Computers in Industry 133 (2021) 103545

Contents lists available at ScienceDirect

Computers in Industry

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

Pixel-level tunnel crack segmentation using a weakly supervised annotation approach

Hanxiang Wang[a], Yanfen Li[a], L. Minh Dang[b], Sujin Lee[a], Hyeonjoon Moon[a,*]

[a] *Department of Computer Science and Engineering, Sejong University, Seoul, Republic of Korea*
[b] *Department of Information Technology, FPT University at Ho Chi Minh city, Vietnam*

ARTICLE INFO

Article history:
Received 6 May 2021
Received in revised form 10 September 2021
Accepted 13 September 2021
Available online 5 October 2021

Keywords:
crack segmentation
deep learning
weakly supervised learning
fully supervised learning
tunnel crack

ABSTRACT

Automatic crack detection plays an essential role in ensuring the safe operation of tunnels, which is also challenging work in reality. In this paper, an innovative framework, which combines the weakly supervised learning methods (WSL) and the fully supervised learning methods (FSL), is presented to detect and segment the cracks in the tunnel images. Firstly, a WSL-based segmentation network Crack-CAM is proposed to annotate the collected data instead of using the traditional manual annotation process. By applying the proposed E-Res2Net101 structure and tuning some hyper-parameters, an FSL-based method named DeepLabv3+ is optimized to enhance the segmentation performance. After the crack segmentation, the risk levels of the detected cracks are judged using a new evaluation metric. In addition, the mean error of the lengths, the mean widths, and the areas are calculated for different types of cracks. A crack dataset in tunnel scenes that contain 3,921,726 sub-images that are cropped from 521 raw images is built to demonstrate the effectiveness of the presented methods. Based on the proposed dataset, the modified DeepLabv3+ achieves the highest MIoU of 0.786 and the best F1 of 0.865. Besides, the proposed framework combining WSL methods (automatic data annotation) and the FSL methods achieved a performance comparable to the framework that is based on manual annotation and the FSL methods, which demonstrates the WSL-based Crack-CAM can label images correctly.

1. Introduction

In recent years, the number and the mileage of railway tunnels rapidly increased due to the large amount of traffic construction investment. At the same time, the cracks caused by improper construction, materials, and maintenance are also followed, which seriously affects the service function and service life of the tunnel. In a tunnel project, it is very momentous to detect and evaluate the surface cracks of tunnels over time. However, the traditional crack detection methods rely on manual inspections, which wastes a lot of human resources and time. Therefore, an effective crack segmentation system is urgently needed to overcome these shortcomings.

Due to the advantages of high efficiency and convenience, image processing technologies have been used more frequently to detect segmentation defects and cracks (Dang et al. 2018; Dorafshan et al., 2018; Li et al. 2021; Su et al. 2011). However, the tunnel surface is different from the general concrete pavement and building, and there is usually insufficient light intensity, low contrast, complex background texture, and more noise. These confounding factors on the tunnel images often lead to the traditional image processing technology not achieving the desired result. Some methods that are based on deep learning (DL) have recently made significant progress in computer vision-related tasks. The DL-based defect and crack segmentation application is mainly divided into two research methods: fully supervised learning method (FSL) (Liu et al. 2019; Ren et al. 2020; Song et al. 2019) and the weakly supervised learning method (WSL) (Chen et al. 2020; Dong et al. 2020; Zhu and Song 2020). Compared with the WSL method, the FSL method has a better segmentation effect, but it needs to spend much time with the process of labeling data. In this study, the WSL and the FSL methods are combined to segment the cracks, and the damage degrees of the detected cracks are evaluated using image post-processing.

The main contributions of this study are listed below.

- A novel network that is based on the WSL method is presented to annotate the dataset, which economizes the time and the cost of human power.

[*] Corresponding author.
E-mail address: hmoon@sejong.ac.kr (H. Moon).

https://doi.org/10.1016/j.compind.2021.103545
0166-3615/© 2021 Elsevier B.V. All rights reserved.

- A crack segmentation model based on the FSL methods is optimized to enhance the final segmentation accuracy.
- A crack dataset in tunnel scenes is established to assess the proposed crack inspection system.
- Finally, a risk assessment metric is introduced to evaluate the damaged condition of the detected cracks by length and mean width.

The rest of the paper is arranged as follows. Section 2 lists some recent studies from three distinct research methods. The detailed information of the established data is described in Section 3. Section 4 shows the proposed crack segmentation framework. In Section 4, some experiments are conducted to demonstrate the proposed methodology, and the advantages and the disadvantages are discussed in Section 5.

2. Related work

Some researchers have presently developed various defect segmentation applications that are based on image processing methods. For example, morphological segmentation based on edge detection (MSED) was proposed and compared with the opening top-hat operation (OTHO) method. The result suggests the MSED can obtain a better performance than the OTHO (Su et al. 2011). Six distinct methods based on edge detection were trained and tested using the same dataset, and experimental results show they can correctly localize most of the cracked pixels. Nevertheless, these methods generated negative noise in the output images. In particular, the methods based on edge detection have poor effectiveness in segmenting low-contrast images (Dorafshan et al., 2018).

Recently, the FSL-based image segmentation has been implemented in various fields especially in tunnel defect diagnosis systems. An automatic crack inspection framework, which is based on an improved DeepLabv3, was proposed to predict the crack segmentation on the tunnel images. Even though the presented method obtained a fast speed of 23 frames per second (FPS), the segmentation accuracy was affected by a small dataset (Song et al. 2019). In contrast, a network named DeepCrack was used for segment crack by learning the features from multiple levels. The network achieved a good Mean Intersection over Union (MIoU) of 85.9%, but the segmentation speed was slow (10fps) (Liu et al. 2019). In another study, U-CliqueNet was proposed to separate the crack sections from the input tunnel images. According to the comparison results, the performance of the presented network was higher than that of the other experimental networks in terms of the MIoU. However, the dataset used in this study only focuses on one type of crack (Li et al. 2020).

Elaborating annotation documents for various defects is time-intensive, and the weakly supervised learning methods were presented to tackle this problem. Zhu et al. detected and segmented the six classes of cracks under complex backgrounds using a weakly supervised network (Zhu and Song 2020). However, the main focus of this study was to compare the visual effects of the different networks, so a detailed analysis of the evaluation metrics for the segmentation was not reported. Chen et al. presented a defect segmentation system based on the weakly supervised learning approach (Chen et al. 2020). The accuracy was improved using the proposed method for both the classification and the segmentation tasks, but the time in the localization process should be further reduced. A patch-based weakly supervised crack segmentation method was introduced to alleviate the time-consuming problem caused by the fully supervised methods (Dong et al. 2020). Different from their work, a new network structure was presented as the backbone of the WSL method, and a K-Means clustering algorithm was added to provide more feature information for enriching the target area in CAM. In addition, we applied different algorithms (dense Conditional Random Field and random walk algorithm) in order to refine the generated synthetic label. In another work, a crack detection model based on the weakly supervised method was used to reduce the workload of the manual labelling process (Xu et al. 2020). The proposed network without pre-training achieved comparable results compared with the fully supervised networks, and the required size of the training samples was very small. However, the defect edge in the output image was not very clear.

In order to further assess the risk level of the detected cracks, some researchers have attempted to calculate the lengths and the widths of cracks. Xincong Yang et al. presented an approach for cracks morphological measurement to obtain the crack width, length, and topology (Yang et al. 2018). It has been verified that the measurement is effective for common cracks, but it has a limited performance on complex cracks. Similarly, crack skeleton extraction was used to obtain the length and the width of cracks, and the results show the predicted crack almost matches with the real cracks (Li et al. 2020). Even though the crack length and the width were predicted accurately in (Li et al. 2020; Yang et al. 2018), the corresponding risk level of each crack was not reported.

By reviewing previous works on the crack segmentation topic, the main research methods are separated into three categories, including conventional edge detection-based methods, the FSL-based methods, and the WSL-based methods. As mentioned above, the traditional methods that use the edge detection schemes have noise that exists in the output images. The FSL methods are limited to the time-consuming labelling process. The WSL methods cannot achieve a higher accuracy than the FSL methods. Therefore, a WSL approach is used to elaborate the annotation data in this research, and a fully supervised network is then used for crack segmentation to enhance the segmentation effect. Finally, the corresponding risk level for each segmented crack is evaluated by calculating some parameters of the cracks.

3. Proposed dataset

In this research, the data is acquired using a deep scanner truck with high-resolution night cameras and LED lights. The collected raw images after image stitching have variant resolutions, which range from 12, 614*2, 922 to 34, 473*2, 956. To better fit a neural network, a total of 521 tunnel images are cut into 3,921,726 sub-images of 224*224 with a step length of 100. After cutting the images, the obtained sub-images are manually validated, and the detailed information is described in Table 1. There are 75,623 sub-images with cracks among all of the cropped images, and the data is collected from four distinct tunnel locations (Bongsan, Deugseong,

Table 1
Detailed information of the proposed dataset from four different locations.

Location	Original image	GT-mask image	Sub-image with crack	Sub-mask with crack	Sub-image without crack
Bongsan	168	168	34,362	34,362	5,000
Deugseong	150	150	26,965	26,965	5,000
Gamcheon	112	112	6,405	6,405	5,000
Hwasan	84	84	7,891	7,891	5,000
Total images	514	514	75,623	75,623	20,000

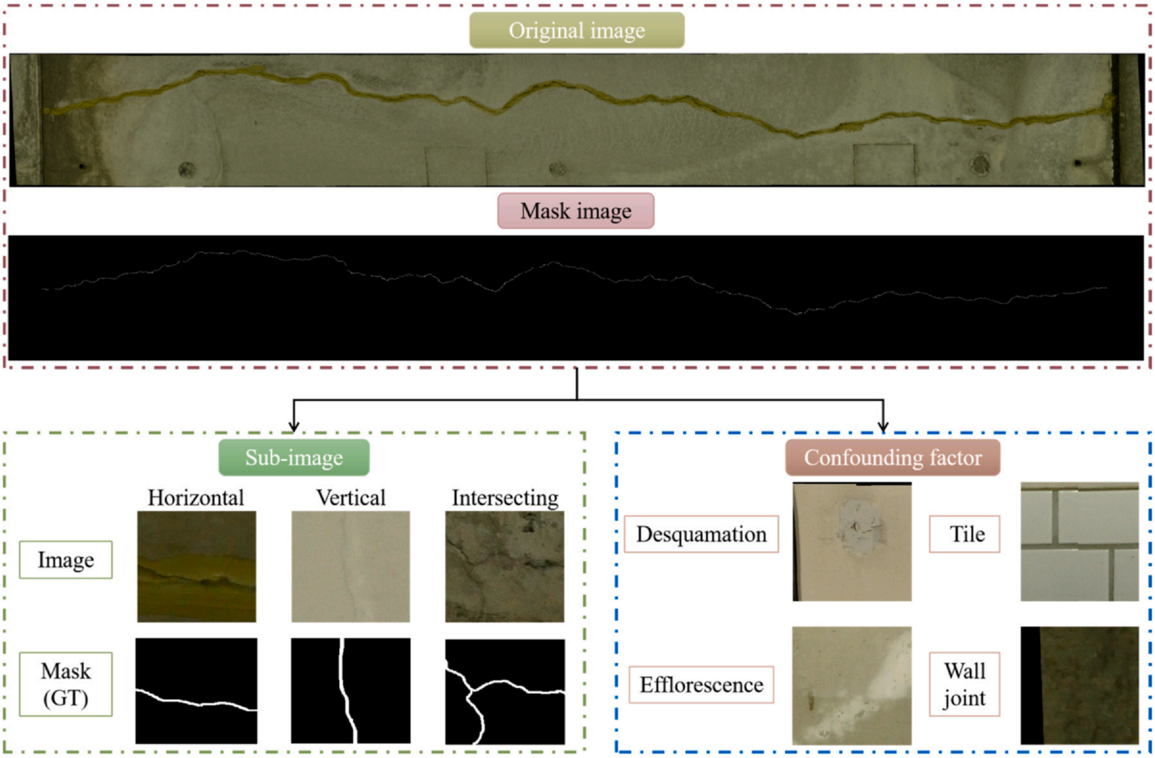


Fig. 1. Sample images (original images, GT-mask images, sub-images, and sub-mask images) and the confounding factors in the tunnel images.

Gamcheon, and Hwasan) in Korea. In addition, the corresponding mask images with the same amounts are considered as the Ground Truths (GTs) in this study, which is manually labeled by the experts from Deep Inspection Co., Ltd. The labelling tool was developed by using Python and PyQt5.

To make it clear, all mask images (synthetic label) generated by WSL methods are represented as *'mask image'*, and all mask images (manual label) annotated by humans are expressed as *'GT-mask image'*. In WSL stage, the original images are used to generate mask images. In FSL stage, GT-mask images and original images are used in the framework that is based on manual data annotation and FSL methods. Mask images and original images are used in the framework that combines WSL methods (automatic annotation) and FSL methods.

Due to the complicated background of the tunnel images, the existing automatic crack inspection systems based on RGB images have a limited performance on the crack detection and segmentation. Fig. 1 shows the original image, the corresponding GT-mask image, some sample images from the cropped sub-images, the corresponding sub-mask images, and common noise factors. In this study, the detected cracks mainly contain three classes: horizontal cracks, vertical cracks, and intersecting cracks. In addition, the four common noise factors (desquamation, tiles, efflorescence, and wall joints) occur in the collected tunnel images, which makes it difficult to detect and recognize the cracks. Desquamation indicates the peeling off of paint from the surface of the wall, and efflorescence refers to a change of the wall's surface after long-term exposure.

4. Proposed method

4.1. Overview of the proposed framework

In this section, a diagram of the proposed framework is illustrated. As presented in Fig. 2, the proposed crack segmentation and the assessment system mainly contain three phases. Firstly, a model that combines the WSL method and the self-supervised learning (SSL) method is proposed to obtain the class activation map (CAM). Also, the dense Conditional Random Field (dCRF) and the random walk algorithm (RW) are used to generate the annotation images based on the acquired CAM (see more details in Section 4.2). Next, a fine-tuned crack segmentation model based on the FSL method is used to train and evaluate the generated data from the first phase. Section 4.3 gives detailed information about the FSL-based crack segmentation model. Finally, a new metric for risk level judgment is used to classify each crack according to the damage severity in the crack risk assessment phase, which is illustrated in Section 4.4.

4.2. Weakly supervised semantic segmentation

Since the WSL-based segmentation network can omit the time-consuming manual annotation process and obtain a synthetic label through the CAM (Božič, Tabernik, and Skočaj 2021), it is used to generate the pixel-level mask to train the fully supervised semantic segmentation network in this framework. The overall structure of the weakly supervised segmentation network (Crack-CAM) is shown in Fig. 2 (Section 4.1). Firstly, a proposed feature extraction network named E-Res2Net101 is trained to generate the CAM, where the input of the network is original sub-images and the corresponding image-level annotation files. Based on the predicted value of each pixel, CAM is transformed into a pixel-level mask. During the transformation process, an RW algorithm is used to refine the mask image, and the dCRF then completes further refinement to obtain the final target area (Ahn and Kwak 2018; Krähenbühl and Koltun 2012). It is difficult to obtain a complete response map only based on the weak supervised semantic segmentation network, because the WSL-based network only detects the discriminative features in the image as opposed to the features of the whole target. To obtain a complete response map, we refer to the idea of combining the SSL method and the WSL method in the SC-CAM algorithm (Chang et al. 2020). A K-means clustering algorithm is used to provide more

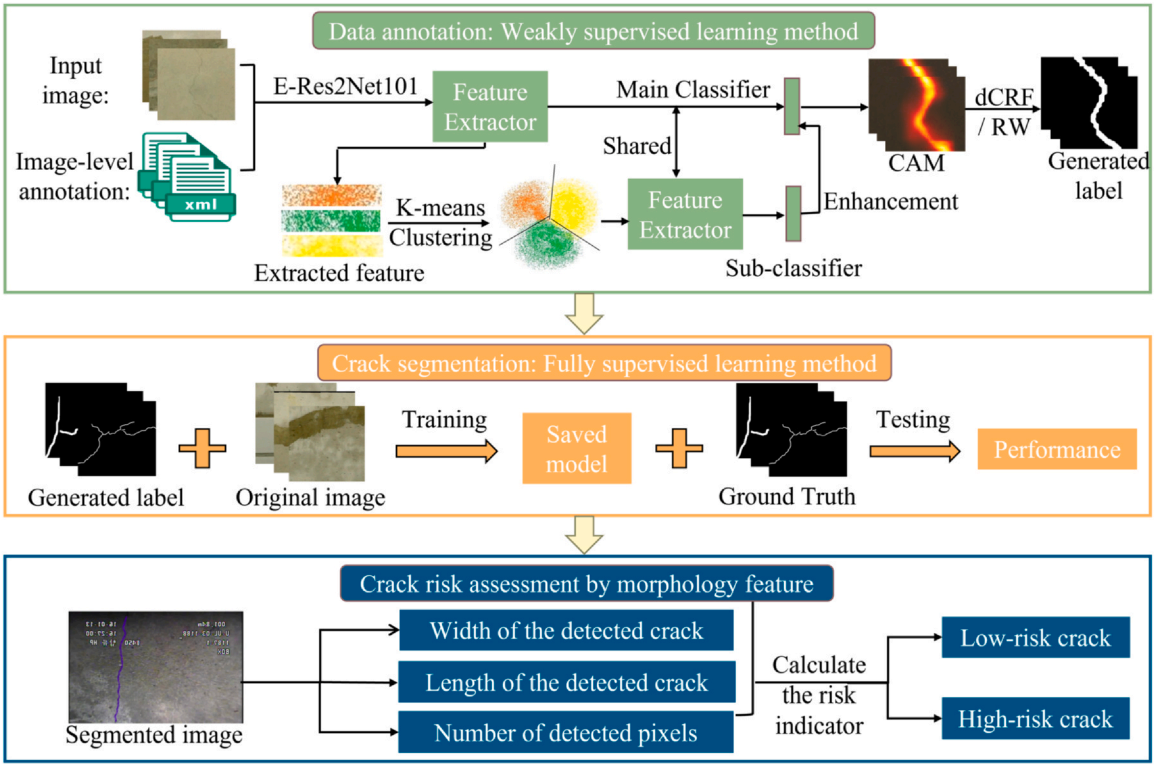


Fig. 2. A diagram of the proposed system, including the data annotation process, the crack segmentation process, and the risk assessment process. The CAM represents the class activation map, the dCRF is the dense Conditional Random Field algorithm, and the RW is the random walk algorithm.

detailed features for the proposed weak supervised segmentation network. In addition, the multi-label soft margin loss function (Lapin, Hein, and Schiele 2017) was used in the proposed Crack-CAM.

In the proposed framework, the segmentation performance of the Crack-CAM is mainly affected by the main classifier. The Res2Net101 has a deep and wide model architecture as a classification network, and it is good at extracting fine-grained features in a weakly supervised segmentation method. Based on Res2Net101 (Gao et al. 2019), an enhanced Res2Net101 (E-Res2Net101) is proposed to enhance the feature extraction ability and expand the width of the Crack-CAM model. As shown in Fig. 3(a) and Fig. 3(b), compared with the Res2Net101, the E-Res2Net101 can connect more small residual blocks to expand the receptive field of the network to obtain more fine-grained features. Furthermore, the activation function in the network is replaced by the Mish instead of the ReLU, because the Mish function has better training stability and accuracy (Misra 2019). After each convolution operation, the feature maps first go through the activation layer, and then perform the normalization operation (Chen et al. 2019; Dang et al. 2021). In addition, the pooling layer and the activation layer are added to the shortcut structure of the E-Res2Net101 network to obtain more effective features, as shown in Fig. 3(c).

Moreover, the RW algorithm and the dCRF algorithm are necessary for the final performance of the Crack-CAM model. The specific

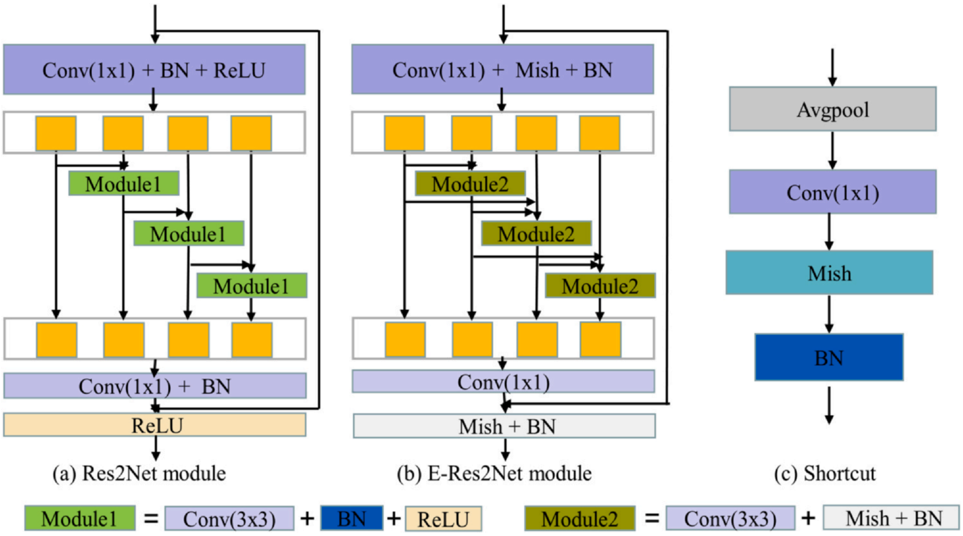


Fig. 3. Network module structures: (a) the module in the Res2Net101, (b) the proposed module in E-Res2Net101, and (c) the Shortcut block of the E-Res2Net101. The BN represents a batch normalization layer, Conv means a convolutional layer, and Avgpool is an average pooling layer.

process is as follows. The Crack-CAM generates the CAM and the probability matrix based on the input image, and the RW algorithm can learn the semantic boundary information of the target region from the obtained probability matrix. Based on the learned boundary information, RW calculates the affinity of the adjacent features by using their L1 distance within certain radius circles (the radius was set *5* in this study) to generate the boundary transition probability matrix that is then multiplied by CAM. After several iterations, RW spreads the activation scores of CAM to adjacent areas to improve the quality of CAM. As a probability graph model, the dCRF extracts the pixel characteristics by adopting a unary potential function on each pixel, and the pairwise potential function was used to obtain the global information. After that, dCRF represents the correlation between the pixels. Furthermore, the dCRF prompts similar pixels to be named the same label, whereas the pixels with large variances are allocated different labels. The dCRF evaluates the effects of all pixels on the current pixel, which includes color, shape, texture, and position information, so it can make the image divided as far as possible at the boundary.

4.3. Fully supervised semantic segmentation

To reach accurate pixel-level segmentation results, the mask images that are generated by the WSL method are sent to an FSL method along with the corresponding original image for the next round of training. The training of the FSL-based method is supervised by the synthetic labels in the framework that combines the WSL and the FSL methods. However, the manually annotated labels are used to supervise the training process in the framework that are only based on the FSL methods. As a fully supervised segmentation model, the DeepLabv3+ (Chen et al. 2018) is fine-tuned and modified to obtain a better segmentation accuracy in this research.

As shown in Fig. 4, the DeepLabv3+ consists of two modules: encoder and decoder. The proposed E-Res2Net101 structure with atrous convolution is applied to the DeepLabv3+ encoder module as a backbone to extract the useful features. The advantage of atrous convolution is that it can control the size of the receptive field through the dilation rate parameter without changing the size of the feature maps compared to normal convolution. The larger the dilation rate is set, the larger the receptive field. The features from the E-Res2Net101 are separated into high-level and low-level features. The high-level features are inputted into the Atrous Spatial Pyramid Pooling (ASPP) module and the decoder module. The ASPP module adopts the atrous convolutional layers with different dilation rates to extract and fuse the multi-scale features. After fusing the features, the feature maps are pooled and inputted into the decoder. An extra convolutional layer is added to process the high-level features and merge the features from the ASPP module to improve the feature utilization and obtain more boundary features. The low-level features are directly sent to the decoder module, and they are combined with other features to provide more details for the final segmentation. Besides, the modified DeepLabv3+ model uses the cross-entropy (Dong et al. 2021) to calculate loss in the training stage.

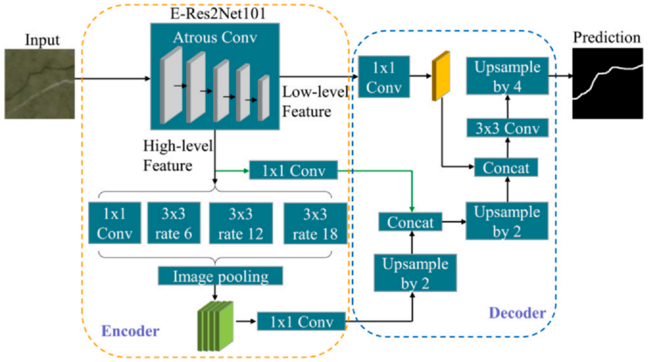


Fig. 4. The overall encoder-decoder structure of the modified DeepLabv3+, including the proposed E-Res2Net101 with atrous convolution and the ASPP module. Conv represents a convolutional layer, and rate is a dilation rate parameter.

In addition, some hyper parameters of the model are adjusted to improve the performance on the proposed tunnel dataset. In the fine-tuning process, two common optimizers (SGD and Adam) and initial learning rates (0.001 and 0.0001) are tested and compared. Moreover, image enhancement methods such as cutout, random jitter (contrast, brightness, and saturation) are added to improve the generalization ability of the model.

4.4. Risk assessment

In the crack risk assessment phase, the area, length, and width of the crack are obtained using the skeleton extraction method (Li et al. 2020). Next, the corresponding risk level is evaluated using a customized evaluation method. The segmentation area can be easily calculated by the total pixels in the predicted image. Based on skeletonized cracks, the length (L) can be calculated by adding all the pixels of a single-pixel wide crack. Since the width of each crack is uneven, the mean width ($\bar{W}$) is calculated according to the length and the area, which is shown in Eq. (1).

$$\bar{W} = \frac{Total\ Number\ of\ detected\ pixels}{L} \tag{1}$$

In addition, a crack damage severity for risk level judgment can be obtained as follows.

$$R = \lambda \times \bar{W} + (1 - \lambda) \times L, \begin{Bmatrix} if \quad R > 60, high \\ if \quad R \leq 60, low \end{Bmatrix} \tag{2}$$

where R represents the risk level, and there are two levels (high and low) in this study. The width parameter is extremely important for the damage severity of the cracks, and the influence of the length parameters cannot be ignored. Considering the different effects on the crack risk, the weight coefficients of the width and the length are set to λ and $1 - \lambda$. The threshold is empirically set to 60 by observing the cracks with different damage severities.In this study, the morphological features of the predicted image are obtained and compared with the morphological features of the ground truth. The mean error (ME) between the prediction and the ground truth can be obtained by using the equation below.

$$ME = \frac{\sum_{i=1}^{n} |F_{G_i} - F_{P_i}|}{n} \tag{3}$$

where F represents three morphological features, including length, the mean width, and the area. F_{G_i} represents the features of the ground truth, and F_{P_i} represents the features of the prediction. n is the total number of cracks. The skeleton extraction algorithm performs well on the proposed dataset. The minimum mean error between the predictions and the GTs is obtained from the horizontal cracks by testing the model with different types of images.

5. Experimental results

In this research, all the experiments were performed using a Linux machine that is pre-installed with Ubuntu 18.04. It is equipped with four Tesla V100 PCIe 32 GB GPUs, an Intel® Xeon® E5-2698 v4 processor, and 256 GB of DDR4 RAM. Section 5.1 illustrates the effectiveness of the proposed automatic pixel-level annotation network based on WSL methods. Then, Section 5.2 explains how the FSL-based segmentation model is optimized and fine-tuned. After that, some state-of-the-art (SOTA) approaches are evaluated and compared in Section 5.3. Finally, a novel metric is introduced to assess the risk level of each detected crack in Section 5.4.

5.1. Evaluation protocols

Some standard evaluation protocols for pixel-level segmentation are explained and applied to present a comprehensive assessment of the experimental methods, which are recommended by the researchers in (Dong et al. 2020; Zhang et al. 2019). In this crack segmentation task, the performance is evaluated using Precision, Recall, F1, and the Mean Intersection over Union (MIoU), which are evaluated using the following equations.

$$\text{Precision} = \frac{\#\text{ of correctly predicted pixels}}{\text{all predictions}} = \frac{TP}{FP + TP} \tag{4}$$

$$Recall = \frac{\#\ of\ correctlty\ predicted\quad pixels}{all\ GTs} = \frac{TP}{FN + TP} \tag{5}$$

$$F1 = \frac{2}{1/Precision\ + 1/Recall} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{6}$$

$$MIoU = \left(\frac{TN}{TN + FN + FP} + \frac{TP}{TP + FP + FN}\right)/2 \tag{7}$$

where TP, TN, FP, and FN refer to True Positives, True Negatives, False Positives, and False Negatives. Besides, the symbol # represents the exact number of pixels.

5.2. Weakly supervised semantic segmentation

As mentioned in Section 4.2, both Res2Net101 and E-Res2Net101 are crucial for the feature extraction process in WSL segmentation methods. The Res2net101 has a strong capability of multi-scale representation, and it can accurately locate the target area in class activation mapping (Gao et al. 2019). Besides, the E-Res2Net101 is proposed by modifying the model structure of Res2Net101. In order to compare the effect of two models on the performance of Crack-CAM, both Res2Net101 and E-Res2Net101 are trained on the same dataset and evaluated using various metrics. As shown in Table 2, the overall performance of Crack-CAM with E-Res2Net101 is better than the Crack-CAM without Res2Net101. In particular, the MIoU value of Crack-CAM with Res2Net101 is 0.34 higher than Crack-CAM without E-Res2Net101.

In the proposed crack segmentation framework, an effective WSL-based network (Crack-CAM) is trained to generate the synthetic labels. The obtained labels and the corresponding original images are then sent to the FSL-based method for the next round of training. Thus, the quality of the generated labels is a vital factor in the whole process. Fig. 5 shows some results of the proposed Crack-CAM in different phases. The first line is some sub-images that are cropped from the original images, and the second line is the corresponding GT-mask images, which are used as the ground truth of the proposed dataset. The third line shows heatmaps from the CAM, and the obtained labels by the dCRF are shown in the final line. As shown in Fig. 5, the proposed Crack-CAM method has a robust performance to locate the cracks with different structures and widths. Nevertheless, the edges of some segmented cracks in the final stage are not smooth.

Table 2
Performance evaluation of the proposed Crack-CAM method based on the proposed dataset.

Evaluation metrics	MIoU	Precision	Recall	F1
Crack-CAM with Res2Net101	0.583	0.924	0.605	0.732
Crack-CAM with E-Res2Net101	**0.617**	**0.942**	**0.631**	**0.755**

5.3. Fully supervised semantic segmentation

In this study, the FSL-based segmentation method named DeepLabv3+ (Chen et al. 2018) was modified by applying the proposed E-Res2Net101 structure. Also, the modified DeepLabv3+ was then fine-tuned by changing the different initial learning rates (lr) and optimizers. Fig. 6(a) represents the MIoU curves of DeepLabv3+ using the same initial learning rate (0.0001) and optimizer (SGD). The blue and orange lines depict the DeepLabv3+ before and after modification, respectively. The MIoU of the modified DeepLabv3+ is 0.04 higher than that of the original DeepLabv3+. Moreover, the performances of the modified model under different hyper-parameter settings are recorded in Fig. 6(b) and Fig. 6(c). When the initial learning rate is 0.0001, the modified model with the Adam optimizer converges faster, but the obtained MIoU is lower than the model with the SGD optimizer. Besides, the MIoU metric of the modified model achieves the best value of 0.786 when the initial learning rate is 0.0001, and the optimizer is set to SGD.

Moreover, the segmentation results of the modified DeepLabv3+ based on the FSL method are compared with the proposed Crack-CAM based on the WSL method in this section. As shown in Fig. 7, the first line is the sub-images that are cropped from the raw images, and the second line is the corresponding GT-mask images labeled by humans. The segmentation results of the proposed Crack-CAM and the modified DeepLabv3+ are in the third line and the fourth line, respectively. Even though the outputs of the Crack-CAM have some noise and the edge of cracks are not smooth, it can accurately locate the detected cracks. Besides, the output images from the FSL-based method have a better segmentation effect than the WSL-based Crack-CAM.

There are various types of confounding factors in the collected tunnel images, such as desquamation, tiles, efflorescence, and wall joints. The proposed methods are performed on the challenging images with the mentioned noises to verify the robustness of the methods used in this study. Fig. 8 shows the original images with four kinds of noises, the corresponding GT-mask images, the outputs of the WSL method, and the outputs of the FSL method. According to the output images, the experimental methods can precisely locate and segment the cracks under different confounding factors. Especially, it is difficult to distinguish the wall joints with cracks due to the similar colors and shapes, but the proposed methods correctly segmented the cracks under the noise of the wall joint in the second column. In addition, the FSL method obtained better segmentation effects than the WSL method.

5.4. Comparison with other methods

In this section, the presented methods are compared with some existing SOTA approaches based on the elaborated database. There are five experimental methods, including SC-CAM (Chang et al. 2020), U-Net (Ronneberger, Fischer, and Brox 2015), DeepLabv3+ (Chen et al. 2018), DeepCrack (Liu et al. 2019), DeepCrack (Zou et al. 2018), Crack-CAM (proposed), and the modified DeepLabv3+. The performance of each method is evaluated using various metrics, which are shown in Table 3. In this study, the multi-label soft margin loss function (Lapin, Hein, and Schiele 2017) was used in the WSL models, and the cross-entropy loss function (Dong et al. 2021) was used for FSL methods. Compared with the other WSL methods, the proposed Crack-CAM obtained better performance from the aspects of segmentation accuracy (MIoU, F1, and loss) and computational complexity (training and testing time). In the experimental FSL methods, the MIoU value of the modified DeepLabv3+ is 0.057 higher than that of the U-Net. However, the U-Net is 0.018 s faster than the modified DeepLabv3+ in terms of the segmentation speed. The FSL methods have a better performance compared with the WSL methods, but they require a complicated annotation process. Thus,

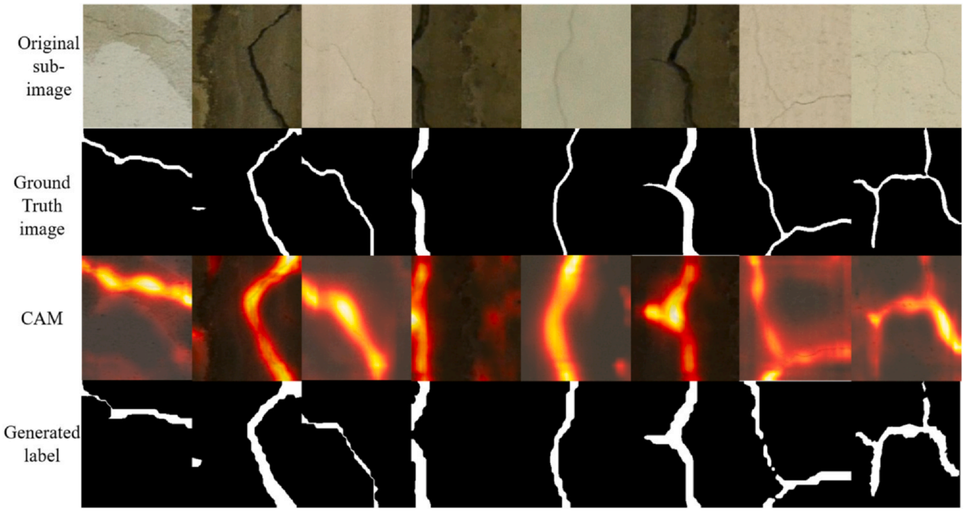


Fig. 5. Output images of the proposed weakly supervised learning network (Crack-CAM) in different stages.

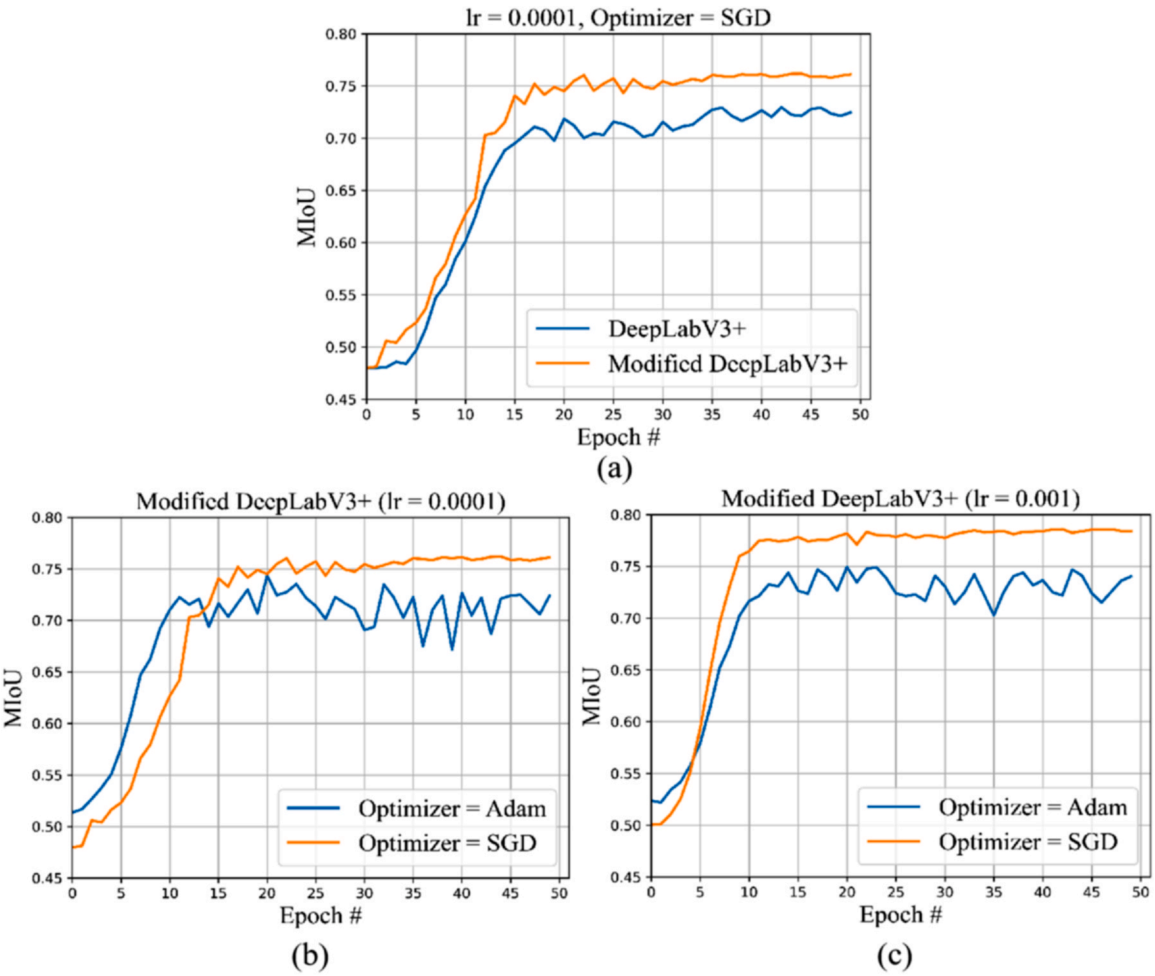


Fig. 6. MIoU curves of the validation process for different models: (a) DeepLabv3+ and modified DeepLabv3+, (b) modified DeepLabv3+ using the same initial learning rate (0.0001) and different optimizers (Adam and SGD), and (c) modified DeepLabv3+ using the same initial learning rate (0. 001) and different optimizers (Adam and SGD).

the proposed Crack-CAM is used for the data annotation in the first stage, and then FSL is used for the segmentation task in the proposed framework. Experimental results indicate the proposed framework combining the WSL method (automatic data annotation) and FSL method (crack segmentation) can obtain a performance that is comparable to the approach based on an FSL method and manual annotation. Besides, the proposed framework contains the time of data annotation, whereas the other methods only calculate the time of crack segmentation, so the time of the proposed framework cannot be compared with the other methods.

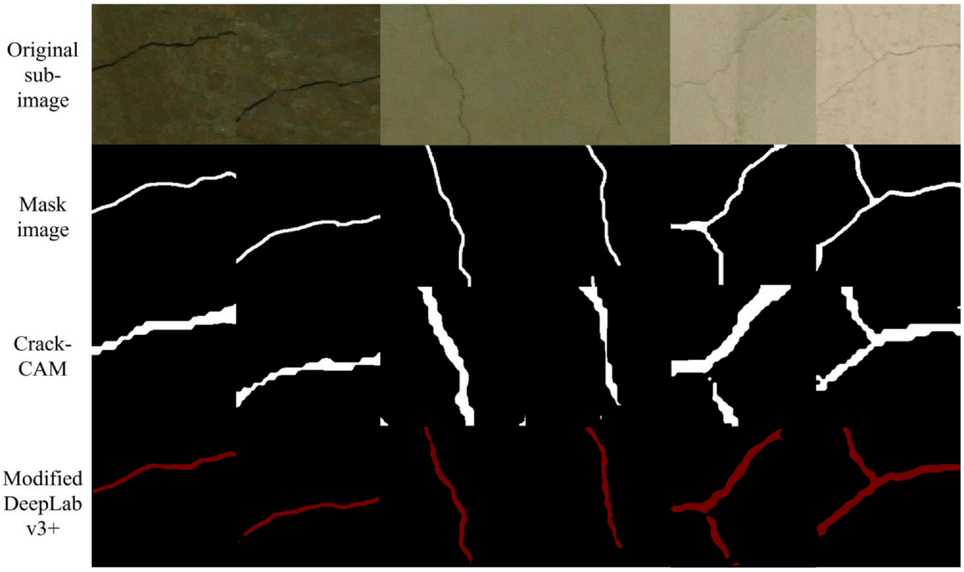


Fig. 7. Output images of the proposed Crack-CAM and the modified DeepLabv3+.

In the crack segmentation process (FSL stage), our modified DeepLabv3+ model is evaluated on a publicly available CrackForest dataset (CFD) (Shi et al. 2016) that contains 118 crack images captured in the urban road scene. Table 4 shows the evaluation results of the presented algorithm in comparison to other recent studies that are considered to be SOTA in the context of the defect segmentation literature, including CrackForest (Shi et al. 2016), MFCD (Li et al. 2019), and DeepLabv3+ (Chen et al. 2018). Experimental results show that our modified DeepLabv3+ obtained the best performance on different metrics (Precision, Recall, and F1). The proposed model has a complex structure, which is capable of generalizing well on various datasets (He et al. 2016). As a result, it showed good performance (F1: 0.936, Precision: 0.928, Recall: 0.945) on the CFD dataset. Although the performances of the two models proposed in (Fan et al. 2020; Inoue and Nagayoshi 2019) were 0.021 and 0.017 higher than the proposed model in terms of F1 score, they only reported the performance for only small datasets, and those models can potentially achieve poor performance on other datasets due to the disadvantage of weak generalization ability (Zhu et al., 2018).

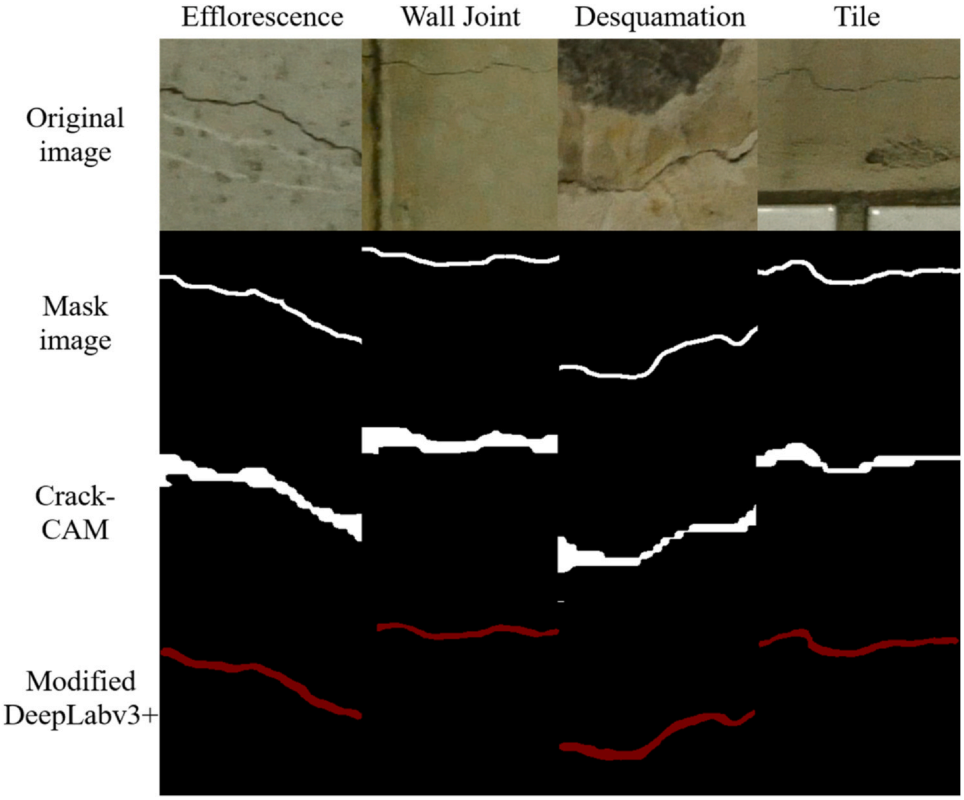


Fig. 8. Segmentation results of the WSL method and the FSL method on the images with different noises, such as desquamation, tiles, efflorescence, and wall joints.

Table 3
Evaluation results of the different methods on the proposed dataset.

Methods		MIoU	F1	Loss	Training time	Testing time
WSL methods	SC-CAM (Chang et al. 2020)	0.566	0.708	0.14	18 h 21 m 06 s (20 epochs)	0.071 s / image
	Crack-CAM (proposed)	**0.617**	**0.755**	**0.11**	**17 h 56 m 55 s (20 epochs)**	**0.063 s / image**
FSL methods	DeepCrack (Liu et al. 2019)	0.665	0.541	0.08	23 h 35 m 12 s (50 epochs)	0.083 s / image
	DeepCrack (Zou et al. 2018)	0.704	0.625	0.07	25 h 16 m 43 s (50 epochs)	0.096 s / image
	U-Net (Ronneberger, Fischer, and Brox 2015)	0.632	0.763	0.08	24 h 10 m 41 s (50 epochs)	**0.018 s / image**
	DeepLabv3+ (Chen et al. 2018)	0.729	0.824	0.09	**14 h 21 m 37 s (50 epochs)**	0.020 s / image
	Modified DeepLabv3+	**0.786**	**0.865**	**0.06**	15 h 13 m 49 s (50 epochs)	0.036 s / image
Combination of WSL and FSL (proposed method)		0.706	0.802	0.13	–	–

Table 4
Evaluation results of the different methods on CFD dataset.

Methods	F1	Precision	Recall
CrackForest (Shi et al. 2016)	0.857	0.822	0.894
MFCD (Li et al. 2019)	0.880	0.899	0.894
DeepLabv3+ (Chen et al. 2018)	0.910	0.898	0.922
Modified DeepLabv3+	**0.936**	**0.928**	**0.945**

5.5. Risk assessment

To confirm a suitable value for the weight coefficient λ, an experiment with 200 cracks is designed in this section. By using the proposed framework, three scatter diagrams of crack risk assessment under different settings of λ are shown in Fig. 9. The solid lines in the three diagrams represent that the risk score R calculated by prediction is equal to the R calculated by the GT. Most points are close to the solid lines, which indicates the proposed framework performed well on the proposed dataset. Besides, the points are very scattered when the λ is equal to 0.3 or 0.6. On the contrary, the points gathered in a certain range when λ is set to 0.9. Thus, it is easy to set a threshold value of the risk metric when λ=0.9.

The objective of the risk assessment is to judge a proper risk level for each crack using the acquired morphological features. Fig. 10 shows the ground truth, the prediction, and the respective skeleton images for each type of crack. Among three morphological features, the obtained areas of predictions are discrepant with that of GTs, but the calculated lengths and mean widths of predictions are quite close to that of GTs. Besides, it can be observed the prediction skeleton images coincide with the ground truth skeleton, which verifies both the proposed segmentation method and the skeleton extraction method have a good effect on the proposed dataset. By calculating the risk parameter (R) as defined in Eq. (2), the maximum deviation between the GT and the prediction result is less than 3.

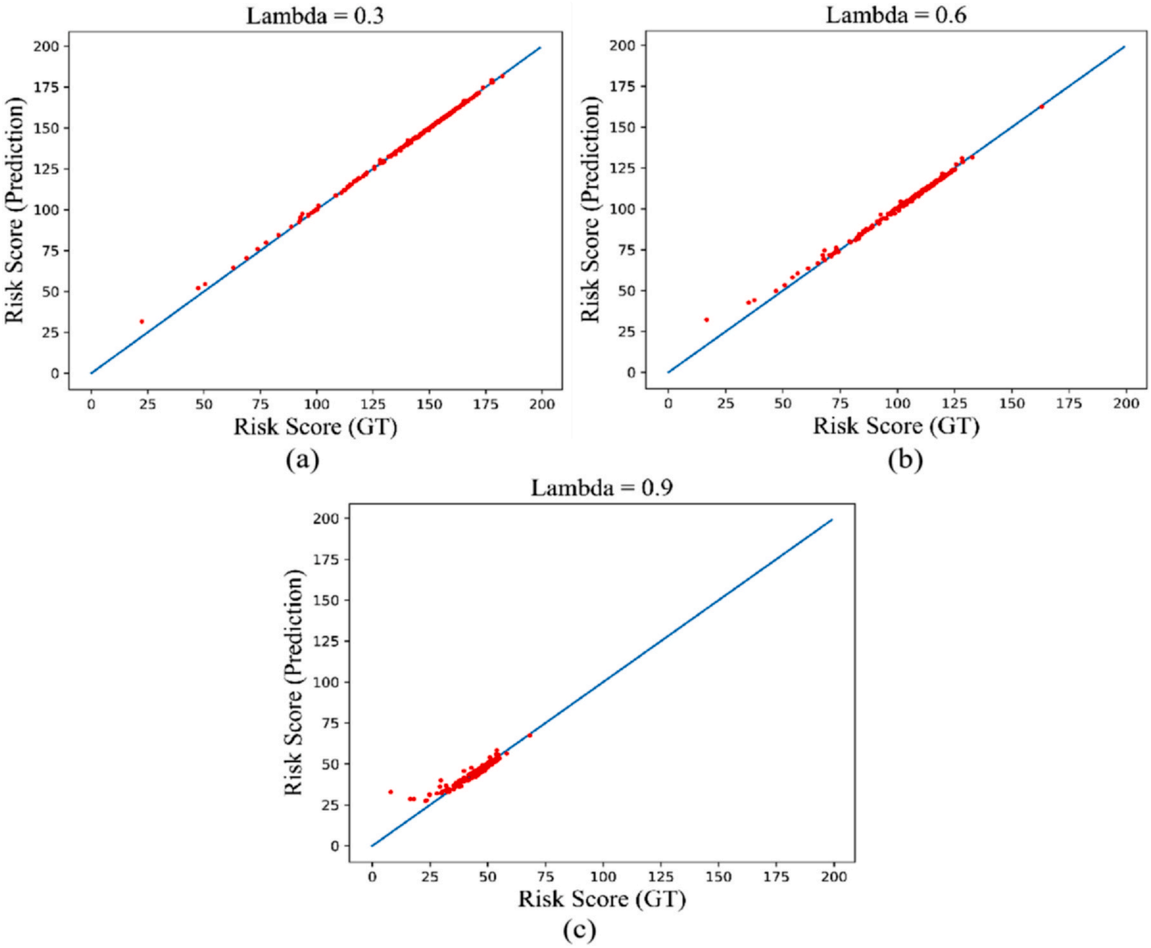
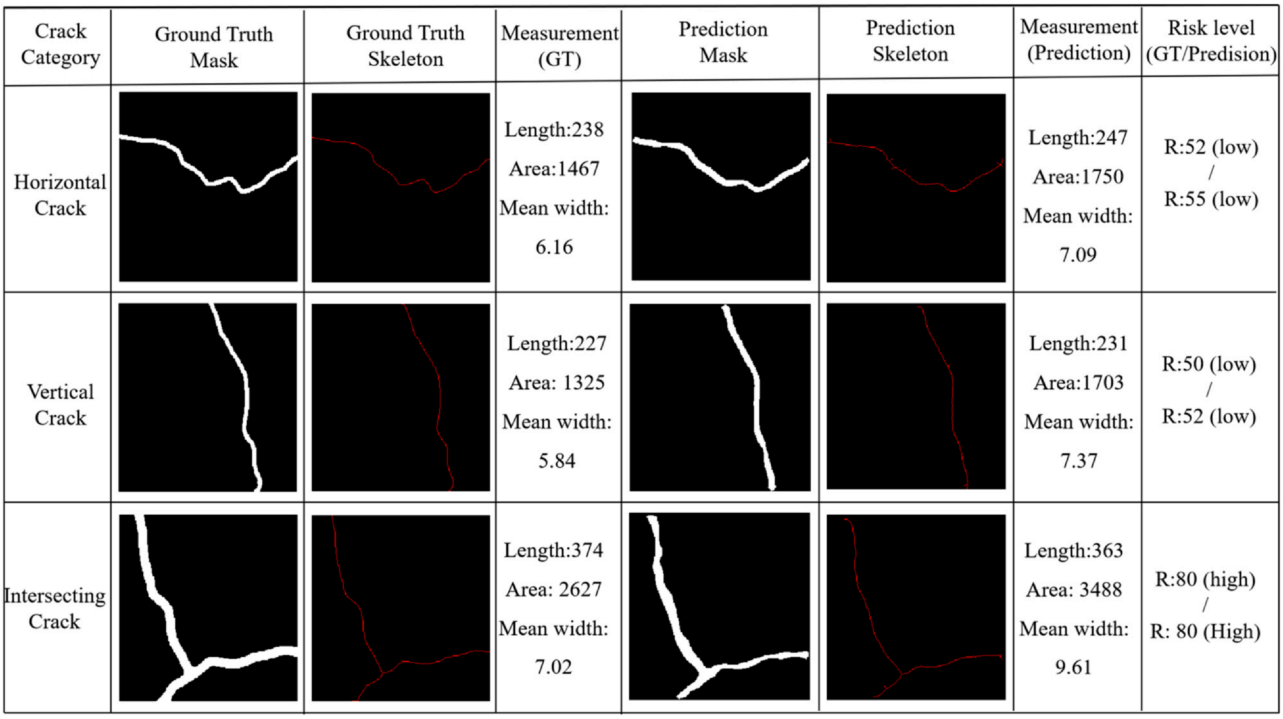


Fig. 9. Scatter diagrams of the crack risk assessment using the proposed framework under different settings of λ: (a) λ=0.3, (b) λ=0.6, and (c) λ=0.9.



Crack Category	Ground Truth Mask	Ground Truth Skeleton	Measurement (GT)	Prediction Mask	Prediction Skeleton	Measurement (Prediction)	Risk level (GT/Predision)
Horizontal Crack			Length:238 Area:1467 Mean width: 6.16			Length:247 Area:1750 Mean width: 7.09	R:52 (low) / R:55 (low)
Vertical Crack			Length:227 Area: 1325 Mean width: 5.84			Length:231 Area:1703 Mean width: 7.37	R:50 (low) / R:52 (low)
Intersecting Crack			Length:374 Area: 2627 Mean width: 7.02			Length:363 Area: 3488 Mean width: 9.61	R:80 (high) / R: 80 (High)

Fig. 10. Risk assessment of ground truth images and prediction results.

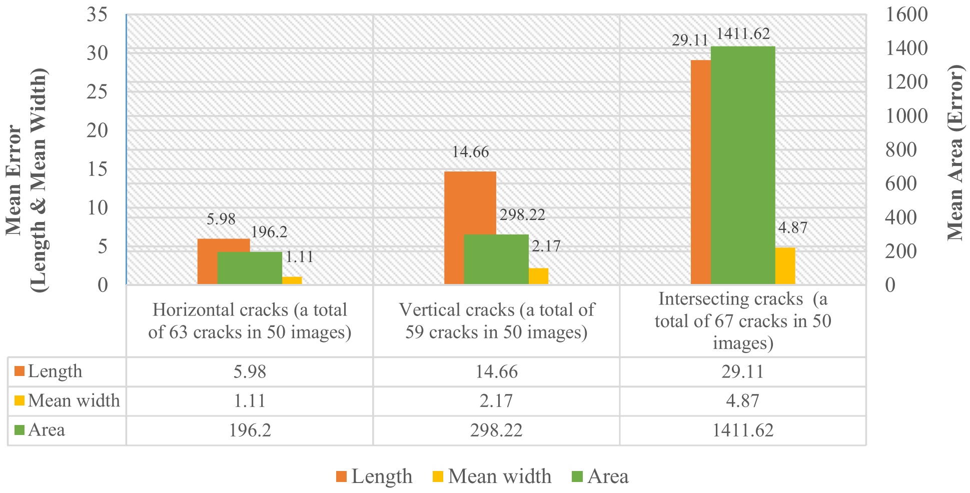


Fig. 11. Mean error of the morphological features, including the length, the mean width, and the area.

To provide a more convincing result, an experiment with randomly selected images from the testing set was conducted to calculate the mean error (ME) of each crack. In this section, a total of 150 images are tested, including 50 images from each category. As mentioned in Eq. (3), the ME of the length, the mean width, and the area are computed. Correspondingly, the detailed information of each category is described in Fig. 11. According to statistics, the MEs (length, mean width, and area) of horizontal cracks are the lowest among all the obtained ME values of three categories. Especially, the mean width error is 1.11 pixels, which means the mean widths of GTs are incredibly similar to predictions. Due to the complex structures and the small number of intersecting cracks, the MEs of intersecting cracks are the highest, which can be improved by enlarging the number of images with intersecting cracks.

6. Conclusion

This paper presents a novel idea of automatic data annotation using the WSL-based method Crack-CAM, which can save a lot of effort and time. Based on the generated labels from Crack-CAM, an FSL-based DeepLabv3+ model is fine-tuned and trained by changing the architecture and the hyper-parameters to enhance the overall performance. A total of 521 tunnel images were collected from four different locations in Korea, and the corresponding GT-mask images were manually made by experts. In order to better fit the segmentation network, the original tunnel images and the corresponding GT-mask images were cut into sub-images of 224*224 pixels. Moreover, the proposed methods are compared with the state-of-the-art approaches on the proposed dataset. The Crack-CAM obtained the highest MIoU of 0.617 among the WSL-based methods, and the modified DeepLabv3+ achieved the highest MIoU of 0.786 among all the experimental methods. The performance of the proposed framework combining Crack-CAM and the modified DeepLabv3+ is similar to that of the method based on the modified DeepLabv3+ and the manual annotation images, which demonstrates the WSL-based Crack-CAM can annotate data correctly. In addition, a new metric (R) is introduced to evaluate the risk levels of the detected cracks, and the ME is calculated to obtain the error between GTs and predictions. Experiments show the ME values of the intersecting cracks are the largest due to the complex structures and the small number of images.

In the future, more images with intersecting cracks will be collected and added in the proposed dataset to improve the overall segmentation results. In addition, the structure of Crack-CAM will be simplified to reduce the computational complexity and save training time.

CRediT authorship contribution statement

Hanxiang Wang, **Yanfen Li:** Conceptualization, Methodology, Data curation. **L. Minh Dang**, **Sujin Lee:** Visualization, Investigation. **Hyeonjoon Moon:** Supervision. **Yanfen Li:** Writing-Reviewing and Editing

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.

Acknowledgement

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540) and National Research Foundation of Korea (NRF) grant funded by the Korea government, Ministry of Science and ICT (MSIT) (2021R1F1A1046339).

References

Ahn, Jiwoon, Kwak, Suha, 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 4981–4990.

Božič, Jakob, Tabernik, Domen, Skočaj, Danijel, 2021. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. Computers in Industry 129, 103459.

Chang, Yu-Ting, Wang, Qiaosong, Hung, Wei-Chih, Piramuthu, Robinson, Tsai, Yi-Hsuan, Yang, Ming-Hsuan, 2020. Weakly-supervised semantic segmentation via sub-category exploration. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 8991–9000.

Chen, Guangyong, Chen, Pengfei, Shi, Yujun, Hsieh, Chang-Yu, Liao, Benben, Zhang, Shengyu, 2019. Rethinking the usage of batch normalization and dropout in the training of deep neural networks. arXiv preprint arXiv *1905, 05928*.

Chen, Haiyong, Hu, Qidi, Zhai, Baoshuo, Chen, He, Liu, Kun, 2021. Perovskite nanoparticles@N-doped carbon nanofibers as robust and efficient oxygen electrocatalysts for Zn-air batteries. Journal of colloid and interface science 581, 374–384.

Chen, Liang-Chieh, Zhu, Yukun, Papandreou, George, Schroff, Florian, Adam, Hartwig, 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV) 833–851.

Dang, L.Minh, SeonJae, Kyeong, Li, Yanfen, Wang, Hanxiang, Nguyen, Tan N., Moon, Hyeonjoon, 2021. Deep learning-based sewer defect classification for highly imbalanced dataset. Computers & Industrial Engineering 161, 107630.

Dang, L.Minh, Syed Ibrahim, Hassan, Suhyeon, Im, Irfan, Mehmood, Moon, Hyeonjoon, 2018. 'Utilizing text recognition for the defects extraction in sewers CCTV inspection videos'. Computers in Industry 99, 96–109.

Dong, Yafen, Shen, Xiaohong, Jiang, Zhe, Wang, Haiyan, 2021. Recognition of imbalanced underwater acoustic datasets with exponentially weighted cross-entropy loss. Applied Acoustics 174, 107740.

Dong, Zhiming, Wang, Jiajun, Cui, Bo, Wang, Dong, Wang, Xiaoling, 2020. Patch-based weakly supervised semantic segmentation network for crack detection. Construction and Building Materials 258, 120291.

Dorafshan, Sattar, Thomas, Robert J., Maguire, Marc, 2018. 'Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete'. Construction and Building Materials 186, 1031–1045.

Fan, Zhun, Li, Chong, Chen, Ying, Mascio, Paola Di, Chen, Xiaopeng, Zhu, Guijie, Loprencipe, Giuseppe, 2020. Ensemble of Deep Convolutional Neural Networks for Automatic Pavement Crack Detection and Measurement. Coatings 10, 152.

Gao, Shanghua, Cheng, Ming-Ming, Zhao, Kai, Zhang, Xin-Yu, Yang, Ming-Hsuan, Philip, H.S.Torr, 2021. Res2Net: A New Multi-Scale Backbone Architecture. IEEE transactions on pattern analysis and machine intelligence 43, 652–662.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, Sun, Jian, 2016. "Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition 770–778.

Inoue, Yuki, Nagayoshi, Hiroto, 2019. Deployment conscious automatic surface crack detection. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) 686–694.

Krähenbühl, Philipp, Koltun, Vladlen, 2012. Efficient inference in fully connected crfs with gaussian edge potentials. arXiv preprint arXiv *1210, 5644*.

Lapin, Maksim, Hein, Matthias, Schiele, Bernt, 2017. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. IEEE transactions on pattern analysis and machine intelligence 40, 1533–1554.

Liu, Yahui, Yao, Jian, Lu, Xiaohu, Xie, Renping, Li, Li, 2019. 'DeepCrack: A deep hierarchical feature learning architecture for crack segmentation'. Neurocomputing 338, 139–153.

Li, Gang, Ma, Biao, He, Shuanhai, Ren, Xueli, Liu, Qiangwei, 2020. Automatic Tunnel Crack Detection Based on U-net and a Convolutional Neural Network with Alternately Updated Clique. Sensors (Basel, Switzerland) 20, 717.

Li, Haifeng, Song, Dezhen, Liu, Yu, Li, Binbin, 2019. 'Automatic pavement crack detection by multi-scale image fusion'. IEEE Transactions on Intelligent Transportation Systems 20, 2025–2036.

Li, Dawei, Xie, Qian, Gong, Xiaoxi, Yu, Zhenghao, Xu, Jinxuan, Sun, Yangxing, Wang, Jun, 2021. Automatic defect detection of metro tunnel surfaces using a vision-based inspection system. Advanced Engineering Informatics 47, 101206.

Misra, Diganta, 2019. 'Mish: A self regularized non-monotonic neural activation function'. arXiv preprint arXiv 4 *1908.08681*.

Ren, Yupeng, Huang, Jisheng, Hong, Zhiyou, Lu, Wei, Yin, Jun, Zou, Lejun, Shen, Xiaohua, 2020. Image-based concrete crack detection in tunnels using deep fully convolutional networks. Construction and Building Materials 234, 117367.

Ronneberger, Olaf, Fischer, Philipp, Brox, Thomas, 2015. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.

Shi, Yong, Cui, Limeng, Qi, Zhiquan, Meng, Fan, Chen, Zhensong, 2016. 'Automatic road crack detection using random structured forests'. IEEE Transactions on Intelligent Transportation Systems 17, 3434–3445.

Song, Qing, Wu, Yingqi, Xin, Xueshi, Yang, Lu, Yang, Min, Chen, Hongming, Liu, Chun, Hu, Mengjie, Chai, Xuesong, Li, Jianchao, 2019. 'Real-time tunnel crack analysis system via deep learning'. IEEE Access 7, 64186–64197.

Su, Tung-Ching, Yang, Ming-Der, Wu, Tsung-Chiang, Lin, Ji-Yuan, 2011. 'Morphological segmentation based on edge detection for sewer pipe defects on CCTV images'. Expert Systems with Applications 38, 13094–13114.

Xu, Liang, Lv, Shuai, Deng, Yong, Xiuxi, Li, 2020. 'A weakly supervised surface defect detection based on convolutional neural network'. IEEE Access 8, 42285–42296.

Yang, Xincong, Li, Heng, Yu, Yantao, Luo, Xiaochun, Huang, Ting, Yang, Xu, 2018. 'Automatic pixel-level crack detection and measurement using fully convolutional network'. Computer-Aided Civil and Infrastructure Engineering 33, 1090–1109.

Zhang, Jianming, Lu, Chaoquan, Wang, Jin, Wang, Lei, Yue, Xiao-Guang, 2019. Screening of Fungi for Potential Application of Self-Healing Concrete. Scientific reports 9, 2075.

Zhu, Hong, Eric, C.C.Tsang, Zhu, Jie, 2018. 'Training an extreme learning machine by localized generalization error model'. Soft Computing 22, 3477–3485.

Zhu, Jinsong, Song, Jinbo, 2020. 'Weakly supervised network based intelligent identification of cracks in asphalt concrete bridge deck'. Alexandria Engineering Journal 59, 1307–1317.

Zou, Qin, Zhang, Zheng, Li, Qingquan, Qi, Xianbiao, Wang, Qian, Wang, Song, 2018. 'Deepcrack: Learning hierarchical convolutional features for crack detection'. IEEE Transactions on Image Processing 28, 1498–1512.