



OPEN

Optimized deep learning model with integrated spectrum focus transformer for pavement distress recognition and classification

Wenlin Wu¹, Fenghua Zhu^{2,3}, Zheng Li¹✉, Xue Li¹, Xiaowei Li¹ & Jinwen Wang¹

In the task of pavement distress recognition and classification, the complexity of the pavement environment, the small proportion of distresses in images, significant variation in distress scales, and the influence of features such as vehicles and traffic signs in the data make distress feature extraction challenging. This paper proposes a spectrum focus transformer (SFT) layer, which processes the signal spectrum and focuses on important frequency components. Initially, by thoroughly analyzing the frequency domain characteristics of image data, frequency value distribution information is obtained to achieve fine-tuning of different frequency components. Subsequently, frequency information and images are learned and weighted in the frequency domain, thereby enhancing the capability to capture pavement distress regions. Experiments conducted on the road pavement distress dataset revealed through heatmap analysis that distress regions received increased attention, achieving an accuracy of 97.73%. This performance demonstrates a higher accuracy compared to other models.

With the rapid development of transportation infrastructure, pavement distress issues have become increasingly prominent, posing serious threats to traffic safety and road lifespan¹. Traditional pavement distress detection methods primarily rely on manual inspections and periodic maintenance. These methods are not only inefficient and costly but also susceptible to human factors, leading to missed detections and false positives²⁻⁴. With the rapid advancement of technologies such as machine learning, the accuracy and efficiency of pavement distress detection can be improved, and detection costs can be reduced. Traditional machine learning algorithms, including support vector machines⁵ and random forests⁶, have been applied in pavement distress recognition. Currently, significant progress has been made in the application of deep learning for pavement distress detection. However, the varying sizes of different pavement distresses and complex backgrounds in images pose challenges. Researchers are working to improve deep learning algorithms for processing and analyzing pavement images to achieve automated distress recognition and classification.

For pavement distress recognition and classification, researchers proposed a two-step sequential automation process: detecting cracks in asphalt pavements and classifying their severity⁷. Due to the difficulty of collecting a large number of images with pavement distress, researchers introduced a novel few-shot pavement distress detection method based on metric learning. This method effectively learns new classes from a limited number of labeled samples and incorporates an attention mechanism to focus on the features required by the model⁸. However, the accuracy performance of the above-mentioned research is relatively poor. Given the complexity of pavement distress scenes, the researchers designed a target detection network named crackyolo⁹. The detection includes three types of common cracks found on rural roads: transverse cracks, longitudinal cracks, and alligator cracks. This network enhances crack localization and identification capabilities through custom convolutional and sampling operations, with the channel and spatial mixed attention mechanism introduced to enhance crack weight. However, this study has limitations, particularly in detecting targets in various scenarios, as well as efficiency issues. In Ref.¹⁰, the researchers utilized three-dimensional ground-penetrating radar and deep convolutional neural networks to achieve automatic classification of pavement distress. Liu et al.¹¹ proposed a feature-enhanced multiscale vision transformer for road distress classification from ground penetrating radar images. The model used the feature-enhanced feature pyramid network and feature enrichment module to extract the distress better features on ground penetrating radar images. Although these studies focus on road distress classification from ground-penetrating radar images, their methods can also be extended to broader applications,

¹School of Rail Transportation, Shandong Jiaotong University, Jinan 250357, China. ²Institute of Automation, Chinese Academy of Sciences, Beijing 100000, China. ³Qingdao Academy of Intelligent Industries, Qingdao 26600, China. ✉email: waiwaivictory@163.com

providing valuable insights for road distress fields. Yuan et al.¹² proposed an innovative road distress classification model that capitalizes on sparse perception. Their method introduced a sparse feature extraction module using dilated convolution, tailored to capture and combine sparse features of different scales from the image. However, the model's generalization ability and efficiency still leave room for improvement. In Ref.¹³, this study aimed to apply infrared thermography and deep learning to multiple -type pavement distress detection, providing a novel combined approach. However, the infrared image dataset is limited, which reduces its practical applicability. In Ref.¹⁴, researchers showed that an attention-based multi-scale convolutional neural network improved the automated classification of common distress and non-distress objects in pavement images, increasing robustness and further enhancing performance. One limitation of this study is that the approach provides a patch-level segmentation mask for pavement images, meaning that it only classifies small sections of the image, rather than providing a full-scale understanding of the entire pavement. Cai et al.¹⁵ conducted a comparison of the residual network(ResNet) with other classical models in terms of their effectiveness in classifying interlayer distresses in pavement. In comparison, the residual network demonstrated superior generalization performance. The various techniques proposed by the researchers have contributed to addressing road distress safety issues and have also opened up new avenues for further research.

Frequency domain analysis has long been a powerful tool in signal processing. Recently, applications integrating frequency analysis into deep learning have emerged. Yang et al.¹⁶ proposed a vibration-based method for identifying transverse cracks in asphalt pavements using a dual relative energy process with fast fourier transform(FFT) and discrete wavelet transform (DWT) for crack detection. In Ref.¹⁷, the authors applied wavelet transform and Radon neural networks, utilizing wavelet modulus calculations for pavement distress classification.

In Ref.¹⁸, the authors reconsidered global average pooling(GAP)from the frequency domain perspective. To address the limitation of insufficient feature information in existing channel attention methods, they extended GAP to a more general two-dimensional discrete cosine transform form, incorporating more frequency components to fully utilize the information. From this literature, we derived the idea of integrating frequency information from pavement distress data into the attention module. By applying FFT to transform the data into the frequency domain for analysis, we can obtain a global receptive field. This approach dynamically adjusts the frequency weights of the image, guiding the attention to focus on the most diagnostically significant frequency ranges, thereby enhancing the model's ability to recognize pavement distresses. We named this design the spectral focus transformer (SFT). Using ResNet34 as the base framework, we integrated SFT into it for pavement distress recognition.

Methods

Dataset

The pavement distress dataset for the competition comes from the 2021 Global Open Data Application Innovation Competition. This dataset includes data captured by vehicle-mounted cameras, comprising a total of 14,000 pavement distress images. The training set provides annotated labels (distress types and bounding box locations), while the test set does not provide labels. In this study, to ensure consistency and simplify the classification task, we selected six distinct types of pavement distress from the training set, excluding images containing multiple distress types. These selected distress types are: Crack, Manhole, Net(Alligator Cracking), Pothole, Patch-Crack(Repaired Crack), and Patch-Net (Repaired Alligator Cracking). After data processing, a total of 5,325 images were obtained. The image annotation process was automated using the original dataset, and the dataset was then divided into training and validation sets in an 80:20 ratio. Table 1 provides the data distribution for different pavement distress categories.

Model structure

As shown in Fig. 1, the framework of the proposed method consists of three main steps. Step one involves performing frequency domain processing on the input images to obtain spectral focus information. This information is incorporated into the Spectral Focus Transformer (SFT) as an additional input by computing the distribution of frequency values across intervals. Step two integrates SFT with ResNet34 for feature extraction. Both the input image information and the spectral focus information are fed into the neural network integrated with SFT. In the SFT, the frequency distribution information guides the model to focus on the important features of the image. Step three involves pavement distress classification. After feature extraction through the final residual block, global average pooling is used to reduce dimensions before inputting the data into the linear layer (i.e., the fully connected layer). Finally, the cross-entropy loss function automatically applies the softmax

Class label	Disease type	Number
0	Crack	800
1	Manhole	2981
2	Net (alligator cracking)	123
3	Pothole	96
4	Patch-crack (repaired crack)	1120
5	Patch-Net (repaired alligator cracking)	205

Table 1. Data category.

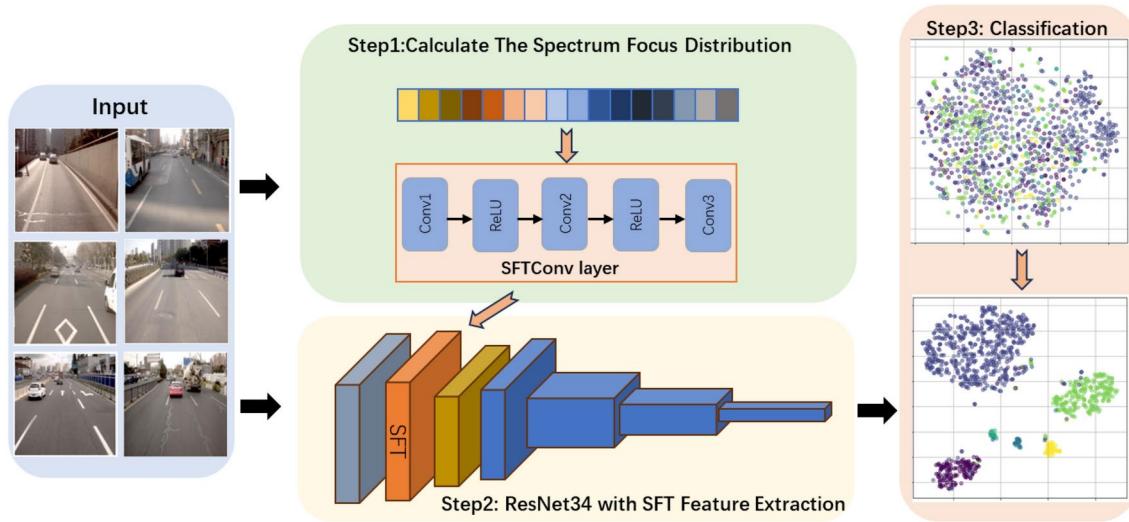


Fig. 1. Overview of the overall method.

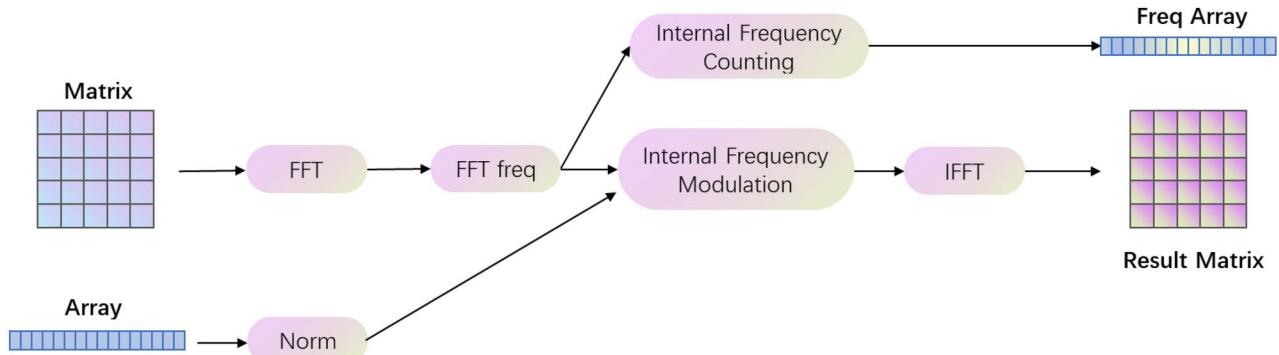


Fig. 2. SF block.

function to calculate class probabilities, determining the pavement distress category. For additional details on the meanings of certain elements in the figure, please refer to the subsequent sections.

Spectrum focus block

FFT is an efficient and rapid algorithm for domain conversion. As a discrete Fourier transform algorithm, it was first introduced by Cooley and Tukey in 1965¹⁹. FFT facilitates the conversion of images between the Fourier (i.e., frequency) domain and the spatial domain. Spectral domain transformation is performed by updating all pixels simultaneously, enabling FFT to help the model extend its receptive field to cover all pixels in the image, thereby achieving a global receptive field and enhancing the information correlation among all pixels.

The key component of the model is the frequency domain processing section, called the Spectrum Focus Block(SF Block). As shown in Fig. 2, the frequency domain processing involves two inputs: an Height*Width (HxW) matrix and a 1x16 array. It ultimately returns two outputs: one is the result matrix after frequency domain processing, and the other is the frequency array containing the number of frequency elements.

The process to obtain the freq array involves performing an FFT operation on the input matrix along the width dimension to obtain its frequency domain representation. Subsequently, the fast fourier transform frequency array generator operation is carried out to obtain the corresponding frequency values. Then, in the internal frequency counting operation, the frequency range is divided into 16 equally spaced intervals, and the number of frequency elements in each interval is computed. This results in the freq array.

The calculation of the result matrix involves applying the FFT operation to the input matrix. The input array is then normalized, and both results are passed into the interval frequency modulation process. Specifically, the process involves taking the matrix transformed into the frequency domain and obtaining values corresponding to the divided frequency intervals. These values are then multiplied element-wise with the normalized array to achieve the final result. The modulation is done along the final dimension, which is the width dimension, ensuring that only the elements in the corresponding intervals are modified. The modified Fourier coefficients are then subjected to an inverse FFT (IFFT) operation, and the real part of the inverse-transformed result is extracted to obtain the result matrix.

Spectrum focus transformer

The channel attention mechanism works by assigning weights to each feature channel, thereby emphasizing or suppressing the importance of information from different channels²⁰. Drawing on the principles of channel attention mechanisms, we designed the SFT, which integrates elements of channel attention and frequency domain processing to create a frequency domain attention mechanism. This mechanism dynamically adjusts feature responses by learning the weights of historical frequency domain channels. This attention mechanism allows the network to focus on the most diagnostically significant frequency intervals, thereby enhancing the model's ability to identify pavement defects.

As shown in Fig. 3, feature extraction that includes frequency information is achieved through two applications of the SF Block. In the figure, FA and RM represent the two output results of the SF Block from Fig. 2: the frequency array and the result matrix, respectively.

Let the input feature map matrix with dimensions $H \times W$ be denoted as X , and let the input random array with dimensions 1×16 be denoted as R . After processing through the SF Block, only the FA output is selected to obtain the frequency value distribution information, denoted as F . The formula for this process is as follows:

$$F = FAs(SF(X, R)) \quad (1)$$

Where s represents the selection symbol.

After obtaining the frequency value distribution information F , feature extraction, nonlinear transformation, and feature mapping of the frequency domain information are performed using three 1×1 convolutions and two ReLU activation functions (as shown in Step 1 of Fig. 1, SFTConv layer). This process generates a frequency domain representation Q that is compatible with the input image. The feature map X and the frequency domain representation Q are then reintroduced into the SF Block for further processing. This time, only the RM output result is selected to obtain the final feature extraction map, denoted as Y . The calculation formula for this process is as follows:

$$Q = w3 * r(w2 * r(w1 * F)) \quad (2)$$

$$Y = RM(s(SF(Q, X))) \quad (3)$$

Among them, $w1, w2, w3$ represent the weights of the first, second, and third convolutional layers respectively, and r represents the ReLU activation function.

The ResNet34 structure integrated with spectrum focus transformer

The experiment chooses to use the ResNet34 model as the basic framework structure²¹. The proposed SFT is integrated into this framework, referred to as the SFT layer. The detailed structure of Step 2 in Fig. 1 is shown in Fig. 4, which represents the overall model of ResNet-34 with SFT. The SFT layer is integrated between the Batch Norm layer and the Max Pool layer, following the initial convolution operation. This integration allows for immediate enrichment of the original feature maps with spectral information, enabling the network to prioritize and refine frequency components crucial for visual pattern recognition. This is beneficial for more advanced feature learning in subsequent network layers.

Results

Experimental details

The experimental framework version used is PaddlePaddle 3.0.0beta0, Python 3.10.10, with the runtime environment including a Tesla V100 GPU, 32 GB of RAM, a 4-core CPU, and 32 GB of video memory. During training, Cross-Entropy Loss is used as the criterion to measure the difference between the model's output

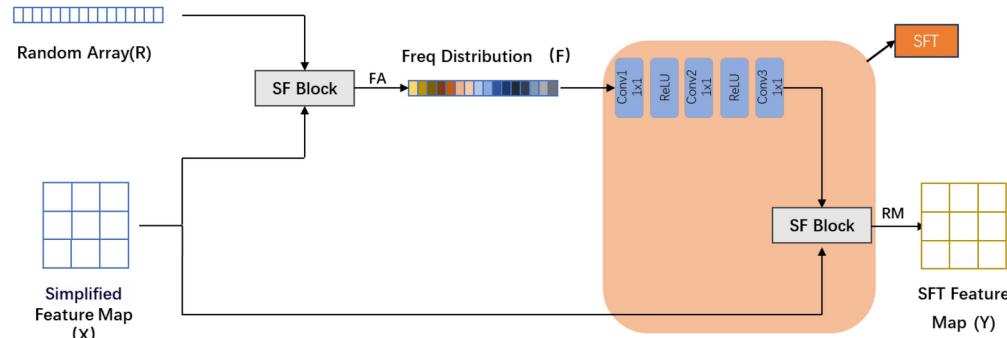


Fig. 3. SFT.

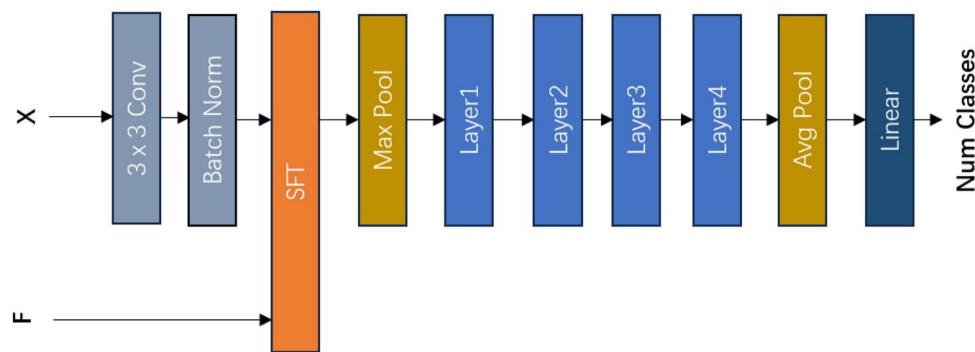


Fig. 4. Model structure of ResNet34 with SFT.

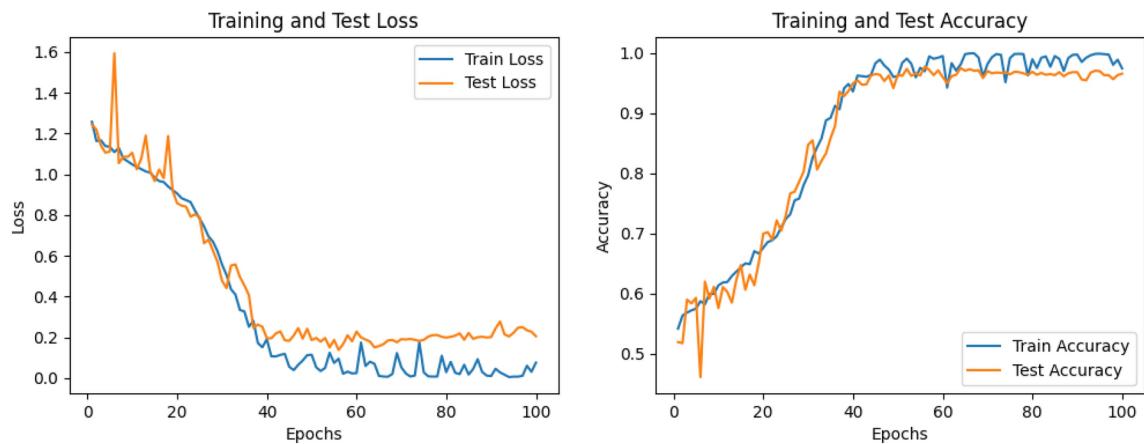


Fig. 5. Loss and accuracy of training and testing.

probability distribution and the true label distribution. The Adam optimizer is chosen for model parameter optimization, with a learning rate set to 0.001. The number of epochs is set to 100, and the batch size is set to 32.

Test results

As shown in Fig. 5, the model reaches stability around the 45th epoch, with an accuracy of approximately 96%. After training for 100 epochs, the accuracy on the test set reaches 97.73%, with a loss of about 0.1374. This performance is considered excellent for the recognition and classification tasks of different pavement defects in complex background road surface data.

T-Distributed Stochastic Neighbor Embedding(T-SNE) is a nonlinear dimensionality reduction technique used to visualize high-dimensional data by mapping it into a lower-dimensional space two-dimensional(2D) or three-dimensional(3D) while preserving relationships between similar data points. In our study, we applied T-SNE to visualize the clustering of different distress classes, helping us assess how well the model separates them. Figure 6 is the 2D visualization of T-SNE features, corresponding to step three of Fig. 1, and Fig. 7 shows the confusion matrix of pavement distress classification results. Analysis of the two figures shows that the model performs well for both multi-sample and few-sample categories. Using T-SNE technology²² for dimensionality reduction and visualization of the original and feature-extracted data, it is clear that the scattered original data points become more concentrated into distinct clusters after feature extraction.

Ablation experiments

We conducted several ablation experiments to evaluate the effectiveness of the SFT layer. Specifically, we compared the following models:ResNet-34, the baseline model without any attention mechanism; ResNet-34 with SE Attention, the model enhanced with Squeeze-and-Excitation (SE) attention; and ResNet-34 with SFT, the model enhanced with the proposed Spectrum Focus Transformer (SFT) layer. Table 2 shows the results of Accuracy, Weighted average F1-score, Weighted average Recall, and Weighted average Average Precision obtained after evaluating each model after 100 epochs of training²³.

Through the analysis of the results from the ablation experiments, it is evident that the inclusion of SFT significantly enhances performance. The SE attention module also has a positive effect. Specifically, the ResNet-34 model with SFT shows an accuracy improvement of 0.98% compared to the ResNet-34 model without SFT.

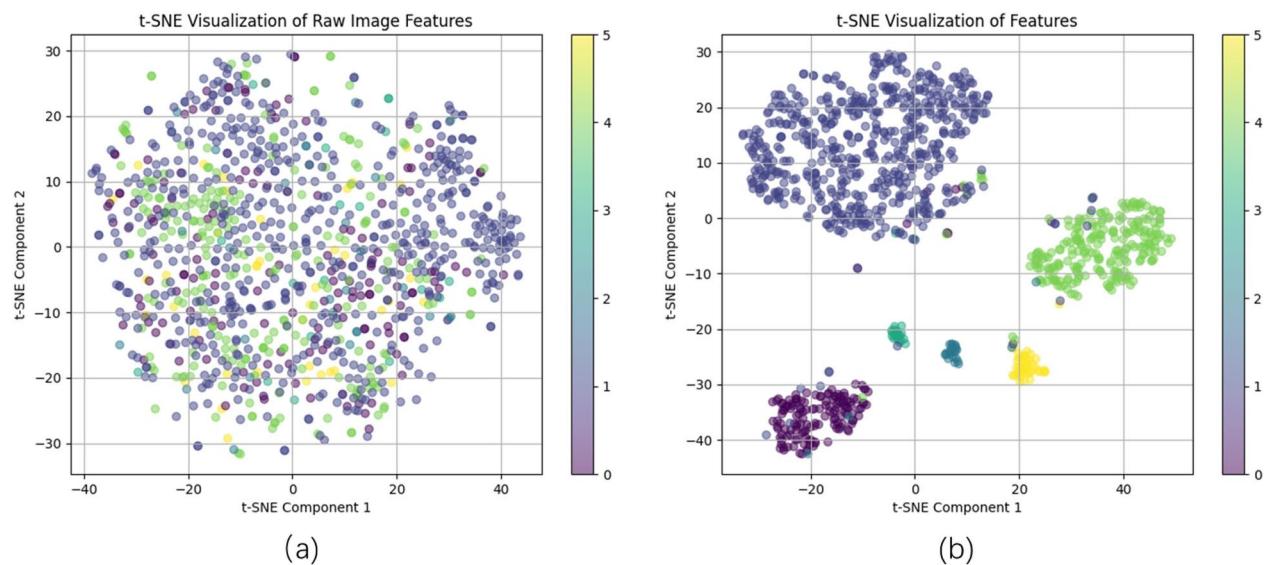


Fig. 6. T-SNE visualization. **(a)** Original data. **(b)** Model feature extraction results.

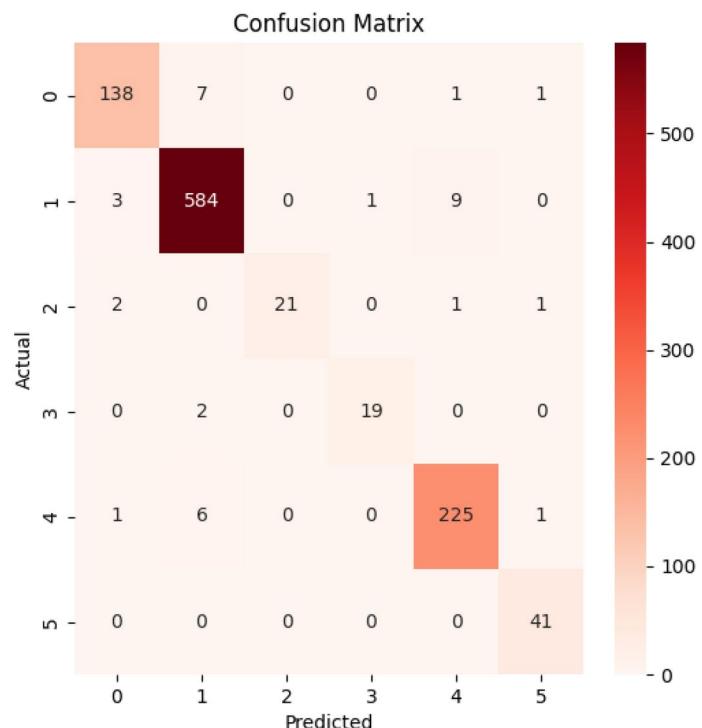


Fig. 7. Confusion matrix.

Model	Accuracy	F1-score	Recall	Precision
Resnet34	0.9675	0.9650	0.9651	0.9858
Resnet34+SE	0.9724	0.9653	0.9657	0.9861
Resnet34+SFT	0.9773	0.9679	0.9680	0.9937

Table 2. Ablation experiments.

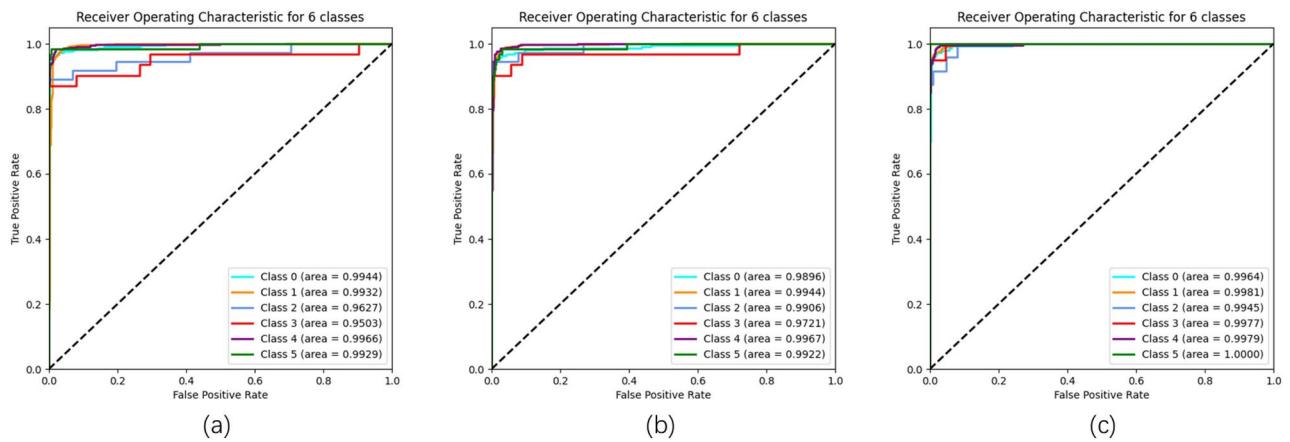


Fig. 8. The ROC curve. (a) Resnet34. (b) Resnet34+SE. (c) Resnet34+SFT.

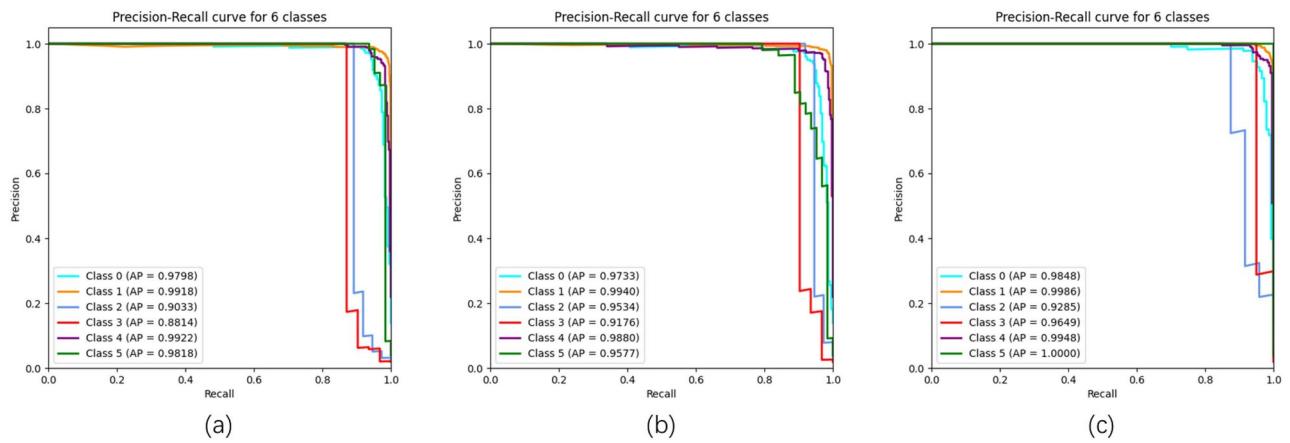


Fig. 9. The Precision-Recall curve. (a) Resnet34. (b) Resnet34+SE. (c) Resnet34+SFT.

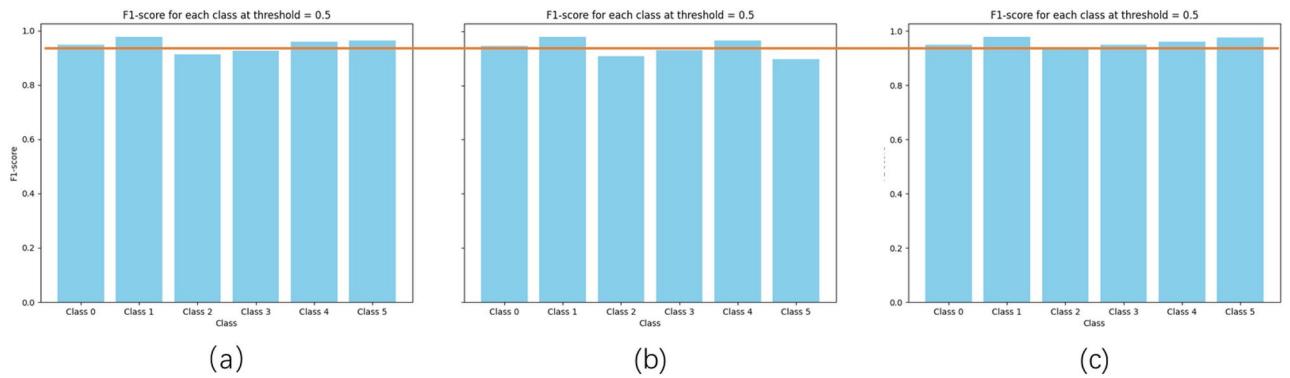


Fig. 10. The F1-score bar chart. (a) Resnet34. (b) Resnet34+SE. (c) Resnet34+SFT.

Next, Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves were plotted to observe the performance of the model, as shown in Figs. 8 and 9. Additionally, F1-scores for each class at a threshold of 0.5 were visualized in Fig. 10. The analysis reveals that the ROC curve AUC values for all classes in our model exceed 0.99, demonstrating excellent discrimination performance. However, the PR curves show that the average precision (AP) values for class 2 and class 3 are slightly lower compared to other classes. The ROC curves perform better than the PR curves, suggesting that the model is encountering class imbalance issues. Therefore, the F1-score bar chart was included to assess the performance across different classes, providing a clearer view of the model's ability to handle class imbalance. By comparing the bar chart, we observe that our

model performs more evenly across classes^{24,25}. To ensure that the improvement in our model's performance is not due to random fluctuations, we performed Bootstrap analysis²⁶. By resampling 1000 times, we obtained the Bootstrap F1-score mean for the three ablation models as follows: 0.9650, 0.9653, and 0.9681, respectively. These results indicate that the improvements are effective and not random.

Feature map visualization

We visualized the frequency domain processing of feature maps by comparing images processed with FFT and those guided by frequency information, as shown in Fig. 11a,b. To facilitate observation, the images used FFTshift to move the low-frequency components to the center of the spectrum. Analysis of the image results shows that in the spectrum before frequency domain processing, the bright areas representing amplitude are concentrated in the central low-frequency part, indicating that low-frequency signals are predominant. This matches the original image information. After processing, it was observed that the energy in the amplitude spectrum spreads towards higher frequencies, indicating that high-frequency information, such as edge details, has been enhanced. This is consistent with the information in the original image. In the phase spectrum, we found that while the shape did not change significantly, the color distribution became more concentrated and clearer. This suggests that the processing method also enhanced the phase information, making structural details and edges in the image more prominent^{27,28}. Figure 11c displays the feature maps of various channels after SFT processing, reflecting that image features are extracted with guidance from different frequency distribution information.

To observe the model's attention distribution on the input image, we used the Class Activation Mapping (CAM) visualization technique²⁹. The generated CAM maps are overlaid on the original image, with adjusted transparency to visually demonstrate the areas of focus for the model. This visualization was performed for models incorporating SE and SFT. As shown in the results of Fig. 12, SFT focuses more intently on the disease-affected regions compared to SE, where the areas of attention are more concentrated. Due to the complexity of the background, both methods inevitably are influenced by background features. However, it can be observed that SFT is less affected by the background than SE.

Comparative experiment

We conducted a simple comparative experiment using the pavement damage dataset, comparing it with other commonly used classification models, including ResNet50²¹, DenseNet, and MobileNet^{30,31}. All three models achieved high accuracy through techniques such as transfer learning. The results are shown in Table 3. In this study, we innovatively introduced the SFT frequency domain processing module, which effectively guides the model to focus more on pavement damage recognition by integrating additional frequency information, thereby further improving the model's accuracy. The proposed model is approximately 81.3MB in size, similar to the model before modification, achieving improved accuracy with fewer parameters.

In practical applications, this model can be integrated with unmanned devices, in-vehicle systems, or road monitoring systems. Captured pavement images can be processed by the model for pavement damage identification and classification. This approach helps to alleviate the burden of manual inspection through automation and intelligent means^{32,33}.

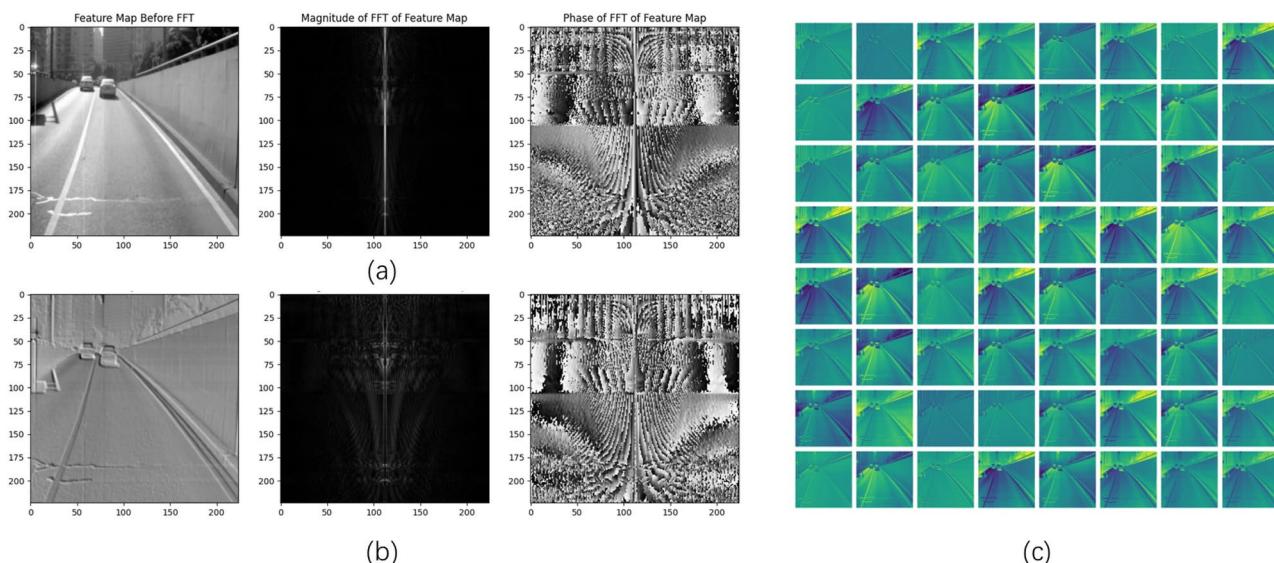


Fig. 11. Feature map and spectrogram. (a) Before frequency domain processing. (b) After frequency domain processing. (c) Feature maps of each channel in SFT feature extraction.



Fig. 12. Heatmaps. **(a)** Original images. **(b)** SE heatmaps. **(c)** SFT heatmaps.

Model	DenseNet	Resnet50	MobileNet	Resnet34+SFT
Accuracy	97.00%	97.45%	97.67%	97.73%

Table 3. Comparison results.

Discussion

Pavement distress classification plays a crucial role in road maintenance and protection. In this paper, we propose the integration of an SFT into ResNet34 for pavement distress classification, and validate the effectiveness of the model through a series of experiments.

In the test results, we used the t-SNE visualization of the pavement distress classification, which provides valuable insights into the clustering performance of the proposed model. Observing the results before and after the clustering, it is clear that the data points corresponding to different distress categories are well-separated, with minimal overlap between clusters. This indicates that the model has successfully learned distinct features for each type of distress, leading to clear and consistent classification. The effective clustering also suggests that the learned representations are robust and that the model is capable of distinguishing between various distress types with high accuracy. Combining the results with the confusion matrix, we observe that Class 1 and Class 5 show the highest accuracy, performing the best among all classes. This indicates that the model has a particularly strong ability to classify these distress types. The confusion matrix also reveals that while certain distress categories are easier to classify, others show some misclassification, highlighting areas for further model improvement.

In the ablation study, we compared the results using ROC curves, PR curves, and F1-score bar charts. The PR curves, particularly for certain classes (such as class 2 and class 3), revealed noticeable performance discrepancies, indicating the presence of class imbalance within the dataset. In the presence of class imbalance within the data, the F1-score serves as a more comprehensive evaluation metric. To this end, we included bar charts illustrating the F1-score performance across different classes. The results demonstrate that our model achieves more balanced performance across all categories, with notable improvements, highlighting the effectiveness of the proposed method in pavement distress classification. We also performed Bootstrap analysis on the three ablation models to ensure that the improvements were not due to random fluctuations. After 1000 resampling iterations, the Bootstrap F1-score means confirmed that the performance gains are statistically reliable and not random.

In the feature map visualization experiment, we compared the results of the feature maps before and after frequency domain processing. The results indicate that incorporating frequency information enhances high-frequency and phase details, thereby improving edge details and other fine-grained features of the image. Additionally, by comparing the heatmaps generated by SE and SFT models, we observed that both models are influenced by background information. However, our proposed model demonstrates a stronger focus on the distress areas, with reduced sensitivity to background noise. This suggests that our model provides more reliable guidance for pavement distress classification.

Conclusion

In this paper, we propose an SFT method that incorporates frequency distribution information and then integrates it into the ResNet34 model specifically for pavement damage classification. This approach achieved an accuracy of 97.73% on the pavement damage dataset.

Ablation experiments were conducted to compare ResNet34 with SFT against ResNet34 and ResNet34 with SE. The effectiveness of the SFT method was validated through evaluation metrics such as ROC curves, PR curves and F1-score bar chart. Visualized spectral and heatmaps were produced. The spectral diagrams of SFT were compared before and after processing in the frequency domain. Compared the feature heatmaps with added SE attention and those with added SFT. Both comparisons demonstrated that SFT plays a guiding role in focusing on road surface defect features. Comparative experiments were conducted with several common classification models to validate the effectiveness of the proposed model.

In future work, we plan to collect training data from more complex scenarios and apply transfer learning to reduce training complexity and optimize the model for better pavement damage classification.

Data availability

The data sources used in this article have been specified within the content. It can be accessed at <https://aistudiobaidu.com/datasetdetail/296925>. If you want to request the complete dataset and code, please email the corresponding author.

Received: 23 September 2024; Accepted: 28 January 2025

Published online: 30 January 2025

References

- Chan, C. Y., Huang, B., Yan, X. & Richards, S. Investigating effects of asphalt pavement conditions on traffic accidents in Tennessee based on the pavement management system (pms). *J. Adv. Transp.* **44**, 150–161. <https://doi.org/10.1002/atr.129> (2010).
- Dong, H., Song, K., Wang, Y., Yan, Y. & Jiang, P. Automatic inspection and evaluation system for pavement distress. *IEEE Trans. Intell. Transp. Syst.* **23**, 12377–12387. <https://doi.org/10.1109/TITS.2021.3113802> (2022).
- Shtayat, A., Moridpour, S., Best, B., Shroff, A. & Raol, D. A review of monitoring systems of pavement condition in paved and unpaved roads. *J. Traffic Transp. Eng. (Engl. Ed.)* **7**, 629–638. <https://doi.org/10.1016/j.jtte.2020.03.004> (2020).
- Cafiso, S., D'Agostino, C., Delfino, E. & Montella, A. From manual to automatic pavement distress detection and classification. In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 433–438 (2017).
- Lin, J. & Liu, Y. Potholes detection based on svm in the pavement distress image. In *2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science*, 544–547, <https://doi.org/10.1109/DCABES.2010.115> (2010).
- Shi, Y., Cui, L., Qi, Z., Meng, F. & Chen, Z. Automatic road crack detection using random structured forests. *IEEE Trans. Intell. Transp. Syst.* **17**, 3434–3445. <https://doi.org/10.1109/TITS.2016.2552248> (2016).
- Tran, T. S., Tran, V. P., Lee, H. J., Flores, J. M. & Le, V. P. A two-step sequential automated crack detection and severity classification process for asphalt pavements. *Int. J. Pavement Eng.* **23**, 2019–2033 (2022).
- Dong, H., Song, K., Wang, Q., Yan, Y. & Jiang, P. Deep metric learning-based for multi-target few-shot pavement distress classification. *IEEE Trans. Indust. Inf.* **18**, 1801–1810. <https://doi.org/10.1109/TII.2021.3090036> (2022).
- Li, Y., Sun, S., Song, W., Zhang, J. & Teng, Q. Crackyolo: Rural pavement distress detection model with complex scenarios. *Electronics*[SPACE]<https://doi.org/10.3390/electronics13020312> (2024).
- Liang, X., Yu, X., Chen, C., Jin, Y. & Huang, J. Automatic classification of pavement distress using 3d ground-penetrating radar and deep convolutional neural network. *IEEE Trans. Intell. Transp. Syst.* **23**, 22269–22277. <https://doi.org/10.1109/TITS.2022.3197712> (2022).
- Liu, Z. et al. Intelligent assessment of pavement structural conditions: A novel femvit classification network for gpr images. *IEEE Trans. Intell. Transp. Syst.* **25**, 13511–13523. <https://doi.org/10.1109/TITS.2024.3403144> (2024).
- Yuan, Y., Chen, J., Lang, H. & Lu, J. J. Exploring the efficacy of sparse feature in pavement distress image classification: A focus on pavement-specific knowledge. *Appl. Sci. Basel*[SPACE]<https://doi.org/10.3390/app13189996> (2023).
- Liu, F., Liu, J., Wang, L. & Al-Qadi, I. L. Multiple-type distress detection in asphalt concrete pavement using infrared thermography and deep learning. *Autom. Constr.*[SPACE]<https://doi.org/10.1016/j.autcon.2024.105355> (2024).
- Eslami, E. & Yun, H.-B. Attention-based multi-scale convolutional neural network (a plus mcnn) for multi-class classification in road images. *Sensors*[SPACE]<https://doi.org/10.3390/s21155137> (2021).
- Cai, W., Li, M., Jin, G., Liu, Q. & Lu, C. Comparison of residual network and other classical models for classification of interlayer distresses in pavement. *Appl. Sci. Basel*[SPACE]<https://doi.org/10.3390/app14156568> (2024).
- Yang, Q. & Zhou, S. Identification of asphalt pavement transverse cracking based on vehicle vibration signal analysis. *Road Mater. Pavement Des.* **22**, 1780–1798. <https://doi.org/10.1080/14680629.2020.1714699> (2021).
- Moghadas Nejad, F. & Zakeri, H. An expert system based on wavelet transform and radon neural network for pavement distress classification. *Expert Syst. Appl.* **38**, 7088–7101. <https://doi.org/10.1016/j.eswa.2010.12.060> (2011).
- Qin, Z., Zhang, P., Wu, F. & Li, X. Fcanet: Frequency channel attention networks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 763–772. <https://doi.org/10.1109/ICCV48922.2021.00082> (2021).
- He, Y., Chen, H., Liu, D. & Zhang, L. A framework of structural damage detection for civil structures using fast fourier transform and deep convolutional neural networks. *Appl. Sci.*[SPACE]<https://doi.org/10.3390/app1199345> (2021).
- Hu, J., Shen, L., Albanie, S., Sun, G. & Wu, E. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 2011–2023. <https://doi.org/10.1109/TPAMI.2019.2913372> (2020).
- He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778, <https://doi.org/10.1109/CVPR.2016.90> (2016).
- van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
- Du, H. et al. Mf-mner: Multi-models fusion for mner in Chinese clinical electronic medical records. *Interdiscipl. Sci. Comput. Life Sci.* **16**, 489–502. <https://doi.org/10.1007/s12539-024-00624-z> (2024).
- Sahiner, B., Chen, W., Pezeshk, A. & Petrick, N. Comparison of two classifiers when the data sets are imbalanced: The power of the area under the precision-recall curve as the figure of merit versus the area under the roc curve. In Kupinski, M. & Nishikawa, R. (eds.) *Medical Imaging 2017: Image Perception, Observer Performance, And Technology Assessment*, vol. 10136 of *Proceedings of SPIE*, <https://doi.org/10.1117/12.2254742> (SPIE; Alpin Med Syst, 2017). Conference on Medical Imaging - Image Perception, Observer Performance, and Technology Assessment, Orlando, FL, FEB 12-13, 2017.
- Rainio, O., Teuho, J. & Klén, R. Evaluation metrics and statistical tests for machine learning. *Sci. Rep.* **14**, 6086. <https://doi.org/10.1038/s41598-024-56706-x> (2024).
- Efron, B. *Bootstrap Methods: Another Look at the Jackknife* 569–593 (Springer, 1992).
- Uno, K., Shimizu, I., Murakami, H. & Umezu, N. Detection of image differences by Fourier transformed magnitude subtraction. *Optik* **123**, 1779–1785. <https://doi.org/10.1016/j.jleo.2012.01.014> (2012).

28. Johnson, T. H., Lhamo, Y., Shi, L., Alfano, R. R. & Russell, S. Fourier spatial frequency analysis for image classification: training the training set. In Farkas, D., Nicolau, D. & Leif, R. (eds.) *Imaging, Manipulation, And Analysis Of Biomolecules, Cells, And Tissues IX*, vol. 9711 of *Proceedings of SPIE*, <https://doi.org/10.1117/12.2212934> (SPIE, 2016). Conference on Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues IX, San Francisco, CA, FEB 16-17, 2016.
29. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2921–2929, <https://doi.org/10.1109/CVPR.2016.319> (2016).
30. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269, <https://doi.org/10.1109/CVPR.2017.243> (2017).
31. Howard, A. G. *et al.* Mobilenets: Efficient convolutional neural networks for mobile vision applications. <http://arxiv.org/abs/1704.04861> (2017).
32. Elghaish, F. *et al.* Deep learning for detecting distresses in buildings and pavements: a critical gap analysis. *Constr. Innov. Engl.* **22**, 554–579. <https://doi.org/10.1108/CI-09-2021-0171> (2022).
33. Manjusha, M. & Sunitha, V. A review of advanced pavement distress evaluation techniques using unmanned aerial vehicles. *Int. J. Pavement Eng.* **24**, 2268796. <https://doi.org/10.1080/10298436.2023.2268796> (2023).

Author contributions

All the authors contributed extensively to the manuscript. W.W. contributed to study design, software experiments and manuscript drafting. Z.L. and F.Z. provided scientific guidance and modification suggestions. Xue Li, Xiaowei Li, and J.W. contributed to the data analysis, proofreading, and editing of the manuscript. All authors approved the manuscript.

Funding

This work was supported by the Key-Area Research and Development Program of Guangdong Province under Grant 2021B0101420001.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025