IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 16, 2021, accepted August 2, 2021, date of publication August 16, 2021, date of current version August 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3105279

A Deep Learning-Based Fine Crack Segmentation Network on Full-Scale Steel Bridge Images With Complicated Backgrounds

ZHIHANG LI, (Graduate Student Member, IEEE), HUAMEI ZHU, AND MENGQI HUANG
Department of Civil Engineering, Monash University, Clayton, VIC 3800, Australia

Corresponding author: Mengqi Huang (mengqi.huang@monash.edu)

The work of Mengqi Huang was supported by Monash University through the scholarships and the high-performance computation platform.

ABSTRACT Automatic defect detection of steel infrastructures in structural health monitoring (SHM) is still challenging because of complicated background, non-uniform illumination, irregular shapes and interference in images. Conventional defects detection mainly relies on manual inspection which is time-consuming and error-prone. In this study, a deep learning-based fine crack segmentation network, termed as FCS-Net was proposed in light of ResNet-50 and fully convolutional network (FCN). Structural modifications including Batch Normalization (BN) and Atrous Spatial Pyramid Pooling (ASPP) were made. In full-scale steel girder images with complicated background and fine foreground, the proposed FCS-Net achieves a MIoU of 0.7408, outperforming benchmark algorithms such as LinkNet, DeepLab V3, and CrackSegNet. Moreover, the ablation experiments were performed that justified the contribution and necessity of each modification.

INDEX TERMS Deep learning, fine crack, complicated background, semantic segmentation.

I. INTRODUCTION

After infrastructure construction is completed and starts to be utilized, the quality of infrastructures will gradually be challenged by problems such as erosion [1] and damage resulted from external forces and natural factors [2], causing hidden dangers [3]. Defects of a civil infrastructure requires regular inspections that are often relied on human labor [2], [4]–[8] and are sometimes susceptible to strong subjectivity, in addition to drawbacks of low accuracy, high labor cost, and dangerous working environments. With the rapid development of computer technology, methods for infrastructure defects detection based on computer vision and image processing are gradually emerging and stimulating continuous research [1], [6], [9]–[13]. The existing machine learning-based algorithms have reasonable recognition performance on the simple crack image while it requires intervention and empirical judgment by experts [4], [14]–[19]. Moreover, the underground infrastructures have inadequate working conditions such as high temperature, cold weather, high humidity for electronic devices and the image collected from the site may experience uneven illumination causing significant amount of noise appearing on the images of complex surface textures. Those factors result in the difficulties of achieving higher accuracy of defects recognition as well as meeting industrial requirements for field application. Recently, deep learning represents the state of the art of artificial intelligence, which has successfully been applied to different aspects such as image recognition, text translation, natural language processing and achieves huge success. The conventional machine learning based image detection and recognition methods have been gradually replaced by more intelligent and effective deep learning algorithms [20]–[23]. Deep learning is a further development of the artificial neural network. By pre-training the neural network layer by layer, the feature expression of different levels can be learned, and the feature expression of each layer is obtained through the previous expression propagation, then all the layers are combined to form a deep convolutional neural network. Compared with the conventional machine-learning and image classification algorithms, deep convolutional neural networks have demonstrated superior performance in parameter prediction and image classification.

In this paper, the atrous spatial pyramid pooling (ASPP) and batch normalization (BN) modules are used to collaborate with the original ResNet-50. Meanwhile, various

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

loss functions are analyzed to determine the best fit for the model so that the segmentation performance can be improved. The remainder of this paper is organized as follows. In Section 2, a literature review is performed to review the previous techniques on crack segmentation. Section 3 is dedicated to elucidate the specific architecture of improved FCS-Net and the loss function of the network. In Section 4, the generation of datasets, the quantitative metrics used to evaluate the prediction results on benchmark datasets and the comparison of proposed network with other methods are explained. In Section 5, the ablation experiment is performed to evaluate the improvement of the modules. Section 6 summarizes the study and discuss limitations requiring further improvements.

II. RELATED WORK

This section reviews studies related to deep learning-based segmentation of cracks in civil infrastructures. Segmentation frameworks enabled by deep learning can be roughly classified as either two-step or one-step. In two-step streamline, an object detection network aiming at localizing region of interests (ROIs) is usually followed by digital image processing (DIP) algorithms to extract crack pixels. For example, a modified tubularity flow field (TuFF) was applied to segment crack within bounding boxes proposed by a trained Faster R-CNN and obtained high performance in concrete structure images [24]. Although DL-DIP frame works rapidly, rule-based DIP method is applicable to images with relatively clean background and may generate noises when the ROI contains interferences. Double-DL framework was proposed to fill this gap, which replaces the DIP post processing by a segmentation network, such as the sequential use of Faster R-CNN and U-Net [25]. Whereas, it would be labor-intensive to train two deep networks separately and time-consuming when conducting object detection and crack segmentation tasks individually.

Hence, some efforts were made to one-step segmentation network, which is normally enabled by FCN and implicitly integrates detection and segmentation tasks into one shot. One of initial trials includes the eight-layer CNN network to localize crack regions with small bounding patches to approximate segmentation results in large-scale images [6]. Fine detection at pixel level of pavement crack was realized by CrackDet, a five-layer network which has been a benchmark study in crack segmentation [26]. With development of computation facilities and progress in DL algorithms, more studies were initiated to update previous baselines. For instance, In light of FCN structure, the segmentation precision outperformed CrackNet [10]. SDD-Net was well-designed for real-time crack segmentation [27], and CrackNet was modified to CrackNet II for more rapid segmentation [28]. However, the above studies were based on pavement or concrete structure surface, which was less contaminated by noises like handwritings and welding joints in steel structure. Moreover, cracks in steel materials are more challenging to be captured with steeper foreground-background rate, especially in large-resolution photograph. Restricted Boltzmann machine was applied to locate crack in steel infrastructures with high accuracy [29]. A deep fusion CNN was proposed to segment fine cracks in steel girder and achieved satisfying performance [30], while the sliding window scanning method with sub patch of 65 $\times$ 65 may slow detection speed in a full-scale image with resolution of nearly 5000 $\times$ 4000. In this study, a one-step framework was proposed based on a CNN network specialized for fine crack segmentation in 512 $\times$ 512 patches and get tested on full-scale images.

III. PROPOSED METHOD

A. OVERALL WORKFLOW

The proposed end-to-end method contains a pre-trained CNN model as its core component to perform the crack segmentation on the steel girder images. The method could generate the predicted crack image by obtaining the original image with a series of processing, while the foreground (crack) marked as zero and background as one. Fig. 1 demonstrates the workflow of the proposed method, which can be summarized as below.

(1) The original full-scale image I, with size of 4928 $\times$ 3264 or 5152 $\times$ 3864, is resized to image II with its width and height being multipliers of 512, with the specific size of 4608 $\times$ 3072 or 5120 $\times$ 3584;

(2) The image II is cropped to a batch of 512 $\times$ 512 images;

(3) A trained segmentation FCS-Net model takes the image batch III as inputs and produce predicted mask batch IV with crack labeled as 1 (white) and background labeled as 0 (black);

(4) The mask batch IV is merged and performed with color inversion to generate the mask V;

(5) The size of mask V is recovered to the same size as the original full-scale image to produce mask VI.

FCS-Net is a further improved semantic segmentation network based on FCN [31] and inspired by PSPNet [32] and U-Net [33]. The core idea of FCS-Net is that if more global information is introduced into the segmentation layer, the accuracy of recognition can be improved. The main structure of FCS-Net can be broadly divided into three parts: the backbone ResNet-50 [34] module with Batch Normalization [35], Atrous Spatial Pyramid Pooling [36]–[38], and the FCN output layer. Among them, the batch normalization is used to obtain more feature details in ResNet module while the pyramid pooling module extracts deep and shallow features of image respectively. Then the features are fused to reduce the probability of false segmentation while dilated convolution increases the receptive field. Compared with the original PSPNet and FCN, the modification to the architectural has enhanced feature extraction performance and increased the segmentation accuracy. And the following introduces the individual feature extraction modules and explains their incorporation into the overall model.

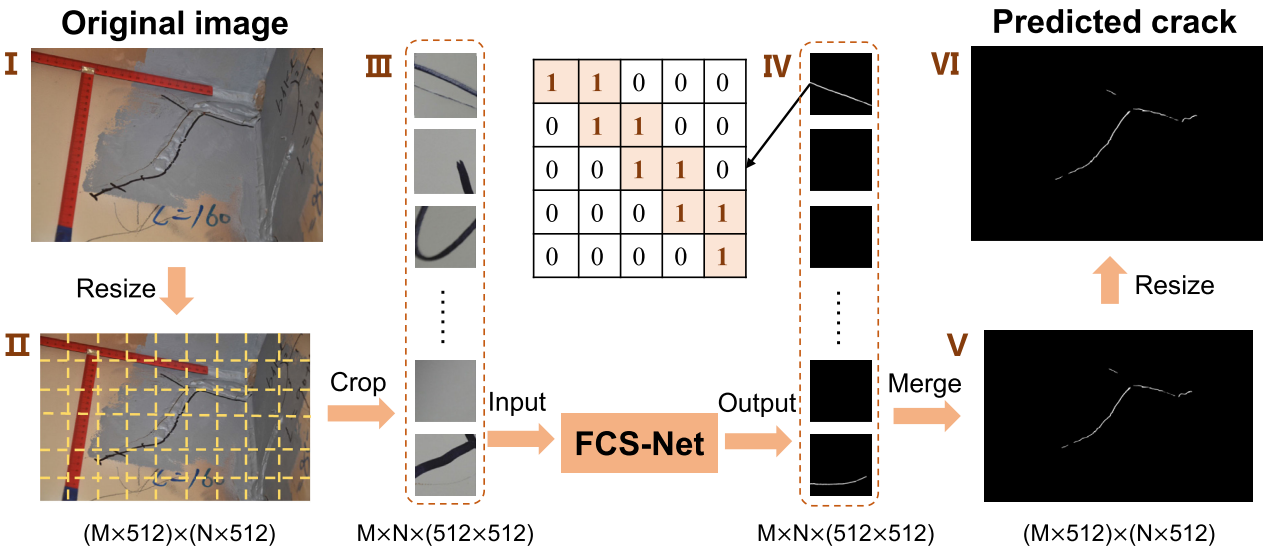


FIGURE 1. Overall workflow of the proposed segmentation method.

B. FCS-NET ARCHITECTURE

Firstly, the feature map is extracted from the network through the pre-trained residual network, and the feature map is transformed into a smaller graph with overall information through the pyramid pooling module. After up-sampling, the smaller graph is restored to the size of the feature graph, and combined with the feature map before pooling, the final output result is obtained after the last fully convolution module. The following introduces the individual feature extraction modules and explains their incorporation into the overall model.

Residual Neural Network [34], referred as ResNet, was proposed to solve the problem of decreased training set accuracy in a deeper network when increasing the layer of the network structure, which is a phenomenon not related to overfit as it did not cause the appearance of extremely high model accuracy or vanishing gradient that results in the stopping of the network training process. The basic unit of ResNet is the residual block. Compared with the conventional plain network structure, residual networks add the skip connections between every two layers, form a residual block so that later layers can learn residuals directly from the previous. This network structure can form a deep residual network and solve the problem of decreasing accuracy during model training. There are two mappings in ResNet, one is the identity mapping, which refers to the input data x itself, and is represented as a curve in the figure, and the other is the residual mapping, which refers to the rest of the network. The advantage of ResNet is the network structure contains skip connection, so that the network could be trained normally while the gradient would not disappear, the layer of the convolutional neural network can be deeper and the error rate of network training will not increase. Spatial Pyramid Pooling Network (SPP-Net) is an algorithm proposed by He [36] to address the problem of repetitive operation in R-CNN architecture. By adding a spatial pyramid pooling structure between the convolutional layer and the fully connected layer to replace method in the R-CNN algorithm which the candidate blocks were clipped and scaled to make the size of the image sub-blocks consistent before the input of the convolutional neural network. The Atrous Spatial Pyramid Pooling [32], [39] is a further improvement of dilated convolution. It combines dilated convolution with Spatial Pyramid Pooling (SPP) module. The Pyramid Pooling Module is inspired by the success of R-CNN Spatial Pyramid Pooling method, which shows that the region of any scale can be classified accurately and effectively by resampling the convolution features extracted from a single scale. The Spatial Pyramid Pooling Network (SPP-Net) is an algorithm proposed by He [29] to address the problem of repetitive operation in R-CNN architecture. By adding a spatial pyramid pooling structure between the convolutional layer and the fully connected layer to replace method in the R-CNN algorithm where the candidate blocks were clipped and scaled to make consistent the size of the image subblocks before inputting to the convolutional neural network. The spatial pyramid pooling structure can effectively avoid the problem of incomplete clipping and shape distortion caused by the R-CNN algorithm, more importantly, it solves the problem of repetitive feature extraction of images by the convolutional neural network, greatly improves the speed of producing candidate blocks, and reduces the total amount of computation.

In order to avoid a large number of crack-like interferences in the complex background and extract the target features more accurately, the network with deeper layers is adopted. However, due to the intensive feature accumulation operation of the traditional convolution kernel method, there may be overlap between the receptive fields, which increases the complexity of semantic information and results in the waste of computation and efficiency [38]. Therefore, there is a balance between large receptive field and maintaining the resolution of feature map. The dilated convolution

technique uses the Atrous convolution kernel, which only has weights in certain positions, and fills other spaces in the kernel with zero [40]. By geometrically increasing the dilation rate in continuous convolution layers, the receptive field can be extended while ensuring the coverage. Compared with the conventional convolution method, the Atrous convolution kernel seems to be 'dilated', extracting features sparsely but still effectively, which reduces the complexity of semantic information and improves the accuracy of image segmentation.

The Pyramid Pooling Module is inspired by the success of R-CNN Spatial Pyramid Pooling method, which shows that the region of any scale can be classified accurately and effectively by resampling the convolution features extracted from a single scale. In ASPP, different rates of parallel dilated convolution are applied in the input feature mapping and are fused together. Since the same kind of objects may have different proportions in the image, ASPP helps to consider different proportions of objects, which can improve the accuracy. The ASPP is a further improvement of dilated convolution. It combines dilated convolution with Spatial Pyramid Pooling (SPP) module. By using dilated convolution with different sampling rates in the first feature map and adding additional Batch Normalization layer, the model aggregates contextual information from different regions to improve the ability to obtain global information [40], as well as to collect more feature details while remaining the same receptive field, thus the segmentation accuracy is improved.

Batch normalization (BN) [35] is widely used in various aspects of deep learning. It is used to solve the converging problem of the modified network during the process of training, when the mathematical distribution of eigenvectors changed, it would result in covariance shift [35]. meanwhile, batch normalization can also effectively prevent the vanishing gradient and the neuron inactivation problem [35]. The BN layer is the same network layer as the convolution layer and pooling layer deployed in the network. the formula of BN is:

$$y = \frac{\gamma}{\sqrt{Var[X] + \varepsilon}} * X + \left(\beta - \frac{\gamma E[X]}{\sqrt{Var[X] + \varepsilon}}\right) \quad (1)$$

In Eq (1) X is the eigenvector to be normalized; E [x] is the mean value; Var [x] is the standard deviation; γ and ε are scaling coefficient and displacement parameter; ε is a small value to prevent the denominator from fading to zero.

As shown in Fig. 2, after three convolutional layers with batch normalization and activation ReLU, the input features are passed through a max pooling layer before entering sixteen residual layers with different convolution specification for feature extraction. This structure is designed to enable better feature extraction by deepening the network while the skip connection in residual network prevents the phenomenon of the vanishing and exploding gradient [34]. Then the atrous spatial pyramid pooling is applied to obtain the representation of different sub-regions [36] while expanding the receptive field by dilated the convolution [1], [38], [40], [41]. Finally, dense pixel-wise prediction is obtained by feeding the feature maps into two convolutional layers.

C. DATASET PREPARATION

After processing of the raw image, a total number of 9268 images training sets with size of 512 × 512 pixels including 8338 background images and 928 crack images were established, while the test sets consist 274 crack-containing images and 2160 background-only images with the same size. Moreover, the original 40 original full-size images before crop also added to the test sets. Considering the significant imbalance between foreground (cracks) and background samples, and the pixels with cracks only account for a small part of the images, the images with only background are removed in the test set, while all the images with cracks are retained.

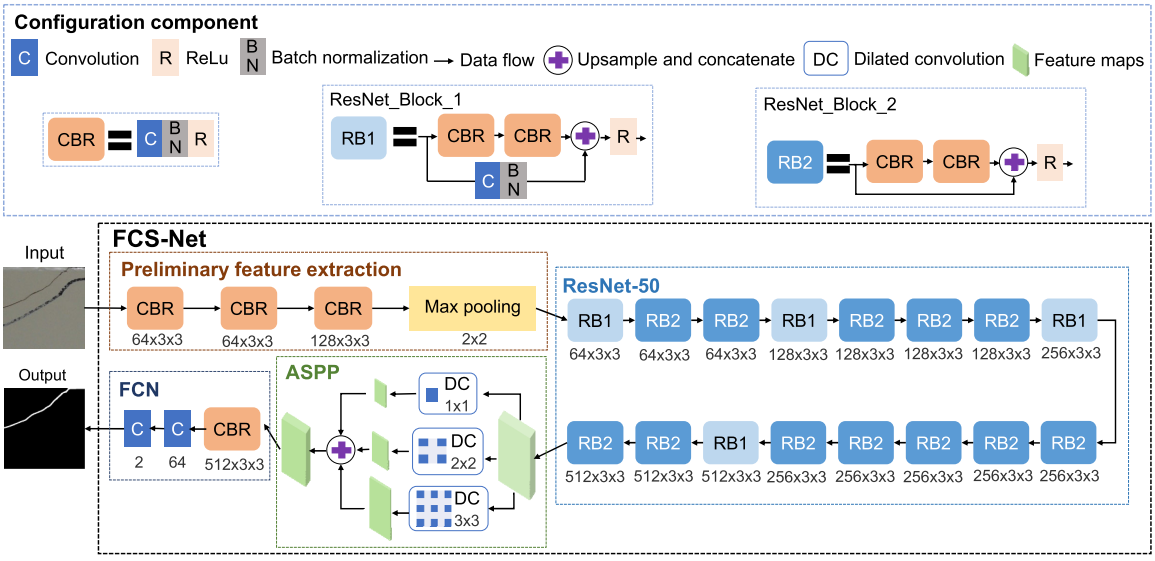


FIGURE 2. Architecture of the proposed FCS-Net.

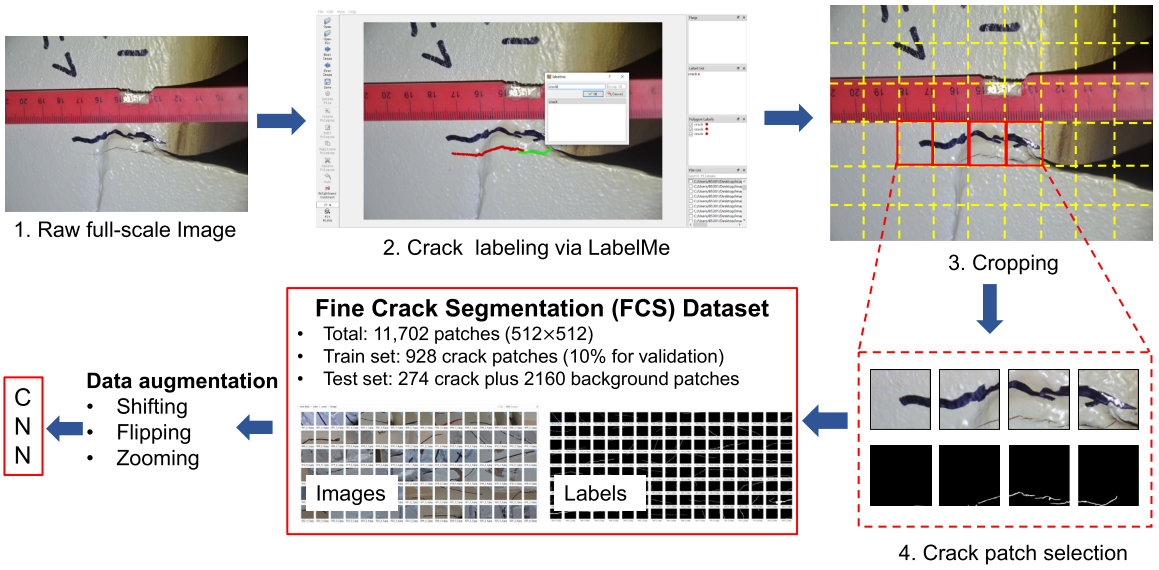


FIGURE 3. Dataset preparation and component.

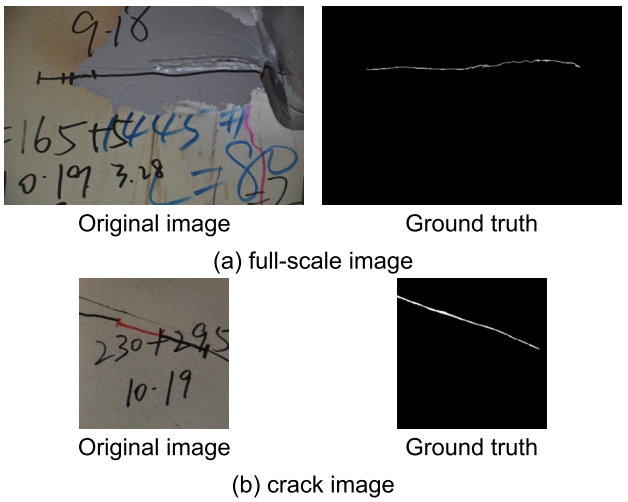


FIGURE 4. Representative (a) full-scale and (b) crack image of the dataset.

Among them, 10% of the images are random selected as the validation set to validate the training of the model. This allows the network in the training process to learn the object features more efficiently with faster convergence, and thus improving the overall network training efficiency. Fig. 3 illustrates the preparation of dataset and its component. And Fig. 4 demonstrates the representative full-scale and crack image of the dataset.

D. TRAINING SPECIFICATIONS

Specifications of the deep learning platform software environment for crack identification framework proposed by this paper are Windows 10, Python 3.7.4, and Keras 2.2.5; platform hardware configuration is CPU Intel i7 9800X with 32GB of memory; configuration for graphics card includes one NVIDIA RTX2080Ti, 11GB of video memory, CUDA10.0, and NVIDIA cuDNN7.4.2 are used for GPU acceleration. Considering the graphic card of the deep learning platform is not designed for deep learning task of large data size, the network structure parameters of proposed model have been adjusted appropriately to avoid the phenomenon of running out of memory. The annotated cracks images are used for training the model by using a backward propagation algorithm. Each deep learning model discussed in this study was trained for 2000 steps and 80 epochs, with a batch size of one. The hyperparameter settings of the model are adjusted according to the validation loss responded after epoch of training meanwhile, the learning rate is adaptively adjusted according to the validation loss from 1E-4 to 1E-6 to achieve higher training efficiently and faster convergence. Table 1 summarized the specific hyperparameters of the model.

TABLE 1. Training details of deep learning models.

PARAMETER	VALUE
Learning rate	Adaptive from 1E-4 to 1E-6
Optimizer	Adam
Loss function	Dice loss
Input size	512×512×3
Step	2000
Epoch	80
Threshold	0.5

IV. RESULTS AND DISCUSSIONS

A. EVALUATION METRICS

There are four kinds of prediction results in a pixel-wide identification task, which are: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). For this case specifically, TP and FP refer to correctly-identified crack and background pixels. FP indicate background pixels are wrongly labeled as crack by the model, while FN are omitted crack samples which are more undesirable than the others from the perspective of safety. Taking proportion of the four results into an integrated consideration, Mean Intersection

over Union (MIoU) was used as an important metric index to measure the accuracy of the models in this study. MIoU can be interpreted as the average ratio of the intersection and union of prediction and ground truth and computes to what degree it has overlapping with the actual one, calculated as:

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \quad (2)$$

where k is the number of class, p_{ii} is the number of TP, p_{ii} is the number of TN, and p_{ii} is the sum of FP and FN.

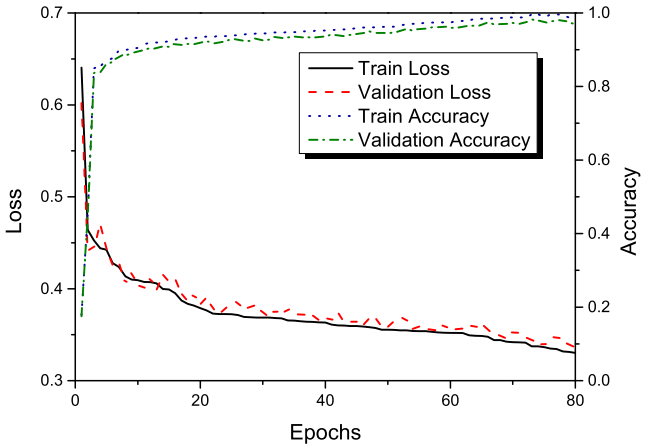


FIGURE 5. Training and validation convergence curves.

B. BENCHMARK PERFORMANCE

The proposed FCS-Net was benchmarked with LinkNet [42], DeepLab V3 Plus [43], and CrackSegNet [1], with regard to segmentation accuracy of fine cracks (Table 2). LinkNet and DeepLab V3 Plus are state-of-the-art semantic segmentation networks, having achieved high performance on multi- and large-scale object datasets like CitySpace or Pascal VOC. However, accuracy of LinkNet and DeepLab V3 significantly declined when detecting fine crack in full-scale images due to the severe data imbalance. In fine crack dataset, the proposed FCS-Net improved MIoU of LinkNet and DeepLab V3 of 9.8% and 15.4%, respectively. Compared with CrackSegNet, which is designed for pixel-wise crack identification in concrete surface, FCS-Net had a performance enhancement of around 2.3%.

Segmentation results in representative crack patches are shown in Fig. 6, in which LinkNet and DeepLab V3 failed to fully extract crack skeleton but output more invalid positive predictions than CrackSegNet and FCS-Net. Being specially modified to identify fine crack, CrackSegNet outperformed LinkNet and DeepLab V3 by successfully recognizing small crack samples. However, this dataset has more complicated backgrounds than concrete surfaces with handwritings and crack-like welding joints, which causes a performance reduction of CrackSegNet in this study.

Confusion matrix results were plotted in Fig. 7, where both LinkNet and DeepLab V3 tend to generate much more noises (FP samples) in background images than the fine

TABLE 2. Benchmark performance comparison of FCS-Net.

MODEL	EVALUATION MODE	MIOU
LinkNet[42]	Crack patch	0.6389
	Full-scale image	0.6594
DeepLab V3 Plus[43]	Crack patch	0.6322
	Full-scale image	0.6274
CrackSegNet[1]	Crack patch	0.7105
	Full-scale image	0.7049
FCS-Net	Crack patch	0.7240
	Full-scale image	0.7408

crack-specialized networks, CrackSegNet and FCS-Net. For example, around 5000 background pixels were improperly classified as crack by CrackSegNet and FCS-Net, respectively. However, the values of LinkNet and DeepLab V3 are nearly 24000 and 19,000, which is the main cause of the performance reductions. Segmentation results of the four networks are almost identical with naked eyes, however, results of LinkNet and DeepLab V3 indicated that both models generate relatively higher proportion of background noises. Whereas, CrackSegNet and the proposed FCS-Net achieved relatively better segmentation performance, which ensured correct identification of the entire cracks without overestimation of mispredictions in background. Moreover, there are discontinuities in crack skeleton extracted by DeepLab V3, which does not conform to the principles of crack generation and propagation. As for details, DeepLab V3 achieve the highest TN, but the highest FN as well, indicating that DeepLab V3 pays more attention to background instead foreground, with the extreme imbalance of positive and negative samples. Compared with CrackSegNet, the proposed FCS-Net recognizes about 2000 more crack pixels than CrackSegNet, which improves the MIoU from 0.7502 to 0.7601 (Fig. 7).

C. ABLATION EXPERIMENTS

The application of ablation experiment was first proposed by Ren [44] in Faster R-CNN, to certify the necessity of different modules in a deep network, by removing each of them and observing variations. In this study, core modules (BN and ASPP) were removed in sequence from the proposed FCS-Net until there was only the ResNet-50 backbone left. With contributions from ASPP and BN, the MIoU of ResNet-50 was improved from 0.6565 to 0.7408 by the proposed FCS-Net (Table 3). It should be noted that the performance of ResNet-50 decreased after adding the ASPP module. This may be caused by enlarged receptive fields and more extracted features enabled by the usage of atrous convolution in ASPP. Details were depicted in Fig. 8, where the original ResNet-50 generated some mispredictions at around the position of the ruler in the input picture.

With existence of BN and ASPP modules, misprediction exist in the location of handwriting, and TP and FP are increased to varied extents, which also verifies the previous inference about the increasing extracted features.

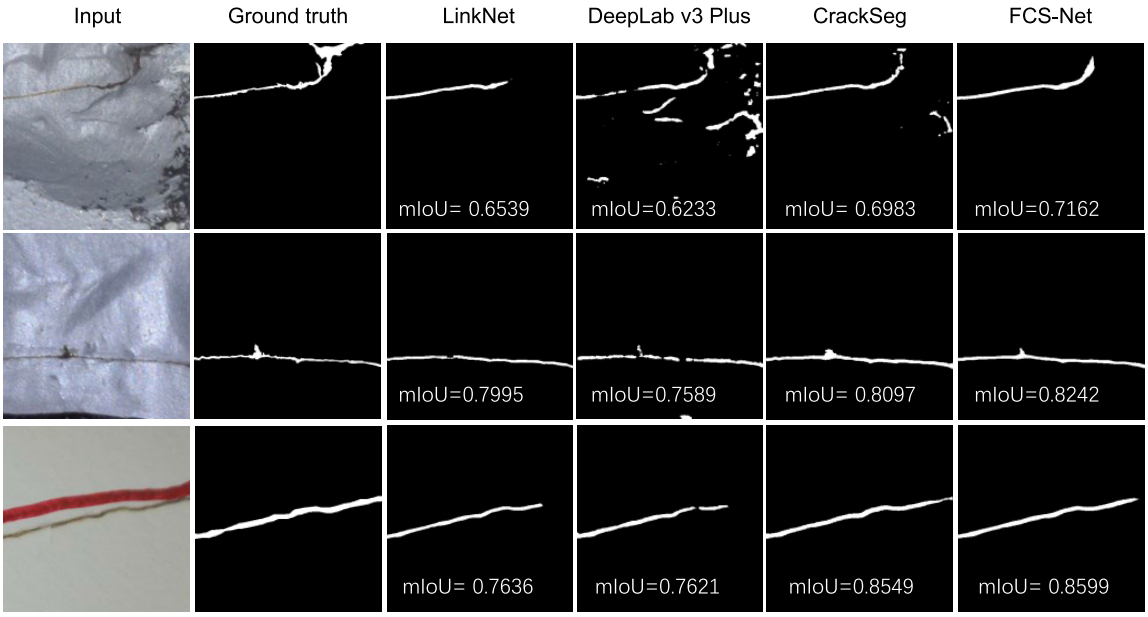


FIGURE 6. Benchmark performance in crack images.

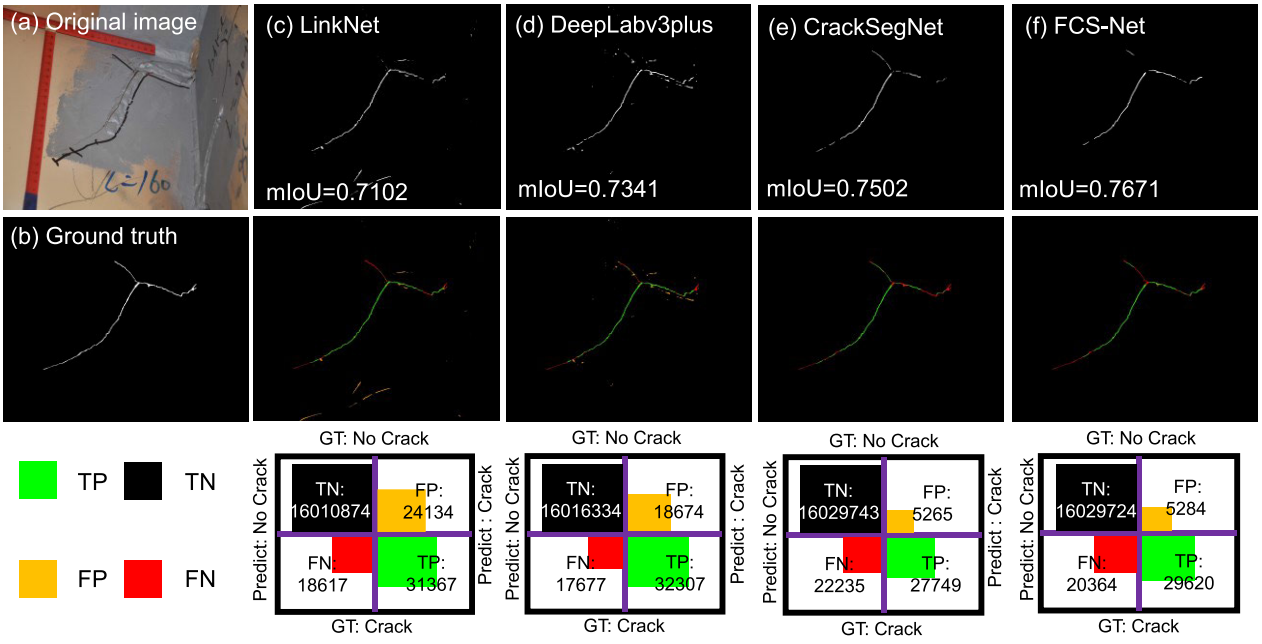


FIGURE 7. Benchmark performance in full-scale image.

Different modules serve the main goal with various functions. BN module can improve the training efficiency of the model, accelerate the convergence of the model, and reduce the cost through normalization. The SPP algorithm can process the input images with different sizes and aspect ratios, which may improve the scale invariance of images and reduce the over-fitting phenomenon during the model training.

With the expansion of receptive field contributed by dilated convolution in ASPP, the model has an enhanced performance in large-scale images. Therefore, the overall MIoU has been further improved when combining BN and ASPP modules, which validate the feasibility of the modifications included in the proposed FCS-Net.

TABLE 3. Ablation experiment of proposed model.

MODEL	MIOU
ResNet-50	0.6565
ResNet-50 + ASPP	0.6548
ResNet-50 + BN	0.6696
ResNet-50 + ASPP + BN (FCS-Net)	0.7408

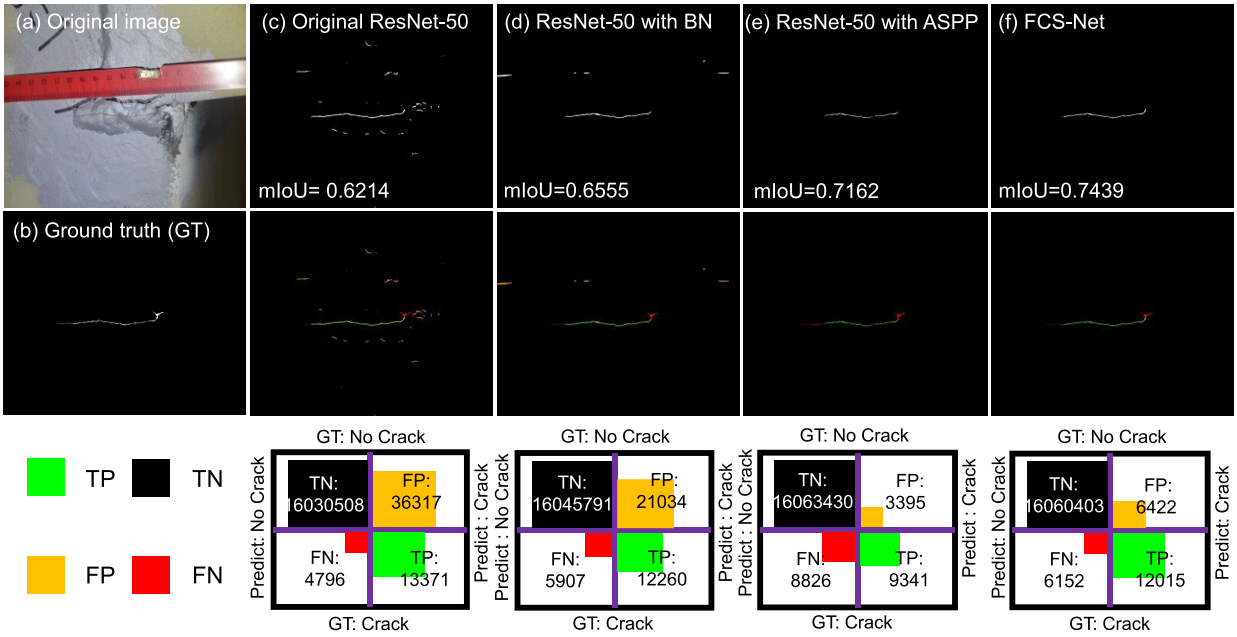


FIGURE 8. Ablation experiment performance in full-scale images.

V. CONCLUSION

To segment fine cracks from complicated large-scale images of steel girder, this study proposed a deep FCN-based network integrating ResNet-50, ASPP, and BN, termed as FCS-Net. The proposed FCS-Net was benchmarked with LinkNet, DeepLab V3, and CrackSegNet with regards to the ability to identify fine cracks with severe background interferences and sample imbalance. Networks specialized for fine crack detection (CrackSegNet and FCS-Net) achieved higher MIoU than LinkNet and DeepLab V3, which are more robust when segmenting multi and large-scale objects. Specifically, the MIoU was enhanced of around 12% by the proposed FCS-Net, compared with LinkNet, indicating its applicability in pixel-wise detection of fine cracks.

CODE AVAILABILITY

The source code and data used in this article can be found in https://github.com/Monash-Civil-CV-Team/FCS-Net.

ACKNOWLEDGMENT

The authors appreciate the organizations of the IPC-SHM 2020 ANCRiSST, Harbin Institute of Technology, China, and the University of Illinois at Urbana-Champaign, Champaign, IL, USA, for their generous preparing and sharing the invaluable dataset (http://www.schm.org.cn/#/IPC-SHM,2020/dataDownload). They would also like to thank the chairs of IPC-SHM 2020 Prof. Hui Li and Prof. Billie F. Spencer Jr., for their leadership on the competition.

REFERENCES

[1] Y. Ren, J. Huang, Z. Hong, W. Lu, J. Yin, L. Zou, and X. Shen, "Image-based concrete crack detection in tunnels using deep fully convolutional networks," *Construct. Building Mater.*, vol. 234, Feb. 2020, Art. no. 117367.

[2] T. Yu, A. Zhu, and Y. Chen, "Efficient crack detection method for tunnel lining surface cracks based on infrared images," *J. Comput. Civil Eng.*, vol. 31, no. 3, May 2017, Art. no. 04016067.

[3] C. M. Yeum and S. J. Dyke, "Vision-based automated crack detection for bridge inspection," *Comput. Aided Civil Infrastruct. Eng.*, vol. 30, no. 10, pp. 759–770, Oct. 2015.

[4] S. Lee, L.-M. Chang, and M. Skibniewski, "Automated recognition of surface defects using digital color image processing," *Autom. Construct.*, vol. 15, no. 4, pp. 540–549, Jul. 2006.

[5] T. H. Dinh, Q. P. Ha, and H. M. La, "Computer vision-based method for concrete crack detection," in *Proc. 14th Int. Conf. Control, Autom., Robot. Vis. (ICARCV)*, Nov. 2016, pp. 1–6.

[6] Y.-J. Cha, W. Choi, and O. Büyüköztürk, "Deep learning-based crack damage detection using convolutional neural networks," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 32, no. 5, pp. 361–378, May 2017.

[7] R. Davoudi, G. R. Miller, and J. N. Kutz, "Data-driven vision-based inspection for reinforced concrete beams and slabs: Quantitative damage and load estimation," *Automat. Construct.*, vol. 96, pp. 292–309, Dec. 2018.

[8] A. Mohan and S. Poobal, "Crack detection using image processing: A critical review and analysis," *Alexandria Eng. J.*, vol. 57, no. 2, pp. 787–798, 2018.

[9] Y.-J. Cha, W. Choi, G. Suh, S. Mahmoudkhani, and O. Büyüköztürk, "Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 33, no. 9, pp. 731–747, 2018.

[10] X. Yang, "Automatic pixel-level crack detection and measurement using fully convolutional network," *Comput. Aided Civil Infrastruct. Eng.*, vol. 33, no. 12, pp. 1090–1109, 2018.

[11] C. V. Dung and L. D. Anh, "Autonomous concrete crack detection using deep fully convolutional neural network," *Autom. Construct.*, vol. 99, pp. 52–58, Mar. 2019.

[12] F. Ni, J. Zhang, and Z. Chen, "Pixel-level crack delineation in images with convolutional feature fusion," *Struct. Control Health Monitor.*, vol. 26, no. 1, Jan. 2019, Art. no. e2286.

[13] B. F. Spencer, V. Hoskere, and Y. Narazaki, "Advances in computer vision-based civil infrastructure inspection and monitoring," *Engineering*, vol. 5, no. 2, pp. 199–222, Apr. 2019.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[15] T. C. Hutchinson and Z. Chen, "Improved image analysis for evaluating concrete damage," *J. Comput. Civil Eng.*, vol. 20, no. 3, pp. 210–216, May 2006.

[16] H. Zhao, G. Qin, and X. Wang, "Improvement of Canny algorithm based on pavement edge detection," in *Proc. 3rd Int. Congr. Image Signal Process.*, Oct. 2010, pp. 964–967.

[17] T. Yamaguchi and S. Hashimoto, "Fast crack detection method for large-size concrete surface images using percolation-based image processing," *Mach. Vis. Appl.*, vol. 21, no. 5, pp. 797–809, Aug. 2010.

[18] F. Bi, "A visual search inspired computational model for ship detection in optical satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 4, pp. 749–753, Feb. 2012.

[19] H. Sun, X. Sun, H. Wang, Y. Li, and X. Li, "Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 1, pp. 109–113, Jan. 2012.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, vol. 1, 2012, pp. 1097–1105.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[23] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017, *arXiv:1704.06857*. [Online]. Available: http://arxiv.org/abs/1704.06857

[24] D. Kang, S. S. Benipal, D. L. Gopal, and Y.-J. Cha, "Hybrid pixel-level concrete crack segmentation and quantification across complex backgrounds using deep learning," *Autom. Construct.*, vol. 118, Oct. 2020, Art. no. 103291.

[25] J. Liu, X. Yang, S. Lau, X. Wang, S. Luo, V. C. S. Lee, and L. Ding, "Automated pavement crack detection and segmentation based on two-step convolutional neural network," *Comput.-Aided Civil Infrastruct. Eng.*, 2020. vol. 35, no. 11, pp. 1291–1305.

[26] A. Zhang, K. C. Wang, B. Li, E. Yang, X. Dai, Y. Peng, Y. Fei, Y. Liu, J. Q. Li, and C. Chen, "Automated pixel-level pavement crack detection on 3D asphalt surfaces using a deep-learning network," *J. Comput.-Aided Civil Infrastruct. Eng.*, vol. 32, no. 10, pp. 805–819, 2017.

[27] W. Choi and Y.-J. Cha, "SDDNet: Real-time crack segmentation," *IEEE Trans. Ind. Electron.*, vol. 67, no. 9, pp. 8016–8025, Sep. 2020.

[28] A. Zhang, K. C. P. Wang, Y. Fei, Y. Liu, S. Tao, C. Chen, J. Q. Li, and B. Li, "Deep learning–based fully automated pavement crack detection on 3D asphalt surfaces with an improved CrackNet," *J. Comput. Civil Eng.*, vol. 32, no. 5, Sep. 2018, Art. no. 04018041.

[29] Y. Xu, S. Li, D. Zhang, Y. Jin, F. Zhang, N. Li, and H. Li, "Identification framework for cracks on a steel structure surface by a restricted Boltzmann machines algorithm based on consumer-grade camera images," *Struct. Control Health Monitor.*, vol. 25, no. 2, Feb. 2018, Art. no. e2075.

[30] Y. Xu, Y. Bao, J. Chen, W. Zuo, and H. Li, "Surface fatigue crack identification in steel box girder of bridges by a deep fusion convolutional neural network based on consumer-grade camera images," *Struct. Health Monitor.*, vol. 18, no. 3, pp. 653–674, May 2019.

[31] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[32] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[37] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[38] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1091–1100.

[39] G. Yang, G. Li, T. Pan, Y. Kong, J. Wu, H. Shu, L. Luo, J.-L. Dillenseger, J.-L. Coatrieux, L. Tang, and X. Zhu, "Automatic segmentation of kidney and renal tumor in CT images based on 3D fully convolutional neural network with pyramid pooling module," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3790–3795.

[40] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2016, *arXiv:1511.07122*. [Online]. Available: https://arxiv.org/abs/1511.07122

[41] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 182–186.

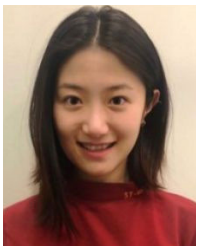
[42] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[43] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[44] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

ZHIHANG LI (Graduate Student Member, IEEE) received the bachelor's (Hons.) degree from Monash University, Australia, in 2018, where he is currently pursuing the M.Eng.Sc. (Research) degree in civil engineering. He is also researching on the TBM performance prediction with neural network. He is also interested in the application of neural network in tunneling.

HUAMEI ZHU received the bachelor's degree in civil engineering from Hefei University of Technology, China, in 2016, and the master's degree in transportation engineering from Southeast University–Monash University Joint Graduate School, in 2019. She is currently pursuing the Ph.D. degree with Monash University, Melbourne, VIC, Australia. Her research interests include employing computer vision and deep learning techniques to automate structural health monitoring in underground infrastructures.



MENGQI HUANG received the bachelor's degree (Hons.) in mining engineering from Monash University, in 2018, where she is currently pursuing the Ph.D. degree. Her current research interests include smart underground construction with building information modeling and deep learning.