Hindawi
Structural Control and Health Monitoring
Volume 2023, Article ID 7552337, 15 pages
https://doi.org/10.1155/2023/7552337

Research Article

Automated Pixel-Level Detection of Expansion Joints on Asphalt Pavement Using a Deep-Learning-Based Approach

Anzheng He,[1] **Zishuo Dong**,[1] **Hang Zhang**,[1] **Allen A. Zhang**,[1] **Shi Qiu**,[2] **Yang Liu**,[3] **Kelvin C.P. Wang**,[3] **and Zhihao Lin**[4]

[1]*School of Civil Engineering, Southwest Jiaotong University, Chengdu 610031, China*
[2]*School of Civil Engineering, Central South University, Changsha 410075, China*
[3]*School of Civil and Environmental Engineering, Oklahoma State University, Stillwater 74078, OK 74078, USA*
[4]*Sichuan Shudao New Energy Technology Development Co., Ltd., Chengdu 610041, China*

Correspondence should be addressed to Shi Qiu; sheldon.qiu@csu.edu.cn

Received 13 March 2023; Revised 8 April 2023; Accepted 27 April 2023; Published 23 May 2023

Academic Editor: Ka-Veng Yuen

Pixel-level detection of expansion joints on complex pavements is significant for traffic safety and the structural integrity of highway bridges. This paper proposed an improved HRNet-OCR, named as expansion joints segmentation network (EJSNet), for automated pixel-level detection of the expansion joints on asphalt pavement. Different from the high-resolution network (HRNet), the proposed EJSNet modifies the residual structure of the first stage by conducting a Conv. + BN + ReLU (convolution + batch normalization + rectified linear unit) operation for each shortcut connection, which can avoid the network degradation. The feature selection module (FSM) and receptive field block (RFB) module are incorporated into the proposed EJSNet model to learn and extract the contexts at different resolution levels for enhanced latent representations. The convolutional block attention module (CBAM) is introduced to enhance the adaptive feature refinement of the network. Moreover, the shared multilayer perceptron (MLP) architecture of the channel attention module (CAM) is also modified in this paper. Experimental results demonstrate that the *F*-measure and intersection-over-union (IOU) attained by the proposed EJSNet model on 500 testing image sets are 95.14% and 0.9036, respectively. Compared with four state-of-the-art models for semantic segmentation (i.e., SegNet, DeepLabv3+, dual attention network (DANet), and HRNet-OCR), the proposed EJSNet model can yield higher detection accuracy on both private and public datasets.

1. Introduction

In the field of highway and bridge engineering, expansion joints are generally used as a special connection device to accommodate the structural deformation of highway bridges under multiple loads [1]. However, due to the repetitive effects of heavy vehicle loads and environmental conditions, expansion joints are generally one of the weakest structures of highway bridges. Once damages, such as the restriction of free movements or unexpected movements occur, significant secondary stress will be imposed on the structure, which will lead to the performance degradation of the expansion joints faster than expected [2]. The common types of expansion joint damage include "subsidence of expansion joints," "cracking/lack of concrete around the expansion joints," and "narrowing of expansion joint gaps." When a vehicle passes through a damaged expansion joint, it causes additional stress on the structure due to jumping behaviour, which affects the service life of bridges and roads, reduces the quality of road service, and even causes serious traffic accidents [3, 4]. Therefore, timely and accurate acquisition of damage information of expansion joints is significant for the structural integrity and traffic safety.

With the development of structural health monitoring (SHM) techniques, many effective solutions for expansion joint damage detection and evaluation based on long-term monitoring data have been developed with various successes. For instance, based on the temperature and displacement data of the Ting Kau Bridge, Ni et al. [5] presented an expansion joint condition assessment procedure and

suggested using the cumulative displacement to assess the health status of expansion joints. Miao et al. [6] developed a technique for the expansion joints damage alarm and utilized an X-bar control chart to detect the abnormal changes in the displacement in expansion joints. Ni et al. [7] proposed a Bayesian-based probabilistic method for condition assessment and damage alarm of bridge expansion joints with the regression between bridge longitudinal displacement and temperature. However, Zhou et al. [8] noted that due to the randomness of temperature variation and three-dimensional characteristics of temperature distributions in the bridge, unreliable damage warnings may be issued when the temperature-displacement relationship is described by a linear or nonlinear regression model. In addition, some other methods that combine intelligent algorithms are also proposed, which show great potential in modeling temperature-induced displacement. Ding et al. [9] implemented improved backpropagation neural networks (BPNNs) to model the correlations between modal frequencies of the Runyang Suspension Bridge and environmental conditions. On the basis of the LSSVM technique, Chen et al. [10] established a temperature-displacement learning model integrated with the PCA and HMFA (an improved firefly algorithm), which was used to detect damages in expansion joints of the road-rail cable-stayed bridge. However, due to the limitations of traditional intelligent algorithms, there are still many challenges to achieve high precision and better robustness concerning the damage detection and evaluation of expansion joints.

Intelligent pixel-level pavement survey has gained breakthroughs with the aid of advanced deep-learning models in recent years. Particularly, deep-learning-based pavement distress (e.g., cracks) detection has attracted much attention [11–13]. However, current studies have paid insufficient attention to the intelligent detection of surface design features (e.g., expansion joints), especially expansion joints. Kim et al. [4] developed an automatic image recognition-based survey system for highway bridges that combines the machine vision (M/V) technology and deep learning models. Such a survey system can monitor the expansion joints gap abnormalities through image analysis while driving at high speed. Based on the integrated architecture of pyramid scene parsing network (PSPNet) [14] and U-Net [15], Wen et al. [16] proposed an end-to-end pavement distress segmentation network (PDSNet) for simultaneous pixel-level detection of multiple asphalt pavement distresses (e.g., cracks, patches, and bridge joints). Moreover, Zhang et al. [17] proposed a deep-learning model named ShuttleNet for simultaneous pixel-level detection of multiple distresses (e.g., cracks and patches) and surface design features (e.g., expansion joints and markings) on asphalt pavements. The ShuttleNet model can perceive the global context as well as finer details many times by repeating the encoding-decoding round.

Pixel-level detection accuracy can provide precise geometric features of damages for the quantitative evaluation of expansion joint conditions on highway bridges. In the field of computer vision, the meaning of semantic segmentation is to assign a class label to each image pixel, so semantic segmentation can be regarded as pixel-level detection. Presently, numerous robust deep-learning networks have been developed for semantic segmentation and applied in various industries, including the U-Net, PSPNet, SegNet [18], DeepLabv3+ [19], and high-resolution network (HRNet) [20]. To solve the loss of edge information problem, the encoder-decoder architecture was developed and progressively became a prevalent solution for semantic segmentation. It has been adopted in many state-of-the-art models, such as U-Net, U2-Net [21], and DeepLabv3+. Encoder-decoder networks normally use a skip connection or similar shortcut structures to fuse low-level semantic features with high-level semantic features to improve edge segmentation [22]. Unfortunately, Xu et al. [23] pointed out that the fusion method of skip connection can destroy high-level semantic representation, resulting in an over-segmentation problem. In view of the over-segmentation problem, some deep learning solutions for constructing context information have been developed, including multiscale context networks (e.g., PSPNet, DeepLabv3+) and relational context networks (e.g., dual attention network (DANet) [24], HRNet-OCR [25]). HRNet-OCR represents a recent study on relational context networks, which augment the representation of a pixel by exploiting the representation of the object region of the corresponding class. Compared to PSPNet, DeepLabv3+, and HRNet-OCR significantly enhances the contribution of pixels from the same class of object, resulting in obtaining a more targeted object context and better segmentation performances.

This paper uses HRNet-OCR as the baseline and proposes an improved HRNet-OCR model to detect expansion joints with pixel-level accuracy. Considering that the OCR ignores the correlation of channel features [26], a convolutional block attention module (CBAM) [27] is also adopted in the paper to calculate the weights of the channel and spatial features to enhance the adaptive feature refinement. Furthermore, the feature selection module (FSM) [28] and receptive field block (RFB) [29] module are also incorporated into the proposed modification to learn and extract the contexts at different resolution levels for enhanced latent representations. In summary, the primary contributions of this paper can be included as follows:

(1) An approach to pixel-level detection of the expansion joints on asphalt pavements based on 2D pavement images that can accomplish robust recognition without misidentifying other noise patterns such as pavement background, cracks, patches, and markings.

(2) A modified residual structure of the first stage of the HRNet that can avoid the network degradation.

(3) A modified channel attention module (CAM) that can reduce noises resulting from invalid channel features.

2. Methodology

As shown in Figure 1(b), the proposed expansion joints segmentation network (EJSNet) is an end-to-end deep learning model with a modified HRNetV2-W32 as the

backbone. The proposed EJSNet first extracts multiscale features and summarizes latent representations through the encoder, and then uses the decoder to retrieve object details and output the final segmentation result.

For better adaptivity in pixel-level detection of expansion joints, the encoder-decoder architecture of HRNet-OCR is modified in this paper, and the modifications to the encoder architecture can be summarized in three aspects. First, the original residual structure of the first stage of HRNet is modified by conducting a 1×1 Conv. + BN + ReLU (convolution + batch normalization + rectified linear unit) operation for each shortcut connection to avoid the network degradation. Second, before rescaling the low-resolution representations through bilinear upsampling to the high resolution, the feature selection module (FSM) is added to enhance multiscale feature aggregation (see stem2 section in Figure 1(a)). Last, the receptive field block (RFB) is applied at the end of feature extraction to expand the receptive field and summarize latent representations. In addition, the convolutional block attention module (CBAM) is incorporated into the original decoder architecture of HRNet-OCR, and the representation of RFB is first fed into CBAM. The noises from invalid channels and spatial features have been successively reduced by using channel and spatial feature maps with different weights rationally and refining the intermediate features adaptively. The representation of CBAM is applied to predict the coarse segmentation result (soft object regions) and used as an input of OCR. Also, the output of the representation of CBAM to go through a 3×3 Conv. + BN + ReLU operation is taken as another input of OCR. In this case, the output of OCR means the augmented representation of features. Afterwards, the number of channels of the OCR output is adjusted by a 1×1 convolution and then the output is restored to the original size by a bilinear upsampling operation with a factor of 4. Ultimately, the final prediction result is obtained by nonlinear activation using the sigmoid function. The following sections will analyse the backbone and each module specifically.

2.1. Backbone: HRNetV2-W32. HRNetV2-W32 is employed as the backbone of the proposed EJSNet. As illustrated in Figure 1(a), the main body of HRNetV2-W32 contains four stages, and the number of the corresponding parallel convolution streams are 1, 2, 3, and 4, respectively [25]. As shown in Figure 1(a) and Figure 2, each convolution stream contains four residual units. The residual unit of the first stage is formed by a bottleneck which contains two 1×1 convolutions and one 3×3 convolution, and the residual unit of other stages contains two 3×3 convolutions. Batch normalization (BN) and rectified linear unit (ReLU) are applied sequentially after each convolution to normalize hidden features and address nonlinear activation. In addition, the number of channels of the four parallel branches are C, 2C, 4C, and 8C, respectively, and the corresponding scales are 1/4, 1/8, 1/16, and 1/32, respectively. In particular, channel number C is 32 in this paper. Given the input image, the original resolution is reduced to 1/4 of the original resolution, while the number of channels of the convolutional layer is increased to 64 through two 3×3 Conv. + BN + ReLU operations (see stem1 section in Figure 1(a), and Conv. stands for convolution). In the same stage, the high-to-low resolution convolution streams are connected in parallel. Between adjacent stages, the branch expansion and feature fusion of the high-to-low resolution convolution streams are performed. By connecting the high-to-low resolution convolution streams in parallel and repeatedly fusing the multiscale resolution information, the high-resolution representation can be well boosted with the help of the low-resolution representations, and vice versa.

The proposed EJSNet model modifies the residual architecture of the first stage. As illustrated in Figure 3, the 1×1 Conv. + BN + ReLU is performed for each shortcut connection to avoid network degradation. Furthermore, at the end of the modified HRNetV2-W32 model, a bilinear upsampling operation with a factor of 2 is performed 1, 2, and 3 times, respectively, for three low-resolution representations from top to bottom, to let the feature map size of the three low-resolution representations be the same as the feature map size of the high-resolution representation (see stem2 section in Figure 1(a)). Moreover, as the unlearnable nature of bilinear upsampling and the repeated bilinear upsampling operations will cause a redundant feature map, the FSM is introduced before each bilinear upsampling operation to enhance multiscale feature aggregation. Ultimately, fusing the feature map outputs of four resolution representations through channel connections yields the backbone output.

2.2. Feature Selection Module. Skip connection can avoid any particular channel responses to be over-amplified or over-suppressed [28]. As discussed previously, the repeated bilinear upsampling operation will result in a redundant feature map. The feature selection module (FSM) is thus introduced before each bilinear upsampling operation in this paper to extract the feature maps containing more spatial details and enhance multiscale feature aggregation (see stem2 section in Figure 1(a)). General architecture of the FSM is illustrated in Figure 4.

First, the global information of each input feature map K_i is extracted by a global average pooling operation, followed by a 1×1 Conv. + BN + sigmoid (convolution + batch normalization + sigmoid activation function) operation to model the importance of each feature map and output an importance vector u. Second, scaling the original input feature maps K_i with the importance vector u, and adding the scaled feature maps to the K_i using a skip connection, named as rescaled feature maps. Last, a feature selection layer $f_s(\cdot)$ (i.e., a 1×1 convolution) is applied on the rescaled feature maps to reserve important feature maps.

2.3. Receptive Field Block Module. Abundant object-contextual information can be effectively obtained by expanding the receptive field. The receptive field of the network is generally increased by adopting larger convolutional kernels or greater pooling strides. However, the

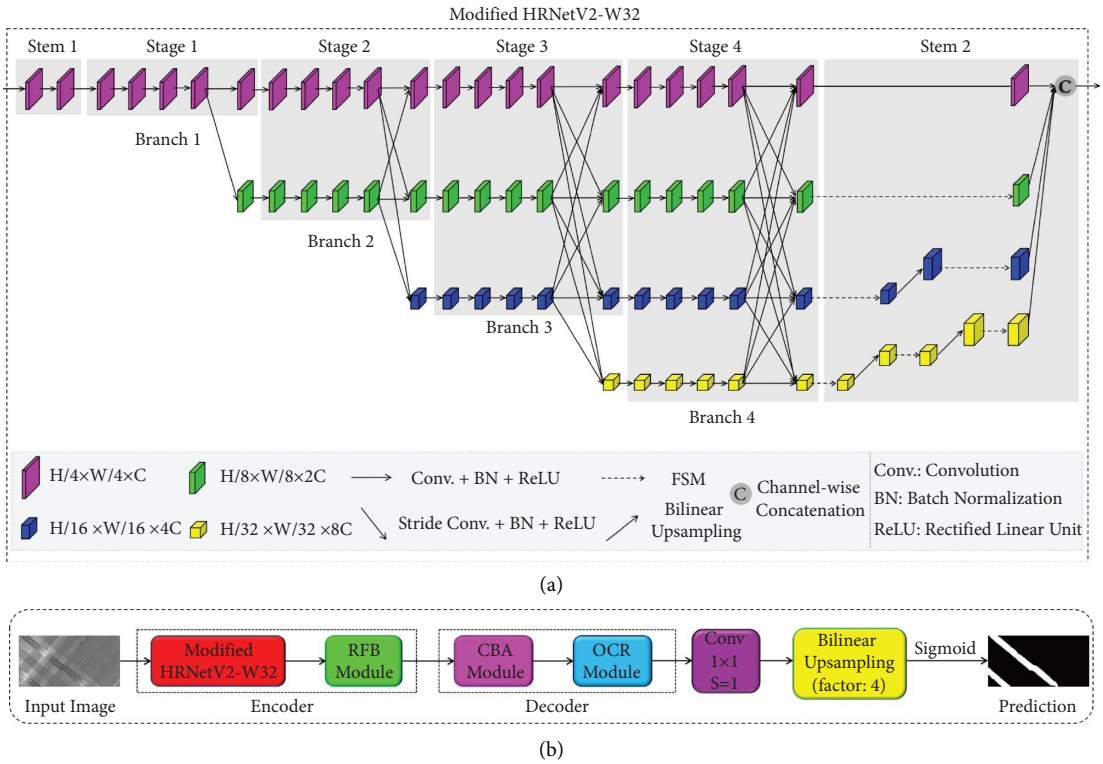

Figure 1: General architecture of the network. (a) The modified HRNetV2-W32. (b) The proposed expansion joints segmentation network (EJSNet).

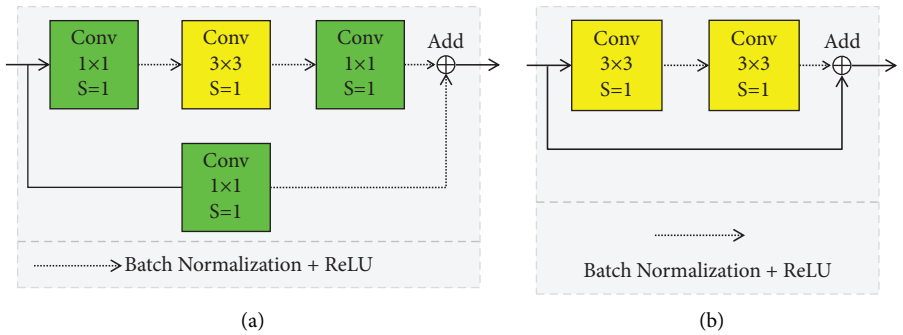

Figure 2: Residual units used in each stage. (a) The first stage. (b) The other stages.

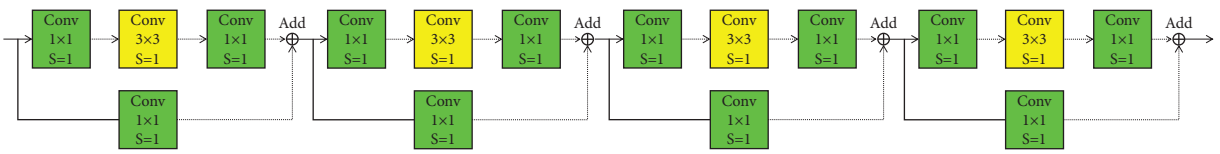

Figure 3: Modified residual architecture of the first stage.

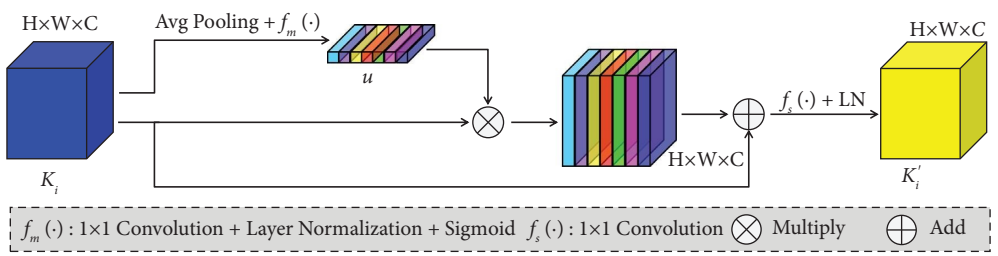

Figure 4: General architecture of the feature selection module.

former increases the computational cost, and the latter loses information. To solve such problems, the dilated convolution was developed and applied in some receptive field modules with various successes (e.g., atrous spatial pyramid pooling (ASPP) [30] and receptive field block (RFB) [29]). By comparing the relevant experiments, the RFB module is introduced in this paper to capture more contexts while minimizing the information loss. The general architecture of the RFB module is illustrated in Figure 5. RFB is a multibranch convolutional block consisting of a multibranch convolution layer with different kernels and a trailing dilated pooling with different dilation rates. To begin with, the number of channels of the input feature map is decreased by a 1×1 convolution to aggregate information, and then a series of convolution and dilated convolution operations are performed in parallel for the first three branches. Next, the feature maps of the above three-branch representations are fused through channel connection, followed by a 1×1 convolution to restore the channel number of the feature map. Last, scaling the above output by a factor of 1.0 and adding the scaled feature map to the shortcut output using an element-wise addition. Ultimately, the feature map goes through nonlinear activation ReLU to yield the final output.

2.4. Convolutional Block Attention Module. Relevant studies have proven the effectiveness of the attention mechanism in improving the feature representation ability of the network [27, 31, 32]. Given that the object-contextual representations (OCR) module ignores the correlation of channel features, a convolutional block attention module (CBAM) is introduced in this paper to help the network focus on important features and enhance adaptive feature refinement. As illustrated in Figure 6, the general architecture of the CBAM consists of two submodels: channel attention module (CAM) and spatial attention module (SAM). First, the channel attention feature M_c of each input feature map F is extracted with the CAM, and then the M_c and F are operated by element-wise multiplying to output the new attention feature F_1. Then, F_1 is fed into SAM, and the spatial attention feature M_s is obtained with the SAM. Last, M_s and F_1 are operated by element-wise multiplying to output the final refined feature map.

Moreover, the proposed EJSNet model also modifies the shared multilayer perceptron (MLP) architecture of CAM. Specifically, as illustrated in Figure 7, an extra dense layer (this layer operates the same as the original first dense layer) is added to the middle of the MLP such that the noise effect of invalid channel features can be further decreased.

2.5. Object-Contextual Representations Module. The fundamental idea of object-contextual representations (OCR) is to augment the representation of a pixel by utilizing the representation of the object region of the corresponding class [25]. The general architecture of the OCR module is illustrated in Figure 8. The OCR mainly consists of soft object regions, object region representations, and object-contextual representations. In addition, the OCR is normally implemented in three steps. First, the contextual pixels are divided into a set of soft object regions with each corresponding to a class. Particularly, the soft object region is a coarse semantic segmentation output that is utilized as an input in OCR, where a loss named as Loss1 is introduced to monitor the convergence direction of the network and assist in the completion of the final semantic segmentation result. Second, k sets of vectors are calculated according to the coarse semantic segmentation results and the pixel representations (the number of k is 2 in this paper). Finally, the relationship matrix between the pixel representations and the object region representation is computed, and then the object-contextual representation with the weighted summation is obtained according to the value of each pixel and the object region features expressed in the relationship matrix.

3. Data Preparation

The 3,229 sets of expansion joints data used in the paper are acquired by the PaveVison3D system, which can collect 2D and 3D pavement images at a maximum speed of 60 MPH with full coverage for a 4-meter-wide lane. In this paper, the 3229 data sets are randomly split into 2129 training data sets, 600 validation data sets, and 500 testing data sets, serving as the source data to train and evaluate different deep-learning networks.

Figure 9 illustrates several representative matched sets of 2D images, 3D images, and ground-truth images. It can be observed that both 2D and 3D images can well reflect the gap features of expansion joint. However, compared with 2D images, the features of the expansion joint on 3D images are relatively unapparent (shown in the dashed circles in Figure 9), especially serrated expansion joints, and the elevation representation of expansion joints on 3D images is invisible (shown in the dashed rectangles in Figure 9). Considering that the purpose of this research is to accurately segment the expansion joint on asphalt pavement; therefore, the 2D pavement images are used to train each network model.

In addition, the size of 2D images is 256×512 ($H \times W$). Each expansion joints data set contains a 2D pavement image and a ground-truth image that are matched in a pixel-to-pixel manner, and all the ground-truth images were manually labelled using the GIMP tool (https://www.gimp.org/downloads/). Specifically, importing a 2D pavement image in GIMP, first creating a single channel black layer (bottom layer) and a single channel transparent layer, then labelling all expansion joint features on the transparent layer, and setting the pixel value of the labelled area to 255. Finally, the annotated ground-truth image that matches the 2D pavement image in a pixel-to-pixel manner is saved in PNG format.

4. Training

4.1. Training Details. To ensure the fairness of evaluating the overall performances of all trained networks, all the networks are treated identically with the training setting elaborated in Table 1 and trained under the TensorFlow 2.5 environment using the aforementioned training data and validation data. In addition, the optimal parameters are saved by monitoring the network performances on validation data. Specifically, during the whole training process,

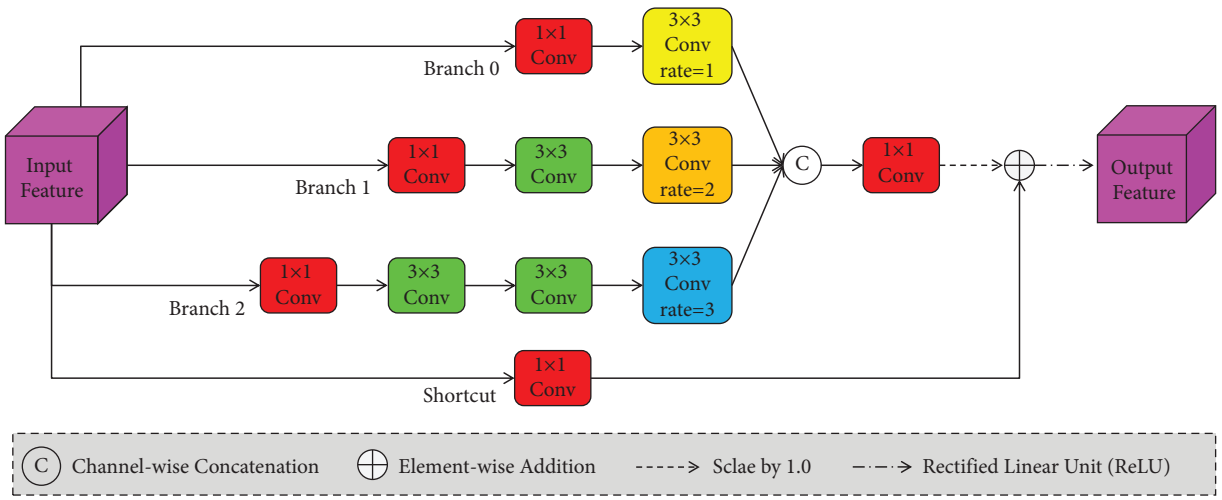

Figure 5: General architecture of the receptive field block module.

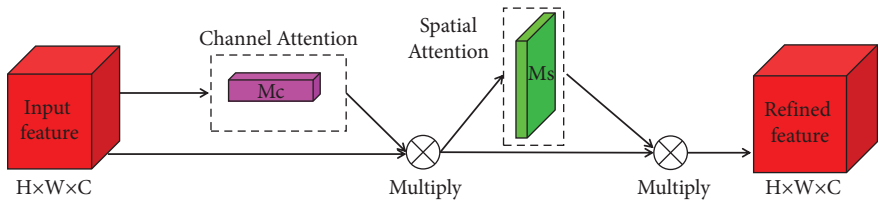

Figure 6: General architecture of the convolutional block attention module.

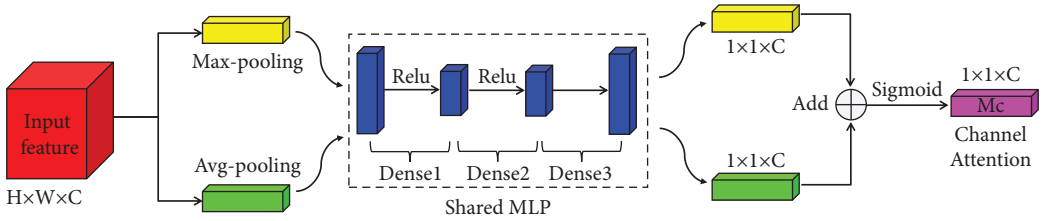

Figure 7: General architecture of the channel attention module.

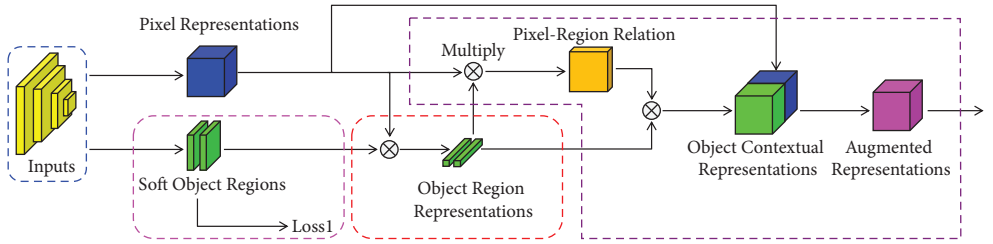

Figure 8: General architecture of the object-contextual representations module.

when the value of intersection-over-union (IOU) on validation data reaches a higher value, the corresponding parameters will thus be saved dynamically to obtain the optimal parameters eventually.

4.2. Performance Metrics. This paper adopts recall, precision, *F*-measure, and intersection-over-union (IOU) as the fundamental indicators of network performance. The four indicators can be expressed in the following equations.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{2}$$

$$F - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{3}$$

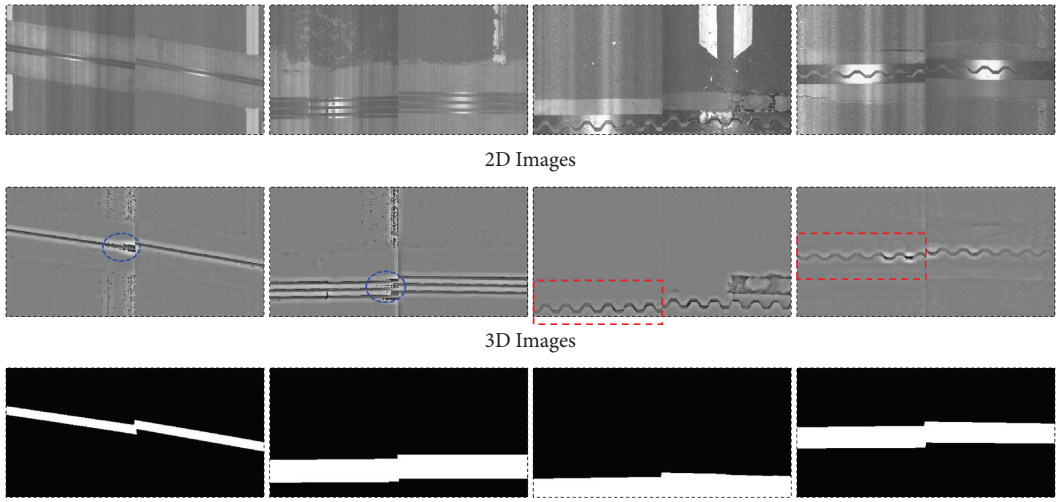

Figure 9: Representative matched sets of 2D images, 3D images, and ground-truth images.

Table 1: Hardware and hyper-parameter settings for the training.

Hardware	Hyper parameters			
	Optimizer	Batch size	Epochs	Learning rate
GPU: RTX 3060 (6 GB) CPU: i7-10870H RAM: 16.0 GB	Adam	3	60	0.0001

$$\text{IOU} = \frac{TP}{\text{TP} + \text{FN} + \text{FP}}, \tag{4}$$

where TP, FP, and FN are the numbers of true positives, false positives, and false negatives corresponding to the pixel-level prediction, respectively.

The recall represents the ratio of relevant instances retrieved to all relevant instances, serving as an important indicator of false-negative errors. The precision stands for the ratio of relevant instances retrieved to all retrieved instances, serving as an important indicator of false-positive errors. The *F*-measure means the harmonic mean of precision and recall, reflecting the balanced accuracy of an algorithm. The IOU represents the ratio of the intersection of prediction and ground truth to the union of prediction and ground truth, defining the accuracy of an algorithm in a unified manner [17]. The recall and precision are contradictory in most cases, while both the *F*-measure and IOU can accurately reflect the accuracy of the algorithm. For balanced evaluations, the *F*-measure and IOU are adopted as two primary indicators in this paper to evaluate the overall performances of trained networks.

4.3. Loss Function Selection. Binary cross-entropy (BCE) loss is employed as the loss function in HRNet-OCR [25], which is expressed in (5). Compared with the background pixels, the expansion joint pixels generally account for a smaller proportion in pavement images, resulting in the quantity imbalance between positive (expansion joint pixels) and negative (background pixels) samples. Milletari et al. [33] introduced a dice loss to help the network avoid getting trapped in the local minima of the loss function during the learning process so that the network prediction will not be strongly biased towards the background pixels. This is undoubtedly beneficial to solving the quantity imbalance problem. Hence, this paper adopts the dice loss as the loss function, which can be expressed in equation (6).

$$\text{Loss}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i), \tag{5}$$

where y_i and $\hat{y}_i$ are the true value and predicted value of the pixel i, respectively. N is the final output size ($H \times W$) used to calculate the loss for the network.

$$\text{Loss}_{\text{Dice}} = 1 - \frac{2 \times \sum_{i=1}^{N} y_i \cdot \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i}, \tag{6}$$

where y_i and $\hat{y}_i$ are the true value and predicted value of the pixel i, respectively. N is the total number of pixels of a mini-batch of images.

To further verify the rationality of the choice of dice loss as the loss function, both dice loss and BCE loss are utilized to train the proposed EJSNet, respectively. The results are illustrated in Table 2 and Figure 10.

Figure 10 illustrates the *F*-measure and intersection-over-union (IOU) on validation data observed during the training process. Compared to dice loss, BCE loss can result in comparatively stable network performances throughout the whole training. However, the optimal network performance that BCE loss can achieve is inferior to that yielded by

Table 2: Network performances on training data and validation data.

Loss functions	Training data		Validation data	
	F-measure (%)	IOU	*F*-measure (%)	IOU
$\text{Loss}_{\text{Dice}}$	**97.16**	**0.9447**	**94.96**	**0.9019**
Loss_{BCE}	96.87	0.9422	94.65	0.8984

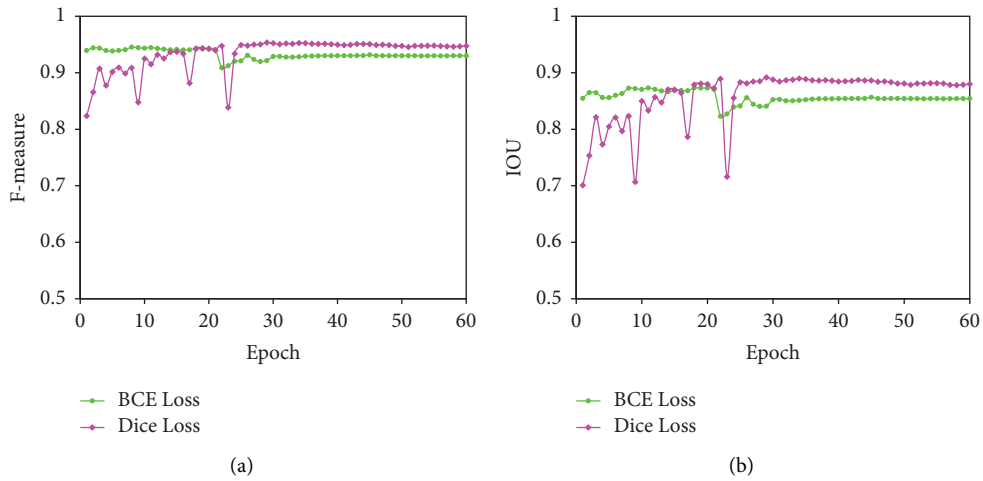

Figure 10: Network performances with Dice loss and BCE loss. (a) *F*-measure. (b) IOU.

Dice loss. Furthermore, as illustrated in Table 2, the dice loss outperforms BCE loss in both *F*-measure and IOU on training data and validation data. BCE loss generally handles positive and negative samples fairly. When positive samples account for a relatively small proportion, they will be overwhelmed by numerous negative samples. Dice loss is a region-related loss, which tends to help the network focus on the foreground area during training. This is undoubtedly beneficial to solving the quantity imbalance between positive and negative samples. In summary, the dice loss is more suitable as a loss function for binary-classification tasks compared to the BCE loss. Therefore, the dice loss is used as the loss function for all trained networks in this paper.

4.4. Loss1 Auxiliary Effect Analysis. HRNet-OCR has two outputs, namely, soft object regions of the coarse segmentation output and final representations of the fine segmentation output. The two outputs correspond to loss1 and loss2, respectively. Particularly, loss1 and loss2 are, respectively, multiplied by scale factors of 0.4 and 1.0, and their sum is used in backpropagation to guide the optimization of network parameter training. Yuan et al. [25] pointed out that the introduction of loss1 in soft object regions can improve network performance to some extent. However, for the binary-classification task with small positive samples, the result may be contrary to the above. Hence, loss1 is multiplied by different scale factors (i.e., 0, 0.3, 0.4, and 0.5) in this paper to explore whether loss1 has a positive auxiliary effect in improving the segmentation accuracy. The results are shown in Table 3.

As illustrated in Table 3, compared to only using loss2 in the backpropagation, using both loss1 and loss2 simultaneously indeed can improve network performance. When the scale factor is set as 0.4, the proposed EJSNet attains the highest performance metrics. The *F*-measures achieved by the EJSNet on training data and validation data are 97.16% and 94.96%, respectively, while the IOUs are 94.47% and 90.19%, respectively. However, when the scale factor is set to be greater (0.5) or less (0.3) than 0.4, the performance of the EJSNet will be degraded, implying that loss1 can have auxiliary impacts on the segmentation accuracy of the network, and the scale factor of loss1 is set as 0.4 in this paper.

4.5. Ablation Experiments. To verify the validity of modules used in the proposed network, Table 4 shows the different combinations of modules used for the ablation experiments, and all models are trained under the same training setting shown in Table 1, and the dice loss is used to train all networks. Moreover, for simplicity, the modified residual structure of the first stage of the HRNet, the convolutional block attention module (CBAM) contained original CAM, and the convolutional block attention module (CBAM) contained modified CAM are, respectively, referred to as Stage1_M, CBAM_O, and CBAM_M.

Figure 11 illustrates the *F*-measure and IOU achieved by all networks with different combinations of modules on training data and validation data. It can be observed in Figures 11(a) and 11(c) that the HRNet-OCR_S outperforms the original HRNet-OCR, indicating the success of the modified residual structure of the first stage of the HRNet. Compared to the HRNet-OCR_S, the networks obtained by incorporating each module individually into the HRNet-OCR_S have better performance, and the effect of incorporating FSM is the best. It illustrates that reducing the

Table 3: Network performances on training data and validation data.

Scale factors of loss1	Training data		Validation data	
	F-measure (%)	IOU	*F*-measure (%)	IOU
0	96.02	0.9234	94.41	0.8941
0.3	96.34	0.9293	94.60	0.8975
0.4	**97.16**	**0.9447**	**94.96**	**0.9019**
0.5	96.25	0.9277	94.52	0.8960

Table 4: Networks with different combinations of modules.

Models	Code names	Module selection				
		Stage1_M	FSM	RFB	CBAM_M	CBAM_O
HRNet-OCR	HO					
HRNet-OCR_S	HO-S	√				
HRNet-OCR_F	HO-F	√	√			
HRNet-OCR_R	HO-R	√		√		
HRNet-OCR_CM	HO-CM	√			√	
HRNet-OCR_CO	HO-CO	√				√
HRNet-OCR_F_R	HO-F-R	√	√	√		
HRNet-OCR_F_CM	HO-F-CM	√	√		√	
HRNet-OCR_R_CM	HO-R-CM	√		√	√	
HRNet-OCR_F_R_CM	HO-F-R-CM	√	√	√	√	

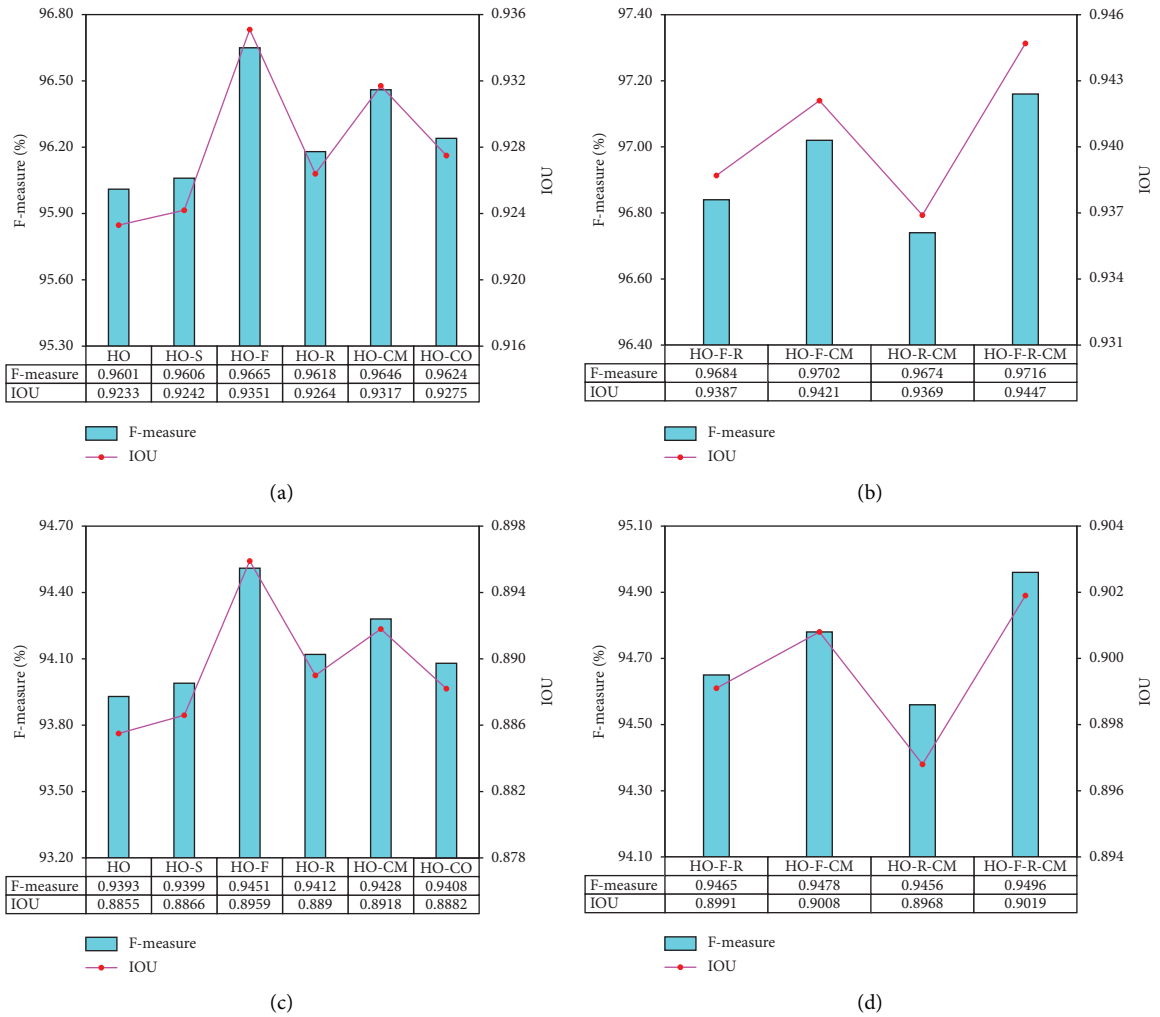

Figure 11: Performances of networks with different combinations of modules. (a) Performances of networks integrating a single module on training data. (b) Performances of networks integrating multiple modules on training data. (c) Performances of networks integrating single module on validation data. (d) Performances of networks integrating multiple modules on validation data.

redundant feature map of the model may be better than increasing the receptive field of the model and improving the attention of the model. In addition, the *F*-measure and IOU achieved by HRNet-OCR_CM on training data and validation data are 96.46%, 0.9282, 94.33%, and 0.8918, respectively, which are higher than those yielded by HRNet-OCR_CO, which validates that an extra dense layer is added to the middle of the MLP of the CAM can decrease the noise effect of invalid channel features. It can be seen from Figure 11 that the networks obtained by incorporating each module in pairs into HRNet-OCR_S perform better than combining them individually. Figures 11(b) and 11(d) illustrate the performance ranking of networks with different combinations of two modules on training data and validation data: HRNet-OCR_F_CM > HRNet-OCR_F_R > HRNet-OCR_R_CM. It can be concluded that on the basis of reducing the redundant feature map of the model, both increasing the receptive field of the model and improving the attention of the model can further improve the performance of the network, while the latter has a better effect.

It can be observed in Figure 11 that the *F*-measure achieved by the HRNet-OCR_F_R_CM on training data and validation data is 97.16% and 94.96%, respectively and the IOU is 0.9447 and 0.9019, respectively, which are all higher than other networks with different combinations of modules. It illustrates that the complementary effects can be provided for improving the recognition accuracy of the model by the combination of reducing the redundant feature map of the model, increasing the receptive field of the model, and improving the attention of the model. Therefore, FSM, RFB, and modified CBAM modules are added to the proposed EJSNet model in this paper for automated pixel-level detection of the expansion joints.

4.6. Training Results. Figure 12 illustrates the performance of the proposed EJSNet and HRNet-OCR on validation data during the training process. For the proposed EJSNet and HRNet-OCR, the values of the loss-value, *F*-measure, and IOU of them all present small fluctuations in the early stage and tend to stabilize in the later stage during the training process. In the early stage, the network learns fewer object features, and the correct convergence direction is difficult to be obtained, resulting in floating network performance. When the network learns sufficient object features, it finds and adaptively converges in the optimal gradient direction, resulting in increased detection accuracy. Particularly compared with HRNet-OCR, the proposed EJSNet ultimately attains smaller loss-function values and higher performance metrics.

It can be observed from Table 5 that although the processing speed (frames per second (FPS)) of EJSNet is slightly slower than HRNet-OCR, the overall performance of EJSNet is better. The *F*-measure achieved by the EJSNet on training data and validation data are 1.15% and 1.03% higher than those yielded by HRNet-OCR, respectively, and the IOU are 2.14% and 1.64% higher than those produced by HRNet-OCR, respectively. The processing speed of EJSNet is slightly slower than HRNet-OCR (11.16FPS vs. 11.92FPS). It can be concluded that the proposed EJSNet achieves better accuracy/FPS trade-offs compared to HRNet-OCR.

5. Evaluation

5.1. Evaluation Using Testing Data. The proposed EJSNet is compared with four state-of-the-art models (SegNet [18], DeepLabv3+ [19], DANet [24], and HRNet-OCR [25]) in this paper. All the four models are trained under the same training setting shown in Table 1, and the Dice loss is used to train all networks. In this paper, DeepLabv3+ and HRNet-OCR employ ResNet-101 and HRNet-W32, respectively, as the backbone networks for feature extraction.

Table 6 illustrates the network performances of SegNet, DeepLabv3+, DANet, HRNet-OCR, and the proposed EJSNet on 500 testing image sets. Compared to other networks, the processing time of the SegNet is the fastest while the proposed EJSNet is the slowest. It can be observed in Table 6 that DeepLabv3+ behaves better than SegNet, DANet, and HRNet-OCR. However, the proposed EJSNet outperforms all other networks noticeably. The *F*-measure and IOU achieved by the proposed EJSNet on 500 testing image sets are 95.14% and 0.9036, respectively. In addition, the *F*-measure and IOU of the EJSNet are 1.11% and 1.63% higher than those attained by HRNet-OCR, respectively, which validates the success of the proposed method of the modification of the HRNet-OCR.

Figure 13 illustrates the typical segmentation results of all the five networks. It can be perceived that all the five networks can perform efficiently on easy images (i.e., without noise objects on pavement surfaces). Nevertheless, as can be seen from the last three comparison result images (counting the images from left to right in landscape orientation) in Figure 13, only the proposed EJSNet can yield superior detections on complex images while the other four networks yield more detection errors, which reveals the proposed EJSNet has a stronger capability in comprehending global context than the other four networks. Compared with other networks, the proposed EJSNet seems to detect expansion joints in a more reasonable way.

5.2. Evaluation Using Public Dataset. To further validate the generalization performance of the proposed EJSNet, a public dataset CRACK500 [34] is used to retrain SegNet, DeepLabv3+, DANet, HRNet-OCR, and the proposed EJSNet. The CRACK500 dataset consists of 3368 pavement crack images, including 1896 training images, 348 validation images, and 1124 testing images. The size of the original crack images is mostly 360×640 ($H \times W$). To better match the expansion joint image size (256×512) used in this paper, both 3368 crack images and ground-truth images that are matched in a pixel-to-pixel manner are center-cropped to a fixed size of 320×640.

The SegNet, DeepLabv3+, DANet, HRNet-OCR, and the proposed EJSNet are trained under the same training setting shown in Table 1, and the dice loss is used to train all networks. Table 7 illustrates the network performances of all

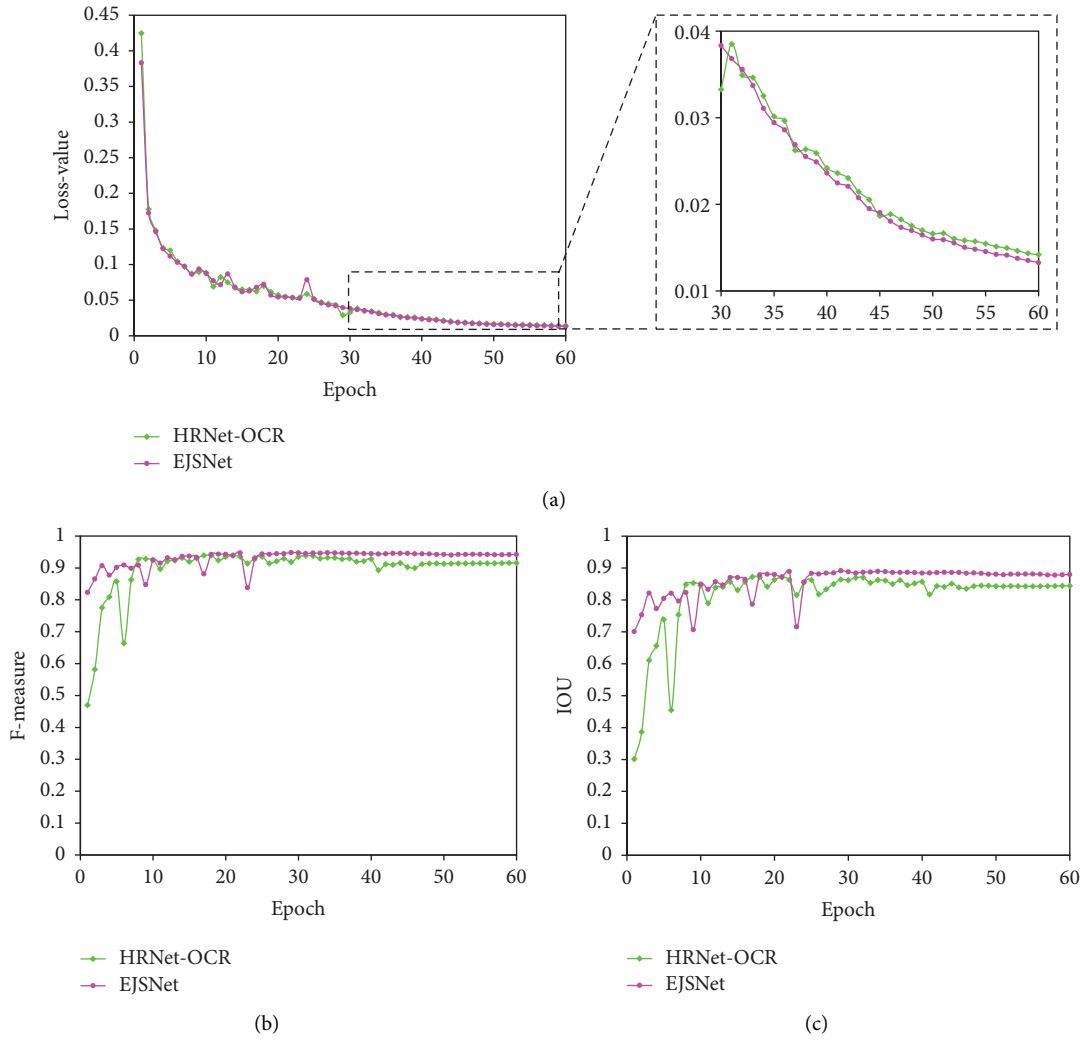

Figure 12: Network performances on validation data during training. (a) Loss-value. (b) *F*-measure. (c) IOU.

Table 5: Network performances on training data and validation data.

Models	Training data		Validation data		FPS
	F-measure (%)	IOU	*F*-measure (%)	IOU	
HRNet-OCR (HRNetV2-W32)	96.01	0.9233	93.93	0.8855	**11.92**
EJSNet (HRNetV2-W32)	**97.16**	**0.9447**	**94.96**	**0.9019**	11.16

Table 6: Network performances on testing data.

Models	Number of parameters	Precision (%)	Recall (%)	*F*-measure (%)	IOU	FPS
SegNet	11, 549,379	93.38	94.60	93.99	0.8866	26.14
DeepLabv3+ (ResNet-101)	36, 812, 071	93.78	94.85	94.31	0.8924	16.63
DANet	11, 229, 763	93.82	94.62	94.22	0.8907	23.56
HRNet-OCR (HRNetV2-W32)	32, 998, 946	95.79	92.34	94.03	0.8873	11.92
EJSNet (HRNetV2-W32)	34, 854, 404	95.27	95.01	**95.14**	**0.9036**	11.16

the five networks on 1124 testing image sets. It can be observed that the *F*-measure and IOU achieved by the proposed EJSNet on 1124 testing image sets are 71.62% and 0.5579, respectively, which outperform all other networks. In addition, Figure 14 illustrates the typical segmentation results of all the five networks on the CRACK500 dataset. It can be seen that compared to other networks, the proposed EJSNet can better capture the details of crack features and

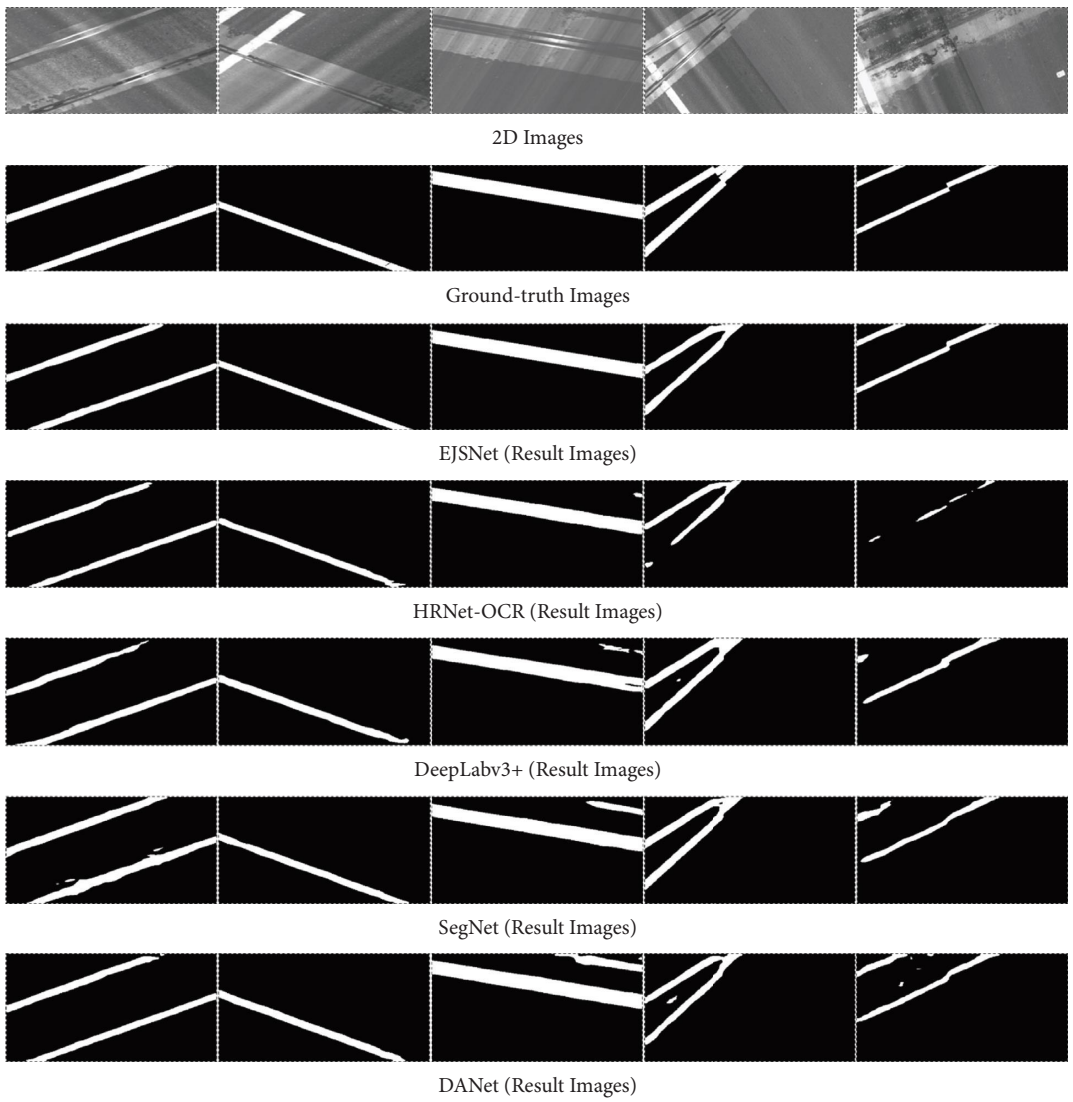

Figure 13: Typical performances of all five networks in detecting the expansion joints.

Table 7: Network performances on testing images from the CRACK500 dataset.

Models	Precision (%)	Recall (%)	*F*-measure (%)	IOU
SegNet	67.55	69.24	68.39	0.5196
DeepLabv3+ (ResNet-101)	69.41	73.59	71.44	0.5556
DANet	66.46	68.59	67.51	0.5095
HRNet-OCR (HRNetV2-W32)	68.12	71.67	69.85	0.5367
EJSNet (HRNetV2-W32)	69.68	73.68	**71.62**	**0.5579**

yield superior detections. These results reveal the proposed EJSNet has a good generalization performance.

6. Discussion

As shown in Table 6, compared with lightweight networks (e.g., SegNet [18], DANet [24], etc), the processing speed (frames per second (FPS)) of the proposed EJSNet is the lowest, indicating that the EJSNet has certain challenges in supporting real-time detection of pavement expansion joints. In addition, Figure 15 shows several representative detection errors yielded by the proposed EJSNet. The false-positive errors are highlighted in the dashed rectangles, whereas the false-negative errors are indicated in the dashed circles. The false-negative errors generally are caused by the local and complex noises (e.g., uneven illumination and markings), while the false-positive errors normally occur at noise patterns similar to the expansion joints. It is demonstrated in Figures 13 and 15 that there are still considerable challenges to perfectly detect the expansion joints

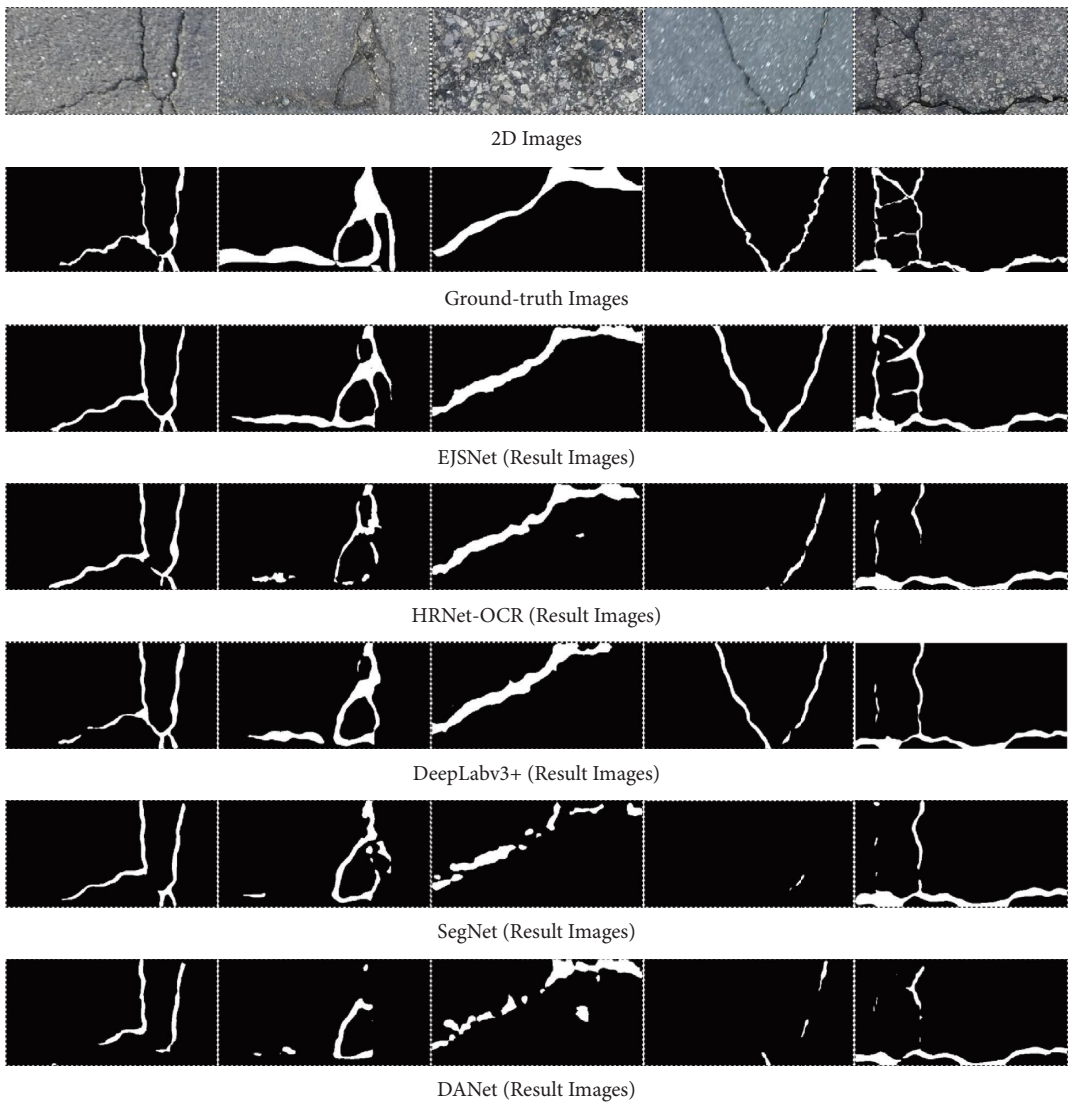

Figure 14: Typical performances of all five networks on the CRACK500 dataset.

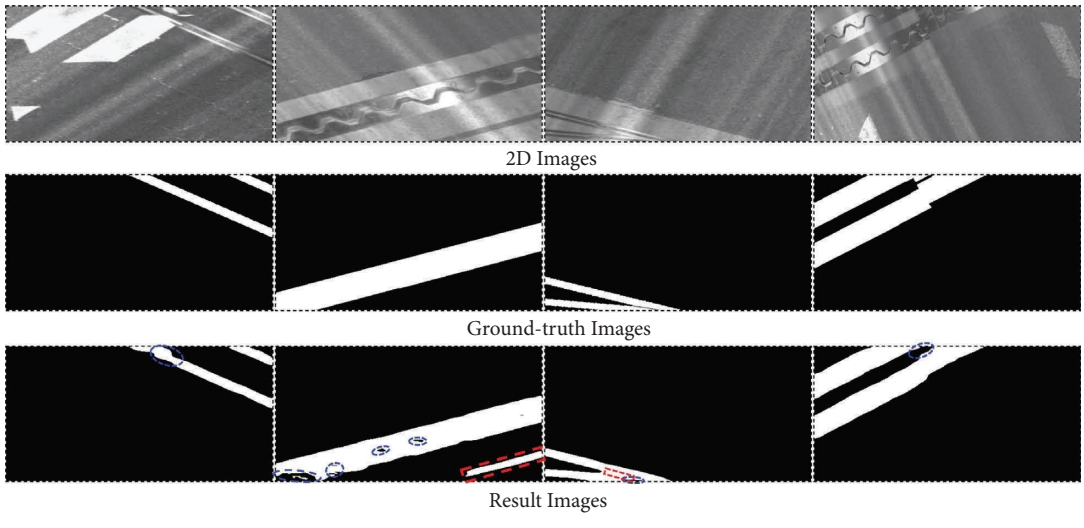

Figure 15: Representative detection errors of the proposed EJSNet on testing data.

with high accuracy due to the complexity of the pavement surfaces. For the above challenges, in the future, the research team will focus on model structure simplification, while also planning to combine self-attention mechanisms to achieve a higher detection accuracy and a faster processing speed.

7. Conclusions

Aiming at pixel-level detection of the expansion joints on asphalt pavements, this paper proposes an improved HRNet-OCR model named as expansion joints segmentation network (EJSNet). In short, the encoder-decoder architecture of the original HRNet-OCR was optimized and adjusted such that more contexts at different scales could be learned and the network segmentation capability can be improved. Specifically, the residual structure of the first stage of the high-resolution network (HRNet) was modified to avoid the network degradation; the feature selection module (FSM) and receptive field block (RFB) were introduced to enhance the feature extraction and summarize latent representations; and the modified convolutional block attention module (CBAM) was also introduced to help the network retrieve more object details.

The experimental results show that compared to four state-of-the-art models for semantic segmentation (i.e., SegNet, DeepLabv3+, dual attention network (DANet), and HRNet-OCR), the proposed EJSNet can yield superior performances and higher detection accuracy on both private and public datasets. Specifically, the F-measure and IOU achieved by the proposed EJSNet on 500 private testing image sets are 95.14% and 0.9036, respectively, and those achieved by the proposed EJSNet on 1124 public testing image sets (CRACK500) are 71.62% and 0.5579, respectively, both of which outperform four other networks. These results indicate that the proposed EJSNet has a better capability of feature refinement, superior performances in semantic segmentation, and good generalization performance.

In addition, although the processing time of the proposed EJSNet is relatively slow and cannot detect deformation of expansion joints, it can segment the features of expansion joints with pixel-level accuracy, providing a data basis for subsequent detection of expansion joint deformation. In the future, the encoder architecture of the proposed EJSNet will be further optimized to achieve better accuracy/processing speed trade-offs, and the collected 3D image data will be used to train the network to develop a method that can segmentate expansion joint features and detect expansion joint deformation.

Data Availability

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Network conception and design was performed by Anzheng He, Zishuo Dong, and Hang Zhang. Yang Liu, Kelvin C.P. Wang, and Zhihao Lin prepared the data. Anzheng He and Allen A. Zhang were responsible for experiment design and analysis of results. Anzheng He and Shi Qiu prepared the manuscript.

Acknowledgments

The study presented in this article was partially supported by the National Natural Science Foundation of China (grant no. 51208419) and the Shudao Investment Group Science and Technology Program.

References

[1] Y. Otsuki, M. Kurata, K. A. Skalomenos, Y. Ikeda, and M. Akazawa, "Fragility function development and seismic loss assessment of expansion joints," *Earthquake Engineering and Structural Dynamics*, vol. 48, no. 9, pp. 1007–1029, 2019.

[2] X. H. He, T. Wu, Y. F. Zou, Y. F. Chen, H. Guo, and Z. Yu, "Recent developments of high-speed railway bridges in China," *Structure and Infrastructure Engineering*, vol. 13, no. 12, pp. 1584–1595, 2017.

[3] H. D. Li, "The cause analysis on the highway bridge expansion joints and the maintenance of construction management," in *Proceedings of the 2018 3rd International Conference on Smart City and Systems Engineering (ICSCSE)*, pp. 267–269, Xiamen, China, December 2018.

[4] I. B. Kim, J. S. Cho, G. S. Zi, B. Cho, S. Lee, and H. Kim, "Detection and identification of expansion joint gap of road bridges by machine learning using line-scan camera images," *Applied System Innovation*, vol. 4, no. 4, p. 94, 2021.

[5] Y. Q. Ni, X. G. Hua, K. Y. Wong, and J. M. Ko, "Assessment of bridge expansion joints using long-term displacement and temperature measurement," *Journal of Performance of Constructed Facilities*, vol. 21, no. 2, pp. 143–151, 2007.

[6] C. Q. Miao, Y. Deng, Y. L. Ding, and A. Li, "Damage alarming for bridge expansion joints using novelty detection technique based on long-term monitoring data," *Journal of Central South University*, vol. 20, no. 1, pp. 226–235, 2013.

[7] Y. Q. Ni, Y. W. Wang, and C. Zhang, "A Bayesian approach for condition assessment and damage alarm of bridge expansion joints using long-term structural health monitoring data," *Engineering Structures*, vol. 212, Article ID 110520, 2020.

[8] G. D. Zhou, T. H. Yi, B. Chen, and H. Zhang, "Analysis of three-dimensional thermal gradients for arch bridge girders using long-term monitoring data," *Smart Structures and Systems*, vol. 15, no. 2, pp. 469–488, 2015.

[9] Y. L. Ding, Y. Deng, and A. Q. Li, "Study on correlations of modal frequencies and environmental factors for a suspension bridge based on improved neural networks," *Science China Technological Sciences*, vol. 53, no. 9, pp. 2501–2509, 2010.

[10] Z. H. Chen, X. W. Liu, G. D. Zhou, H. Liu, and Y. X. Fu, "Damage detection for expansion joints of a combined highway and railway bridge based on long-term monitoring data," *Journal of Performance of Constructed Facilities*, vol. 35, no. 4, 2021.

[11] A. Zhang, K. C. P. Wang, B. X. Li et al., "Automated pixel-level pavement crack detection on 3D asphalt surfaces using

a deep-learning network," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 10, pp. 805–819, 2017.

[12] A. Zhang, K. C. P. Wang, Y. Fei et al., "Deep-Learning based fully automated pavement crack detection on 3D asphalt surfaces with an improved CrackNet," *Journal of Computing in Civil Engineering*, vol. 32, no. 5, 2018.

[13] A. Zhang, K. C. P. Wang, Y. Fei et al., "Automated pixel-level pavement crack detection on 3d asphalt surfaces with a recurrent neural network," *Computer-Aided Civil and Infrastructure Engineering*, vol. 34, no. 3, pp. 213–229, 2018.

[14] H. S. Zhao, J. P. Shi, X. J. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239, Honolulu, HI, USA, July 2017.

[15] O. Ronneberger, P. Fischer, and B. Thomas, "U-Net: convolutional networks for biomedical image segmentation," 2015, https://arxiv.org/abs/1505.04597.

[16] G. Wen, L. Yao, Y. Hao et al., "Bilirubin ameliorates murine atherosclerosis through inhibiting cholesterol synthesis and reshaping the immune system," *Journal of Translational Medicine*, vol. 20, pp. 1–14, 2022.

[17] A. Zhang, K. C. P. Wang, Y. Liu et al., "Intelligent pixel-level detection of multiple distresses and surface design features on asphalt pavements," *Computer-Aided Civil and Infrastructure Engineering*, vol. 37, no. 13, pp. 1654–1673, 2022.

[18] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[19] L. C. Chen, Y. K. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation," 2018, https://arxiv.org/abs/1802.02611.

[20] J. D. Wang, K. Sun, T. H. Cheng et al., "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2021.

[21] X. B. Qin, Z. C. Zhang, and C. Y. Huang, "U2-Net: going deeper with nested U-structure for salient object detection," *Pattern Recognition*, vol. 106, Article ID 107404, 2020.

[22] B. Y. Chen, M. Xia, and J. Q. Huang, "MFANet: a multi-level feature aggregation network for semantic segmentation of land cover," *Remote Sensing*, vol. 13, no. 4, p. 731, 2021.

[23] Z. Y. Xu, W. C. Zhang, T. X. Zhang, and J. Li, "HRCNet: high-resolution context extraction network for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 13, no. 1, p. 71, 2020.

[24] J. Fu, J. Liu, and H. J. Tian, "Dual attention network for scene segmentation," in *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3141–3149, Long Beach, CA, USA, June 2019.

[25] H. Y. Yuan, X. K. Chen, and X. L. Chen, "Segmentation Transformer: Object-Contextual Representations for Semantic Segmentation," 2019, https://arxiv.org/abs/1909.11065.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[27] S. Woo, J. Park, J. Y. Lee, and I S. Kweon, "CBAM: Convolutional Block Attention Module," 2018, https://arxiv.org/abs/1807.06521.

[28] S. H. Huang, Z. C. Lu, R. Cheng, and C. He, "FaPN: Feature-Aligned Pyramid Network for Dense Image Prediction," 2021, https://arxiv.org/abs/2108.07058.

[29] S. T. Liu, D. Huang, and Y. H. Wang, "Receptive Field Block Net for Accurate and Fast Object Detection," 2017, https://arxiv.org/abs/1711.07767.

[30] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," 2017, https://arxiv.org/abs/1706.05587.

[31] L. Chen, H. W. Zhang, J. Xiao et al., "SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6298–6306, Honolulu, HI, USA, April 2017.

[32] F. Wang, M. Jiang, C. Qian et al., "Residual attention network for image classification," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6450–6458, Honolulu, HI, USA, April 2017.

[33] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the 2016 Fourth International Conference on 3D Vision 3DV*, pp. 565–571, Stanford, CA, USA, 2016.

[34] F. Yang, L. Zhang, S. Yu, D. Prokhorov, X. Mei, and H. Ling, "Feature pyramid and hierarchical boosting network for pavement crack detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1525–1535, 2020.