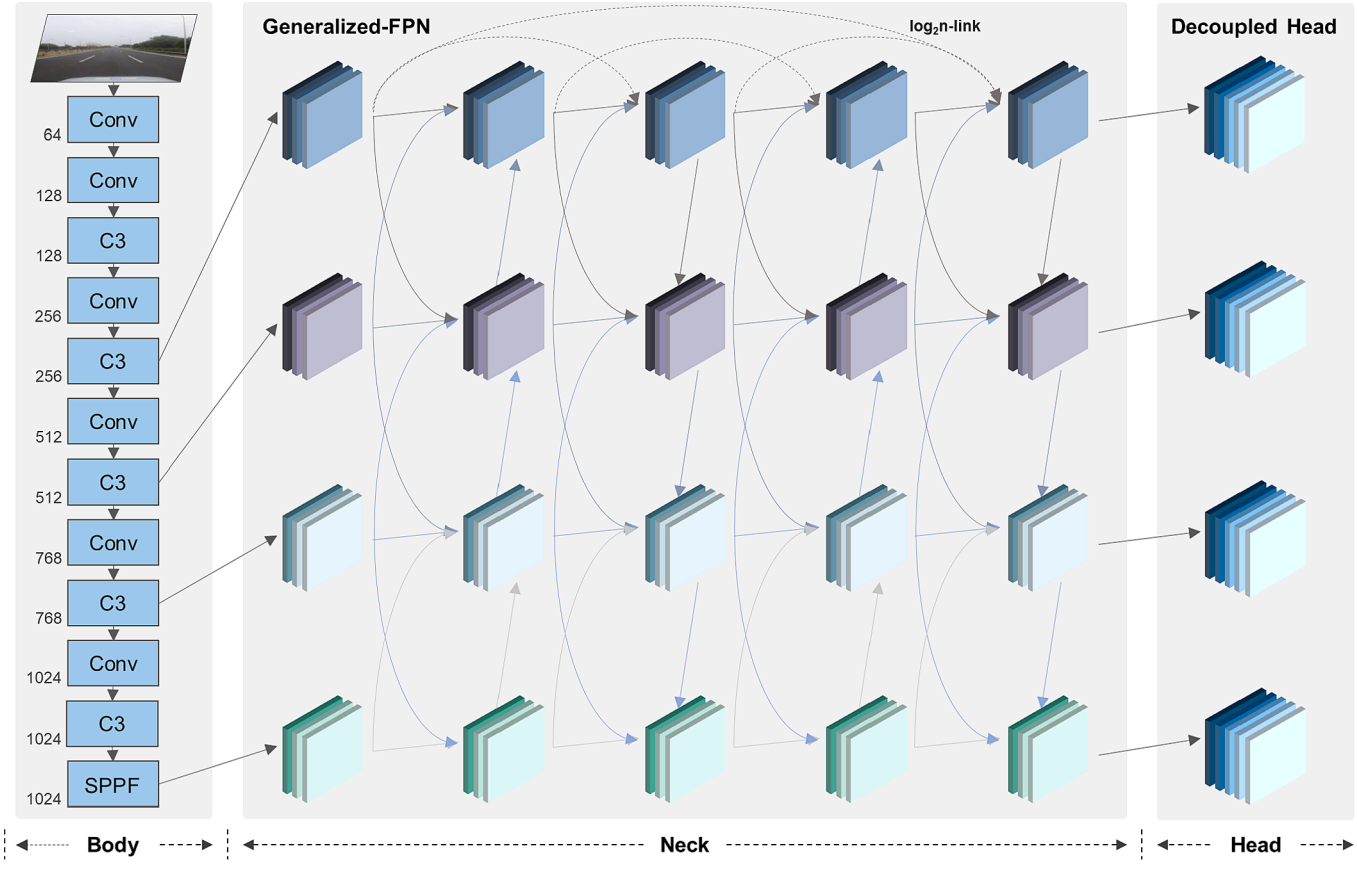


Fig. 2. The framework of the proposed YOLOv5s-M model: *Body* module contains an *Input* module and a *Backbone* module; An additional large-scale detection layer for large distress objects is added in the *Backbone* module; both cross-layer and cross-scale feature fusion is achieved by using the Generalized-FPN structure in the *Neck* module; Decoupled Head in the *Head* module to predict the bounding box and class label of objects.

2.2.2. Improvement of backbone module

YOLOv5 is detected at three scales, which is achieved by down-sampling the input image size by 32, 16, and 8, respectively. However, YOLOv5 is trained based on the COCO dataset, and its image input size is set to 640 × 640. In this study, the size of the street-view images is regularly 1024 × 1024 as mentioned above, showing an increase in image size compared to the original network images. Moreover, some of the street-view images have a large distress object (*e.g.* alligator cracks), thus occupying a large area of the whole image. In addition, the three original scales of YOLOv5 have poor applicability. Therefore, an additional large-scale detection layer was added in this study, which is obtained by down-sampling the input image size by 64. To create a feature map with a large receptive field that is suitable for identifying a sizable area of damage in street-view images, this large-scale detection layer can extract the semantic features from high layers and further fuse the obtained features with the spatial features from low layers.

2.2.3. Improvement of neck module

The PANet network, used in the Neck module of the YOLOv5 model, achieves feature fusion at different scales in adjacent stages but it lacks cross-layer fusion of features in non-adjacent stages and internal block connections. Therefore, the Generalized-FPN (GFPN) (Jiang et al., 2022) was applied to improve the Neck module outcome in our study (Fig. 3). In addition, GFPN consists of two types of connections, that is, skip-layer connection and cross-scale connection. The skip-layer connection is log_2n-link (Fig. 4(a)). In each level k, the l^{th} layer performs feature mapping from at most $log_2 l+1$ previous layers and these input layers show a power-of-2 distribution with depth i, as shown in Eq. (1).

$$P_k^l = C3\left(Concat\left(P_k^{l-2^n}, \cdots, P_k^{l-2^1}, P_k^{l-2^0}\right)\right) \tag{1}$$

where $l-2^n \geq 0$, $Concat()$ refers to the concatenation of the feature maps generated in all previous layers, and $C3()$ denotes the C3 module operation applied in YOLOv5. As for the log_2n-link, it avoids the information redundancy present in the densely connected structure of DenseNet (Huang et al., 2017), allowing thus for more efficient information transmission. This feature can be also extended to deeper networks. As for the cross-scale connection, queen fusion considers the features of the same and adjacent levels, and each node accepts not only the input from the previous node, but also the input from the nodes on the upper and lower inclined sides (Fig. 4(b)). Besides, the up-sampling and down-sampling methods used in the Neck module of YOLOv5 are also used at this level, i.e., the bilinear interpolation and convolution are used as the up-sampling and down-sampling functions, respectively.

To sum up, the Neck module of the improved YOLOv5s-M network uses a 5-layer GFPN structure, which extends the number of layers of the original Neck module and enables a deep and large Neck module. By doing so, the model has sufficient information exchange between the high and low layers, and it enhances the fusion of weak feature information among the feature layers at different scales, yielding to high accuracy.

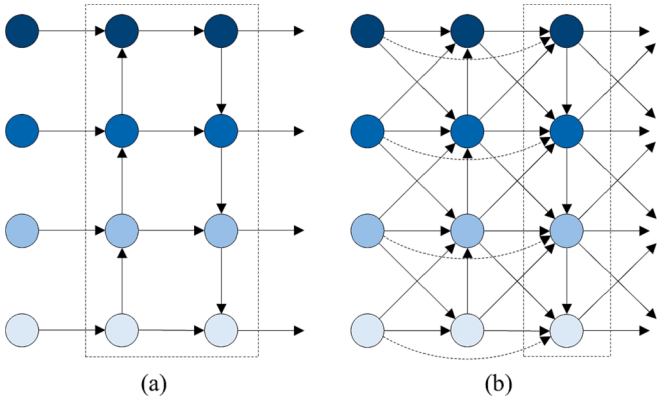


Fig. 3. Modification of the neck module (a) PANet, (b) GFPN.

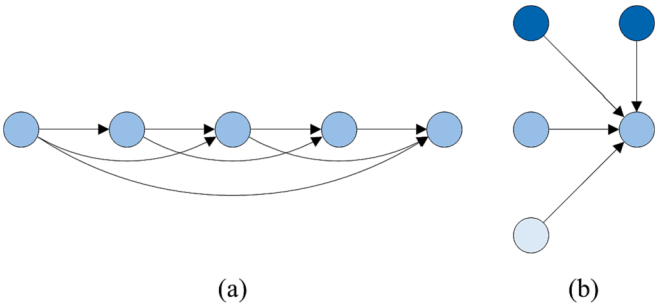


Fig. 4. (a) log_2n-link: the skip-layer connection, (b) Queen-fusion: the cross-scale connection.

2.2.4. Improvement of head module

The Head module of the YOLOv5s model uses a fusion of classification and regression branches to detect pavement damages in street-view images, although the classification and the regression tasks do not focus on the same features (Song et al., 2020), and the expression capability is insufficient. Hence, the decoupled head (Ge et al., 2021) was used to improve the Head module in our study, and its structure is shown in Fig. 5. This approach decouples the classification and regression branches to improve the performance of the model. Decoupling the detection head will undoubtedly increase the complexity of the operation, so the decoupled head is simplified to balance the speed and performance. For each level of feature, the decoupled head first adopts a 1 × 1 conv layer to reduce the feature channel to 256, and two parallel branches with two 3 × 3 conv layers are added, and each for classification and regression tasks, respectively. Finally, by implementing the decoupled head module, the proposed model achieves better performance with only a few added parameters.

2.2.5. Improvement of IoU loss

The loss function of YOLOv5 consists of three parts: the bounding box regression loss, the confidence loss, and the classification loss. Intersection over union (IoU) loss can use the boundary box regression to predict the target detection box more accurately, and YOLOv5 uses Complete IoU (CIoU) loss (Zheng et al., 2020), which considers the overlap area, the center distance, and the aspect ratio of the boundary box. Given the predicted box B and the real box B^{gt}, CIoU loss is defined as follows:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v, \tag{2}$$

where b and b^{gt} denote the centers of B and B^{gt}, respectively, $\rho(\bullet) = \|b - b^{gt}\|_2$ denotes the Euclidean distance between the two centers, c is the diagonal length of the minimum bounding rectangle of the two boxes, and $v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2$ and $\alpha = \frac{v}{(1-IoU)+v}$ measure the difference of the aspect ratio, where w and w^{gt} denote the width of B and B^{gt}, respectively, and h and h^{gt} represent the diagonal length of B and B^{gt}, respectively.

In this study, due to the multi-scale nature of the road pavement damages in the street-view images, the boundary boxes vary significantly in size. Since the targets we focus on are mostly linear, such as transverse crack, longitudinal crack, transverse patch, and longitudinal patch, the aspect ratios of the boundary boxes of these targets tend to be large. The accuracy of the diagonal length is more important for the linear pavement damage targets since they are typically situated on the diagonals of the detection boxes, which may be used to quantify the length of the linear pavement damage. Therefore, the CIoU loss method

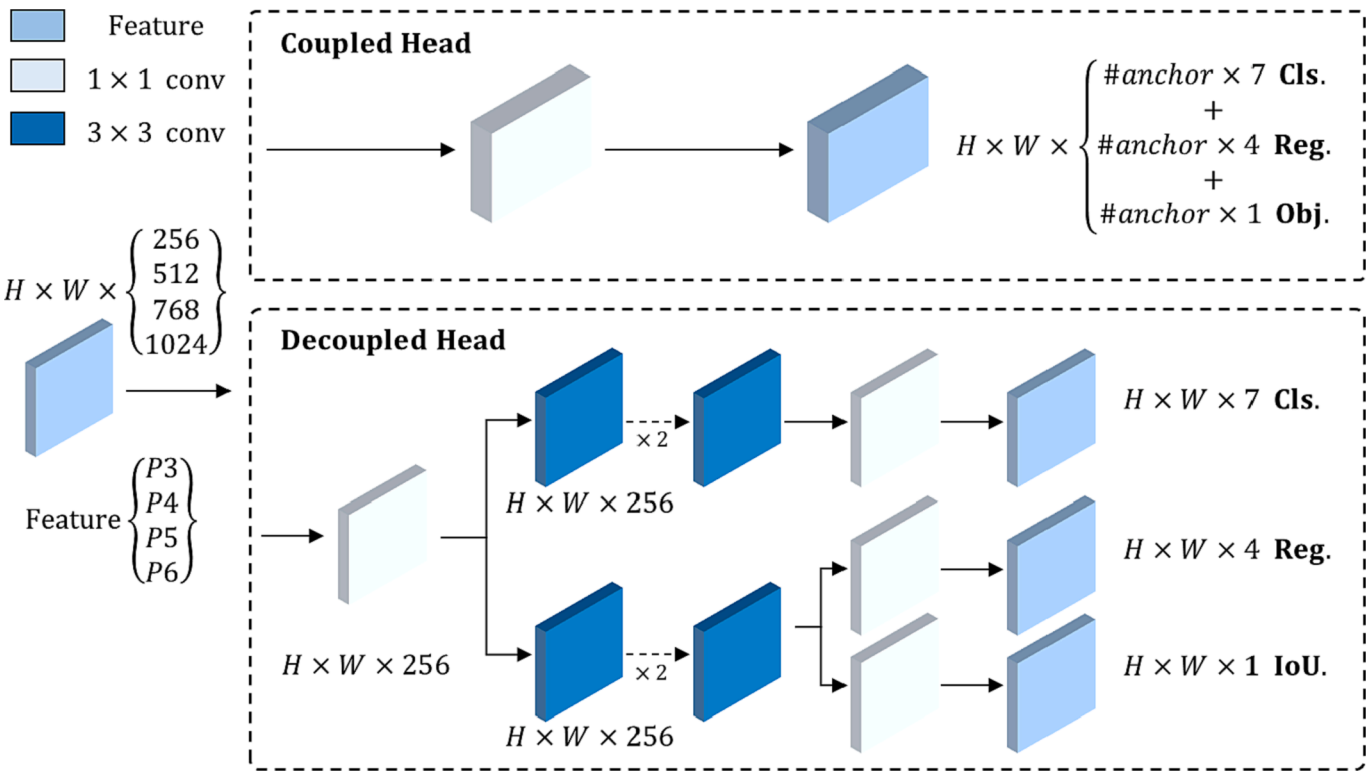


Fig. 5. Illustration of the coupled head and the decoupled head.

has been extended in our study. By considering the difference between the diagonal lengths of the two boundary boxes, diagonal IoU (LIoU) loss was proposed, which is defined as follows:

$$L_{LIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{\rho^2(l, l^{gt})}{c^2} + \alpha v, \tag{3}$$

where b and b^{gt} denote the centers of B and B^{gt}, respectively, $\rho(\bullet) = \|b - b^{gt}\|_2$ denotes the Euclidean distance between the two centers, c is the diagonal length of the minimum bounding rectangle of the two boxes, and $v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2$ and $\alpha = \frac{v}{(1-\mathrm{IoU})+v}$ measure the difference of the aspect ratio, where w and w^{gt} denote the width of B and B^{gt}, respectively, h and h^{gt} represent the diagonal length of B and B^{gt}, respectively, and l and l^{gt} denote the diagonal lengths of B and B^{gt}, respectively.

3. Experiments and results

3.1. Experiments setup

3.1.1. Training parameters

The hardware configuration includes *Intel*(R) *Xeon*(R) *Silver* 4116 CPU, 128 GB RAM, and 4 *NVIDIA* GeForce 1080Ti graphics cards. As for the software environment, it consists of Windows Server 2012 R2 Standard operating system in which *Python* 3.7.10 is used as the programming language, *CUDA*10.2 is used as the GPU computing platform, and *Pytorch* 1.7.1 serves as the deep learning framework. Concerning the network training hyperparameters, the number of epochs is set to 500, the batch size is set to 16, the learning rate is optimized during training using the Adam method, with the initial learning rate set to 0.001 and the momentum set to 0.9, and the learning rate is being warmed up using the warm-up method.

3.1.2. Performance evaluation

To evaluate the performance of the YOLOv5s-M model in the detection of road pavement damage, Precision (P), Recall (R), F1-Score, Average Precision (AP), mean Average Precision (mAP), parameters, FLOating Point operations (FLOPs) and Frames Per Second (FPS) were used for quantitative analysis. The precision and the recall are calculated as shown in Eqs. (4) and (5), respectively.

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

where True Positive (TP) is a positive sample predicted as positive by the model, False Positive (FP) is a negative sample predicted as positive by the model, True Negative (TN) is a negative sample predicted as negative by the model, and False Negative (FN) is a negative sample predicted as positive by the model. Here, the IoU ratio is used to reflect whether the prediction is correct or not as it represents the intersection ratio between the predicted boundary box and the real boundary box. Therefore, if IoU is greater than the critical level (0.5 in this study), then the detection is correct (TP); otherwise, it is incorrect (FP).

Considering the balance between Precision and Recall, the $F1$-Score was calculated as the Harmonic Mean of Precision and Recall values, as shown in Eq. (6).

$$F1 - Score = 2 \times \frac{P \times R}{P + R} \tag{6}$$

The average precision is calculated based on the precision and the recall values. The value of average precision is equal to the area under the P-R curve, and the higher the average precision, the higher the accuracy of the network. The calculation is shown in Eq. (7):

$$AP = \int_0^1 P(R)dR \tag{7}$$

In a multi-type target detection task, the detection precision of the model is evaluated by calculating the mAP of all types whose expression

is expressed in Eq. (8):

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i \tag{8}$$

where C represents the number of target types.

Parameters and FLOPs are usually used to measure the complexity of the model where the former refers to the number of parameters in the model and the latter represent the number of floating-point operations in the same model.

In addition to the detection precision, running speed is another important performance metric for target detection algorithms. A common metric for evaluating speed is FPS, which represents the number of images that can be processed per second. In our study, FPS was examined on a single NVIDIA GeForce 1080Ti graphics card.

3.2. Results and analysis

3.2.1. Results on the testing set

The results of the proposed YOLOv5s-M model on the testing set are shown in Table 2 where the *precision* is 0.782, the *recall* is 0.721, the *F1-Score* is 0.750, the average Precision is 0.798 for a critical IoU of 0.5, and it is equal to 0.509 when the critical IoU gets to the 0.5 ~ 0.95 range. For different categories of pavement damages, YOLOv5s-M has the highest Precision in detecting alligator cracks due to its obvious image features and large area; since transverse and longitudinal patches have much darker grey levels compared to transverse and longitudinal cracks, and the they usually has a bigger width than the cracks, consequently the detection precision of patches is significantly higher. As for the potholes, although the manhole cover is specifically labeled in our study to help learn the features so as to improve the capability of detecting potholes, the detection precision is totally the lowest due to the difficulty of recognition.

In more details, Fig. 6 shows the detection results of the YOLOv5s-M network from some example images of the testing set. Fig. 6(a) displays transverse and longitudinal cracks showing that the model can detect pavement damage in a good way and has a high confidence level. As for Fig. 6(b), a transverse patch with vehicle shadow blocking, a longitudinal patch with a large detection box, and a longitudinal patch with a small detection box at a distance were accurately detected. Moreover, Fig. 6(c) displays a longitudinal crack, a transverse patch, and a pothole. The proposed YOLOv5s-M achieved high confidence in both longitudinal crack and transverse patch and successfully detected the pothole of a small area. Finally, Fig. 6(d) presents mainly a large alligator crack and two potholes included in the alligator crack. The proposed model could detect them all but had low confidence in the detection of the smaller pothole in the experiment. As observed, the model performed well for targets of different scales as the pavement distress targets in the images could be accurately detected even in the case of vehicle shadow blocking.

3.2.2. Comparative experiments

To demonstrate the effectiveness of the YOLOv5s-M network in our study, comparative experiments were designed and conducted between the state-of-the-art detection models, including the single-stage target detection algorithms such as Scaled-YOLOv4 (Wang et al., 2021a), YOLOR (Wang et al., 2021b), and YOLOv7 (Wang et al., 2022) and the two-stage target detection algorithms such as Cascade R-CNN (Cai and Vasconcelos, 2021), and Sparse R-CNN (Sun et al., 2021). Referring to Table 3, the proposed YOLOv5s-M network has significant advantages in terms of parameters, *FLOPs*, *F1-Score*, *mAP*, and *FPS*, and it shows high performance on the pavement damage detection dataset of this study compared to the previously listed methods.

Moreover, the experimental results show that the YOLOv5s-M network has significant advantages in pavement damage detection from street-view images. The Scaled-YOLOv4 and YOLOR are only better than the proposed model in terms of *recall* but worse in terms of *precision* and have a lower final *average precision*. In addition, YOLOv7 is comparable to the proposed model in terms of the number of parameters as it achieves the highest accuracy and *FPS* among the other compared models, but the *precision* and *FPS* are still lower than those obtained using the proposed YOLOv5s-M model. Both Cascade R-CNN and Sparse R-CNN have a great number of parameters, high FLOPs, and a low FPS; however, they didn't achieve good precision in pavement damage detection.

In summary, the proposed YOLOv5s-M model has 36.5 M parameters and 39.8G FLOPs, indicating a relatively low complexity. The detection speed of the proposed model is fast, which makes it easy to be implemented in practice. Compared to the current state-of-the-art detection models, the proposed model achieves higher average precision and is more suitable for pavement damage detection in street-view images.

Table 2
Performance metrics of the proposed network on the test images.

Class	Precision	Recall	F1-Score	AP@0.5	AP@0.5:0.95
All	0.782	0.721	0.750	0.798	0.509
Longitudinal crack	0.683	0.608	0.643	0.691	0.426
Transverse crack	0.797	0.661	0.723	0.767	0.461
Alligator crack	0.869	0.849	0.859	0.899	0.616
Pothole	0.708	0.601	0.650	0.688	0.419
Manhole cover	0.810	0.783	0.796	0.860	0.565
Longitudinal patch	0.782	0.791	0.786	0.844	0.556
Transverse patch	0.821	0.752	0.785	0.834	0.521

3.2.3. Ablation study

To further analyze the effectiveness of the improvements in the proposed YOLOv5s-M network, ablation study was conducted on the pavement damage detection dataset from the street-view images in Fengtai, Beijing. In YOLOv5s-M, YOLOv5s was used as the baseline model, and the large target detection layer (P6 module), the GFPN module, the Decoupled Head module, and the LIoU loss method were integrated into the baseline network to create the improved one. Here, the influence of each factor is analyzed. The results obtained by YOLOv5s-M without each component were compared with those of the complete YOLOv5s-M to directly demonstrate its effectiveness. The "YOLOv5s-M−P6″ denotes the simplified form of the YOLOv5s-M without the P6 module. The other variants of the YOLOv5s-M in Table 4 are denoted in similar forms.

It can be seen in Table 4 that when the P6 module was removed from YOLOv5s-M, the model suffered loss in the large object detection, and lost the capacity to extract high-level semantic features. This impacted the effectiveness of subsequent feature fusion across various scales. Consequently, the performance of YOLOv5s-M without P6 module was attenuated and led to a decrease in *mAP* scores. As shown in the third row in Table 4, The *AP@0.5* score of YOLOv5s-M−GFPN is 0.676, which is considerably lower than those of the YOLOv5s-M and other variants. This demonstrated that GFPN module had a significant influence on the detection capability of the YOLOv5s-M model by affecting the fusion of the features in various scales in the feature layers. Furthermore, the results of YOLOv5s-M without the LIoU loss and Decoupled module indicate that the LIoU loss and the Decoupled Head module can improve the detection capability without substantially increasing its parameters and FLOPs.

4. Discussion

To meet the specific needs of pavement damage target detection in the street-view images, this study proposed an improved YOLOv5s network by improving and optimizing the *Backbone* module, the *Neck* module, the *Head* module, and the *IoU* loss function of the YOLOv5s network model. The proposed YOLOv5s-M network was validated and verified on the labeled street-view image dataset which we created in Fengtai District, Beijing City, China, and the precision and performance

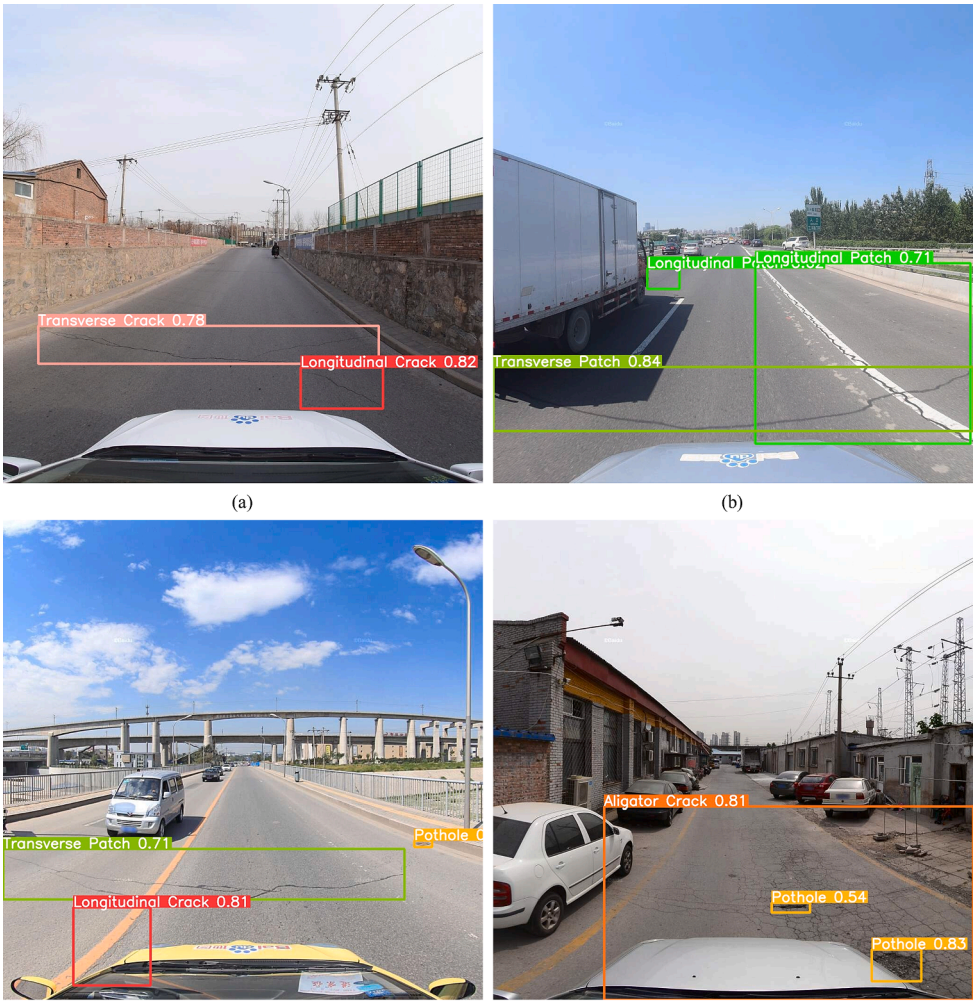


Fig. 6. Some examples of pavement damage detection in the test images. (a) cracks; (b) patches; (c) patch, crack and pothole; (d) alligator crack and potholes.

Table 3
Performance comparison of the YOLOv5s-M and different state-of-the-art models.

Methods	Parameters	*FLOPs*	*Precision*	*Recall*	*F1-Score*	*AP*@0.5	*AP*@0.5:0.95	*FPS*
Scaled-YOLOv4	52.5 M	119.9G	0.391	**0.840**	0.534	0.746	0.440	11.2
YOLOR	52.5 M	119.7G	0.470	0.825	0.599	0.754	0.480	18.1
YOLOv7	36.5 M	103.0G	0.773	0.677	0.722	0.758	0.481	25.5
Cascade R-CNN	68.9 M	239.1G	0.753	0.674	0.711	0.633	0.357	11.0
Sparse R-CNN	106.0 M	153.3G	0.741	0.658	0.697	0.672	0.402	8.7
YOLOv5s-M	**36.5 M**	**39.8G**	**0.782**	0.721	**0.750**	**0.798**	**0.509**	**42.0**

Table 4
Ablation study of the proposed YOLOv5s-M model.

Methods	P6	GFPN	LIoU	Decoupled	Parameters	FLOPs	AP@0.5	AP@0.5:0.95
YOLOv5s	–	–	–	–	7.03 M	15.8G	0.759	0.450
YOLOv5s-M−P6	–	√	√	√	21.44 M	39.7G	0.781	0.495
YOLOv5s-M−GFPN	√	–	√	√	12.33 M	16.2G	0.767	0.474
YOLOv5s-M−LIoU	√	√	–	√	36.49 M	39.8G	0.792	0.501
YOLOv5s-M−Decoupled	√	√	√	–	36.49 M	39.8G	0.794	0.496
YOLOv5s-M	√	√	√	√	36.49 M	39.8G	0.798	0.509

of the model in pavement damage detection were improved after applying the proposed approach. As a result, the YOLOv5s-M network more closely matches the specific features of the street-view images and thus it provides a novel technical method for damage detection on urban roads. In the following section, the effect of IoU loss on the model precision, the inclusion or exclusion of manhole cover training samples, and the application for large-scale pavement damage detection will be discussed and analyzed in detail.

4.1. Comparative analysis of IoU loss

It has been demonstrated that IoU loss affects the model precision, including the proposed LIoU loss itself (Eq. (3)), the CIoU loss (Eq. (2)) used in YOLOv5, and other mainstream IoU losses such as the distance IoU (DIoU) loss (Rezatofighi et al., 2019) (Eq. (9)), and the Efficient IoU (EIoU) loss (Zhang et al., 2022) (Eq. (10)).

$$L_{DIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2}, \tag{9}$$

$$L_{EIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{(w^c + h^c)^2} + \frac{\rho^2(w, w^{gt})}{(w^c)^2} + \frac{\rho^2(h, h^{gt})}{(h^c)^2}, \tag{10}$$

where w^c and h^c refer to the width and the height of the minimum bounding rectangle of the two boxes, respectively.

To ensure rational precision, the other structures of the YOLOv5s-M model remain constant in the comparative experiments with different IoU losses. The generated model precision results are shown in Table 5. As observed, LIoU loss has a high *mAP* and an improved *AP@0.5* by 0.009, 0.006, and 0.015 compared with DIoU loss, CIoU loss, and EIoU loss. Additionally, despite the fact that EIoU loss—an enhanced CIoU loss—extends the aspect ratio difference to the difference in length and width and also improves the ability to identify the shape of the rectangular box—it did not perform well in the proposed dataset. By adding the difference in the diagonal length to the difference in the aspect ratio as the case in CIoU loss, LIoU loss could determine the shape of the rectangular box better than EIoU loss, and achieved the best results in the proposed dataset.

4.2. Training with or without manhole cover

For the purposes to investigate the effect of adding manhole cover on pothole detection, comparative experiments were conducted on the dataset while removing the manhole cover, and the results are displayed in Table 6. To eliminate the effect of the manhole cover on the model precision, the average precisions of all categories of damage were calculated after the removal of manhole cover. As it can be seen in Table 6, the precision of the pothole detection is significantly improved by adding the manhole cover, where *AP@0.5* is improved by 4.7% and *AP@0.5:0.95* is improved by 5.3%; except for pothole, the detection precision of the pavement damage of other categories is also slightly improved.

The confusion matrix was plotted for the two methods (Fig. 7). The columns of the matrix represent the predicted categories and the rows represent the ground truth. The confusion matrix was normalized in the direction of the columns. The sum of the values in each column is equal to the unit, which indicates the recall of the corresponding category, and the values in each row show the proportion of predictions in the corresponding category. As observed, although the *recall* of pothole does not differ significantly, the *recall* of the model with manhole cover training increases from 0.662 to 0.682. Moreover, a significant discrepancy in the pothole detection precision is caused by the incorrect detection of 11.6% of the manhole covers without manhole cover training. Therefore, pothole and manhole covers are prone to be confused due to their similar geometric features, and adding the labeled data of manhole cover could effectively improve the detection precision of pothole.

Table 5
Comparison of different IoU loss functions.

IoU Loss	AP@0.5	AP@0.5:0.95
DIoU	0.789	0.502
CIoU	0.792	0.501
EIoU	0.783	0.499
LIoU	**0.798**	**0.509**

Table 6
Comparison of the training strategies without or with manhole cover.

Methods	Class	AP@0.5	AP@0.5:0.95
Without Manhole Cover	Pothole	0.657	0.398
	Others	0.799	0.510
	All	0.779	0.494
With Manhole Cover	Pothole	0.688	0.419
	Others	0.807	0.516
	All	0.787	0.500

4.3. Evaluation in practical application

The trained YOLOv5s-M model was afterwards used to detect the street-view images captured in Fengtai District, Beijing, and the number of detected pavement damages is shown in Fig. 8. It can be seen that relatively little pavement damage was detected in the western part and the motorway of Fengtai District, whereas more damage is encountered in the inner part of the city. Moreover, some pavement damage is serious and should be urgently maintained or restored.

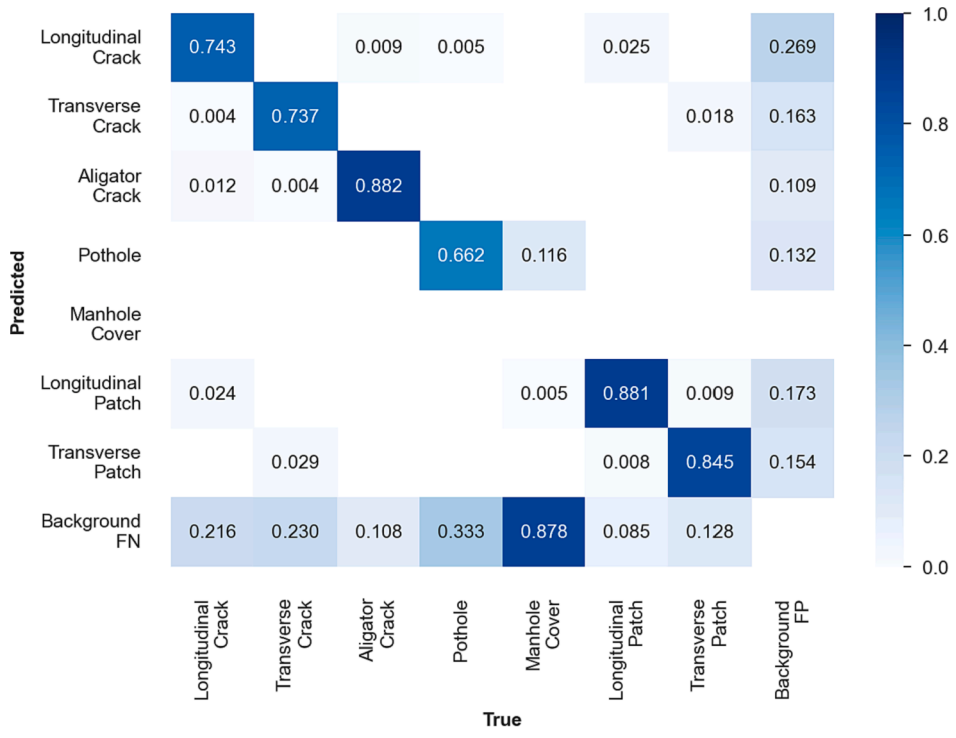
The average number of pavement distresses at the different grades of the roads in Beijing was counted as shown in Table 7. It can be seen that the motorways had the least number of damage due to its high grade and good quality, and its damage had been restored in time. The Trunk and primary roads are the two types of road grades with the largest number of pavement damage due to the heavy traffic, but their pavement damage had been also restored extensively. The number of crack and pothole is significantly higher in low-grade roads that are of low quality and aging. Among the six pavement damages, longitudinal and transverse cracks were the most encountered and they were formed relatively in the early stage of pavement damage; however, the number of alligator cracks and potholes was slightly lower although their severity is the highest, basically in the late stage of road aging. In terms of the number of transverse and longitudinal patches, the overall number was relatively small, but timely restoration of the pavement damage is still required to improve driving performance and to ensure traffic safety.

In this study, the Xiaoyue Road in Fengtai District, Beijing was employed as an example for pavement damage detection. The street-view images of this road in 2017 and 2019 were acquired and screened to obtain 408 images of the same location for each year. According to public data, roads in this area were overhauled in 2018. It can be seen in Fig. 9 that most of the road had been improved by the maintenance in 2018 and the number of pavement damage was reduced significantly. However, there still exist some areas where the road was further deteriorated due to the fact that the road was not maintained completely in 2018 although the traffic load was important.

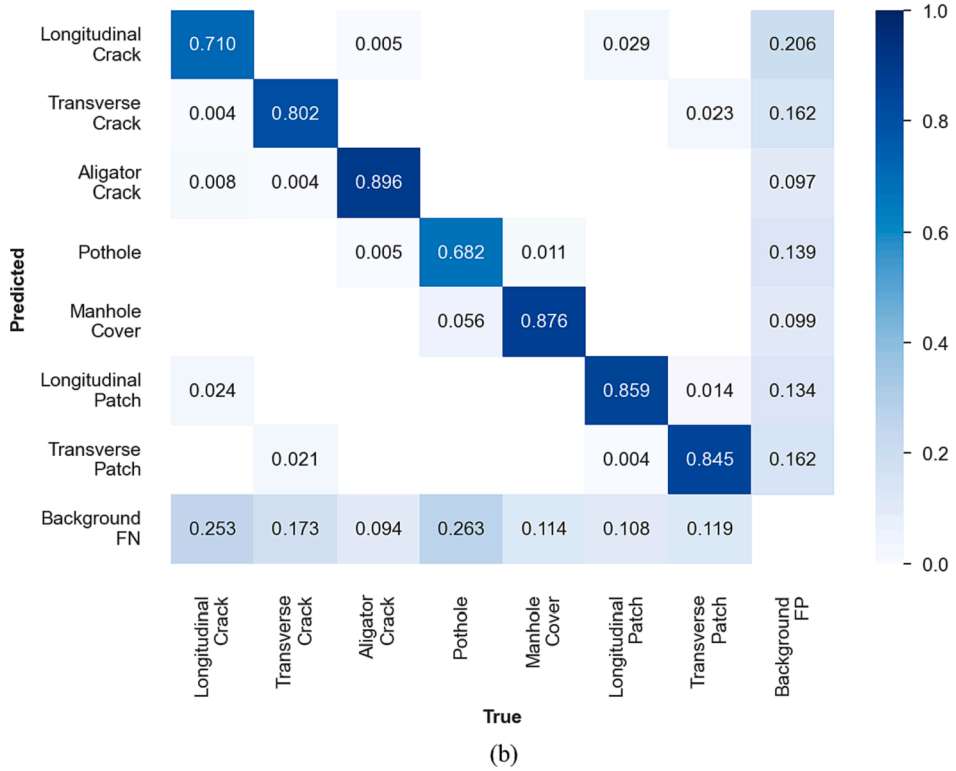
4.4. Further test on the public dataset

To further validate the performance of the proposed YOLOv5s-M model, it is further tested on a public dataset, the Global Road Damage Detection challenge 2020 (RDD2020) dataset (Arya et al., 2021). The RDD2020 dataset was collected in Japan, India and Czech using smartphones installed on the windshield of the vehicle, which has a similar perspective to street-view imagery. It provides a training set including 21,041 images with four types of road damage labels (longitudinal cracks, reverse cracks, complex cracks and potholes) and two test sets without annotations, including 2631 and 2664 images, respectively. The image size of the dataset collected in Japan and India is 600 × 600 pixels, and 720 × 720 pixels in Czech. The detection results of both test sets can be submitted to the evaluation server, and the *F1*-score can be calculated for each submission. The training set in the RDD2020 dataset is divided into training and validation sets in the ratio of 4:1, and the images are scaled to 1024 × 1024 pixels and then input into the network model for training.

The YOLOv5s-M model achieved good *F1-scores* of 0.6709 and



(a)



(b)

Fig. 7. Confusion matrix of the model training (a) without manhole cover and (b) with manhole cover.

0.6601 on the two test datasets, respectively. The results and rankings in RDD2020 (Arya et al., 2020) as well as our results, are shown in Table 8. The term "*Ensemble*" indicates that the method uses *ensemble* learning, and "*Single*" represents that the method is a single network without *ensemble* learning. Our proposed method belongs to a single network. As can be seen in Table 8, among the single-model methods, our method outperformed the other single-model methods on both test datasets. The improvements on the test dataset #1 is larger than 15.39%, and 14.78% on test dataset #2. It should be noted that although the ultralytics-YOLO achieved best results on the test datasets, it adopts *ensemble* prediction approach and uses three models for the model *ensemble* (Hegde et al., 2020), and consequently a significant cost of training resource is

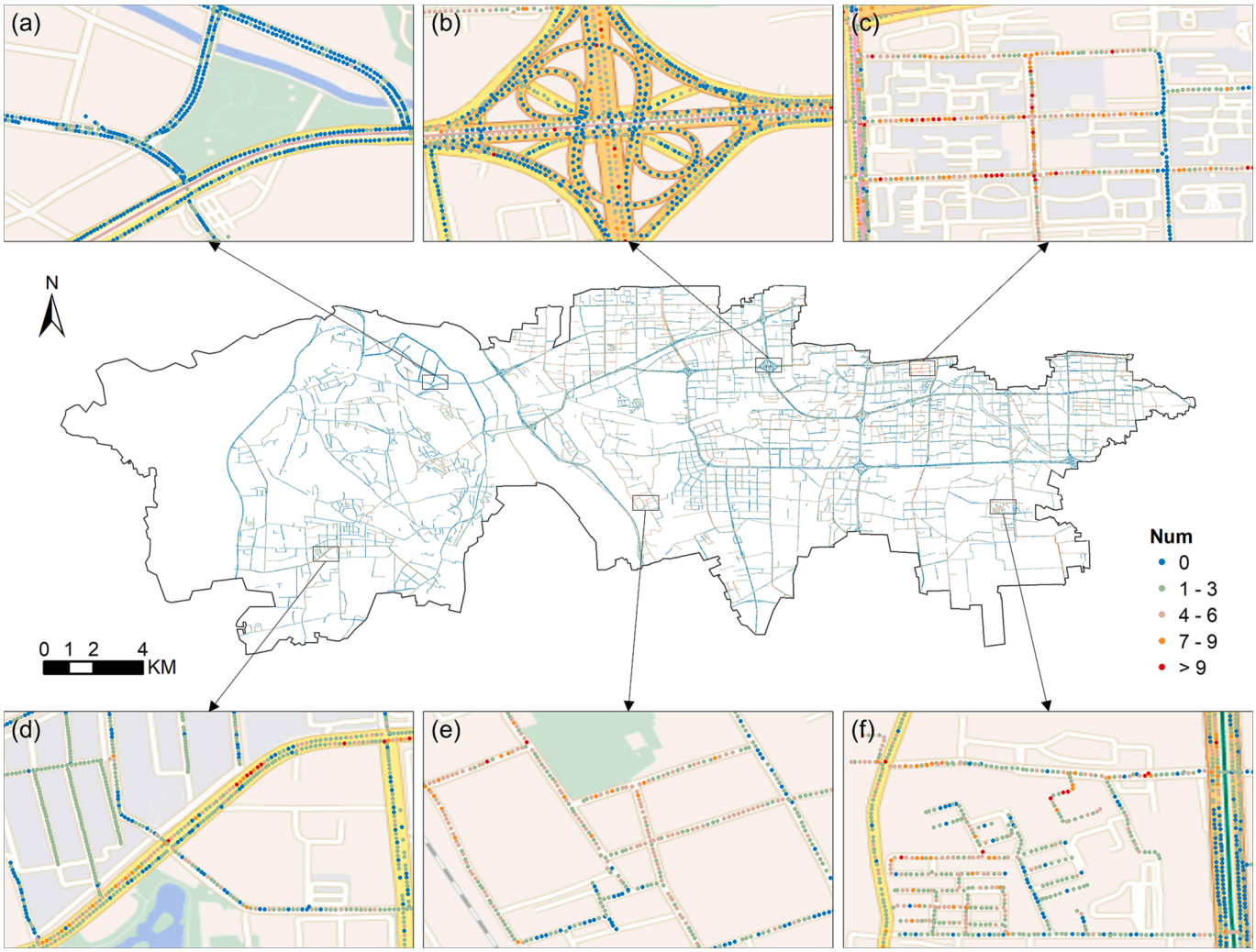


Fig. 8. Distribution of pavement diseases in Fengtai District, Beijing. (a), (b), (c), (d), (e), and (f) represents the results of some typical areas (zoom in by 16 times).

Table 7
Average number of detected pavement damages at different road levels in Fengtai District, Beijing.

Class	Images	Longitudinal Crack	Transverse Crack	Alligator Crack	Pothole	Longitudinal Patch	Transverse Patch	All
Motorway	7941	0.274	0.108	0.027	0.075	0.539	0.293	1.315
Trunk	13,565	0.524	0.228	0.074	0.071	0.797	0.526	2.220
Primary	9143	0.855	0.376	0.137	0.119	0.315	0.209	2.010
Secondary	21,392	0.682	0.417	0.188	0.113	0.110	0.088	1.598
Tertiary	49,539	0.706	0.460	0.197	0.140	0.086	0.066	1.656
Residential	46,107	0.687	0.571	0.200	0.263	0.010	0.011	1.742
Unclassified	8617	0.698	0.604	0.339	0.295	0.012	0.013	1.960
All	156,304	0.668	0.452	0.182	0.171	0.161	0.110	1.742

necessary. Our YOLOv5s-M method, as a single network method, achieved similar results to the ultralytics-YOLO, only 0.58% lower on test dataset #1 and 0.92% on test dataset #2. This test on RDD2020 dataset indicates that the proposed YOLOv5s-M network is applicable to different road pavement datasets, and achieve good identification of pavement damages.

5. Conclusions

A novel deep learning network YOLOv5s-M for pavement damage detection from street-view images is presented in our study. The proposed YOLOv5s-M network enhances the detection of large-sized damage targets (*i.e.* alligator cracks) by adding large-scale detection layers, employs the Generalized-FPN to implement a model structure with a very deep and large *neck* module, creates the improved LIoU loss to regress the boundary box, and adopts the Decoupled *Head* to achieve the decoupling detection for classification and regression. This approach significantly improved the precision and the adaptability of the YOLOv5 network in the detection of pavement damage targets from street-view images. Moreover, the experiments conducted in Fengtai District, Beijing City, China, indicated that the proposed YOLOv5s-M model could outperform the existing advanced target detection models (*e.g.*, Scaled-YOLOv4, YOLOR, YOLOv7, Cascade R-CNN, and Sparse R-CNN, etc.) with a *precision* of 0.782, *recall* of 0.721, *F1-Score* of 0.750, and an average *precision* of 0.798 at a critical IoU of 0.5 on the testing set. Furthermore, the ablation study shows that the YOLOv5s-M model has

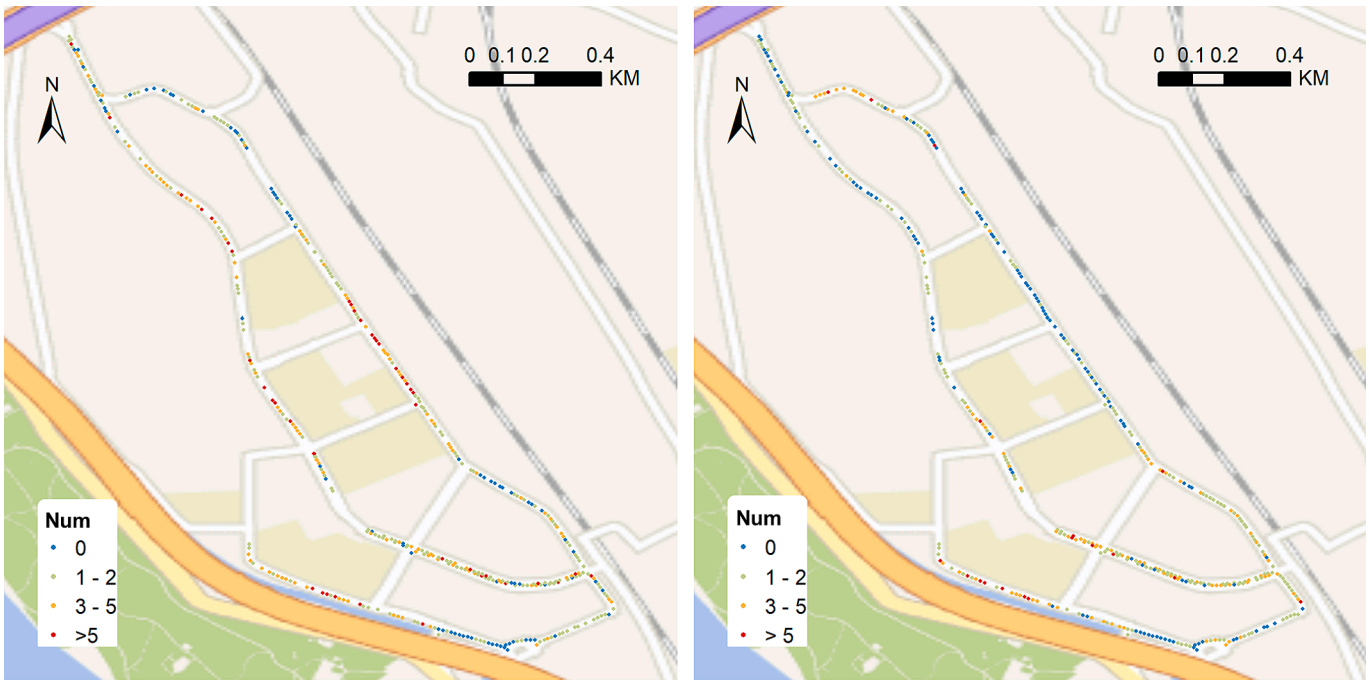


Fig. 9. Results of pavement damage detection on Xiaoyue Road, Fengtai District, Beijing (a) 2017, (b) 2019.

Table 8
Comparison of experimental results in RDD2020 dataset.

Rank	Methods	Solution	Test1 F1-Score	Test2 F1-Score
8	YOLOv4 (Zhang et al., 2020)	Single	0.5538	0.5412
7	EfficientDet (Naddaf-Sh et al., 2020)	Single	0.5650	0.5470
6	YOLOv4 + Faster-RCNN (Liu et al., 2020)	Ensemble	0.5636	0.5707
5	YOLOv5x (Jeong, 2020)	Single	0.5683	0.5710
4	CSPDarknet53-YOLO (Mandal et al., 2020)	Single	0.5814	0.5751
3	Cascade-RCNN (Pei et al., 2020)	Ensemble	0.6290	0.6219
2	YOLOv4 (Doshi and Yilmaz, 2020)	Ensemble	0.6275	0.6358
1	ultralytics-YOLO (Hegde et al., 2020)	Ensemble	0.6748	0.6662
–	YOLOv5s-M (ours)	Single	0.6709	0.6601

effectively improved the model's capability of detecting pavement damage in the street-view images compared to the baseline model. Additionally, the proposed method was applied to detect several typical pavement damages from 156,304 street-view images in the entire Fengtai District, Beijing. The statistical analysis of pavement damages of different roads shows that the proposed network can detect longitudinal and transverse cracks, longitudinal and transverse patches, alligator crack, and pothole very well as they can provide basic data support for road maintenance decision-making.

The proposed YOLOv5s-M network only considers the features of pavement damage in the street-view images of Fengtai District, Beijing. Thus, street-view images will be collected from different cities to improve the model's robustness and stability in future work. In the meanwhile, the pavement damage detection in the cases where some areas are blocked by obstacles, such as vehicles and trees, in the street-view images, needs to be further investigated. Additionally, the transfer learning capability of the YOLOv5s-M network on street-view images and other vertically captured image datasets (*e.g.*, Unmanned Aerial Vehicle (UAV) images and vehicle on-board videos) is also an interesting research direction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This work was financially supported by the National Natural Science Foundation of China under Grant No. 42171327, and the International Research Center of Big Data for Sustainable Development Goals, China under Grant No. CBAS2022GSP06. The authors would like to thank the anonymous reviewers for their valuable comments and questions, which help improve this manuscript.

References

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., Weaver, J., 2010. Google street view: capturing the world at street level. Computer 43, 32–38.

Arya, D., Maeda, H., Ghosh, S.K., Toshniwal, D., Omata, H., Kashiyama, T., Sekimoto, Y., 2020. Global Road Damage Detection: State-of-the-art Solutions. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE, pp. 5533–5539.

Arya, D., Maeda, H., Ghosh, S.K., Toshniwal, D., Mraz, A., Kashiyama, T., Sekimoto, Y., 2021. Deep learning-based road damage detection and classification for multiple countries. Autom. Constr. 132, 103–935.

Asian Development Bank 2003. *Road Funds and Road Maintenance: An Asian Perspective*, Manila: Asian Development Bank. https://www.adb.org/publications/road-funds-and-road-maintenance-asian-perspective.

Cai, Z., Vasconcelos, N., 2021. Cascade R-CNN: high quality target detection and instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 43, 1483–1498.

Camilleri, N., Gatt, T. 2020. Detecting road potholes using computer vision techniques. In: 2020 IEEE 16th International Conference on Intelligent Computer

Communication and Processing (ICCP). IEEE, pp. 343–350. https://doi.org/10.1109/ICCP51029.2020.9266138.
Chacra, D.B.A., Zelek, J.S. 2017. Fully Automated road defect detection using street view images. In: 2017 14th Conference on Computer and Robot Vision (CRV). IEEE, pp. 353–360. https://doi.org/10.1109/CRV.2017.50.
Coenen, T.B.J., Golroo, A., 2017. A review on automated pavement distress detection methods. Cogent Eng. 4, 1374822.
Doshi, K., Yilmaz, Y., 2020. Road Damage Detection using Deep Ensemble Learning. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE, pp. 5540–5544.
Fassmeyer, P., Kortmann, F., Drews, P., Funk, B., 2021. Towards a Camera-Based Road Damage Assessment and Detection for Autonomous Vehicles. In: Applying Scaled-YOLO and CVAE-WGAN. 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall). IEEE, pp. 1–7.
Fei, Y., Wang, K.C.P., Zhang, A., Chen, C., Li, J.Q., Liu, Y., Yang, G., Li, B., 2020. Pixel-level cracking detection on 3D asphalt pavement images through deep-learning-based CrackNet-V. IEEE Trans. Intell. Transp. Syst. 21, 273–284.
Feng, H., Li, W., Luo, Z., Chen, Y., Fatholahi, S.N., Cheng, M., Wang, C., Junior, J.M., Li, J., 2022. GCN-based pavement crack detection using mobile LiDAR point clouds. IEEE Trans. Intell. Transp. Syst. 23, 11052–11061.
Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J. 2021. YOLOX: Exceeding YOLO Series in 2021. arXiv preprint arXiv:2107.08430. https://arxiv.org/abs/2107.08430.
Hascoet, T., Zhang, Y., Persch, A., Takashima, R., Takiguchi, T., Ariki, Y., 2020. FasterRCNN Monitoring of Road Damages: Competition and Deployment. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE, pp. 5545–5552.
Hegde, V., Trivedi, D., Alfarrarjeh, A., Deepak, A., Kim, S.H., Shahabi, C., 2020. Yet Another Deep Learning Approach for Road Damage Detection using Ensemble Learning. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE, pp. 5553–5558.
Hoang, N.-D., Nguyen, Q.-L., 2019. A novel method for asphalt pavement crack classification based on image processing and machine learning. Eng. Comput. 35, 487–498.
Hou, Y., Li, Q., Zhang, C., Lu, G., Ye, Z., Chen, Y., Wang, L., Cao, D., 2021. The state-of-the-art review on applications of intrusive sensing, image processing techniques, and machine learning methods in pavement monitoring and analysis. Eng. 7, 845–856.
Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 4700–4708.
Jeong, D., 2020. Road Damage Detection Using YOLO with Smartphone Images. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE, pp. 5559–5562.
Jiang, Y., Tan, Z., Wang, J., Sun, X., Lin, M., Li, H., 2022. GiraffeDet: A Heavy-Neck Paradigm for Target detection. International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=cBu4ElJfneV.
Jiang, H., Li, Q., Jiao, Q., Wang, X., Wu, L., 2018. Extraction of wall cracks on earthquake-damaged buildings based on TLS point clouds. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 11, 3088–3096.
Kortmann, F., Talits, K., Fassmeyer, P., Warnecke, A., Meier, N., Heger, J., Drews, P., Funk, B. 2020. Detecting Various Road Damage Types in Global Countries Utilizing Faster R-CNN. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE, pp. 5563-5571. https://doi.org/10.1109/BigData50022.2020.9378245.
Lei, X., Liu, C., Li, L., Wang, G., 2020. Automated pavement distress detection and deterioration analysis using street view map. IEEE Access 8, 76163–76172.
Li, Y., Che, P., Liu, C., Wu, D., Du, Y., 2021. Cross-scene pavement distress detection by a novel transfer learning framework. Comput.-Aided Civ. Infrastruct. Eng. 36, 1398–1415.
Liu, S., Qi, L., Qin, H., Shi, J., Jia, J., 2018. Path Aggregation Network for Instance Segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 8759–8768.
Liu, Y., Zhang, X., Zhang, B., Chen, Z., 2020. Deep Network For Road Damage Detection. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE, pp. 5572–5576.
Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T., Omata, H., 2018. Road damage detection and classification using deep neural networks with smartphone images. Comput.-Aided Civ. Infrastruct. Eng. 33, 1127–1141.
Majidifard, H., Adu-Gyamfi, Y., Buttlar, W.G., 2020. Deep machine learning approach to develop a new asphalt pavement condition index. Constr. Build. Mater. 247, 118513.
Mandal, V., Uong, L., Adu-Gyamfi, Y., 2018. Automated Road Crack Detection Using Deep Convolutional Neural Networks. In: 2018 IEEE International Conference on Big Data (Big Data). IEEE, pp. 5212–5215.
Mandal, V., Mussah, A.R., Adu-Gyamfi, Y., 2020. Deep Learning Frameworks for Pavement Distress Classification: A Comparative Analysis. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE, pp. 5577–5583.
Mei, Q., Gül, M., 2020. A cost effective solution for pavement crack inspection using cameras and deep neural networks. Constr. Build. Mater. 256, 119397.
Mokhtari, S., Wu, L., Yun, H.-B., 2016. Comparison of supervised classification techniques for vision-based pavement crack detection. Transp. Res. Rec. 2595, 119–127.
Naddaf-Sh, S., Naddaf-Sh, M.M., Kashani, A.R., Zargarzadeh, H., 2020. An Efficient and Scalable Deep Learning Approach for Road Damage Detection. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 5602–5608.
Ng, J.R., Wong, J.S., Goh, V.T., Yap, W.J., Yap, T.T.V., Ng, H., 2019. Identification of road surface conditions using IoT sensors and machine learning. Lect. Notes Electr. Eng. 259–268.
Nguyen, S.D., Tran, T.S., Tran, V.P., Lee, H.J., Piran, M.J., Le, V.P., 2022. Deep learning-based crack detection: a survey. Int. J. Pavement Res. Technol. 1–25.
Pan, Y., Zhang, X., Tian, J., Jin, X., Luo, L., Yang, K., 2017. Mapping asphalt pavement aging and condition using multiple endmember spectral mixture analysis in Beijing. China. *J. Appl. Remote Sens.* 11, 016003.
Pan, Y., Chen, X., Sun, Q., Zhang, X., 2021. Monitoring asphalt pavement aging and damage conditions from low-altitude UAV imagery based on a CNN approach. Can. J. Remote Sens. 47, 432–449.
Pei, Z., Lin, R., Zhang, X., Shen, H., Tang, J., Yang, Y., 2020. CFM: A Consistency Filtering Mechanism for Road Damage Detection. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 5584–5591.
Ragnoli, A., De Blasiis, M.R., Di Benedetto, A., 2018. Pavement distress detection methods: a review. Infrastructures 3, 58.
Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 658–666.
Singh, J., Shekhar, S. 2018. Road Damage Detection And Classification In Smartphone Captured Images Using Mask R-CNN. arXiv preprint arXiv:1811.04535. https://arxiv.org/abs/1811.04535.
Song, G., Liu, Y., Wang, X., 2020. Revisiting the Sibling Head in Object Detector. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11563–11572.
Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., Luo, P. 2021. Sparse R-CNN: End-to-End Target detection with Learnable Proposals. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14454-14463. https://openaccess.thecvf.com/content/CVPR2021/html/Sun_Sparse_R-CNN_End-to-End_Object_Detection_With_Learnable_Proposals_CVPR_2021_paper.html.
Tedeschi, A., Benedetto, F., 2017. A real-time automatic pavement crack and pothole recognition system for mobile Android-based devices. Adv. Eng. Inf. 32, 11–25.
Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.-M., 2021a. Scaled-YOLOv4: Scaling Cross Stage Partial Network. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13029–13038.
Wang, C.-Y., Yeh, I.-H., Liao, H.-Y.M. 2021b. You Only Learn One Representation: Unified Network for Multiple Tasks. arXiv preprint arXiv:2105.04206. https://arxiv.org/abs/2105.04206.
Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y. M. 2022. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696. https://arxiv.org/abs/2207.02696.
Zhang, Y.-F., Ren, W., Zhang, Z., Jia, Z., Wang, L., Tan, T., 2022. Focal and efficient IOU loss for accurate bounding box regression. Neurocomputing 506, 146–157.
Zhang, X., Xia, X., Li, N., Lin, M., Song, J., Ding, N., 2020. Exploring the Tricks for Road Damage Detection with A One-Stage Detector. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 5616–5621.
Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D., 2020. Distance-IoU loss: faster and better learning for bounding box regression. Proc. AAAI Conf. Artif. Intell. 34, 12993–13000.