Automation in Construction 158 (2024) 105192

Contents lists available at ScienceDirect

Automation in Construction

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

CNN-based pavement defects detection using grey and depth images

Peigen Li [a], Bin Zhou [d], Chuan Wang [e], Guizhang Hu [b], Yong Yan [b], Rongxin Guo [b], Haiting Xia [a,c,*]

[a] Faculty of Civil Engineering and Mechanics, Kunming University of Science and Technology, Kunming 650500, China
[b] Yunnan Key Laboratory of Disaster Reduction in Civil Engineering, Faculty of Civil Engineering and Mechanics, Kunming University of Science and Technology, Kunming 650500, China
[c] Faculty of Civil Aviation and Aeronautics, Kunming University of Science and Technology, Kunming 650500, China
[d] Yunnan Jiaofa Consulting Co.Ltd., Kunming 650100, China
[e] Yunnan Jiantou Boxin Engineering Construction Center Test Co., Ltd., Kunming 650217, China

ARTICLE INFO

Keywords:
Pavement defect detection
3D laser profiling technology
Convolutional neural networks
Attention mechanism

ABSTRACT

This paper introduces a method for detecting pavement defects based on convolutional neural networks. First, grey and depth image data were acquired using a 3D pavement information collection system, followed by pre-processing and labelling of the data. Subsequently, two network structures were developed to accommodate the image data characteristics: classic U-shaped and double-headed structures. Attention modules were integrated into the models to enhance the accuracy of defect detection. Finally, a quantitative analysis of four types of pavement defects was conducted. The numerical evaluation results demonstrated that training the network with a combination of grey and depth images significantly improves the detection accuracy, resulting in a 10% enhancement in mean intersection over union (MIoU). The proposed model attained a global pixel accuracy (GPA) of 97.36% and an MIoU of 80.28%. The proposed network model was found to have an increased focus on the pavement defect areas, making it highly effective.

1. Introduction

Various defects can occur in pavements due to factors such as the physical environment, load conditions, structural configurations, materials, building methods, technical proficiency, and external impacts. The construction of extensive road networks in recent years has led to increased maintenance requirements. Therefore, accurate, efficient, and safe testing and assessments have become essential for initiating maintenance work. Over time, detection methods have evolved from manual to fully automatic processes. Digital image processing (DIP) technology is extensively employed for inspecting pavement defects. Commonly used equipment includes RGB cameras, infrared (IR) thermal imagers, and 3D laser scanners. In recent studies, digital cameras have been utilized for analysing pavement surface texture [1,2] and crack detection [3–5]. These cameras offer the advantages of low cost, ease of installation, and user-friendly operation. However, they are susceptible to issues such as bleeding, shadows, and uneven lighting, which can affect their effectiveness in complex detection environments [6,7]. Consequently, RGB image-based methods may not be suitable for such scenarios. IR thermal technology, on the other hand, is unaffected by light intensity and is useful for rapidly identifying cracks. It has been increasingly employed for crack detection and evaluation [8–10]. However, there is a need for improved resolution in quantitative pavement distress studies.

Ideally, 2D image camera measurement techniques should be employed for pavement distress detection and assessment. However, it is necessary to address environmental disturbances and provide accurate 3D information about pavements to enhance intelligent pavement distress management. 3D laser profiling technology fulfils the requirements of modern pavement inspection technology by measuring elevation changes on road surfaces. Laser scanning, based on trigonometric triangulation, is a well-established imaging technology for accurate 3D road data acquisition [11]. James et al. [12] employed a 3D laser scanner to measure cracks in asphalt pavements, demonstrating that the technique was unaffected by variations in brightness and contrast. Guan et al. [13] proposed a detection framework for automatic extraction of cracks from high-density point clouds collected using a mobile laser-scanning system. Zhou and Song [14] employed laser-

* Corresponding author at: Faculty of Civil Engineering and Mechanics, Kunming University of Science and Technology, Kunming 650500, China.
E-mail address: haiting.xia@kust.edu.cn (H. Xia).

https://doi.org/10.1016/j.autcon.2023.105192
Received 6 June 2023; Received in revised form 3 November 2023; Accepted 7 November 2023
Available online 11 November 2023
0926-5805/© 2023 Elsevier B.V. All rights reserved.

scanned range images to classify roadway cracks. Li et al. [15] utilized high-speed 3D transverse scanning technology for road shoving and pothole detection. Zhang et al. [16] evaluated the skid resistance of asphalt pavements using a 3D dense-point scanning technique. Three-dimensional (3D) laser imaging technology has emerged as the primary approach for collecting data from pavements [17,18].

Image segmentation is performed to identify distress once pavement data has been collected. Common methods for image segmentation include threshold-based segmentation, multi-scale segmentation, and deep learning-based methods. Lu et al. [19] proposed a double-threshold technique for precise crack number and width detection. Peng et al. [20] introduced a triple-threshold pavement crack-detection system based on random structured forest. To overcome non-uniform illumination noise, Zhou et al. [21] employed mean and standard deviation values as threshold parameters. Koch and Brilakis [22] used a histogram-based threshold segmentation method to identify potholes in the road. Additionally, multi-scale segmentation methods based on spatial filtering and wavelet analysis have also been explored [23–26]. In recent years, research in pavement distress detection has shifted towards deep learning-based techniques, with researchers increasingly applying deep learning methods for image processing. Augustauskas et al. [27] proposed a U-Net-based approach for pixel-level crack segmentation in pavement quality assessment. Ji et al. [28] employed Deeplabv3+ and a pixel-level quantization technique for crack identification. Liu et al. [29] developed the Deepcrack network, which incorporates deep supervision and facilitates crack segmentation in various scenarios. Pereira et al. [30] trained a U-Net model to segment colour road images for pothole detection. Tong et al. [31] proposed a framework that integrates full convolutional networks and multiple methods for pavement defect detection. However, previous studies have mainly focused on processing 2D images using DIP or directly utilizing 3D information for recognition. Hence, there is a research gap in the field of pavement inspection regarding the integration of 2D and 3D data and whether these two types of data can complement each other in the identification process. Furthermore, most studies have primarily concentrated on single pavement distress, with limited research on simultaneous extraction of multiple types of distress.

To address this issue, pavement information extraction studies have utilized grey images containing 2D information and depth images containing 3D information. In this research, we constructed datasets of four defects for pavement defect detection: potholes, cracks, bleeding, and patches. Two network structures based on U-shaped convolutional neural networks were developed. The impact of different input data on the model training process was examined, and a comparative study of the proposed methods was conducted. Additionally, a quantitative analysis was performed for different types of defects.

The remaining sections of this paper are structured as follows: In the second section, an overview of the pavement data collection method, attention mechanism module, network model structure, and dataset generation are provided. The training details and the model training process are elaborated in the third section. This section also presents a comparison of identification results obtained from different methods and provides quantitative results for pavement distress. The fourth section discusses the advantages of the proposed method and demonstrates its superiority through spatial heatmap confusion matrices. In the final section, we briefly summarize the study for detecting multiple types of pavement distresses and present the conclusions drawn.

2. Methodology

The primary objective of this study is to develop a convolutional neural network (CNN) model capable of segmenting multiple pavement detection objects. First, pavement images were captured using a CCD camera, and the pavement profile was measured using a 3D pavement data collection system. The elevation data were obtained and calibrated to match the grey images. Subsequently, semantic segmentation models with different structures were constructed based on CNNs. Finally, datasets were generated for four types of pavement defects.

2.1. Acquisition of pavement image dataset

Two types of image data were obtained during the pavement inspection process: grey images of the road surface and depth images representing surface elevation information. The grey images were captured using two-line scan CCD cameras with an effective detection width of 3.75 m or more. The 3D pavement data acquisition system had a longitudinal sampling interval of 1 mm and a recognition accuracy of 1 mm or higher. The pavement detection speed ranged from 0 to 100 km/h, and the sampling interval was adjusted based on the direction of road travel. The system performed data pre-processing and storage functions. Since the camera provided the relative position of the laser line in the image rather than the actual elevation of the measured profile, the depth image data of the road surface were obtained through conversion and calibration of the image and object coordinates.

Fig. 1 shows the acquired pavement image data, which contains abundant pavement information. The locations of potholes, cracks, bleeding, and patches on the pavement can be visually identified. Fig. 2 presents an example of a pavement profile in a depth image. These images can be adapted to various complex pavement environments. The laser scanning 3D pavement system, with its high precision, frequency, and dynamic features, enables effective direct inspection and maintenance by providing accurate 3D data.

Both grey and depth images of the road surface encompass a wealth of information about the scene, and our objective is to perform segmentation of distinct objects within these images. We specifically focused on evaluating potholes, cracks, bleeding, and patches, considering their relevance to pavement characteristics. In terms of pixel distribution, the grey image exhibits higher gradients in cracks, bleeding, and patches, whereas the depth image displays more pronounced gradients in potholes. To enhance recognition accuracy, both images are utilized in the segmentation task.

2.2. Attention mechanism

The collected data comprised grey and depth images, and the detection task involved categorizing four distinct categories. To enhance feature extraction and prediction by leveraging contextual information in both the spatial and channel dimensions, the attention mechanism was inserted to model dependencies between pixels at different locations within an image. The image data exhibit block-like distributions for potholes, patches, and bleeding, while cracks are distributed as thin strips. In the spatial dimensions of an image, the model may overlook fine stripes of information while focusing on larger blocks. Moreover, in the channel dimension, both grey and depth images were fed into the CNN for training, and the convolutional operation performed feature extraction on all images. As a result, the importance of grey and depth images cannot be determined during the training process. To address this issue, two types of attention modules that have demonstrated effectiveness in the field of computer vision were introduced to the network model: the Squeeze-and-Excitation Network (SE-Net) module [32] and Convolutional Block Attention Module (CBAM) [33]. In this module, the spatial attention mechanism allowed the network to focus on the locations of defects in the image space. In contrast, the channel attention mechanism allows the network to choose whether to focus more on grey or depth images.

2.2.1. Squeeze-and-excitation networks module

SE-Net introduces an attention mechanism to the channel dimensions, enabling the neural network to learn the significance of each feature channel. In this operation, the current channel is multiplied by a weight value specific to it, allowing the network to concentrate on particular feature channels. This amplifies the feature channels that are

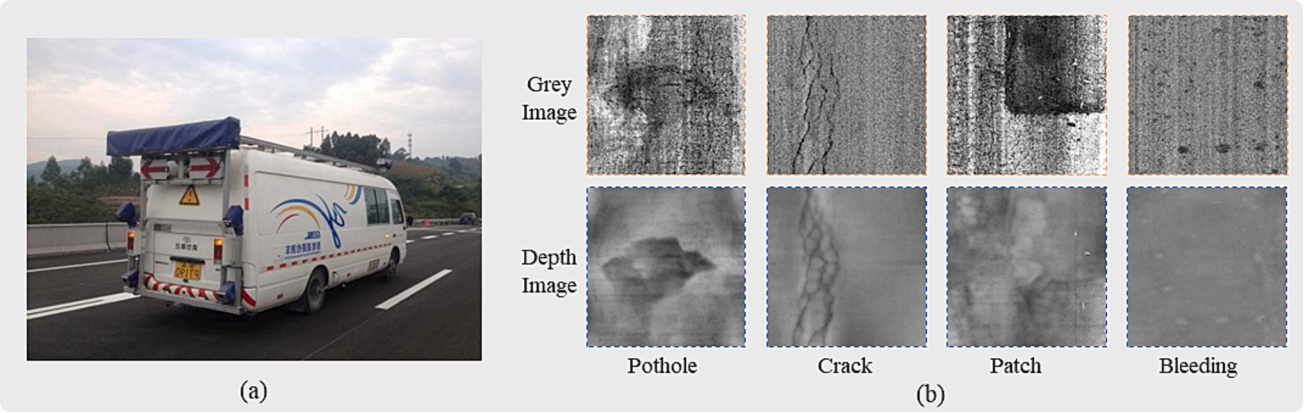

Fig. 1. Process of collecting pavement data. (a) detection vehicle with a 3D pavement data collection system, (b) collected grey and depth images.

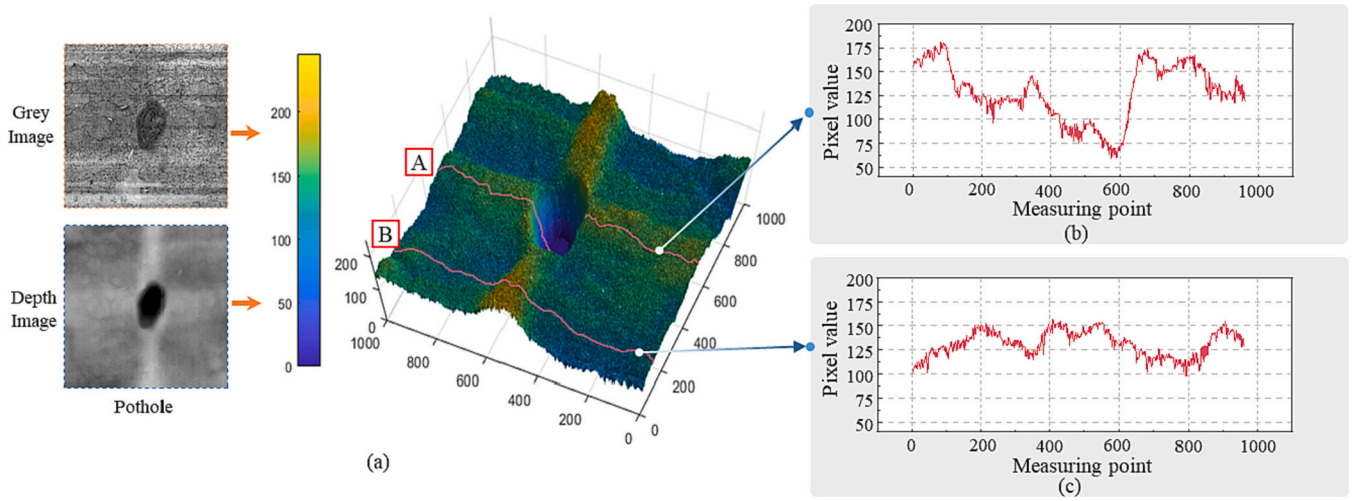

Fig. 2. 3D pavement data with a pothole and some profile examples. (a) 3D pavement data, (b) and (c) are pavement profile examples.

beneficial for the given task, while suppressing those that are irrelevant. The structure detail of the SE-Net module is shown in Fig. 3. $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ is the feature map set entered into the module. This can be expressed as $\mathbf{X} = [x_1, x_2, \ldots, x_C]$. First, features $\mathbf{X}$ are passed through a squeeze operation, $\mathbf{F}_{sq}$, which aggregates the feature maps across their spatial dimensions, and $\mathbf{m} \in \mathbb{R}^{1\times 1\times C}$ is obtained. It is calculated by Eq. (1):

$$m_C = \mathbf{F}_{sq}(x_C) = \frac{1}{H\times W}\sum_{i}^{H}\sum_{j}^{W} x_C(i,j) \tag{1}$$

Subsequently, excitation operation, $\mathbf{F}_{ex}$, is implemented to fully capture channel-wise dependencies, and $\mathbf{n} \in \mathbb{R}^{1\times 1\times C}$ is obtained. It is calculated by Eq. (2):

$$\mathbf{n} = \mathbf{F}_{ex}(\mathbf{m}) = \sigma(\mathbf{F}_2(\delta(\mathbf{F}_1(\mathbf{m})))), \tag{2}$$

where $\mathbf{F}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $\mathbf{F}_2 \in \mathbb{R}^{\frac{C}{r}\times C}$ are fully connected layers, r is 16; $\delta(\bullet)$ is the ReLu function; $\sigma(\bullet)$ is the Sigmoid function. The module outputs a feature map set $\mathbf{X}^* = [x_1^*, x_2^*, \ldots, x_C^*]$, as defined by Eq. (3):

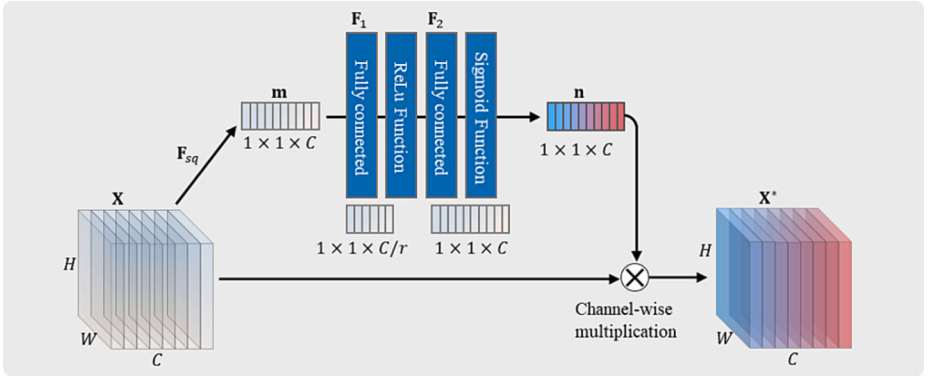

Fig. 3. Squeeze-and-excitation networks (SE-Net) module composition and structure.

$$x_C^* = n_C x_C \tag{3}$$

2.2.2. Convolutional block attention module

CBAM not only learns the significance of each channel in the channel dimension but also dynamically selects the regions to focus on the spatial dimension. The Channel Attention Module (CAM) and the Spatial Attention Module (SAM) constitute CBAM. The structure detail is depicted in Fig. 4.

The CAM compresses the 1D channel attention vector $\mathbf{W}^C \in \mathbb{R}^{1\times1\times C}$ along the spatial dimension to infer the relevance of each channel for arbitrary intermediate feature map $\mathbf{X}$. Specific details were generated using average and maximum pooling operations to generate two different spatial context descriptions, representing the average and maximum pooling features, respectively. The two feature maps are fed to a shared multilayer perceptron (MLP) to generate two-channel attention maps, and the two attention maps were merged using an element-wise summation. Finally, the maps passed through a sigmoid function, and we obtained the channel attention vector $\mathbf{W}^C$. Channel-wise multiplication is used to construct the channel feature map $\mathbf{X}' \in \mathbb{R}^{H\times W\times C}$, and the intermediate result is expressed as $\mathbf{X}' = [x_1', x_2', \ldots, x_C']$. The CAM was computed using Eqs. (4–7):

$$\mathbf{F}_{avg}^C = AvgPool^C(\mathbf{X}) \tag{4}$$

$$\mathbf{F}_{max}^C = MaxPool^C(\mathbf{X}) \tag{5}$$

$$\mathbf{W}^C = \sigma\left(\mathbf{F}_2\left(\delta\left(\mathbf{F}_1\left(\mathbf{F}_{avg}^C\right)\right)\right) + \mathbf{F}_2\left(\delta\left(\mathbf{F}_1\left(\mathbf{F}_{max}^C\right)\right)\right)\right) \tag{6}$$

$$x_C' = \mathrm{w}_C^C x_C \tag{7}$$

where the $\mathbf{F}_{avg}^C \in \mathbb{R}^{1\times1\times C}$ and $\mathbf{F}_{max}^C \in \mathbb{R}^{1\times1\times C}$ are the results of the global average pooling and global max pooling operations, respectively. Fig. 4 (b) shows the CAM structure.

SAM outputs spatial attention weights by utilizing the interspatial relationship of features. Average pooling and max pooling operations are used along the channel axis, and the two results are concatenated to generate efficient feature maps $\left[\mathbf{F}_{avg}^S, \mathbf{F}_{max}^S\right]$. Then, convolution and sig-

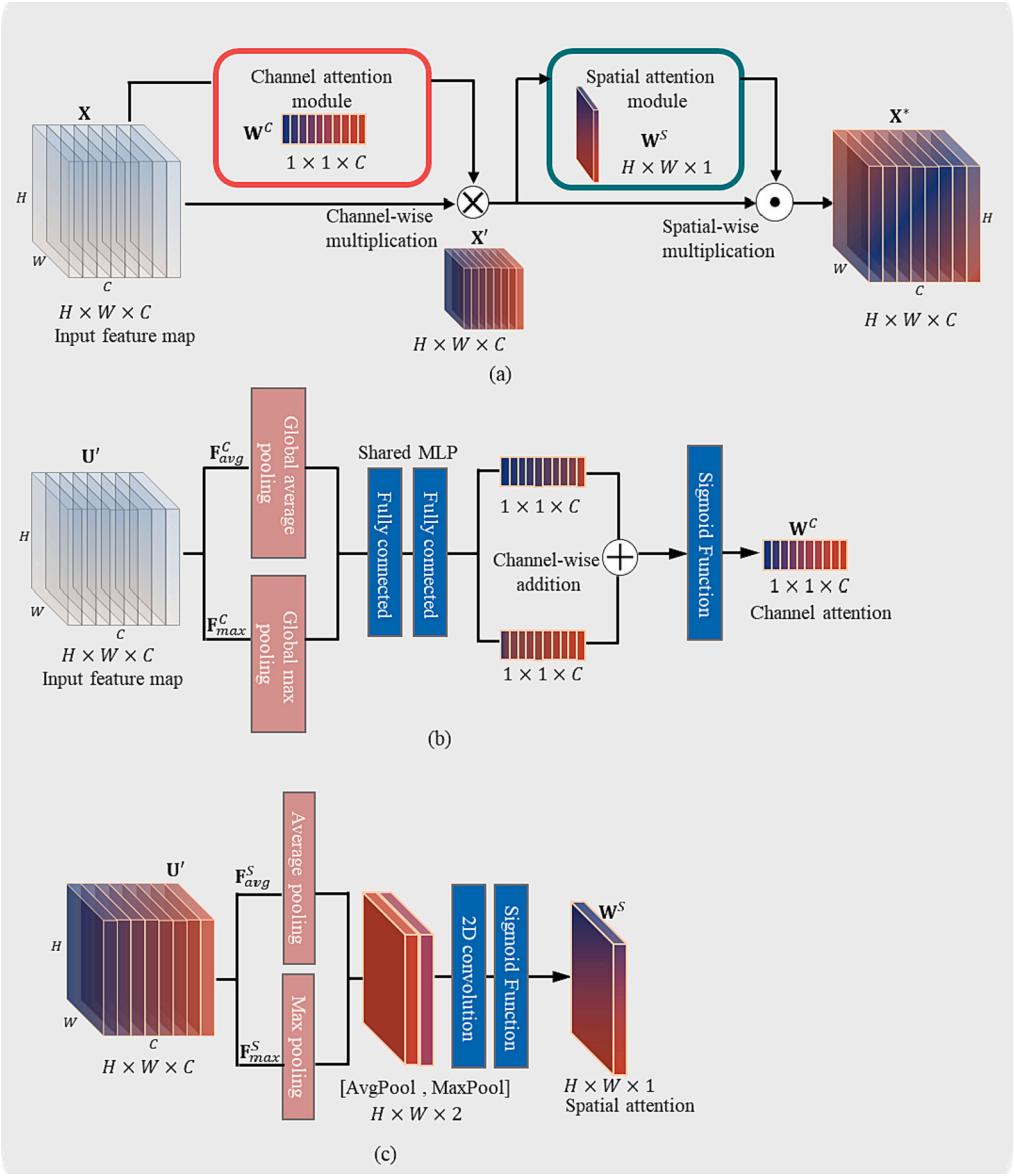

Fig. 4. Structure of the convolutional block attention module (CBAM). (a) the overview figure, (b) the diagram of the channel attention module(CAM), and (c) the diagram of the spatial attention module(SAM).

moid operations are used to produce the spatial attention weight $W^S \in \mathbb{R}^{H\times W\times 1}$. Finally, the output feature maps $\mathbf{X}^*$ are calculated using the Hadamard product. The SAM was computed using Eqs. (8–11):

$$\mathbf{F}^S_{avg} = AvgPool^S(\mathbf{X}), \tag{8}$$

$$\mathbf{F}^S_{max} = MaxPool^S(\mathbf{X}), \tag{9}$$

$$\mathbf{W}^S = \sigma\left(Conv\_1d_7\left(\left[\mathbf{F}^S_{avg}, \mathbf{F}^S_{max}\right]\right)\right), \tag{10}$$

$$x^*_C = x'_C \odot \mathbf{W}^S \tag{11}$$

where $Conv\_1d_7$ denotes a convolutional procedure with a 7×7 kernel size. The $\mathbf{F}^S_{avg} \in \mathbb{R}^{1\times 1\times C}$ and $\mathbf{F}^S_{max} \in \mathbb{R}^{1\times 1\times C}$ represent the results of the average and maximum pooling operations along the channel axis, respectively. Fig. 4(c) shows the structural details of the SAM.

2.3. Overall network model structure

U-Net has exhibited excellent performance in various semantic segmentation tasks [34–37], including pavement detection tasks. It serves as the fundamental structure for extracting pavement information. The network model comprises two components: an encoder and a decoder. The encoder conducts feature extraction, while the decoder takes the extracted feature map, refines it through further learning, and ultimately generates the segmented image.

Initially designed for cell segmentation [38], the U-Net model employed a single-channel input image. Subsequently, it has been adapted for other RGB image segmentation tasks. Fig. 5 illustrates two strategies employed to construct the model based on the U-Net architecture. The first model type adopts the conventional approach, wherein grey and depth images are stacked along the channel dimension and fed into the network for training, as depicted in Fig. 5(a). Due to the significant variations in object morphology and grey features between the grey and depth images, the second model type employs two separate encoders, each performing feature extraction separately on the grey and depth images. The resulting sets of feature maps are stacked, fused, and then passed to the decoder for learning and segmentation, as shown in Fig. 5(b).

Convolution and pooling operations are essential in CNN. The convolution operation extracts the morphological and colour features of the object. The pooling operation reduces the image resolution, facilitating the model to identify more advanced semantic information while reducing the training parameters. The sequence of convolution and pooling operations is referred to as the feature processing stream.

Only a single processing stream exists for the first model type, where the network can decide how to process the image pairs to extract the disease information, as shown in Fig. 5(a). Another model type is to create two separate but identical processing streams for the two images and combine them in the post stage, as shown in Fig. 5(b). With this dual-input header design, the network first needs to generate meaningful feature maps for each of the two images and then combine them at a higher level. However, simply stacking a given feature map set of two images is not enough. Attentional mechanisms are introduced to assign weights to different feature maps during training. Thus, the feature maps of the image pairs are selectively combined by assigning different levels of importance to them.

Incorporating attention mechanisms into the network structure was an additional enhancement to the model. Following every two convolution operations, the feature maps underwent a CBAM module. This module directed attention to spatial regions of interest and acquired the weight distribution along the channel dimension for these feature map sets. Furthermore, the feature map sets were passed through the SE module via skip-layer connections. This module determined the significance of each feature map in the channel dimension. By assigning distinct weights to the grey and depth images, the model aimed to enhance recognition accuracy. Ablation experiments on the choice of attentional mechanisms are described in Section 3.3

Lastly, the eight types of methods that need to be examined are listed in Table 1, where Segnet [39] and Deeplabv3 [40] are used as bench marked models.

The objective of pavement distress recognition is to accurately segment objects in the pavement from the background. In the annotated image, the pixel value of the object area is assigned as 1, while the pixel

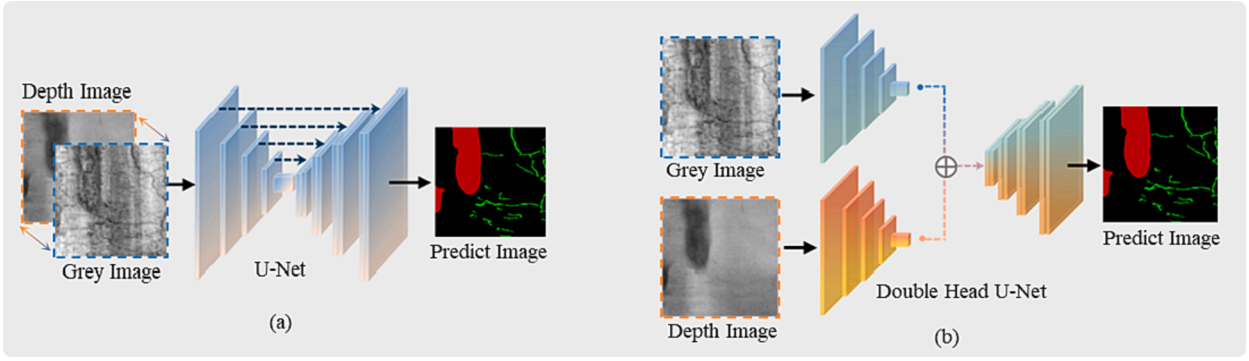

Fig. 5. Diagram of the overall network model structure. (a) the classic U-shaped network structure, (b) network structure with dual encoders.

Table 1
Description of the six models to be trained.

Model name	Description	Dataset for training
U-Net Grey	Ordinary U-shaped network structure.	Grey images
U-Net Depth	Ordinary U-shaped network structure.	Depth images
SegNet [39]	Fully convolutional network.	Grey and depth images
Deeplab v3 [40]	Semantic segmentation model with atrous separable convolution.	Grey and depth images
U-Net Two Channel	Ordinary U-shaped network structure. Abbreviated as U-Net Two.	Grey and depth images
U-Net Two Channel + Attention	Adding attention mechanism to U-Net Two. Abbreviated in the chart as U-Net Two + Attention.	Grey and depth images
Double Head U-Net	U-shaped network structure with dual encoders. Abbreviated in the chart as DHU-Net.	Grey and depth images
Double Head U-Net + Attention	Adding attention mechanism to DHU-Net. Abbreviated in the chart as DHU-Net + Attention.	Grey and depth images

value of the background is set as 0. To create labels, one-hot coding was employed. In the annotated image, each category of objects is represented by a separate channel. The final layer of the network structure utilizes the softmax function layer to generate a categorical probability distribution. This function ensures that the sum of all outputs equals one, as defined in Eq. (12). The categorical cross-entropy function was chosen as the loss function, as depicted in Eq. (13).

$$Softmax(x)_i = f(x)_i = \frac{e^{x_i}}{\sum_j^N e^{x_j}} \tag{12}$$

$$CE(x) = -\sum_{i=1}^{N} y_i log x_i \tag{13}$$

where $f(\bullet)_i$ is the calculation result of the activation function, which depends on all elements x. x_i is the score inferred by the network for each class i in N. y_i is the actual true score value.

2.4. Image pre-processing and annotation

Various types of pavement defects arise due to the combined effects of driving loads and the natural environment. The identification of these intricate data elements poses challenges. A total of two thousand grey and depth images, with a pixel resolution of 960 × 1000 pixels, were collected using a road-detection vehicle. Subsequently, the original images were cropped to a resolution of 480 × 500 pixels. To accommodate the input port of the network model, the images were further resized to 256 × 256 pixels.

For semantic labelling, the open-source tool LabelMe [41] was employed. The labelled information was saved as a JSON file, encompassing details such as the labelled image's name, distress type, marked point location, and other relevant information. From this file, the information was extracted to generate a binary image for training purposes. The entire dataset creation process is illustrated in Fig. 6.

3. Implementation details and experiments results

After constructing the network structure and creating the dataset, the network model was trained using the Python API. Semantic segmentation evaluation metrics were employed throughout the training process to monitor the results obtained from the predictions. After the completion of training, these results were evaluated and compared. Finally, distinct quantification methods were examined for various distress characteristics.

3.1. Training details and evaluation metrics

The training platform utilized a workstation equipped with an NVIDIA 3060, 12G GPU, and an Intel(R) Xeon(R) Platinum 8255C CPU. The network was constructed and trained using TensorFlow, Google's open-source deep learning framework. The software configuration comprised Windows 10, CUDA 10.1, cuDNN-v7.6, TensorFlow-GPU-2.2, and Python 3.7.

Of all the collected pavement images, a total of 200 distress image pairs were selected for this study. A grey image and a depth image form a set of image pairs. Through x-axis flip transformation, y-axis flip transformation, and random segmented affine transformation, the number of pairs was expanded to 1080. The dataset has 1080 grey images and 1080 depth images. The conventional approach involves dividing the dataset into training, validation, and test sets in a 7:2:1 ratio. Thus training set has 756 image pairs, the validation set has 216 image pairs, and the test set has 108 image pairs. However, with a small dataset, the evaluation metrics calculated from the validation set demonstrated a high correlation with the original grouping. This simplistic method of data division does not provide an accurate understanding of the model's performance.

To address this, a k-fold validation method is introduced. It is known as one of the most reliable and non-exhaustive validation techniques for estimating unbiased models [42]. The k-fold validation splits the dataset into K sets of size M/K, where M represents the total number of samples. $K+1$ sets were utilized for training the network, while the remaining set was used for validation. The model was independently repeated K times. The model accuracy was determined by averaging the accuracies obtained from the K training iterations, and the most accurately trained model was employed for prediction. The k-fold validation method also divides the dataset in the ratio of 7:2:1. Three commonly used semantic segmentation metrics were employed to evaluate the output: global pixel accuracy (GPA), mean pixel accuracy (MPA), and mean intersection over union (MIoU). MIoU was also employed to monitor the entire training process. GPA calculates the proportion of accurately predicted pixels among total pixels. MPA calculates the pixel accuracy for each category and then computes the average pixel accuracy across all categories. MIoU is the average of the IoU values for each category on the predicted image. The definitions of each metric are provided in Eqs. (14–16).

$$GPA = \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{i=1}^{N} N_{ii}}{\sum_{i=1}^{N}\sum_{j=1}^{N} N_{ij}}, \tag{14}$$

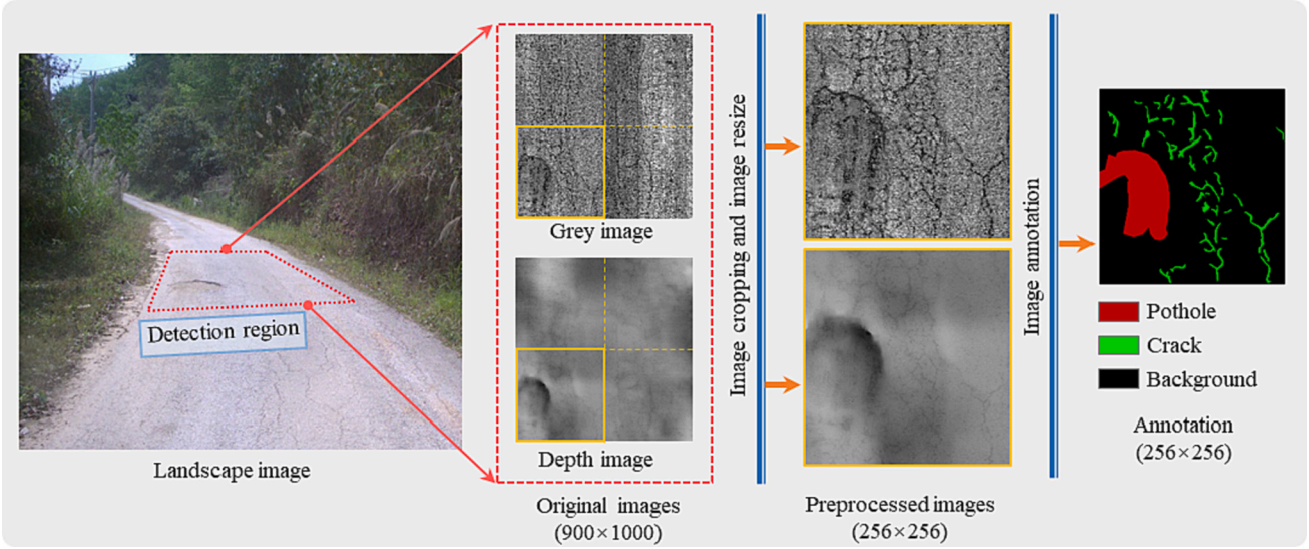

Fig. 6. Schematic of image pre-processing and annotation.

$$MPA = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{c+1}\sum_{i=0}^{c}\frac{N_{ii}}{\sum_{j=0}^{c}N_{ij}}\right), \tag{15}$$

$$MIoU = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{c+1}\sum_{i=0}^{c}\frac{N_{ii}}{\sum_{j=0}^{c}N_{ij}+\sum_{j=0}^{c}N_{ji}-N_{ii}}\right), \tag{16}$$

where $c+1$ is the number of categories to be classified and takes the value of five (categories contain different distress areas and background areas). N_{ii} is the number of correctly predicted pixels. N_{ij} is the number of pixels in which the image of category i is predicted to be in category j, and N is the total number of pixels. K is the number of folds divided in the k-fold validation, taking a value of 10 for K.

The optimiser played a crucial role in adjusting the learning rate throughout the training process, aiming to approach the optimal parameters for the model's trainable parameters. This dynamic adjustment optimised the learning process. The adaptive moment estimation (Adam) optimizer [43] was used. The hyperparameters employed in the training process are detailed in Table 2. During the training process, the accuracy is evaluated once on the validation set for each completed epoch. At the conclusion of training, three distinct metrics were utilized to evaluate both the validation set results and the prediction outcomes for the test set.

3.2. Training process

After setting the training parameters according to the contents in Table 2, the proposed model starts to be trained. The training history files are saved to output training curves and observe the training process. The training curves of the optimal weight model were selected, as depicted in Fig. 7(a) and (c), showcasing the changes in loss value and accuracy for the training set, respectively. Similarly, Fig. 7(b) and (d) illustrate the corresponding curves for the validation set. Fig. 7(a) and (c) show that all models exhibited gradual convergence during training, with the DHU-Net + Attention model showing the fastest convergence. As depicted in Fig. 7(b, d), the inclusion of the attention mechanism resulted in improved detection efficiency for the original model. Furthermore, DHU-Net + Attention achieved the highest accuracy and lowest loss values in the validation set. If there are large fluctuations in the loss curve, it indicates that the training process is unstable, then the robustness of the model is also insufficient. The curve of DHU-Net + Attention remained smooth without significant fluctuations. However, it is important to note that the training process alone does not provide a definitive measure of effectiveness. In Section 3.3, we present the model's prediction results, offering a more comprehensive evaluation.

3.3. Prediction results and comparison

Tensorflow was designed with an API for saving model weights. The model weight is saved once per training epoch throughout the training process. After conducting K-fold validation and determining the best model, the optimal weights of the model were extracted and loaded into the API. The final layer of the network produced a 3D matrix consisting of five channels, each representing one of the identified categories. The segmentation results were obtained by selecting the maximum value from the matrix along the channel dimensions.

Table 2
Hyperparameters used in training process.

	Parameter	Value
Network related	Convolution kernel size	3
	Dropout (rate)	Yes (0.05)
Number of filters	The 1st conv2d block	64
	The 2nd conv2d block	128
	The 3rd conv2d block	256
	The 4th conv2d block	512
	The 5th conv2d block	1024
Training process related	Batch size	4
	Initial learning rate	1×10^{-4}
	epochs	300
Adam algorithm related	Gradient decay factor	0.9
	Squared gradient decay factor	0.99
	Epsilon	1×10^{-7}

The two attentional modules are described above. The SE-Net and CBAM attention mechanisms are added to the U-Net Two and DH U-Net described in Table 1. The results of the ablation experiments for the modules are summarised in Section 3.3.1. The ablation experiments will compare the effect of different modules on the model. In addition, the proposed methods are compared with other models in section 3.3.2.

3.3.1. The ablation study of attentional mechanisms

The ablation experiments were conducted to test the effect of different attention modules on recognition accuracy. The two attentional mechanisms, SE-Net and CBAM, were inserted separately into the basic model for training. The quantitative results of the ablation experiments are given in Table 3. The table shows that both U-Net and DHU-Net have improved their detection accuracy by adding the attention module. For the choice of Attention Module, the boosting effect is about the same when inserted alone or together in U-Net. However, the improvement was significant when both modules were inserted into the DHU-Net model. The MIoU was improved by 3.58% over the initial model. Therefore, in the comparative study in the following section, U-Net and DHU-Net with two attentional mechanisms are chosen for comparison with other models.

3.3.2. Results of the comparative study

The eight methods described in Section 2.3 were compared to assess the proposed network's performance. Table 4 presents the results of these models based on the three evaluation metrics. The recognition accuracy of U-Net Grey and U-Net Depth was relatively low. However, the recognition accuracy significantly improved when both types of images were jointly used for training. The accuracy of all models is improved after data augmentation. The accuracy of the proposed method is higher than that of the two benchmarked models. Notably, when trained with the two-headed structural model, the recognition accuracy further improved. The impact of the attention mechanism in the network is evident as both the traditional U-Net and DHU-Net achieved enhanced accuracy. Consequently, the proposed method is reliable, and the DHU-Net + Attention approach demonstrates optimal recognition performance.

Figs. 8 (a)–(d) illustrate the segmentation results. The results of U-Net Grey and U-Net Depth exhibit incomplete segmentation, indicating the limitations of these models. Although U-Net Two demonstrates good recognition performance, it still misses certain detections. DHU-Net performs well overall, but it exhibits discontinuities when segmenting cracks. The incorporation of the attention mechanism addresses the shortcomings of the original model and reduces the leakage rate. Among all defect categories, the DHU-Net + Attention approach achieves the most accurate segmentation results, closely resembling the Ground Truth. In the presence of water-stain interference, as depicted in Fig. 8 (e) and (f), the DHU-Net + Attention method demonstrates excellent anti-interference capability. Conversely, the other models exhibit missed and false detections under such interference conditions. Therefore, the adoption of a two-headed structure and the inclusion of an attention mechanism effectively improve the recognition accuracy, surpassing the single-input design of the traditional U-Net structure.

The *P*-values were used to determine the results of hypothesis testing to demonstrate that DHU-Net + Attention can improve recognition accuracy. The proposed model is retrained thirty times by using the training parameters and methods mentioned in Sections 3.1 and 3.2. The optimal model weights are saved at each training session's end and evaluated on the test data set. The thirty evaluation results are shown in

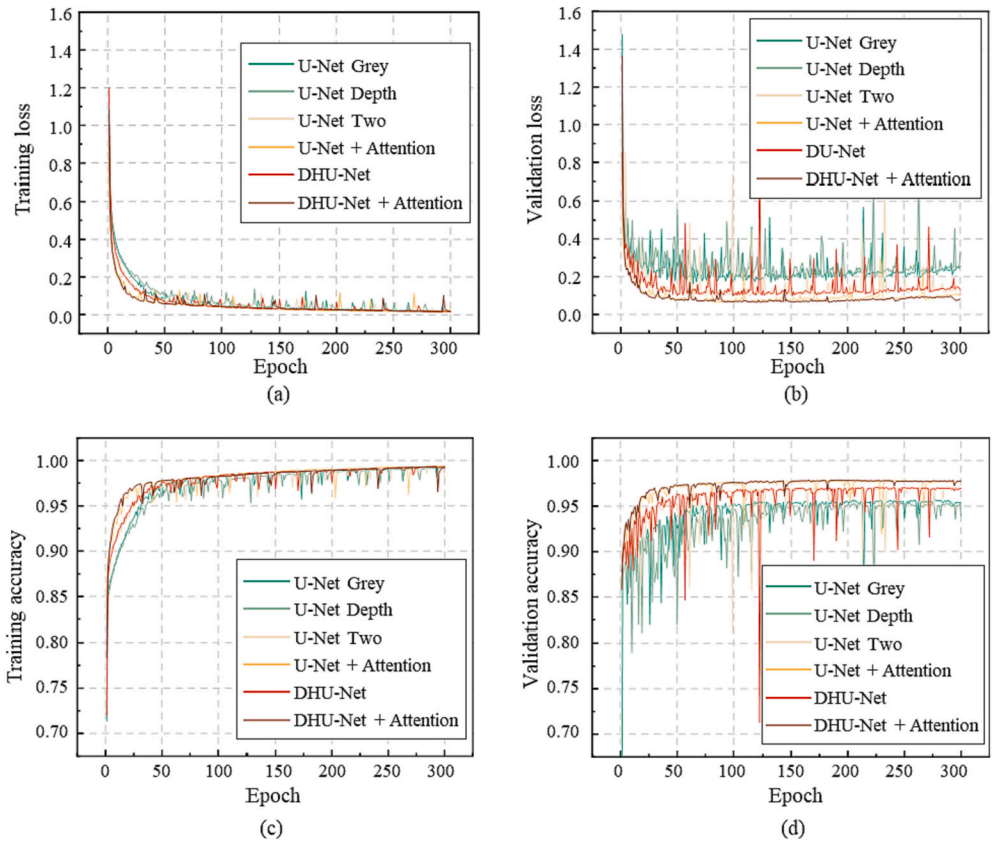

Fig. 7. Training process curve. (a) loss curve under the training set, (b) loss curve under the validation set, (c) accuracy curve under the training set, (d) accuracy curve under the validation set.

Table 3
Effect of attention module on the performance of base models.

Model	Attention module		Metrics		
	SE-Net	CBAM	GPA(%)	MPA(%)	MIoU(%)
U-Net Two	–	–	96.52	82.94	76.28
U-Net Two	√	–	97.03	84.80	77.98
U-Net Two	–	√	97.08	85.29	78.77
U-Net Two	√	√	96.34	84.35	76.61
DHU-Net	–	–	96.55	83.99	76.70
DHU-Net	√	–	97.31	85.96	79.63
DHU-Net	–	√	96.90	84.97	78.33
DHU-Net	√	√	**97.36**	**86.44**	**80.28**

Table 4
Numerical results for the evaluated models.

Models	Data aug.	GPA(%)	MPA(%)	MIoU(%)
U-Net Grey	√	94.16	74.83	66.96
U-Net Depth	√	94.20	71.44	63.51
Segnet	–	94.46	72.99	64.30
Segnet	√	95.14	78.14	69.20
Deeplabv3	–	94.18	73.82	64.05
Deeplabv3	√	96.28	78.58	72.77
U-Net Two	–	95.78	81.11	72.84
U-Net Two	√	96.52	82.94	76.28
U-Net Two + Attention	√	96.34	84.35	76.61
DHU-Net	√	96.55	83.99	76.70
DHU-Net + Attention	√	**97.36**	**86.44**	**80.28**

Table 5. MIoU is used as the primary assessment metric, and models are generally considered excellent when MIoU is >80%. There is a null hypothesis H_0: $\mu \leq 0.8$ and alternative hypothesis H_1: $\mu > 0.8$, α set as 0.05. The P-value was calculated to be equal to 1.87354×10^{-4}, which was calculated from the data in the table. Since the P-value <0.05, the null hypothesis is not valid. Finally, it is concluded that MIoU is significantly >80%, and the proposed model is effective.

3.4. Methods of quantifying pavement defects

Once the image accurately segments the distressed area, the measurement of distress features becomes possible. In the acquired image data, four types of information were measured: cracks, potholes, patches, and bleeding. The cracks' widths and lengths were determined through specific methods. The structural thickness measurements utilized the 3D width detection method [44], which defines the structural thickness as the diameter of the largest sphere that can fit inside the object and contain that point. The method proposed a definition of local structure thickness, as shown in Fig. 9(a). Let $\Omega \subset R^3$ be the set of all points within the geometry, and $\vec{p} \in \Omega$ be an arbitrary point of them. The local thickness $\tau(\vec{p})$ was defined as the diameter of the largest sphere, which contains the point $\vec{p}$. The diameter of the largest sphere that fits inside the object and contains the point. Where $sph(\vec{x}, r)$ is the set of points inside a sphere with center $\vec{x}$ and radius r.

$$\tau(\vec{p}) = 2 \cdot max(\{r | \vec{p} \in sph(\vec{x}, r) \subseteq \Omega, \vec{x} \in \Omega\}). \quad (17)$$

To measure crack width, ImageJ software was employed, as depicted in Fig. 10(a).

The skeleton extraction algorithm uses a fast parallel algorithm [45], The skeleton line has only a single pixel width. Inspired by Yang [46] and Ji [47], The following Eq. (18) can calculate the crack length L:

$$L = \int_C f(x,y)dl \cong \sum f(x,y)dl, \quad (18)$$

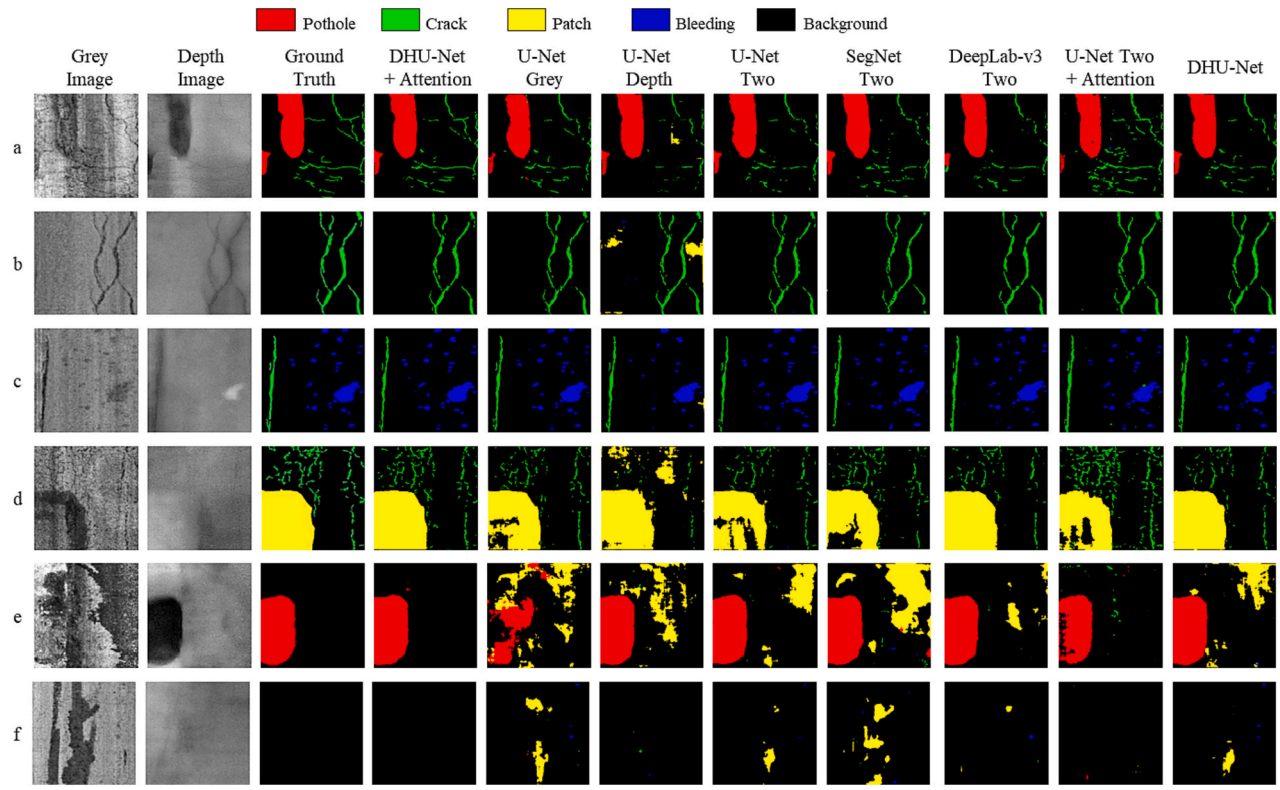

Fig. 8. Visual prediction results and comparison of the models.

Table 5
The assessment results of 30 model weights.

Number	GPA (%)	MPA (%)	MIoU (%)	Number	GPA (%)	MPA (%)	MIoU (%)
# 1	97.28	85.93	80.06	# 16	97.27	86.52	80.03
# 2	97.35	87.02	80.19	# 17	97.27	86.27	79.97
# 3	97.35	86.01	80.11	# 18	97.28	86.17	79.99
# 4	97.29	87.05	80.40	# 19	97.58	86.32	80.73
# 5	97.15	85.89	78.82	# 20	97.41	86.94	80.65
# 6	97.43	86.52	80.52	# 21	97.32	86.26	80.20
# 7	97.41	87.25	80.42	# 22	97.47	86.92	80.77
# 8	97.30	86.36	80.13	# 23	97.42	86.17	80.48
# 9	97.41	86.72	80.40	# 24	97.39	86.38	80.38
# 10	97.45	85.63	80.33	# 25	97.42	86.06	80.28
# 11	97.40	85.73	80.28	# 26	97.40	86.70	80.52
# 12	97.46	85.74	80.37	# 27	97.42	87.15	80.53
# 13	97.56	86.15	80.68	# 28	97.29	87.05	80.40
# 14	97.37	86.28	80.28	# 29	97.36	86.84	80.42
# 15	97.16	86.97	79.50	# 30	97.51	86.57	80.63

where $f(x,y)$ is the geometric calibration index and is simplified to be one. dl represents the finite length of skeleton lines. $f(x,y)dl$ is realised by calculating the equation $\sqrt{(x_{i+1}-x_i)^2+(y_{i+1}-y_i)^2}$. x_i and y_i are coordinate points on the skeleton line, as shown in Fig. 9(b). A skeleton extraction algorithm was used to extract the centreline, as shown in Fig. 10(b).

As for potholes, patches, and bleeding, their defective areas were calculated by determining the smallest external rectangular area, as demonstrated in Fig. 10(c).

4. Discussion

The proposed method detects four types of pavement defects in the images. However, it remains to be determined what types of defects the model is most effective at identifying. This will be discussed in this section, utilizing metrics such as Intersection over Union (IoU), confusion matrices, and heat maps. In addition, it is also necessary to discuss the limitations of the proposed methodology.

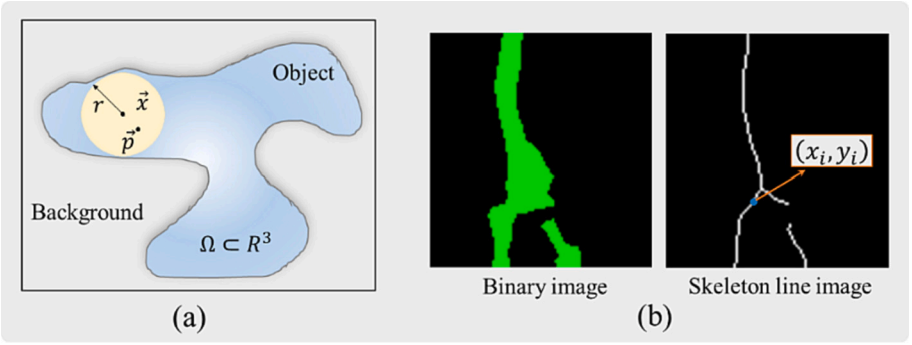

Fig. 9. Quantification of crack width and length. (a) definition and calculation of local thickness for crack width, (b) extraction of skeleton line for crack length.

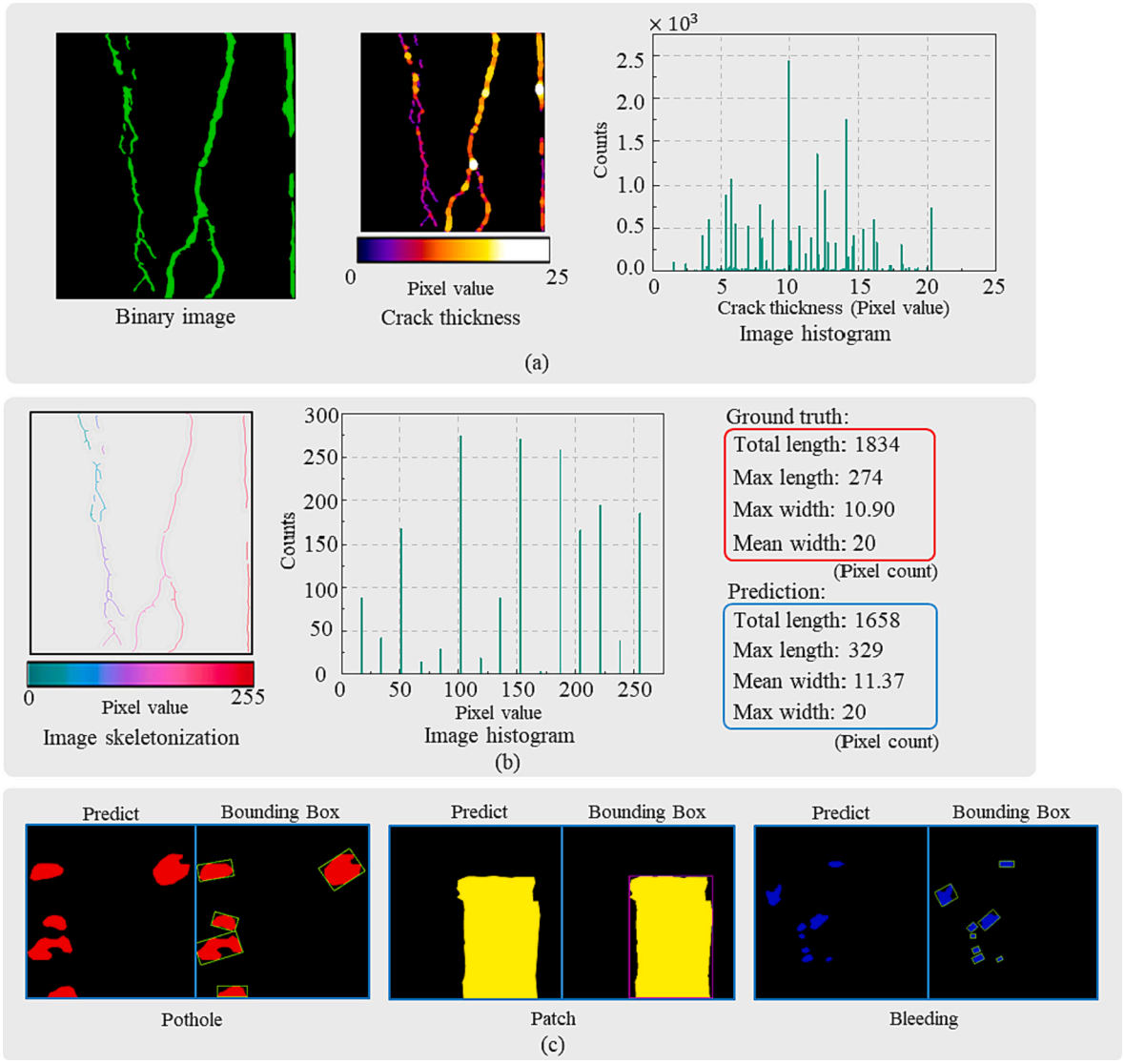

Fig. 10. Quantitative results of pavement defects. (a) measurement results of crack width, (b) measurement results of crack length, (c) measurements of other defects.

4.1. Discussion of the identification results

The evaluation of the networks' effectiveness in identifying different defect types was conducted using the IoU as an evaluation metric. The results achieved by the network models are summarised in Table 6. Comparing U-Net Grey and U-Net Depth, it was observed that the model trained with depth images achieved the highest IoU for pothole areas. Conversely, the model trained with grey images exhibited better segmentation in other areas. Notably, when the grey and depth images were combined as a dual channel, the overall accuracy increased. The double-headed structure showed slightly better performance in segmenting potholes and cracks. Furthermore, the attention mechanism played a significant role, leading to the best recognition results for the double-headed U-Net + attention model. Additionally, referring to the confusion matrix in Fig. 11, resulting in higher values along the main diagonal of the matrix, indicating increased recognition accuracy with the continuous improvement of the original method. Ultimately, the DHU-Net + attention model delivered optimal results.

To gain further insight into the attention mechanisms of the two models, the model weights were loaded and the predicted images were output. Then the spatial attention weights were extracted at corresponding locations. The visualisation results of the spatial attention weights are presented in Fig. 12. Both networks assigned weights to the pothole and crack regions, with the pothole receiving a higher weight and the crack a lower weight. Additionally, it is noticeable that the attention mechanism in DHU-Net encompasses a broader range of regions compared to U-Net. This mechanism demonstrates a greater focus on the crack regions in the grey channel and the pothole regions in the depth channel. Hence, the attention mechanism plays a critical role in enabling the double-headed U-Net to achieve optimal segmentation results.

Table 6
IoU values of models under different categories.

Models	IoU Metrics (%)				
	Back Ground	Pothole	Crack	Patch	Bleeding
U-Net Grey	94.36	61.23	42.77	78.95	57.51
U-Net Depth	93.59	82.79	33.32	71.09	36.78
U-Net Two Channel	96.19	83.10	45.55	86.23	70.33
U-Net Two + Attention	95.84	84.12	45.50	85.78	71.81
DHU-Net	96.17	85.07	46.44	85.76	70.08
DHU-Net + Attention	**96.97**	**87.90**	**48.75**	**93.37**	**74.38**

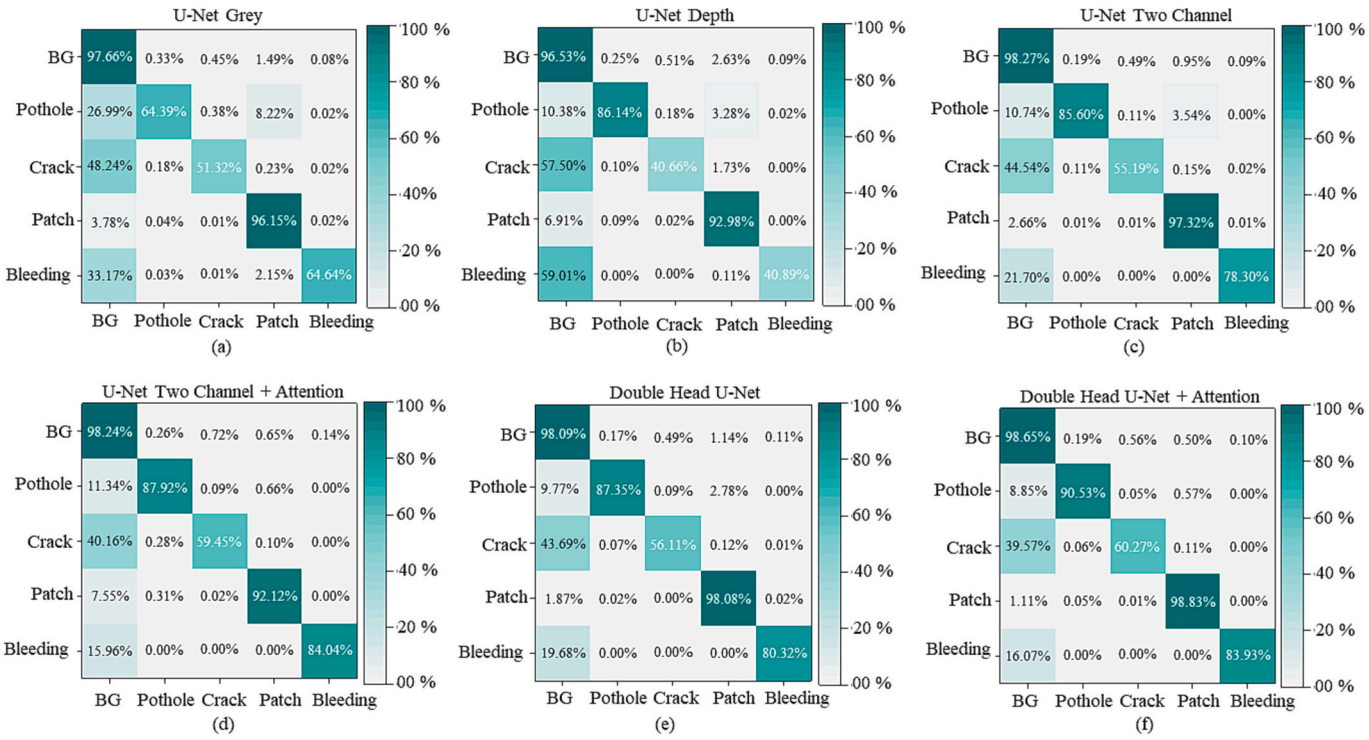

Fig. 11. Comparison of confusion matrix, (a) U-Net Grey, (b) U-Net Depth, (c) U-Net Two, (d) U-Net Two + Attention, (e) DHU-Net, (f) DHU-Net + Attention.

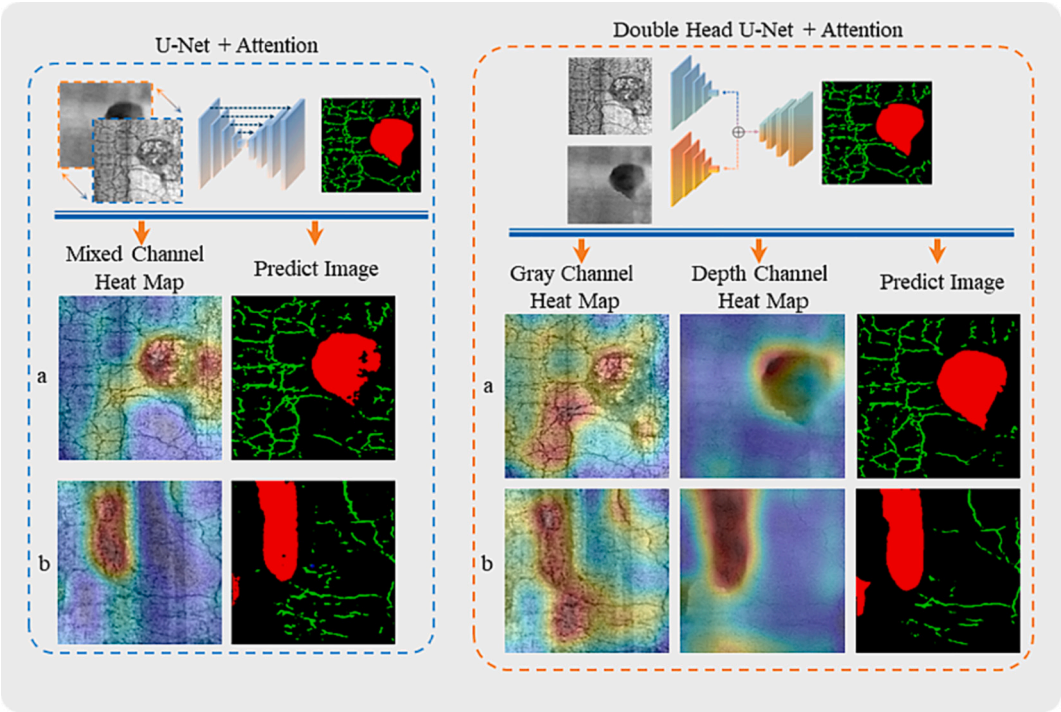

Fig. 12. Visualisation of spatial attention with U-Net + attention and DHU-Net.

4.2. The limitations of the proposed method

The CNN-based method in the paper identified the defects, and the quantification method was practised. However, there are still deficiencies in three aspects: dataset, network model and quantification method.

(1) Dataset: The algorithm is dataset dependent. It is difficult and laborious to manually label the four defects on the depth and grey image. The proposed method is more sensitive to patches and bleeding on grey images and more effective to potholes on depth images. Therefore, our work is instructive for future labelling tasks, but developing richer datasets and more efficient semi-supervised algorithms is still necessary.

(2) Network model: The improved model has a double-headed structure. This design improves the recognition accuracy but increases the network training parameters, which is not conducive to model training and deployment. In the future, it is necessary to investigate how to carry out the lightweight design while ensuring recognition accuracy.
(3) Quantification methods: Although 3D depth images are used for defect identification, the quantification method only measures two-dimensional parameters, and there is room for improvement (e.g. measuring crack and pothole depths).

5. Conclusions

This paper utilizes convolutional neural networks for the detection of pavement defects. The method entails collecting image data through a 3D pavement information collection system, followed by pre-processing and annotation of both grey and depth images. Then two network structures were constructed based on the characteristics of the image data: a classical U-shaped structure and a double-headed structure, with the inclusion of an attention mechanism in the model. The proposed models are comprehensively compared and discussed in conjunction with numerical evaluation and visualisation results. Finally, we provided a quantitative analysis of the four types of pavement information. The main conclusions drawn from our study are as follows:

(1) The proposed network models are capable of segmenting multiple pavement objects. A comparison of several models indicated that the double-headed U-Net + attention model demonstrated the best performance on the test dataset. The model achieved a Global Pixel Accuracy (GPA) of 97.36% and Mean Intersection over Union (MIoU) of 80.28%, with MIoU values improving by up to 16.77%.
(2) Recognition results were unsatisfactory when solely relying on grey or depth images for model training. Our proposed method effectively harnesses the information from both data types, significantly enhancing detection accuracy in both the classical U-shaped network and the innovative two-headed network.
(3) The addition of an attention mechanism slightly improved the recognition performance of the classical U-shaped network. However, it is noteworthy that the designed two-headed network maximized the utilization of the attention mechanism, leading to a 3.47% improvement in MIoU. The reasons behind the model's satisfactory results were analysed by examining the confusion matrix and visualizing spatial attention.
(4) The quantification methods were proposed for different pavement defects, aiming to provide valuable insights for future pavement maintenance decisions.

Currently, there is a scarcity of pavement datasets that encompass both depth and grey images. Furthermore, other types of pavement defects, such as roadway rutting, bulging and subsidence, present challenges in the pavement detection tasks. Our future research endeavours will focus on enriching and expanding existing datasets to achieve even better detection results. Additionally, the development of lightweight design is crucial for real-time detection. It is difficult to reduce the model parameters while ensuring accuracy. Therefore, we will explore the lightweight and precise network models to facilitate real-time detection capabilities.

CRediT authorship contribution statement

Peigen Li: Conceptualization, Methodology, Writing – original draft, Visualization. **Bin Zhou:** Investigation, Data curation. **Chuan Wang:** Software, Validation. **Guizhang Hu:** Validation, Formal analysis. **Yong Yan:** Investigation, Visualization. **Rongxin Guo:** Supervision, Project administration, Funding acquisition. **Haiting Xia:** Resources, Writing – review & editing, Funding acquisition.

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.

Acknowledgments

This project is supported by the National Natural Science Foundation of China (No. 12262015).

References

[1] D. Chen, N.R. Sefidmazgi, H. Bahia, Exploring the feasibility of evaluating asphalt pavement surface macro-texture using image-based texture analysis method, Road Mater. Pave. Design 16 (2) (2015) 405–420, https://doi.org/10.1080/14680629.2015.1016547.
[2] S. Dong, S. Han, C. Wu, O. Xu, H. Kong, Asphalt pavement macrotexture reconstruction from monocular image based on deep convolutional neural network, Comp. Aided Civil Infrastruct. Eng. 37 (2022) 1754–1768, https://doi.org/10.1111/mice.12878.
[3] D. Zhang, An efficient and reliable coarse-to-fine approach for asphalt pavement crack detection, Image Vis. Comput. 57 (2017) 130–146, https://doi.org/10.1016/j.imavis.2016.11.018.
[4] S. Li, Y. Cao, H. Cai, Automatic pavement-crack detection and segmentation based on steerable matched filtering and an active contour model, J. Comput. Civ. Eng. 31 (2017) 04017045, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000695.
[5] A. Ayenu-Prah, N. Attoh-Okine, Evaluating pavement cracks with bidimensional empirical mode decomposition, EURASIP J. Adv. Signal Proc. 2008 (2008), 861701, https://doi.org/10.1155/2008/861701.
[6] Q. Zou, Y. Cao, Q. Li, Q. Mao, S. Wang, CrackTree: automatic crack detection from pavement images, Pattern Recogn. Lett. 33 (2012) 227–238, https://doi.org/10.1016/j.patrec.2011.11.004.
[7] P. Li, H. Xia, B. Zhou, F. Yan, R. Guo, A method to improve the accuracy of pavement crack identification by combining a semantic segmentation and edge detection model, Appl. Sci. 12 (2022) 4714, https://doi.org/10.3390/app12094714.
[8] E. Bauer, P.M. Milhomem, L.A.G. Aidar, Evaluating the damage degree of cracking in facades using infrared thermography, J. Civ. Struct. Heal. Monit. 8 (2018) 517–528, https://doi.org/10.1007/s13349-018-0289-0.
[9] J. Yang, W. Wang, G. Lin, Q. Li, Y. Sun, Y. Sun, Infrared thermal imaging-based crack detection using deep learning, IEEE Access. 7 (2019) 182060–182077, https://doi.org/10.1109/ACCESS.2019.2958264.
[10] F. Liu, J. Liu, L. Wang, Deep learning and infrared thermography for asphalt pavement crack severity classification, Autom. Constr. 140 (2022), 104383, https://doi.org/10.1016/j.autcon.2022.104383.
[11] J. Laurent, M. Talbot, M. Doucet, Road surface inspection using laser scanners adapted for the high precision 3D measurements of large flat surfaces, in: Proceedings. International Conference on Recent Advances in 3-D Digital Imaging and Modeling (Cat. No.97TB100134), Ottawa, ON, Canada, 1997, pp. 303–310, https://doi.org/10.1109/IM.1997.603880.
[12] Y.-C.J. Tsai, F. Li, Critical assessment of detecting asphalt pavement cracks under different lighting and low intensity contrast conditions using emerging 3D laser technology, J. Transp. Eng. 138 (2012) 649–656, https://doi.org/10.1061/(ASCE)TE.1943-5436.0000353.
[13] H. Guan, J. Li, Y. Yu, M. Chapman, H. Wang, C. Wang, R. Zhai, Iterative tensor voting for pavement crack extraction using mobile laser scanning data, IEEE Trans. Geosci. Remote Sens. 53 (2015) 1527–1537, https://doi.org/10.1109/TGRS.2014.2344714.
[14] S. Zhou, W. Song, Deep learning-based roadway crack classification using laser-scanned range images: a comparative study on hyperparameter selection, Autom. Constr. 114 (2020), 103171, https://doi.org/10.1016/j.autcon.2020.103171.
[15] Q. Li, M. Yao, X. Yao, B. Xu, A real-time 3d scanning system for pavement distortion inspection, Meas. Sci. Technol. 21 (2010), 015702, https://doi.org/10.1088/0957-0233/21/1/015702.
[16] X. Zhang, T. Liu, C. Liu, Z. Chen, Research on skid resistance of asphalt pavement based on three-dimensional laser-scanning technology and pressure-sensitive film, Constr. Build. Mater. 69 (2014) 49–59, https://doi.org/10.1016/j.conbuildmat.2014.07.015.
[17] A. Zhang, K.C.P. Wang, Y. Fei, Y. Liu, C. Chen, G. Yang, J.Q. Li, E. Yang, S. Qiu, Automated pixel-level pavement crack detection on 3D asphalt surfaces with a recurrent neural network, Comp. Aided Civil Infrastruct. Eng. 34 (2019) 213–229, https://doi.org/10.1111/mice.12409.

[18] D. Zhang, Q. Zou, H. Lin, X. Xu, L. He, R. Gui, Q. Li, Automatic pavement defect detection using 3D laser profiling technology, Autom. Constr. 96 (2018) 350–365, https://doi.org/10.1016/j.autcon.2018.09.019.
[19] C. Lu, J. Yu, C.K.Y. Leung, An improved image processing method for assessing multiple cracking development in strain hardening cementitious composites (SHCC), Cem. Concr. Compos. 74 (2016) 191–200, https://doi.org/10.1016/j.cemconcomp.2016.10.005.
[20] C. Peng, M. Yang, Q. Zheng, J. Zhang, D. Wang, R. Yan, J. Wang, B. Li, A triple-thresholds pavement crack detection method leveraging random structured forest, Constr. Build. Mater. 263 (2020), 120080, https://doi.org/10.1016/j.conbuildmat.2020.120080.
[21] H. Zhou, S. Yang, J. Zhu, Illumination invariant enhancement and threshold segmentation algorithm for asphalt pavement crack image, in: In: 2010 International Conference on Computational Intelligence and Software Engineering, IEEE, Chengdu City, China, 2010, pp. 1–4, https://doi.org/10.1109/WICOM.2010.5600853.
[22] C. Koch, I. Brilakis, Pothole detection in asphalt pavement images, Adv. Eng. Inform. 25 (2011) 507–515, https://doi.org/10.1016/j.aei.2011.01.002.
[23] F. Liebold, H.-G. Maas, Advanced spatio-temporal filtering techniques for photogrammetric image sequence analysis in civil engineering material testing, ISPRS J. Photogramm. Remote Sens. 111 (2016) 13–21, https://doi.org/10.1016/j.isprsjprs.2015.10.013.
[24] K. Lakshmi, Detection and quantification of damage in bridges using a hybrid algorithm with spatial filters under environmental and operational variability, Structures. 32 (2021) 617–631, https://doi.org/10.1016/j.istruc.2021.03.031.
[25] A. Ghanbari Mardasi, N. Wu, C. Wu, Experimental study on the crack detection with optimized spatial wavelet analysis and windowing, Mech. Syst. Signal Process. 104 (2018) 619–630, https://doi.org/10.1016/j.ymssp.2017.11.039.
[26] K.C.P. Wang, Q. Li, W. Gong, Wavelet-based pavement distress image edge detection with À Trous algorithm, Transp. Res. Rec. 2007 (2024) 73–81, https://doi.org/10.3141/2024-09.
[27] R. Augustauskas, A. Lipnickas, Improved pixel-level pavement-defect segmentation using a deep autoencoder, Sensors. 20 (2020) 2557, https://doi.org/10.3390/s20092557.
[28] A. Ji, X. Xue, Y. Wang, X. Luo, W. Xue, An integrated approach to automatic pixel-level crack detection and quantification of asphalt pavement, Autom. Constr. 114 (2020), 103176, https://doi.org/10.1016/j.autcon.2020.103176.
[29] Y. Liu, J. Yao, X. Lu, R. Xie, L. Li, DeepCrack: a deep hierarchical feature learning architecture for crack segmentation, Neurocomputing. 338 (2019) 139–153, https://doi.org/10.1016/j.neucom.2019.01.036.
[30] V. Pereira, S. Tamura, S. Hayamizu, H. Fukai, Semantic segmentation of paved road and pothole image using U-Net architecture, in: 2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA), IEEE, Yogyakarta, Indonesia, 2019, pp. 1–4, https://doi.org/10.1109/ICAICTA.2019.8904105.
[31] Z. Tong, D. Yuan, J. Gao, Z. Wang, Pavement defect detection with fully convolutional network and an uncertainty framework, Comp. Aided Civil Infrastruct. Eng. 35 (2020) 832–849, https://doi.org/10.1111/mice.12533.
[32] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7132–7141. http://arxiv.org/abs/1709.01507.
[33] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19. http://arxiv.org/abs/1807.06521.
[34] D. Lin, Y. Li, S. Prasad, T.L. Nwe, S. Dong, Z.M. Oo, CAM-guided multi-path decoding U-Net with triplet feature regularization for defect detection and segmentation, Knowl.-Based Syst. 228 (2021), 107272, https://doi.org/10.1016/j.knosys.2021.107272.
[35] L. Pratt, D. Govender, R. Klein, Defect detection and quantification in electroluminescence images of solar PV modules using U-net semantic segmentation, Renew. Energy 178 (2021) 1211–1222, https://doi.org/10.1016/j.renene.2021.06.086.
[36] Q. Zhong, J. Zhang, Y. Xu, M. Li, B. Shen, W. Tao, Q. Li, Filamentous target segmentation of weft micro-CT image based on U-net, Micron. 146 (2021), 102923, https://doi.org/10.1016/j.micron.2020.102923.
[37] S.S. Bangaru, C. Wang, X. Zhou, M. Hassan, Scanning electron microscopy (SEM) image segmentation for microstructure analysis of concrete using U-net convolutional neural network, Autom. Constr. 144 (2022), 104602, https://doi.org/10.1016/j.autcon.2022.104602.
[38] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, 2015, pp. 234–241. http://arxiv.org/abs/1505.04597.
[39] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 2481–2495.
[40] L.C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, 2017, https://doi.org/10.48550/arXiv.1706.05587 arXiv preprint arXiv:1706.05587.
[41] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, LabelMe: a database and web-based tool for image annotation, Int. J. Comput. Vis. 77 (2008) 157–173, https://doi.org/10.1007/s11263-007-0090-8.
[42] Y. Bengio, Y. Grandvalet, No unbiased estimator of the variance of k-fold cross-validation, the, J. Mach. Learn. Res. 5 (2004) 1089–1105.
[43] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, arXiv preprint arXiv:1412.6980, http://arxiv.org/abs/1412.6980, 2017.
[44] T. Hildebrand, P. Rüegsegger, A new method for the model-independent assessment of thickness in three-dimensional images, J. Microsc. 185 (1997) 67–75, https://doi.org/10.1046/j.1365-2818.1997.1340694.x.
[45] T.Y. Zhang, C.Y. Suen, A fast parallel algorithm for thinning digital patterns, Commun. ACM 27 (3) (1984) 236–239, https://doi.org/10.1145/357994.358023.
[46] X. Yang, H. Li, Y. Yu, X. Luo, T. Huang, X. Yang, Automatic pixel-level crack detection and measurement using fully convolutional network, Comput. Aided Civ. Inf. Eng. 33 (12) (2018) 1090–1109, https://doi.org/10.1111/mice.12412.
[47] A. Ji, X. Xue, Y. Wang, X. Luo, W. Xue, An integrated approach to automatic pixel-level crack detection and quantification of asphalt pavement, Autom. Constr. 114 (2020), 103176, https://doi.org/10.1016/j.autcon.2020.103176.