Engineering Applications of Artificial Intelligence and Machine Learning

pixel-level crack detection

[a] *School of computer science and organisationn, VIT University*
Amaravati University,.MAHIMALURU ACHYUTH / 21BCE9535

ARTICLEINFO

Keywords:
Crack detection
Transformer
Pixel-level
Semantic segmentation
Encoder-decoder

1.ABSTRACT

One of the most prevalent structural distresses in pavement is surface cracking, which compromises the pavement's sustainability and usability. The development of computer vision-based models (such as image processing or deep learning-based) for the autonomous detection of pavement surface cracks has received a lot of attention during the last ten years. However, real-world image data captured by a linear array charge-coupled device (CCD) camera, which typically has high resolution and few crack pixels per image, differs significantly from public image data captured by phones or other portable devices. We provide a novel two-stage framework for automatic pavement surface detection at the pixel level, which aims to increase the efficiency and accuracy of pavement surface fracture detection in engineering practice. During phase I, A classification network based on convolutional neural networks (CNNs) is used to identify the last pixel cracks in the photos.

.

2. Introduction

Surface crack is one of the most common pavement distresses, unfavorably impacting pavement structural health and bringing potential safety risks to the public. Thus, the periodic survey is mandatorily ordered by transportation management agencies in the maintenance plan. Currently, pavement surveys still heavily rely on experienced crews, which make pavement surface crack detection tedious and expensive. To efficiently and accurately detect the pavement surface crack without subjective intervention and reduce the high cost on both of the human and time, computer vision-based systems integrating image processing or deep learning-based models have been developed and adopted in practice

Reviewing the development of crack detection techniques, image processing attracts the attention in the early stage since they are computationally efficient and economical . Generally, it includes two categories which are threshold-based and edge-based methods, respectively (). Threshold-based methods (e.g., global or local thresholding) separate pavement surface image pixels into the background and crack by setting proper threshold values. For instance, global thresholding methods apply a single value to separate the intensity distributions of the foreground and the background with significant differences. In most cases, the acquired image data doesn't have manifest contrast, and thus it is impossible to find an optimal threshold to conduct the pixel separation. Alternatively, the local thresholding approaches which can better adapt to the changed road environment have been developed. However, the characteristics of high noise-sensitivity and massive false detections limit its wide application.

Edge detection models, including Sobel edge , Canny edge , Roberts edge , etc., are widely used to detect the pavement surface crack. Taking advantage of intensity changes of color, texture, light, etc., edge detection models focus on the linear features corresponding to crack boundaries to make detection results present

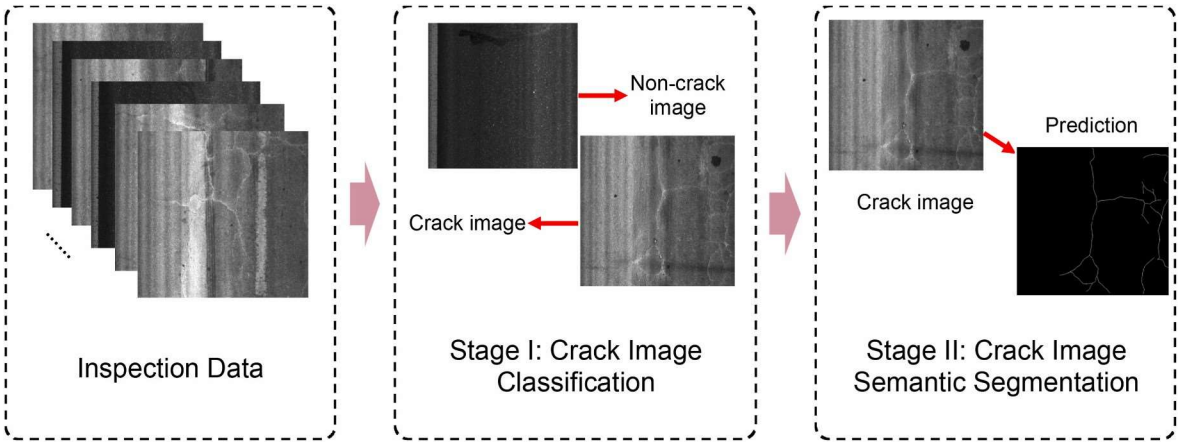


Fig. 1. Workflow of this study.

discontinuities of two regions. In practice, edge detectors cannot detect thin and smooth cracks due to inapparent feature transitions between two regions. In addition, when there are noisy images, complex parameter configuration operations for each image are needed, limiting its engineering application with massive image data.

Inspired by the application of deep learning models such as object detection and semantic segmentation, many studies) have contributed to detecting the pavement surface crack in an end-to-end manner. Unlike image processing-based approaches, deep learning models automatically learn multi-level features and are not sensitive to low-contrast images that can easily cause crack detection failure. Regarding the application using object detection models (e.g., YOLO (), Faster R–CNN), etc.), the crack is classified, and its location is provided. However, the object detection-based model only presents the pavement crack in a bounding box and cannot depict the exact contour of the crack, restricting the calculation of crack indexes (e.
g., width and length). Semantic segmentation models such as fully convolutional network (FCN) (Long et al., 2015) and UNet (Ronneberger et al., 2015) are therefore explored to present the pavement crack at the pixel level. With the architecture of the encoder-decoder, multi-scale feature maps are utilized, and meanwhile, high precision and recall results are generated.

In real-world applications, most computer vision-based pavement crack inspection vehicles are equipped with multiple high-resolution linear array charge-coupled device (CCD) cameras, generating a large volume of image data. For instance, the inspection vehicle applied in this study can generate 20 G image data per kilometer at 60 km/h on the G342 national highway for processing. It needs to be mentioned that most of them are crack-free, or the crack pixels occupy a very small ratio in the image. The inclusion of redundant data, such as images without cracks, not only consumes excessive computational resources but also extends model training time and costs without contributing significantly to model performance. Furthermore, this inclusion may lead the trained model to overly prioritize the crack-free regions, potentially resulting in overfitting and a notable decline in performance during real-world pavement surface crack inspections.

To address this issue, we propose a novel two-stage framework that unifies the advantage of CNN on efficiently classifying crack data and the second version of crack transformer (CTv2) featuring a neck structure for more accurate pavement surface crack detection. Fig. 1 illustrates the research flow of this study. In this work, the proposed two- stage framework is applied to the acquired image dataset collected from in-service sections of the G342 national highway and S318 provincial highway. It presents a superior performance in the classification, localization, and detection of the pavement surface crack. The main contributions can be concluded as follows.

- A novel two-stage framework including the CNN-based and transformer-based subnetwork for accurate and efficient pavement surface crack detection is proposed.
- A novel strategy named separation-combination has been proposed to improve the training efficiency and crack detection accuracy with the high resolution image.
- A comprehensive investigation using three public pavement image datasets has been conducted, showing the superior performance of the proposed framework.

The rest of this paper is organized as follows: Related work presents representative studies of image processing-based and deep learning- based models for pavement crack surface detection. Proposed network describes the details of our developed two-stage framework including the classification network and the semantic segmentation network. Experiments and discussion show the detailed information on the system configuration, model training results, visualization results, and discussion. Concluding remarks give the conclusions of this work.

3. Literature Survey

In this section, we briefly review the development of pavement surface crack detection technologies including image processing-based methods and deep learning-based approaches.

Image processing-based pavement surface crack detection

Image processing-based pavement surface crack detection methods can be mainly classified into two categories: intensity thresholding and edge detector (Zakeri et al., 2017; Kheradmandi and Mehranfar, 2022; Hsieh and Tsai, 2020). Intensity thresholding referred to classifying the pixels of a given pavement surface image into groups of crack pixels and background. To make the prediction smooth and reliable, global thresholding and local thresholding methods were developed. In (Li et al., 2011), the neighboring difference histogram (NDHM) was introduced with the optimized global threshold. Akagic et al. (2018) developed a strategy to compute the histogram and Otsu's thresholding of each sub-image for better performance on crack segmentation. Quan et al. (2019) designed an improved Otsu method through the modification of the probability weighting factor of the gray histogram. However, since collected images often exist of heavy noises, single global thresholding fails in the prediction of complex pavement crack patterns.

Compared to global thresholding methods, local thresholding methods can provide reasonable thresholding values in road environments with large variabilities. Oliveira and Correia (2009) applied double dynamic thresholding for entropy computation and the identification of image blocks containing crack pixels. In (Wang and Tang, 2011), the local optimal threshold with multi-scale features was developed, showing more effective performance than the global thresholding method. Zhang et al. (2017a) proposed an adaptive

thresholding method by taking the spatial distribution, intensities and geometric features of the pavement surface crack into consideration. Beyond that, other studies also combined two thresholding, or triple thresholding approaches for better prediction.

Unlike thresholding methods, Edge detectors can extract crack edges by using intensity changes of light, color, texture, etc., of crack and crack-free regions. In (the Sobel edge detector was applied after the image pre-processing by using the bidimensional empirical model decomposition (BEMD) for pavement crack detection. Zhu (2016) utilized the median filter to smooth the crack edge on the gray level image and applied the image enhancement to improve the prediction performance. With the Mallat wavelet transform and quadratic optimization of the genetic algorithm, Zhao et al. (2010) improved the Canny edge detector on the detection of weak edges of cracks. Especially, to achieve satisfactory performance on the detection of the pavement surface crack, the optimization process undoubtedly introduced an amount of effort.

2.2. Deep learning-based pavement surface crack detection

Benefiting from the great success of deep learning techniques, object detection or semantic segmentation-based models were employed to detect the pavement crack at the block or pixel level). The issues including low contrast and intensity difference which make image processing suffering can be addressed in the "black box" of the deep learning models, greatly improving the detection accuracy. At the block level, YOLO family models were preferred to detect the pavement surface crack due to its balanced performance on accuracy and inference speed). reported prediction results using the YOLO network and its variants. Even they can provide accurate location and category information of different cracks, the crack indexes such as length and width were difficult to derive.

Motivated by the successful segmentation of biomedical images by using U-Net,) proposed CrackU-Net to detect the pavement surface crack at the pixel level. Based on the evaluation metrics, it showed better performance than the traditional image processing techniques and FCN. developed the ShuttleNet to provide robust semantic segmentation results on pavement surface crack prediction via shortcut connections between successive encoding-decoding rounds. On the basis of CrackNet designed CrackNet-V with invariant spatial sizes for improved prediction results at the pixel level. Sun et al. (2022) put a multi-scale attention module in the decoder of DeepLabv3+, which can assign reasonable weights to different feature maps, improving the crack detection accuracy. Combing the infrared thermography (IRT) and the CNN, trained and evaluated five classical CNN segmentation models and two UNet-based models on RGB images, infrared images, and fused images.

Inspired by the great success of transformers in general computer vision tasks, our previous model, CT), can accurately detect pavement surface cracks by integrating the Swin Transformer as the encoder and the Segformer as the decoder. In CrackFormer (the hybrid-window attentive vision transformer was proposed, employing a hybrid-window-based self-attention scheme and a weighted multi-head self-attention philosophydeveloped LETNet by creating a convolutional stem and a local enhancement module to improve the local crack feature perception ability in transformer architecture.) introduced SwinCrack to perform pavement surface crack detection by replacing linear parts of the Swin Transformer and integrating the convolutional attention-gated skip connection. While transformer-based methods achieved high accuracy and demonstrated superior long-range crack pixel modeling ability, there were few reports on processing large-scale pavement surface crack image data.

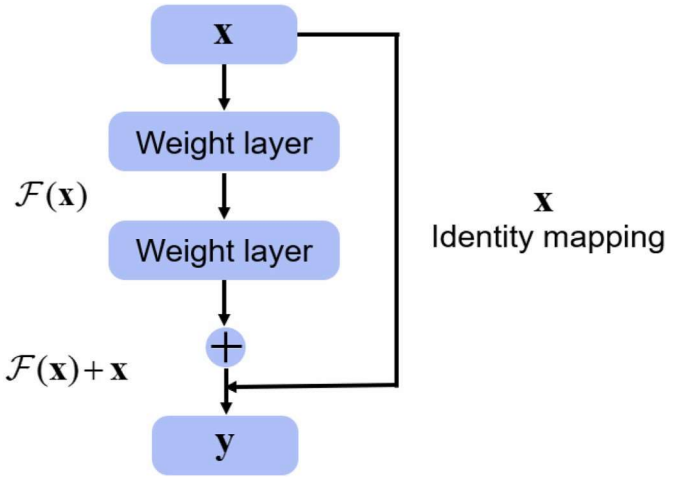


Fig. 2. The structure of the building block in the ResNet.

To sum up, current semantic segmentation-based networks can greatly improve the detection accuracy of pavement surface crack and depict the contour and shape of the crack at the pixel level. However, it needs to be mentioned that the efficient framework to process the large volume of inspection data containing massive crack-free images is still rare in practice.

4. Proposed / My Work :

In this section, we present the detailed design of the proposed two- stage framework for pavement surface crack detection. In the first stage, the survey data is fed into a CNN for separating crack image and crack-free image. Once the image is identified as the crack image, it would be fed into the transformer-based network for pavement surface crack detection at the pixel level. The details of the CNN for classification and the transformer-based semantic segmentation network are successively introduced.

ResNet developed by He et al. (2016) in 2016, is one of the most successful classification networks or backbones in the network design applied in multiple vision tasks. Through residual learning with the shortcut connections design, the issue of performance degradation along with the deep model is solved. To achieve robust classification performance, the ResNet-34 which has 34 layers, is adopted in stage I to conduct the crack image classification. The reason we choose ResNet-34 rather than ResNet-18 or ResNet-50 is that it has better classification accuracy, and the training performance is reported in the section of Experiments and Discussion.

Unlike the VGGNet (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al., 2015), etc., ResNet learns residuals rather than the original features. The motivation of learning from residuals is that it can be easier to optimize the residual learning rather than the original learning. Meanwhile, since the proposed identity shortcut connections simply conduct addition operation, no superfluous computational complexity is added. Equations (1) and (2) present the direct mapping and identity mapping, respectively. Fig. 2 illustrates the building block in the ResNet.

$$\mathbf{y}=\mathbf{F}(\mathbf{x},\{W_i\}) \quad (1) \qquad \mathbf{y}=\mathbf{F}(\mathbf{x},\{W_i\})+\mathbf{x} \quad (2)$$

where $\mathbf{x}$ and $\mathbf{y}$ are the input and output vectors of a layer, $\mathbf{F}(\mathbf{x},\{W_i\})$ denotes the residual learning operation.

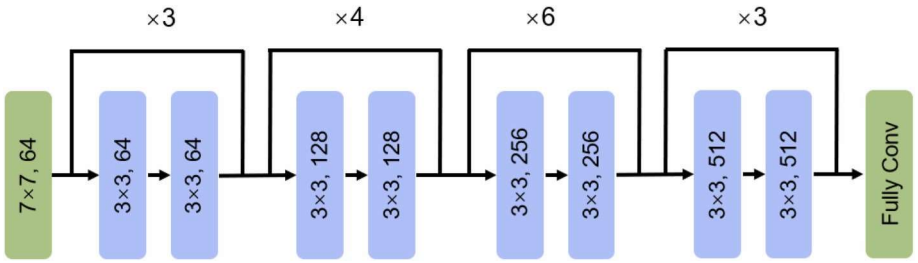


Fig. 3. The architecture of ResNet-34.

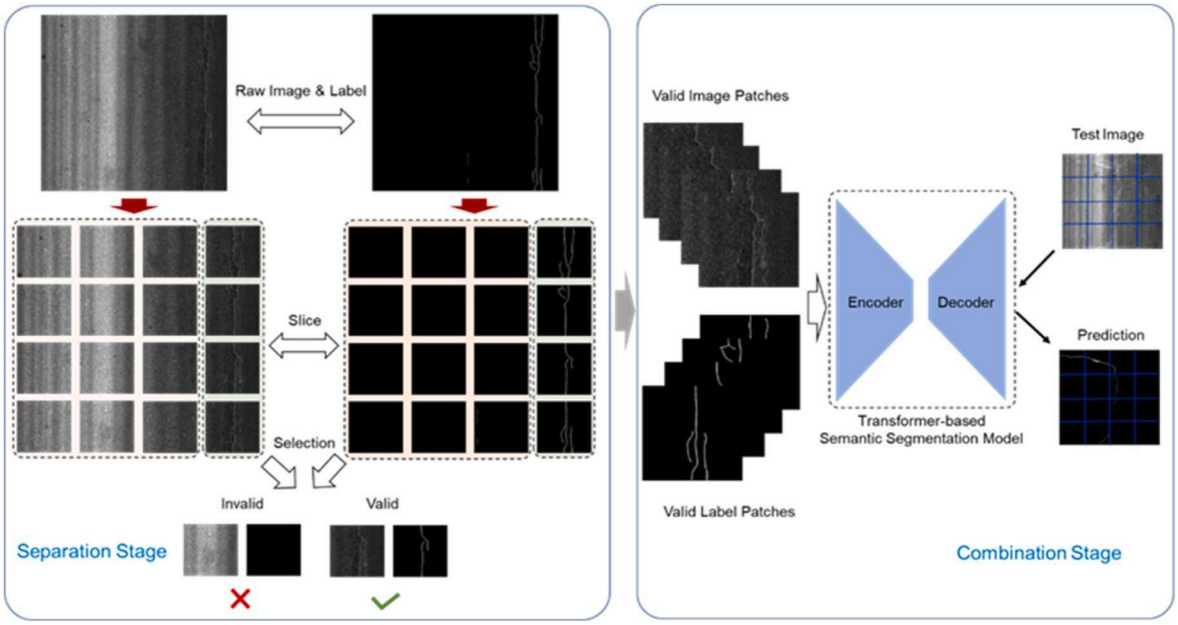


Fig. 4. Illustration of the proposed separation-combination strategy.

The acquired image has a resolution of 2000 (H) × 2048 (W) which is cropped to 224 × 224 for training convenience. There are five convolutional stages in the ResNet-34. The convolutional layer in the first stage contains the 7 × 7 filter with 64 channels. After the first stage, the output size is 112 × 112. From the second stage to the fifth stage, the residual blocks are repeated 3, 4, 6, and 3 times, respectively. Each convolutional layer contains the 3 × 3 filter. The channel numbers are 64, 128, 256, and 512, respectively. The output size is 56 × 56, 28 × 28, 14 × 14, 7 × 7, respectively. Fig. 3 shows the architecture of ResNet-34.

3

Compared to traditional CNN models, the transformer-based network (Yuan et al., 2021; Zhang et al., 2022b) is notable for modeling the long-range dependencies by using the attention mechanism in the high-resolution image. In general, the attention module allows the network to focus on the pavement crack by assigning different weights to different parts of the pavement crack image. With a high relevance to the crack, the weight score is higher than the background part and thus benefits the pavement crack detection. In this study, we design CTv2 using the Swin Transformer (Liu et al., 2021) as the encoder, feature pyramid network (FPN) (Yang et al., 2019) as the neck

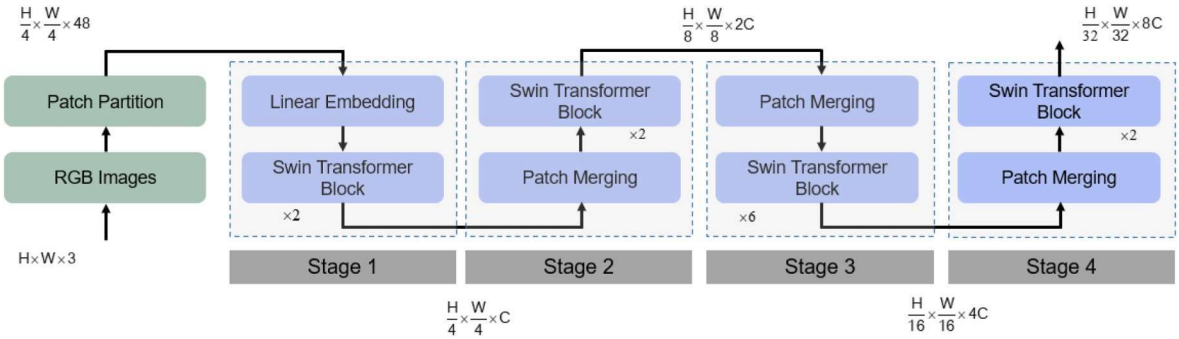


Fig. 5. The structure of encoder.

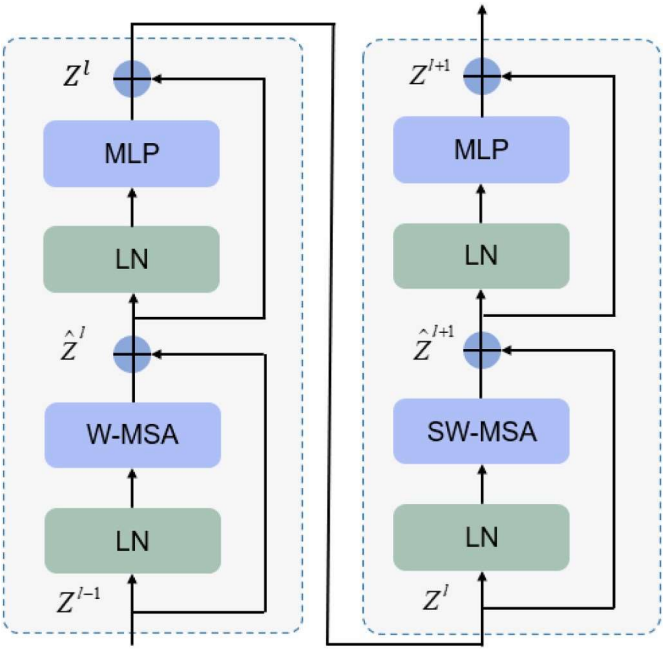


Fig. 6. The structure of the Swin Transformer block.

and the Segformer (Xie et al., 2021) head as the decoder to detect the pavement surface crack in the complicated survey data. To alleviate the impact brought by the severe data imbalance, we develop a separation-combination strategy throughout the semantic segmentation training process. The detailed steps of this proposed strategy are explained in Fig. 4. Specifically, in the separation stage, the acquired images are classified into the valid image (with crack) and the invalid image.

(without crack). Since the raw image is high resolution, the slice operation aiming to save the computational cost is performed. With the trained classification model, the valid image and invalid image can be separated. In the combination stage, the valid imaged are fed into the proposed encode-decoder structure to complete the model training. With the trained semantic segmentation model, newly acquired images can be tested for crack detection. Notably, the test images are sliced first, and all patches are numbered for recovering the whole image after the crack detection. Since the training data is still highly limited, the trained model is only proven effective on the inspection images acquired in the sections with similar conditions.

In the encoder, as shown in Fig. 5, the input image is split into patches, and each patch size is 4 × 4. Similar to the channels in CNN, the concept of the feature dimension is developed and is denoted as C. Thus, the feature dimension of each patch is 4 × 4 × 3 = 48. A total of four stages are included in the architecture. Following the image input, it is the patch partition operation which splits the raw image into patches. Since the patch size is 4 × 4, a total of H/4 × W/4 patches are generated. With the feature embedding layers, the feature dimensions become H/4 × W/4 × 48. Unlike the rest of stages, stage I contains linear embedding and other stages include the patch merging layer. The Swin Transformer blocks successively repeat 2, 2, 6 and 2 times in each stage, respectively. In our implementation, C is defined as 96. Thus, with the patch merging and feature transformation, the total feature dimensions of the rest of stages are H/8 × W/8 × 2C, H/16 × W/16 × 4C, H/32 × W/32 × 8C.

Fig. 6 illustrates the structure of the two successive Swin Transformer blocks which play significant roles on feature extraction. In the Swin Transformer, the standard window-based multi-head self-attention (W- MSA) module and the shifted window-based multi-head self-attention (SW-MSA) module are included. Compared to W-MSA, SW-MSA has the better modeling power and can strengthen connections across windows. In addition, the layer norm (LN) and the multilayer perceptron (MLP) are utilized. The computation process in the Swin Transformer block can be referred to from Equations (3)–(6). More details of the Swin Transform can be referred to (Liu et al., 2021)

$$Z_l = \text{W} - \text{MSA}\left(\text{LN}\left(Z_{l-1}\right)\right) + Z_{l-1} \tag{3}$$

$$Z^l = \text{MLP}\left(\text{LN}\left(\hat{Z}^l\right)\right) + \hat{Z}^l \tag{4}$$

$$\hat{Z}^{l+1} = \text{SW} - \text{MSA}\left(\text{LN}\left(Z^l\right)\right) + Z^l \tag{5}$$

$$Z_{l+1} = \text{MLP}\left(\text{LN}\left(Z_{l+1}\right)\right) + Z_{l+1} \tag{6}$$

where $\hat{Z}^l$ is the output feature of the W-MSA module. $\hat{Z}^{l+1}$ is the output feature of SW-MSA module. Z^{l-1} is the input feature. Z^l and Z^{l+1} are the output features of MLP.

In the decoder part, the neck and the decoder head are incorporated. In the neck part, the feature pyramid network (FPN) (Lin et al., 2017) is utilized for feature fusion with multiple scales generated in the encoder. Following our previous design (Guo et al., 2023), the decoder head of Segformer (Xie et al., 2021) is adopted since it can accelerate the inference speed with the simple and lightweight design. The architecture of the neck and decoder can be found in Fig. 7. The detailed information of the decoder can be referred to (Guo et al., 2023). In specific, the neck part follows the top-down fashion. The input channels are 96, 192, 384, and 768, respectively. While, based on the experiments results, all output channels of the neck are set to 96. In the decoder part, the lightweight design with all multilayer perceptron (MLP) is adopted. Under this manner, the output dimension is 56 × 56 × 96. As shown in Fig. 7, after the feature fusion operation gathers global and local information, the background and crack pixels can be classified. Inspired by the application of dice loss (Sudre et al., 2017) in the medical image segmentation, we employ the dice loss in the decoder.

5. Results and Discussion

In this section, the details of the system configuration and training implementation are first introduced. Afterwards, the data preparation, performance indicators, training results and visualization results are described. In the last, we present the crack detection results using different models.

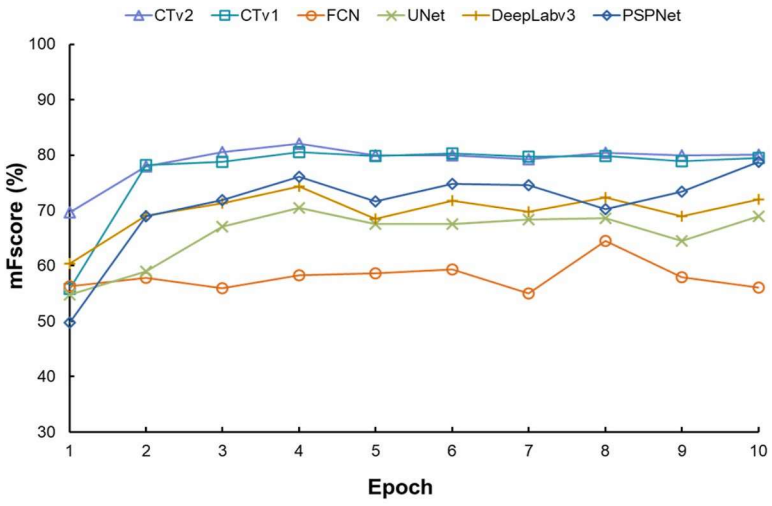
4.1. System configuration and training implementation

The Pytorch library with a version of 1.12.0 is adopted for model training. A deep learning machine with an NVIDIA 3080 Ti graphics processing unit (GPU) and the operating system of Ubuntu 20.04 LTS are prepared for the hardware and software configuration. In specific, the deep learning machine is with 16 G RAM and i7-CPU @ 3.5Hz. The first stage is responsible for the classification of crack and crack-free images from the survey data. Three classification models include ResNet-18, ResNet-34, and ResNet-50 are used for training and evaluation. The second stage implemented on the MMSegmentation (Contributors, 2020) accounts for crack detection at the pixel level. A total of six models including UNet (Ronneberger et al., 2015), FCN (Long et al., 2015), DeepLabv3 (Chen et al., 2017), PSPNet (Zhao et al., 2017), CTv1 (Guo et al., 2023), and the developed CTv2 are included. The hyperparameters and settings of the first stage can be referred to Table 1. To accelerate the training process of semantic segmentation models, each input image is cropped to 256 × 256 with overlapped regions. The hyperparameters of the second stage adopt the default settings.

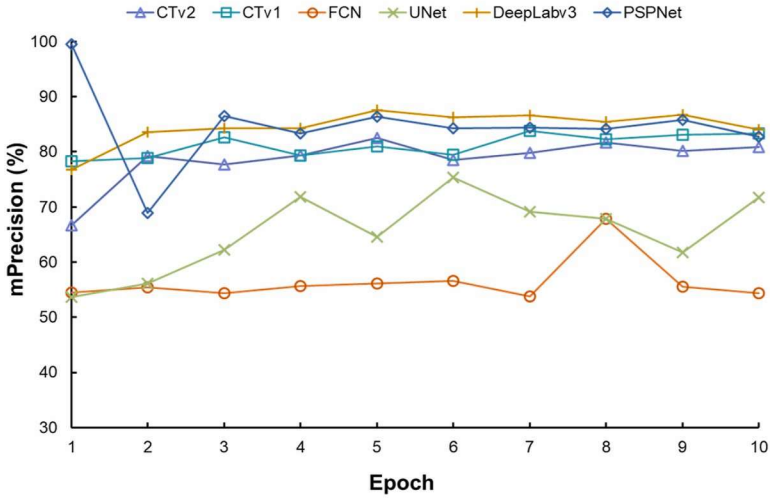
4.2. Data preparation

To fairly compare the detection performance of the proposed

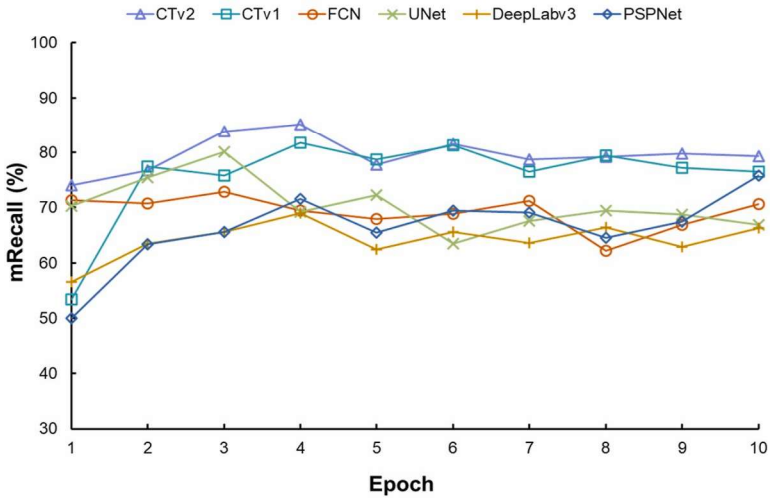
Table 1



(a)



(b)



(c)

Fig. 9. Evaluation results of different models in the second stage using CrackSD dataset (a) mFscore; (b) mPrecision; (c) mRecall.

Table 3 Evaluation results of different models on CFD in the second stage.

Model	mFscore (%)	mPrecision (%)	mRecall (%)
CTv2	**94.69**	95.14	94.24
CTv1 (Guo et al., 2023)	94.60	92.94	**96.41**
FCN	86.13	88.54	84.02
UNet	89.27	96.84	83.86
DeepLabv3	89.15	**97.21**	83.5
PSPNet	90.09	96.36	86.14

Table 4
Evaluation results of different models on CrackSC in the second stage.

Model	mFscore (%)	mPrecision (%)	mRecall (%)
CTv2	**92.23**	92.55	**91.93**
CTv1 (Guo et al., 2023)	90.01	90.09	89.93
FCN	82.09	88.23	79.08
UNet	86.09	91.67	81.88
DeepLabv3	86.50	92.95	81.81
PSPNet	86.04	**93.49**	80.92

Figs. 11 and 12. In the survey data, the cracks are not evident, making the labeling work difficult. With the CTv2 model, we can find its effective modeling performance on long and complicated Cracks. However, in the representation of fine features, it still has room to improve. As for the rest of the models, CTv1 and PSPNet are qualified on the prediction of images 1 and 2. In addition, CTv1 has the similar result on image 3 as CTv2. However, the rest of models fail in the prediction of image 3. FCN, UNet and DeepLabv3 cannot predict successive and long crack patterns. Meanwhile, they have poor performance on image 3. Notably, even with the complex survey data, CTv2 can still predict most of the crack pixels and outperforms other models.

Figs. 11 and 12 present the prediction results on the CFD and CrackSC datasets, respectively. In Fig. 11, with image 1, we find that most of the trained models can predict the long and crossed cracks well. Meanwhile, CTv2 can predict fine details but other models fail to capture them. For instance, at the bottom of image 1, there is a transversal crack connecting the longitude cracks. Only CTv2 and CTv1 can retrieve these cracks, and other models predict nothing or predict protrusions of longitude cracks. With image 2, we can find there are long and thin cracks with bifurcations at the top of the pattern. CTv2 has the closest

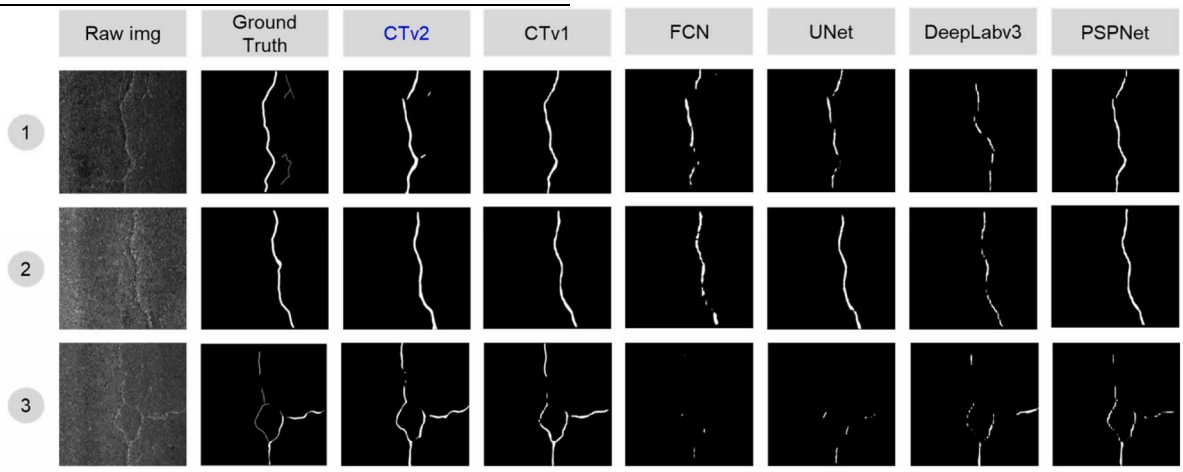


Fig. 10. Visualization results of different models in the second stage using CrackSD dataset.

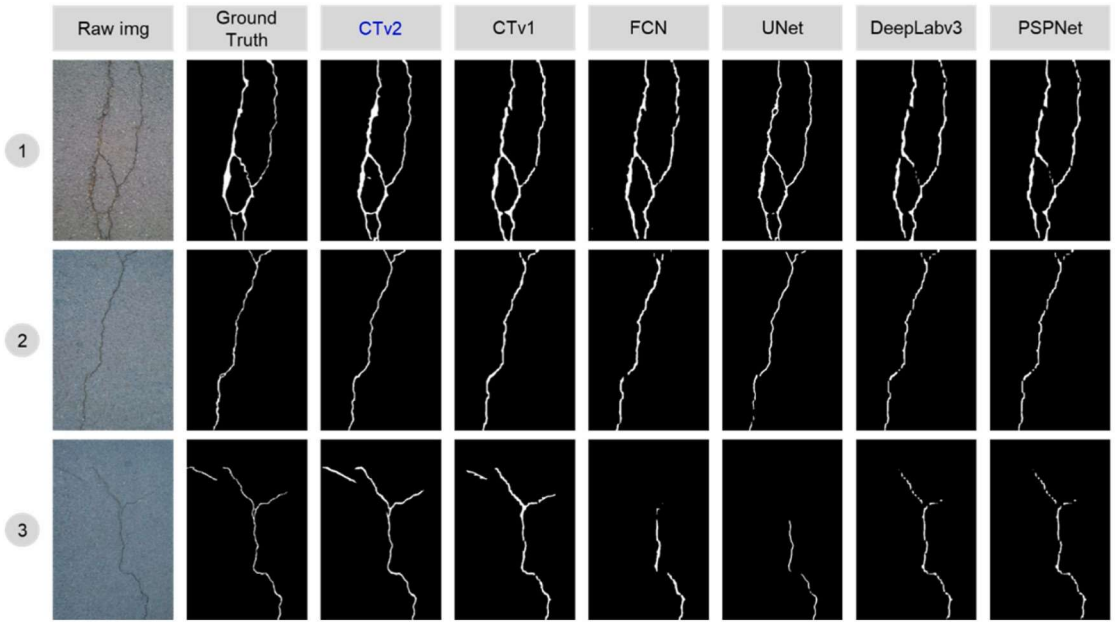


Fig. 11. Visualization results of different models in the second stage using CFD dataset.

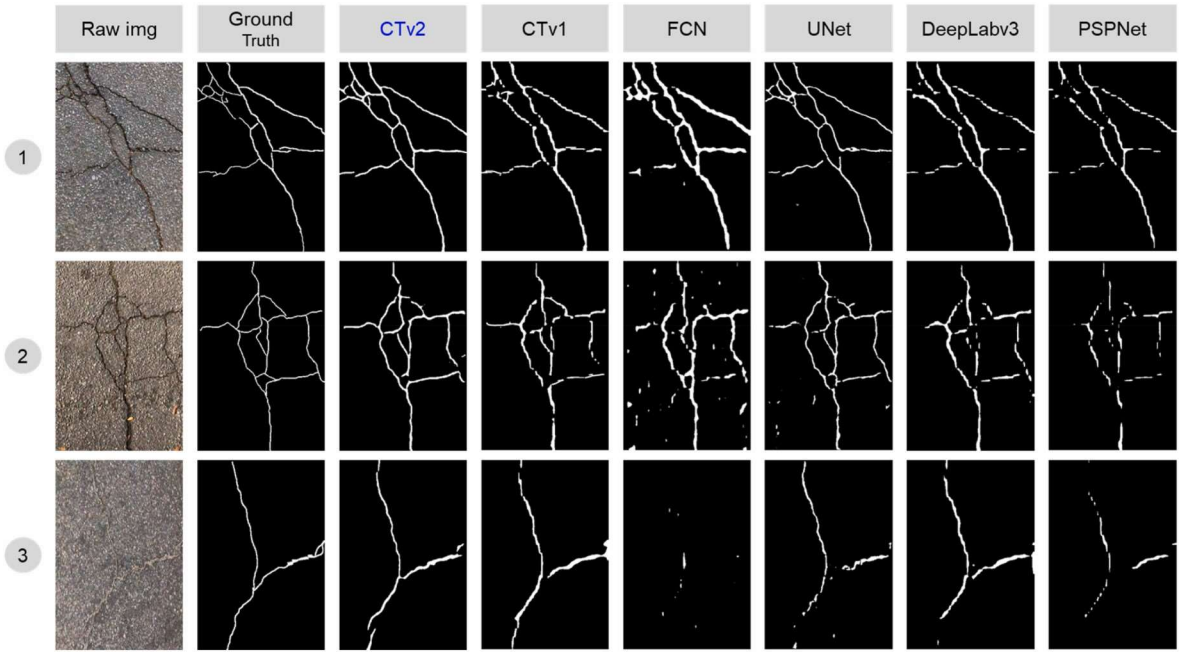


Fig. 12. Visualization results of different models in the second stage using CrackSC dataset.

results to the ground truth label, and other results show discontinuous segments. As for the prediction results of image 3. We can find the raw image and ground truth label have separate cracks. CTv2 and CTv1 can predict them at different locations, but other models only predict part of the whole crack leaving the left part of the crack without getting predicted. Taking advantage of the neck part, it can be found that CTv2 can predict more fine details than CTv1.

Fig. 12 reveals the visualization results of different models in the second stage using the CrackSC dataset. Compared to CFD, CrackSC is more challenging on crack patterns. Images 1 and 2 are connected patterns that make predictions challenging. Image 3 indicates shallow cracks which represent crack features that are similar to the texture of the surroundings. On the one hand, cracks are thin. On the other hand, the depths of these cracks are shallow. In predictions of images 1 and 2, CTv2 still outperforms other models and shows superior modeling performance on complicated cracks. Especially, compared to CTv1, we can find that it can predict the whole crack pattern with fewer disconnections.

With respect to image 2, FCN shows many white dots, indicating it generates numerous predictions of false crack pixels. Meanwhile, other models produce isolated dots rather than a closed pattern. Even though CTv1 has the similar result to CTv2, it fails to connect the isolated crack lines. As for the prediction results of image 3, only CTv1 can compete with CTv2, indicating the superior capacity of the transformer-based model on long crack predictions. Apparently, it can be found that cracks in image 3 are difficult to capture even with human eyes, showing the potential of using the transformer-based network to complete the challenging inspection work in practice. In addition, based on our experimental results, the inference speed of the proposed method can achieve 110 frames per second (FPS) in the pavement surface crack detection with a testing image resolution of 320×480. Compared to the requirement of the real time inference speed (60 FPS), the developed method can save half the time in practice.

6. Concluding / Future works

This work introduces a novel two-stage framework for pavement surface crack detection using public datasets and survey data. Specifically, the first and second stages are responsible for crack image selection and crack pixel prediction, respectively. In the first stage, three classification networks (i.e., ResNet-18, ResNet-34, and ResNet-50) are trained and evaluated on the survey data. With the conduction of the second stage, a strategy named separation-combination is developed to narrow the qualified crack patches to avoid training failure. In the second stage, the encoder-decoder structure is employed. The Swin Transformer is adopted as the encoder with superior performance on the modeling of long dependencies and hierarchical feature representation. The Segformer head is utilized as a decoder that can generate global prior to crack features with the FPN neck.

Two public image datasets (i.e., CFD and CrackSC) and one survey dataset (CrackSD) are trained, evaluated, and tested on six models, which are CTv2, CTv1, FCN, UNet, DeepLabv3, and PSPNet. Specifically, the CrackSD dataset is built based on the acquired images from the G342 national highway and S318 provincial highway sections. First-stage training operations follow the PyTorch version of ResNet. All training procedures for the second stage are completed within the MMSegmentation framework. Based on the first-stage training results, ResNet-34 achieves the maximum training accuracy on the CrackSD dataset. Regarding the second stage training results, the developed CTv2 reaches the best performance on both of the evaluation metrics (i.e., mFscore, mPrecision and mRecall) and visualization results, proving the effectiveness and robustness of our proposed two-stage framework on pavement surface crack detection.

Future work will concentrate on performance improvement of the pavement surface crack detection with the survey data. Image enhancement techniques will be developed to enlarge the differences between the background and the crack pixels. More data will be labeled, and semi-supervised learning models will be developed to reduce the reliance on the labeled data to complete the large-scale engineering application. Overall, the proposed framework shows great potential for pavement surface crack detection and will accelerate the development of transformer-based networks on crack predictions.

.

Acknowledgement

This work is partially by the Shanghai Key Laboratory of Rail Infrastructure Durability and System Safety (K202202), the National Natural Science Foundation of China (52308457) and the Natural Science Foundation of Shandong Province (ZR2023QE220). The opinions presented in this paper

solely reflect those of the authors and do not represent any opinions of funding agencies.

7 .References

Ahmadi, A., Khalesi, S., Golroo, A., 2021. An integrated machine learning model for automatic road crack detection and classification in urban areas. Int. J. Pavement Eng. 1–17.

Akagic, A., Buza, E., Omanovic, S., Karabegovic, A., 2018. Pavement crack detection using Otsu thresholding for image segmentation. 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE, pp. 1092–1097.

Ayenu-Prah, A., Attoh-Okine, N., 2008. Evaluating pavement cracks with bidimensional empirical mode decomposition. EURASIP Journal on Advances in Signal Processing 1–7, 2008.

Bao, P., Zhang, L., Wu, X., 2005. Canny edge detection enhancement by scale multiplication. IEEE Trans. Pattern Anal. Mach. Intell. 27 (9), 1485–1490.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation arXiv preprint arXiv:1706.05587.

Contributors, M., 2020. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. Availabe online: https://github.com/open-mmlab/mmsegm entation. (Accessed 18 May 2022).

Dorafshan, S., Maguire, M., Qi, X., 2016. Automatic Surface Crack Detection in Concrete Structures Using OTSU Thresholding and Morphological Operations.

Du, Y., Pan, N., Xu, Z., Deng, F., Shen, Y., Kang, H., 2021. Pavement distress detection and classification based on YOLO network. Int. J. Pavement Eng. 22 (13), 1659–1672.

Fan, R., Bocus, M.J., Zhu, Y., Jiao, J., Wang, L., Ma, F., Cheng, S., Liu, M., 2019. Road Crack Detection Using Deep Convolutional Neural Network and Adaptive Thresholding. IEEE Intelligent Vehicles Symposium (IV), IEEE, pp. 474–479, 2019.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

Liu, F., Liu, J., Wang, L., 2022. Asphalt pavement crack detection based on convolutional neural network and infrared thermography. IEEE Trans. Intell. Transport. Syst.

Shi, Y., Cui, L., Qi, Z., Meng, F., Chen, Z., 2016. Automatic road crack detection using random structured forests. IEEE Trans. Intell. Transport. Syst. 17 (12), 3434–3445.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition arXiv preprint arXiv:1409.1556.

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer 240–248.