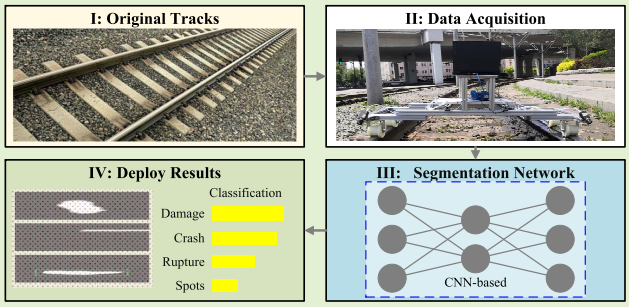
Segmentation of Track Surface Defects Based on Machine Vision and Neural Networks

Hongfei Yang, Yanzhang Wang, Jiyong Hu, Jiatang He, Zongwei Yao, and Qiushi Bi

***Abstract*—Local failure of rail track commonly grow from surface defects. Hence, timely detection of surface defects helps to identify potential hazards on the track and reduce the occurrence of railroad accidents. Since track surface defects are scattered and diverse, and different service life leads to different types of defects, it is crucial to detect surface defects in real time, with efficiency, reliability and robustness. To this end, a pixel-level defects segmentation method is proposed. In this paper, features are tessellated together at the Channel level to form denser features, allowing additional information on surface defects textures to be propagated among high-resolution layers. Dropout is performed on the weak correlations learned during the convolution, so that the convolution blocks can share a uniform weight matrix, reducing computation redundancy and model complexity. Firstly, the track datasets of different ages are categorized into four sets, then, the samples are normalized to grey scale by pre-processing, and fed into the proposed network for training. An evaluation of the proposed model on defective samples was performed to demonstrate the performance of the method, with an Accuracy of 97.47%, a Loss of 0.0061 and an Average Frame Rate of 0.033s. The results of different networks tested on the same dataset show that the proposed model exhibits strong stability, adaptability and robustness. In addition, the proposed model is assessed on two different datasets with distinct challenges, with Mean Intersection over Union yielding 2.13% and 3.77% boosts respectively.**



***Index Terms*—Complex and diverse track surface, defects segmentation, neural networks, track surface defects.**

I. Introduction

TRACK failures trigger severe train accidents and endanger the safety of passengers and property, so prompt intervention at an early stage of failure is essential. Accurate and reliable real-time detection of track surface defects serving as the base for advance involvement to inhibit the progression of defects is an effective method of avoiding rail traffic accidents. The ability to accurately segment defects from a sophisticated background is crucial to the identification and classification of defects. The current widely practiced method, i. e. the manual inspection, is highly subjective and the results are related to the work experience, professional and technical expertise, physical and mental state of the inspector as well as working environment. The knowledge of track inspection takes a long time to build up, so experienced inspectors are generally older and prone to fatigue when operating at high intensity while younger inspectors are often inexperienced, making it difficult to guarantee the credibility of the results. In adverse weather such as rain, snow, fog, frost, strong wind and heavy cloud, even experienced inspectors find it difficult to perform quality work and have to operate during sunny daylight conditions. In addition, the passing trains pose a high risk to inspectors, especially in harsh environments. Therefore, it is essential to develop accurate, efficient, fully automated solutions for the segmentation of track surface defects to assist even substitute manual work.

Benefiting from the rapid development of computer vision, the above challenges are gradually being addressed in the industrial production. An image pyramid-based network model for surface defect detection was proposed in [1]. A deep learning smoothed surface defect detection method based on wrapped phase map was proposed in [2] to classify track defects. Hajizadeh *et al.* [3] improved the accuracy of track squat detection in images with semi-supervised techniques.

Manuscript received November 15, 2021; accepted December 2, 2021. Date of publication December 8, 2021; date of current version January 12, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 51875232, Grant 42074216, and Grant 51875233; and in part by the Special Project on Frontier Guidance for Provincial-Varsity Cooperation under Grant SXGJQY2017-11. The associate editor coordinating the review of this article and approving it for publication was Dr. Andre E. Lazzaretti. *(Corresponding authors: Zongwei Yao; Qiushi Bi.)*

Hongfei Yang and Yanzhang Wang are with the College of Instrumentation & Electrical Engineering, Jilin University, Changchun, Jilin 130061, China (e-mail: yanghf20@mails.jlu.edu.cn; yanzhang@jlu.edu.cn).

Jiyong Hu and Jiatang He are with FAW-Volkswagen Automobile Company Ltd., Changchun, Jilin 130011, China (e-mail: jiyong.hu@faw-vw.com; Jiatang.he@faw-vw.com).

Zongwei Yao and Qiushi Bi are with the School of Mechanical and Aerospace Engineering, Jilin University, Changchun, Jilin 130022, China (e-mail: yzw@jlu.edu.cn; bqs@jlu.edu.cn).

Digital Object Identifier 10.1109/JSEN.2021.3133280

Rizvi *et al.* [4] performed track crack detection based on image processing. Li *et al.* [5] employed a local normalization method to boost the contrast of the track images, followed by defect localization based on the projected profiles. Li *et al.* [6] exploited automatic thresholding methods for track defect segmentation and localization. In order to enhance image contrast, a new anti-Perona-Malik diffusion model was proposed in [7], using the inverse of the gradient as a feature to tune the diffusion coefficients and introducing a unique nearest-neighbour difference scheme in the discretization implementation to target proper defect boundaries. Gan *et al.* [8] proposed a hierarchical detection framework consisting of a coarse extractor and a fine extractor, which leverages features to search for the defect background. Fekri-Ershad *et al.* [9] proposed a noise-resistant and multi-resolution version of local binary patterns to extract color and texture features jointly, and presented a robust algorithm for detecting abnormalities in surfaces. Furthermore, ternary patterns were coded into two binary patterns, and then classified into two uniform/non-uniform groups for bark texture classification with high accuracy [10].

The deep learning-based defect methods can be categorized into image-level classification, area-level detection and pixel-level segmentation, according to the different tasks.

The basic ideology of the region-level methods is to cluster pixels with similar properties around the seed pixel to form a region, where the similarity criterion such as grey level, color, texture, gradient, and spreading criteria can be formulated according to various principles. He *et al.* [11] introduced a semi-supervised convolutional neural network (CNN) for feature extraction and fed the features into a classifier for defect classification. Natarajan *et al.* [12] suggested an SVM classifier overcoming the overfitting issues readily occured on small data sets. Wang *et al.* [13] proposed a segmentation method with CNN as the feature extractor and random forest as the resulting classifier. Masci *et al.* [14] proposed a steel defect classification method based on maximum pooling CNN. Soukup *et al.* [15] suggested a classical CNN trained in a purely supervised manner and explored the influence of regularization methods. This type of method is computationally straightforward, yet requires artificially located seed pixel, making it sensitive to noise and may lead to gaps in the region. Additionally, this type of approach for serial operation is not that efficient for segmentation when the target is large.

The essence of the image-level methods is to remove certain edges and divide the graph into several subgraphs to achieve segmentation. Masciet *et al.* [16] proposed a multi-scale pyramidal ensemble network for the classification of steel defects that does not require all images to be of equal size. He *et al.* [17] proposed a multi-level feature fusion network that combines multi-level hierarchical features extracted from a backbone CNN into a single resolution for steel plate defect detection. Zhang *et al.* [18] employed a hybrid Gaussian model combined with filtering techniques for track defect segmentation. This type of method is based on grey-scale graphs, which are stripped of detail and requires manual labelling of at least one foreground pixel and one background pixel, making it difficult for autonomous applications. The result obtained with a hard border does not take into account the transparency of the edges between 0 and 1.

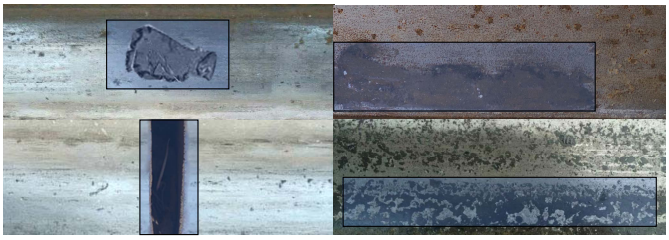


Fig. 1. Typical track surface defects.

Track surface defect segmentation is a concern for safety, the two methods mentioned above are not employed in this paper because of their inherent flaws Currently, one of the effective surface defect detection methods is the fully convolutional network-based approach [19]. A fully convolutional model for cell detection was published in [20]. Yang *et al.* [21] presented a fully convolutional texture surface defect detection method based on multiscale feature clustering. Dong *et al.* [22] introduced a pyramid feature fusion and global attention network for segmentation of rolled steel surface defects. In many cases the pixel-based approaches load higher runtime than region-/image-based approaches, so the network structure requires adaptation to improve segmentation efficiency.

There are also studies on defect detection of critical components on railway systems. Giben *et al.* [23] performed track material classification and semantic segmentation based on deep CNN. Gibert *et al.* [24] improved the detection accuracy of defects in railway ties and fasteners by a multi-task learning framework incorporating multiple detectors. Liu *et al.* [25] posed an automatic fault detection system consisting of an CNN and a Markov random field to detect loose faults in electric wires. Zhong *et al.* [26] presented a network-based method for detecting defects in split pins of high-speed railway suspension lines. Chen *et al.* [27] detected fastener defects in the energy transmission system of high-speed rail based on neural networks.

Considering the advantages in reducing the subjective of inspectors and the widely applications, a multi-level, end-to-end fast track surface defect segmentation method is proposed in this paper based on deep learning and computer vision. Firstly, the images of track surfaces with different service life and weather conditions as shown in Fig.1, are acquired. Secondly, the track samples containing defects are classified into multiple categories and then grey-scale normalized. Finally, the proposed CNN is applied for accurate segmentation of defects with high Recall. The main contributions of this paper are as follows.

1) A deep learning-based method for track surface defect segmentation is introduced, which yields outstanding performance on different defect datasets, proving that the method exhibits strong generalization features and theoretical value.
2) A pixel-level surface defect segmentation method is proposed rather than an image-level or region-level method. Features are tessellated together at the channel level to form denser features, allowing additional information on

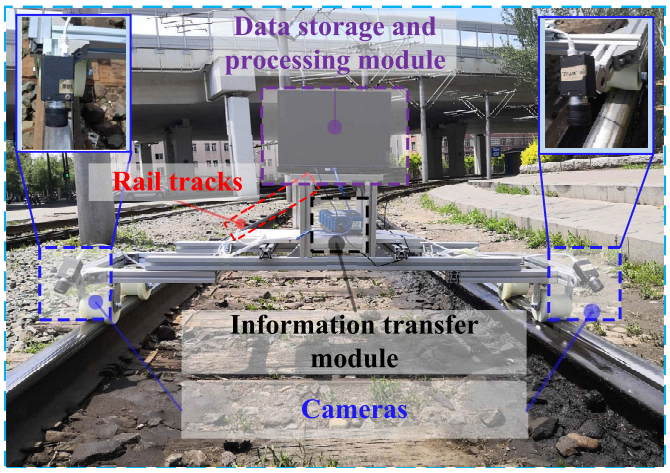


Fig. 2. Typical track surface defects.

surface defects textures to be propagated among high-resolution layers. The model drops the fully connected layer to ensure that the segmentation results are obtained based on contextual features without information missing, improving the accuracy of the segmentation.

3) Dropout is performed on the weak correlations learned during the convolution, so that the convolution modules contribute to a shared set of weights, reducing computation redundancy and model complexity.

The following is the organization of the sections of this paper in detail: The image acquisition system and dataset are introduced in Section II, while the methodology theories in Section III. The comparison and analysis of experimental results are discussed in Section IV with the conclusions in Section V.

II. Image Acquisition System and Dataset

A. Track Samples

The track image dataset is taken from the video with the cameras facing the tracks from the top down and the acquisition device is shown in Fig.2. The lengths of the various types of tracks recorded in the video total approximately 125 km and include tracks that have been in service for about 2, 5 and 10 years. The different service times of the tracks result in different types of noise in the corresponding images. The images are grouped into four different datasets, each containing 800 samples and a total of 3200, according to the type of noise contained in the images. As shown in TABLE I, the tracks in C1 suffer from material chipping and unevenly reflected light on the surface, in C2 show deterioration and material damage due to long service, in C3 present partial or total fracture on the surface, and in C4 feature numerous small damages and noise on the surface and some highlights (taken on rainy days). Some contain one or more of these features in a single image file.

B. Pre-Processing of the Images

As the captured images are taken from tracks with different service times, various types of defects are mixed together.

TABLE I
Rail Track Surface Dataset

Set	Scene	Characteristic
C1	Crash	The defect boundary is well defined and regular in shape and the defect area is small with regard to the track surface.
C2	Damage	The boundary is poorly continuous and irregularly shaped, distributed over a large and relatively uneven area of the track surface.
C3	Rupture	The defect boundary is well defined and regular in shape and the defect area is small with regard to the track surface.
C4	Spot	The defect boundary is long and continuous, irregular in shape and unevenly distributed.

Hence, classification by appropriate pre-processing is preferable for CNN to leverage the strengths of feature extraction. In this paper, each track image is resized to 256×256 followed by a greyscale process. For converting the original RGB image into an easy-to-recognize grayscale image, the grayscale weighting method is used. The weighted gray value $I(i, j)$ at pixel (i, j) can be calculated according to Equation (1) based on the component of red $R(i, j)$, green $G(i, j)$ and blue $B(i, j)$, as follows.

$$I(i, j) = 0.30R(i, j) + 0.59G(i, j) + 0.11B(i, j) \quad (1)$$

Rectangular boxes are employed to mark the ROIs considering the significant colour difference between the track and the subgrade. The mean mu_f and variance σ_f of the greyscale of all pixels within the ROI are calculated, and then the segmentation threshold I_T is derived according to Equation (2), from which the greyscale image is segmented to produce the binary image.

$$I_T = \mu_f + 3\sigma_f \quad (2)$$

With the limited amount of available data, the data augmentation is required in order to intensify the training of the network without collecting additional data. Unsupervised feature learning was performed in [28] with data augmentation, which fully demonstrates the advantages of the technique. The similar approach to data expansion is adopted in this paper, with operations such as horizontal flipping and lateral range shifting being applied to each extracted image.

III. Proposed Method

A neural network for automatic segmentation of neuronal structures was established in [29], which implemented an automatic segmentation technique under electron microscopy, albeit with two downsides. Firstly, the structure of the network is complex with lots of overlapping patches causing significant redundancy, resulting in insufficient real-time performance. Secondly, the accuracy of localisation is compromised by an extensive amount of Maxpooling layers required for large targets. A CNN network was exploited in [19] to produce hierarchical features and to replace the fully connected layers of networks from AlexNet [30], VGG-16 [31], GoogLeNet [32] and ResNet [33] with a convolutional layer that produces dense pixel-level labels from small-step convolutional upsampling

(also known as deconvolution) of the output space mapping, thus transforming the networks into fully convolutional. The architecture of the full convolutional network [19], [20] have been modified and extended in this paper so that it can work with less training images and produce as accurate a segmentation as possible.

A. Overall Network Architecture

The architecture of the network proposed in this paper is shown in Fig.3 with 58 layers consisting of contraction and expansion channels. The green module in the figure represents a combination of a 3×3 convolutional layer (Stride=1, Padding=1) and a Relu activation layer, the red module is a 2×2 max-pooling layer (Stride=2, Padding=0), the bright blue module is a Dropout layer which is set at Dropout=0.5 to prevent overfitting. The module with the thickest orange and yellow combination is a 2×2 Upconv layer (Stride=2) and Relu layer, the bright orange module for the Depth-Concatenation layer, the 3rd countdown layer is a 1×1 convolution layer (Stride=1, Padding=0), the 2nd countdown layer is a Softmax layer and the last layer is a Pixel Classification Layer.

The former half, from the 2nd to the 27th layer, of the network serves for feature extraction and the latter half for upsampling (the Upconv and Relu layers). The network fuses features together in the Channel dimension with a stitching feature fusion (Depth-Concatenation layer) to form denser features, allowing additional information about track defect texture to propagate within the high-resolution layers. The structure of the fully connected layer is fixed, assigning weights to each connection, and cannot learn to filter or modify the connection relationships. The convolution process, on the other hand, trains the structure of the connections to find relationships between targets and pixels, reinforcing the beneficial relationships and weakening the ineffective ones (dropout is performed directly in this paper). As such, the convolutional modules can contribute to a shared set of weights, reducing redundancy and complexity of the model. Therefore, the proposed model drops the fully connected layer (FC) to ensure that the segmentation results are yielded based on the contextual features without information missing, and to boost the accuracy of the segmentation results

B. Convolution Modules

Each layer of data in a convolutional network is a three-dimensional array of dimensions h×w×d, where h and w are spatial dimensions and d is the feature or channel dimension. The first layer is the image, with a size of h×w pixels, and d channels. The positions in the subsequent layers correspond to those in the image to which their paths are connected, these are called the receptive domain.

Convolutional networks are constructed on the basis of translational invariance. Their elementary components (convolution, pooling and activation functions) operate on local input regions and rely only on relative spatial coordinates. Let x_{ij} on a layer be the data vector at coordinates (i, j) and y_{ij} on the following layer, the relationship between the two is:

$$y_{ij} = f_{ks}(\{x_{si+\delta i, sj+\delta j}\}0 \le \delta i, \delta j \le k) \tag{3}$$

where k is the convolution kernel size, s is the step size or down-sampling factor, f_{ks} determines the layer type: a matrix multiplication for convolution or average pooling, a spatial max for max pooling, or an elementwise nonlinearity for an activation function, and so on for other types of layers. Once the convolution kernel size and step size follow the conversion rules, the function is expressed in the following form:

$$f_{ks} \circ g_{k's'} = (f \circ g)_{k'+(k-1)s', ss'} \tag{4}$$

Numerous feature channels are introduced in the up-sampling part, allowing the network to propagate background information to higher resolution layers. The fully connected layers are dropped only saving effective component of each convolution to reduce the computational resources required. Meanwhile, the limitation on image size is breached. The image boundary region pixels are predicted by mirroring the input image.

C. Defect Boundary Extraction Refinement

Softmax maps the outcome of the final feature to the range of [0, 1]. The output of Softmax is calculated as follows.

$$p_\beta(x) = \exp\left[\alpha_\beta(x)\right] \Big/ \sum_{\beta'=1}^{\beta} \exp\left[\alpha_{\beta'}(x)\right] \tag{5}$$

where $\alpha_\beta(x)$ represents the activation value of the pixel at position x at layer β in the feature map, with $x \in \Omega$, $\Omega \subset Z^2$, β means the total number of categories of pixels, and $p_\beta(x)$ is the maximum likelihood function.

By pre-computing the weights of each pixel in the loss function, it is possible to compensate for the different frequencies of each type of pixel in the training data making the network target on learning the edges of track defects. Morphological operation is employed for the boundary segmentation and the feature map can be calculated as follows.

$$\omega(x) = \omega_c(x) + \omega_0 \exp\left\{-\frac{[d_1(x) + d_2(x)]^2}{2\sigma^2}\right\} \tag{6}$$

where ω_c is the weight map used to balance the category frequencies, d_1 and d_2 are the distances from the pixel to the nearest and second nearest defect boundary respectively; ω_0 and σ are constants, 5 and 10 respectively, taken in this paper.

D. Loss Function

The aim of this paper is to segment the defects from the track images. Binary cross entropy is chosen as the segmentation loss function in order to train the model.

$$L = -\frac{1}{M}\sum_{r=1}^{M}\left[y_r \ln y^r + (1 - y_r)\ln\left(1 - y^r\right)\right] \tag{7}$$

where M denotes the number of pixels in each image, y_r is the Groundtruth of the pixel, and y^r is the prediction of the pixel.

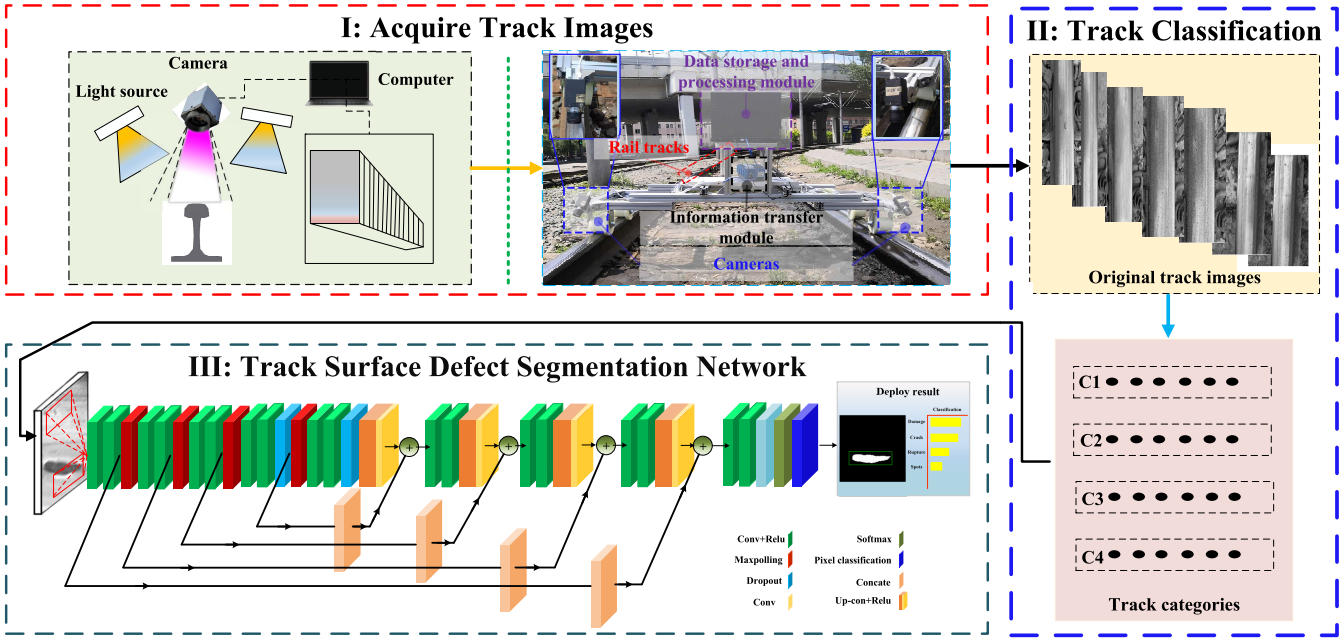


Fig. 3. Overall architecture of the proposed track defect segmentation method.

IV. Experiments and Analysis

SGDM (Stochastic Gradient Descent with Momentum) optimizer was deployed to train the network. Although the high Learning Rate can improve training efficiency, it may cause the gradient of the network explosion, resulting in training failure. Gradient clipping was applied to keep the gradient within a reasonable range by specifying Gradient Threshold=0.05. To speed up the learning process, especially for coping with high curvature or relatively large gradients, set Momentum at 0.9. Set Minbatchsize=10 and Initial Learning Rata=0.01 for optimal utilization of GPU.

The experiments in this paper were all performed on a computer with an Intel(R) Xeon Silver 4210R@2.40GHz CPU, Tesla V100s GPU and 256GB RAM.

Different iterations and learning rates were conducted to find the optimal training parameters in order to obtain stable and robust results. In this paper, a Learning Rate of 10^{-4} and a number of iterations of 1600 were finally chosen, and the average training accuracy and loss are shown in Fig.4. The final Accuracy=97.47% and Loss=0.0061.

Experiments and discussions are conducted for demonstrating the robustness and adaptability of the model. Recall, Sensitivity, Precision, Average Precision (AP) and Mean Intersection over Union (mIoU) are introduced as metrics to evaluate the performance of models.

$$Recall = Sensitivity = \frac{TP}{TP + FP} \tag{8}$$

$$Precision = \frac{TP}{TP + FN} \tag{9}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{10}$$

where TP is True Positives, TN is True Negatives, FP is False Positives and FN is False Negatives. The area enclosed by the Precision-Recall curve is the AP, with higher AP indicating superior performance of the model. The mIoU is the average of the IoU of all classes.

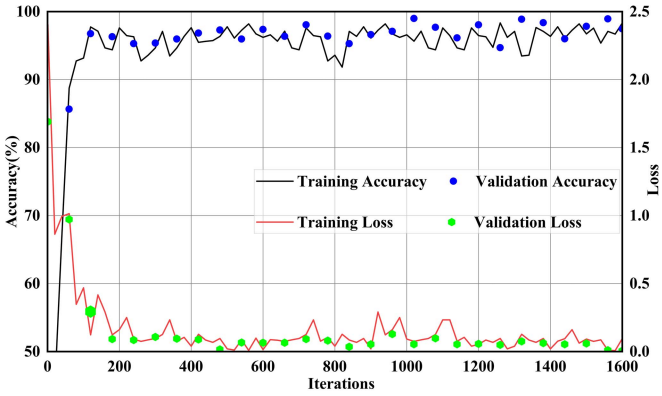


Fig. 4. The training results of the proposed track surface defect segmentation network.

A. Sample-Based Estimation

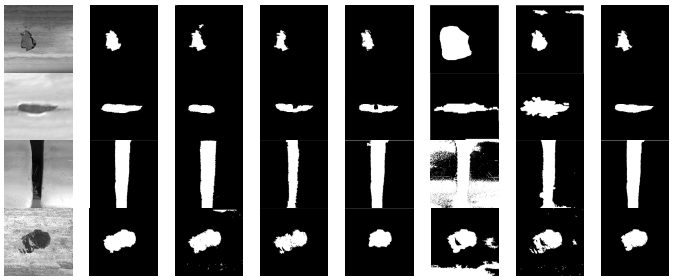
The results of the segmentation experiments of the proposed model with PSPNet [34], FCN [19], SegNet [35], K-means [36] and GraphCut [37] are shown in Fig.5.

PSPNet shows fluctuations in the edges of the defect in C2, under-segmentation in C3 (segmented defects smaller than Ground truth, where Ground truth is the area manually marked by two track defect detection experts), over-segmentation in C1, and numerous over-segmentations in C4 due to the presence of high intensity light. The reason for these issues is that PSPNet parallelly considers target features under multiple perception while sensitive to targets that are either large or undersized.

The FCN generally manages to segment the samples with enough correctness in C1-C4. However, under-segmentation

TABLE II
PERFORMANCE OF DIFFERENT MODELS IN DIFFERENT SAMPLES

Dataset	Method	mIoU (%)	Time (s)	AP	Recall @ Precision=0.9	Precision @ Recall=0.8	Approximate inflection points (Precision, Recall)
C1	FCN	**88.21**	0.0581	0.556	N/A	0.58	(0.55, 0.86)
	SegNet	79.34	0.0434	**0.896**	**0.85**	**0.92**	**(0.88, 0.88)**
	PSPNet	86.31	**0.0157**	0.694	0.20	0.51	(0.73, 0.70)
	K-means	37.42	0.7805	0.291	N/A	0.01	(N/A, N/A)
	GraphCut	55.31	1.427	0.530	N/A	0.25	(0.57, 0.58)
	Proposed	**89.67**	**0.0268**	**0.935**	**0.95**	**0.97**	**(0.93, 0.93)**
C2	FCN	86.41	0.0547	0.529	N/A	0.54	(0.52, 0.84)
	SegNet	**87.62**	0.0395	**0.856**	**0.61**	**0.88**	**(0.82, 0.89)**
	PSPNet	77.97	**0.0311**	0.799	0.45	0.81	(0.78, 0.82)
	K-means	41.39	0.5466	0.302	N/A	0.07	(N/A, N/A)
	GraphCut	50.09	1.5501	0.649	N/A	0.61	(0.66, 0.76)
	Proposed	**86.69**	**0.0297**	**0.918**	**0.88**	**0.95**	**(0.91, 0.86)**
C3	FCN	**82.07**	0.0483	0.485	N/A	0.48	(0.44, 0.85)
	SegNet	75.24	0.0576	**0.728**	N/A	**0.77**	**(0.77, 0.79)**
	PSPNet	71.56	**0.0257**	0.703	**0.23**	0.60	(0.65, 0.75)
	K-means	39.05	1.4358	0.276	N/A	N/A	(N/A, N/A)
	GraphCut	65.84	1.3845	0.605	N/A	0.42	(0.69, 0.60)
	Proposed	**82.74**	**0.0364**	**0.882**	**0.79**	**0.89**	**(0.93, 0.77)**
C4	FCN	**74.58**	0.0559	0.534	N/A	0.56	(0.53, 0.91)
	SegNet	69.69	0.0588	**0.876**	**0.71**	**0.88**	**(0.82, 0.90)**
	PSPNet	72.36	**0.0307**	0.737	0.33	0.58	(0.82, 0.69)
	K-means	43.27	0.8504	0.272	N/A	N/A	(N/A, N/A)
	GraphCut	59.67	1.4571	0.556	N/A	0.31	(0.56, 0.64)
	Proposed	**76.21**	**0.0398**	**0.907**	**0.85**	**0.93**	**(0.87, 0.90)**



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 5. Comparison of the segmentation effects of different models based on the data collected. Note that (a) are Original images, (b) are Ground truths, (c)-(g) are the results of PSPNet, FCN, SegNet, K-means and Grapcut respectively, and (h) are the resluts of the proposed model. From top to bottom are the typical results obtained by each method when processing C1 to C4.

exists in processing of C1 and mis-segmentation in C2. For samples of C3, segmented defects are smaller than Ground truths. The potential cause is that the fusion of deep and shallow information in FCN is processed by adding up the corresponding pixels, which is insensitive to the details in the image and uncoherent in space.

SegNet performs well in C1, mis-segments the upper edge of the defect in C2, and correctly segments the defects in C3 and C4 with inadequate edge accuracy. The possible contributing factor is that the small number of encoder parameters leads to a low confidence level in the segmented boundaries, increasing the uncertainty in segmenting the defects with overlapping and indistinguishable edges.

K-means performs comparatively well in C2, with over-segmentation in C1, C3 and C4. The likely cause is that the defects and background pixels are so similar that the K-means fails to cluster properly.

GraphCut shows over-segmentation in all of C1-C4, possibly due to insufficient contrast between the defects and background.

The proposed network can segment the defects fairly accurately in C1-C4, but there is over-segmentation in C3, and a gap between the segmented defect edges and Ground truth in C4. As similarity and overlap between boundaries of the highlighted region and the defect are both significant, the upsampling part of the model will blur the boundaries between the two regions, leading to mis-segmentation.

Compared to other models, the performance of the network proposed in this paper is intuitively superior in terms of track surface defect segmentation.

For evaluating the robustness of the presented model, the mIoU and processing time of each model in handling the four types of samples have been compared with the results shown in TABLE II. In terms of mIoU, the proposed model is the highest in all three categories of samples except for C2 where it is slightly lower than PSPNet, while the worst performer is K-means. In terms of processing time, the proposed model is the fastest in C2 whereas is slower than PSPNet within the other three categories to varying degrees. It is observed that the proposed model achieves Pareto optimal regarding the indicators of mIoU and time, i.e., no model in a certain category of samples performs better than the proposed model simultaneously. According to the requirements of the use-cases, although the processing time of the presented model is not optimal, it can still cope with the real-time requirements. As the improvement in mIoU can reduce the probability of

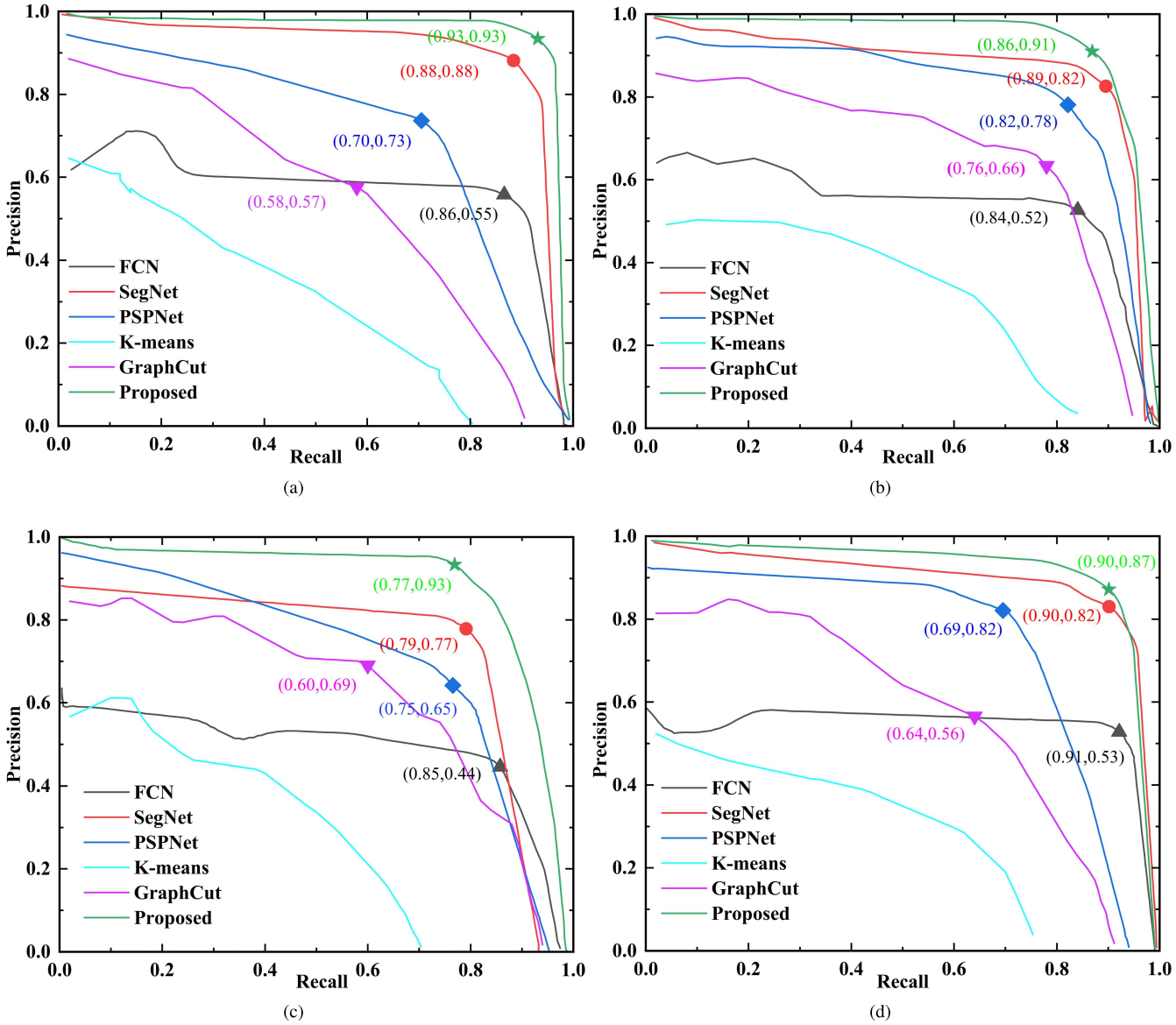


Fig. 6. The Recall and Precision of models when dealing with different samples. Note that (a)-(d) are the results of the models dealing with C1-C4 respectively.

mis-segmentation and thus reduce the risk of accidents, so it is worthwhile to compromise a slight real-time performance.

The curves of Recall and Precision for each model when processing the four types of samples are shown in Fig.6, and the approximate inflection points for each curve are also marked, with the corresponding AP, Recall and Precision at the inflection points listed in TABLE II. From the perspective of AP, Recall and Precision, the performance of the proposed model is optimal. The inflection point of the proposed model in C1 dominates those of the other models, while achieving Pareto optimal in C2-C4.

For evaluating the binary classification performance of the model, 70% of all samples in C1-C4 were used as training set and 30% as test set. Besides, two experts H1 and H2 who have been working on track defect detection for more than 2 years were invited to perform manual detection experiments. Fig.7 and TABLE III show the performance of the model on the training and test sets. The free-response receiver operating characteristic (FROC) curve represents the relationship between Sensitive and FP. The sensitivity of the model in the training set reaches a maximum value of about 0.90 when the FP is around 2, and almost maintains the same value as the FP continues to increase. The sensitivity of the model in the test set reaches a maximum value of about 0.93 at FP of 1.45, which also does not vary with further increasing of FP. As can be seen from TABLE IV, the proposed model outperforms FCN, PSPNet and SegNet significantly, with the highest sensitivity and the smallest Avg FP. It is worth mentioning that despite the lower Avg FP of H1 and H2, they did not perform as well as the neural network model in terms of sensitivity.

The parameter amount and mean time for sample segmentation for the model proposed and cited in this paper are listed in TABLE V. Although the number of parameters in the present

TABLE III
SENSITIVITIES OF THE PROPOSED MODEL IN THE TRAINING AND TEST SETS AT DIFFERENT FP LEVELS

Datasets	0.25	0.5	1	2	4	Avg
Training	0.551	0.672	0.794	0.896	0.901	0.763
Test	0.662	0.753	0.881	0.927	0.927	0.830

TABLE IV
COMPARISON OF THE PERFORMANCE OF DIFFERENT MODELS

Methods	Sensitivities @ FP Levels						Segmentation	
	0.25	0.5	1	2	4	Avg	Sensitivity	Avg FP
FCN	0.613	0.702	0.764	0.857	0.874	0.762	0.874	4.21
SegNet	0.651	0.742	0.824	0.866	0.903	0.797	0.903	3.12
PSPNet	0.621	0.736	0.802	0.872	0.911	0.788	0.911	5.35
Proposed	0.662	0.753	0.881	0.927	0.927	0.830	0.927	1.45
H1	N/A	N/A	N/A	N/A	N/A	N/A	0.742	0.98
H2	N/A	N/A	N/A	N/A	N/A	N/A	0.684	1.12

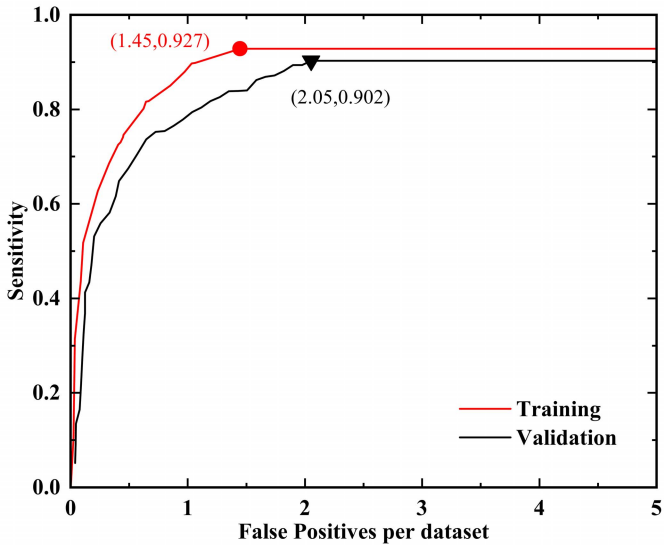


Fig. 7. FROC curves of the proposed model in the training and test sets.

TABLE V
NUMBER OF PARAMETERS AND MEAN SEGMENTATION TIMES FOR EACH MODEL

Model	Number of Parameters	Mean segmentation times (s)
FCN	1.20M	0.0542
SegNet	29.50M	0.0498
PSPNet	68.00M	0.0258
Proposed	7.40M	0.0332

model is larger than that of FCN, the mean operation time is shorter than that of FCN. The possible reason is that the features are fused in a point-by-point manner. In contrast, the present model employs Depth-Concatenation layers to stitch features together directly in the Channel dimension to form thicker features, allowing more track defect texture information to be propagated through the high-resolution layers, which can improve efficiency considerably. In summary, the present model is less complex with higher performance.

Fig. 8. Defect segmentation results from models formed by combining different modules. Note that (a) is the original image, (b) is the ground truth, (c)-(f) are the outputs of Combination #1-#4 respectively.

TABLE VI
RESULTS OBTAINED BY COMBINING DIFFERENT MODULES INTO A MODEL

Combination	Model	Fold=1	Fold=2	Fold=3	Average
#1	CP+FC	62.21	64.36	67.15	64.53
#2	CP+EP+FC	70.35	73.64	75.28	73.09
#3	CP+EP+Conv	74.07	78.67	81.06	77.93
#4	CP+EP+Conv+DC (Proposed)	80.04	84.28	88.79	84.37

It is evident from the above analysis that the model proposed outperforms the other six models in terms of stability and robustness for segmenting the four categories of track samples.

B. Ablation Study of The Proposed Model

Ablation experiments are conducted to evaluate the contribution of each part of the model to the overall performance. The contraction path in the proposed model is denoted as CP, the expansion path as EP, the convolution layer as Conv, the fully connected layer as FC and the feature mapping module as DC. The modules were sequentially combined to form different models. A mixed sample set consisting of the datasets of NEU-DET [11], RSDD [38] and the one collected in this paper was used for cross-validation for a fair comparison. Typical experimental results are shown in Fig.8 and the overall results are shown in TABLE VI.

The comparison between combination #1 and #2 shows that EP upsamples the feature map output by CP to restore it to the same size as the input image, so that it can predict each pixel while retaining the location information of the input image. And thenthe feature map can be classified from EP pixel by pixel, which can improve the pixel segmentation accuracy, as shown in Fig.8c and 8d. The average accuracy is improved by 8.56%.

The comparison between combination #2 and #3 shows that when the model takes FC as the output, the feature map is mapped into a feature vector of fixed length, which is helpful for classification and can distinguish well what classes of objects are contained in an image. However, it is difficult to achieve accurate segmentation because a large number of details are missing, the full information transfer of contextual features cannot be guaranteed, and the precise contour of the object cannot be properly presented, as in Fig.8d and 8e, so the performance of combination #2 is 4.84% lower than that of combination #3.

Since the model is improved from CNN, even if FC is substituted with Conv, it still classifies pixels independently without taking into account the pixel-to-pixel relationship. The

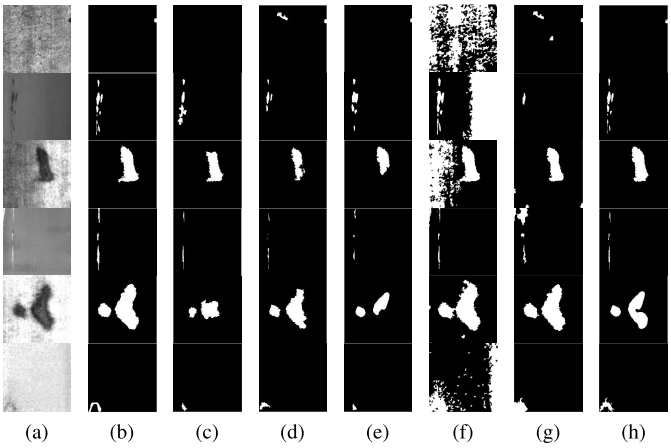


Fig. 9. Segmentation results comparison of the models based on NEU-DET dataset. Note that (a) are Original images, (b) are Ground truths, (c)-(g) are the results of PSPnet, FCN, SegNet, K-means and Grapcut respectively, and (h) are the resluts of the proposed model. From top to bottom are the typical results obtained by each method when processing A1 to A6.

spatial regularization step is omitted, which lacks spatial consistency and results in discontinuous segmentation, as shown in Fig.8e. The DC integrates the track defect texture features, which serves as a complementary information, allowing more information to be propagated in the high-resolution layers, thereby alleviating the information insufficient in the upsampling, and thus boosting the segmentation accuracy, as shown in Fig.8f, so that combination #4 performs the best.

C. Validation Based on Public Datasets

To verify the adaptability of the proposed model, the NEU-DET dataset in [11] have been incorporated for demonstration. A comparison of the segmentation results of the models when handling the NEU-DET dataset are shown in Fig.9.

PSPnet suffers from mis-segmentation in A1, over-segmentation in A2, promising results in A3, and smaller defective areas than Ground truth in A4-A6. FCN is able to segment defects in A1-A6, yet there is over-segmentation in A1, a smaller segmented defect area than Ground truth in A2-A5, and edge mis-segmentation in A6.

K-means shows over-segmentation in all of A1-A6, with severe over-segmentation in A1-A3. GraphCut exhibits over-segmentation and mis-segmentation in A1, A3, A4 and A6, and under-segmentation in A2.

SegNet performs well in A1, with mis-segmentation and over-segmentation of the upper edge of the defect in A2, proper segmentation in A3-A5 with inadequate edge accuracy, and under-segmentation in A6.

The proposed model in this paper properly segments defects in A1, A2 and A4, but suffers from under-segmentation in A3 and A5 and over-segmentation in A6.

The mIoU and processing times of the six models on NEU-DET are shown in TABLE VII. The mIoU is below 80% for K-means, GraphCut, SegNet and PSPNet, while above 80% for the FCN and the proposed models, which is the most advanced. As can be noticed, the performances of each model on NEU-DET appear to be lower than those on the dataset collected in this paper. There are potentially three reasons. Firstly, the dataset contains lots of noise behaving in similar manner with defects, thus making it prone to incorrect segmentation. Secondly, the traditional segmentation models are incapable of discriminating the exact location of defects well for noisy datasets. Thirdly, the deficiency that the poorly performing networks themselves do not handle the dataset well, as mentioned before: each model corresponds to a specific dataset that excels at handling, whereas the dataset in this paper represents a new challenge for them. The proposed model took 0.0215s, the least time consuming.

TABLE VII
PERFORMANCE OF THE MODELS ON NEU-DET

Method	mIoU (%)	Time (s)	AP	Recall@ Precision=0.9	Precision @ Recall=0.8
FCN	83.24	0.0514	0.787	N/A	0.79
SegNet	78.21	0.0397	0.664	N/A	0.63
PSPNet	71.11	0.0238	0.699	N/A	0.73
K-means	38.57	0.7709	0.269	N/A	N/A
GraphCut	53.64	0.9841	0.555	N/A	0.30
Proposed	85.37	0.0215	0.866	0.52	0.87

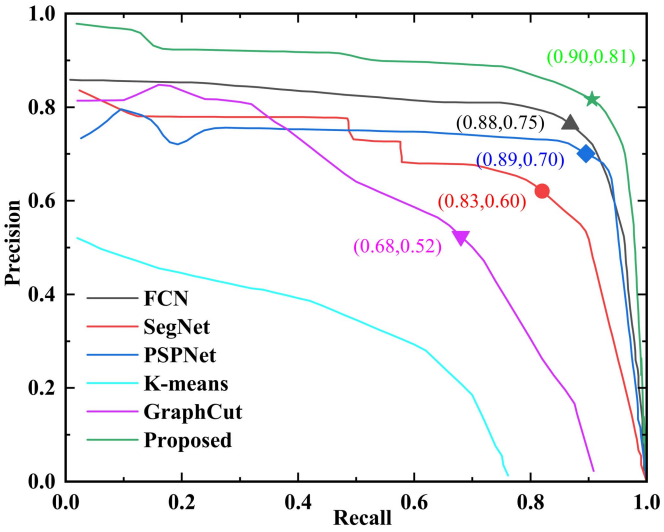


Fig. 10. Performance of each model on NEU-DET.

The curves of Recall and Precision for each model when processing the NEU-DET are shown in Fig.10, and the approximate inflection points for each curve are also marked, with the corresponding value of AP, Recall and Precision at the inflection points listed in TABLE VII.

As can be seen in Fig.10, the Precision of the proposed model fluctuates slightly when the Recall is between 0.1 and 0.2, and is stationary when the Recall is between 0.2 and 0.8. FCN and K-means performed well regarding stability, with GraphCut, SegNet and PSPNet all presenting significant fluctuations. In terms of AP, Precision and Recall, the proposed model performs best, followed by SegNet, while FCN, which performs well against mIoU, and PSPNet, which performs well against time, appear to be underperformed. The inflection point of the proposed model dominates those of the other models.

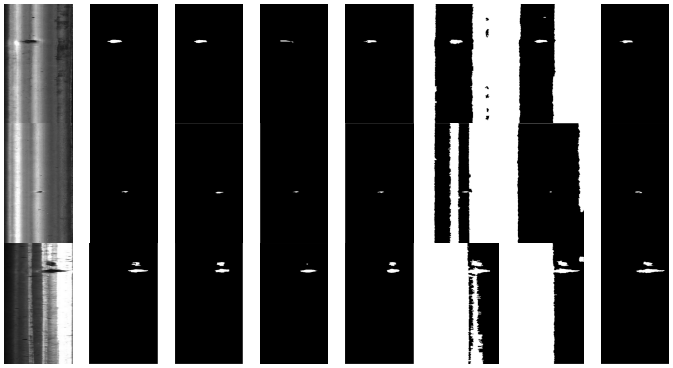


Fig. 11. Segmentation results comparison of the models based on RSSD dataset. Note that (a) are Original images, (b) are Ground truths, (c)-(g) are the results of PSPNet, FCN, SegNet, K-means and Grapcut respectively, and (h) are the resluts of the proposed model. From top to bottom are the typical results obtained by each method when processing B1 to B3.

TABLE VIII
PERFORMANCE OF THE MODELS ON RSSD

Method	mIoU (%)	Time (s)	AP	Recall@ Precision=0.9	Precision@ Recall=0.8
FCN	75.59	0.0962	0.617	N/A	0.51
SegNet	63.17	0.0527	0.430	N/A	0.30
PSPNet	52.34	0.0394	0.412	N/A	0.29
K-means	13.27	0.584	N/A	N/A	N/A
GraphCut	19.36	0.627	N/A	N/A	N/A
Proposed	79.36	0.0405	0.785	0.44	0.75

It can be safely concluded based on the above analysis that although the stability and robustness of the proposed model is not as good as in C1-C4, the comprehensive performance is also the optimal among the six models.

Similar to the previous section, the RSDD dataset from [38] is deployed for validation, and the segmentation results are shown in Fig.11.

The proposed model segments the defect accurately on both B1 and B2, except for a slight over-segmentation on B3.

The mIoU and processing times of the six models on RSDD are shown in TABLE VIII. For mIoU, SegNet and PSPNet are both below 70%, the worst performers are K-means and GraphCut, which are less than 20%, while FCN and the proposed model are above 75% with the proposed model being the highest. The model proposed in this paper is not optimal in terms of detection time, nevertheless is able to satisfy the real-time requirements.

The curves of Recall and Precision for each model when processing the RSDD are shown in Fig.12. As can be seen, the Precision of the proposed model fluctuates slightly when the Recall is about 0.5, and is otherwise stable. Generally, the stability and robustness of the proposed model are optimal among the six models

D. Comparison With Existing Models

TABLE IX presents brief information on the track detection models for track defects proposed in recent papers, where "N/A" denotes not present in the paper. Since the experiments were conducted under different conditions, with varied samples and computer configurations, and each article had a distinct research orientation and focus, a direct comparison in terms of 'Accuracy' and 'Speed' is of little interest and is not the intention of TABLE IX. However, it can still be concluded from the comparison that the proposed model is capable of performing the track defect segmentation.

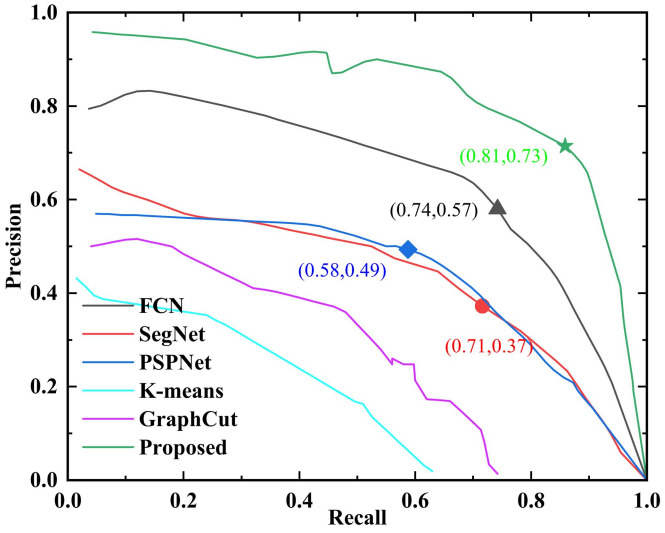


Fig. 12. Performance of each model on RSSD.

TABLE IX
SUMMARY AND COMPARISON OF CURRENT TRACK DETECTION/SEGMENTATION MODELS

Method	Year	Components	Accuracy (%)	Speed (s/frame)
Method in [23]	2015	Concrete ties	93.35	N/A
MT in [24]	2017	Ties / Fastener	96.74	N/A
ILNET in [25]	2017	Wires	96.54	N/A
PVANET++ in [26]	2018	Split pins	95.28	2.325
Method in [27]	2018	Fastener	92.78	0.00157
DM-RIS in [39]	2018	Rail	96.74	0.485
Method in [40]	2020	Rail	90.00	0.173
Method in [41]	2020	Fastener	96.26	N/A
Method in [42]	2021	Fastener	93.50	N/A
Proposed	2021	Rail	97.47	0.033

V. DISCUSSION

A segmentation method is proposed in this paper for detection of track surface defects. The collected track datasets with different service lifetimes and states are classified into four categories and the images are grey-scale normalised and then fed into the proposed network model. The proposed model stitches features together in the Channel dimension to form denser features, allowing additional information about track defect textures to be propagated through the high-resolution layers. The model drops the fully-connected layer to ensure that the segmentation results are based on contextual features that are not deficient, improving the accuracy of the results. The model drops the weak correlations learned during the convolution process so that the convolutional modules can contribute to a shared set of weights, reducing redundancy and complexity of the model.

To evaluate the performance of the method, experiments were conducted on four types of defective samples, yielding Accuracy=97.47%, Lose=0.0061 and an average frame rate of 33.3ms. In addition, the model was also analyzed with two publicly available datasets, NEU-DET and RSDD, on which the mIoU of the proposed model obtained 2.13% and 3.77% improvements relative to other models, respectively. Moreover, the method proposed in this paper has shown excellent stability, adaptability and robustness when tested against different models and can effectively perform track defect segmentation.

Because of the limitations in this research, further work would be accomplished in the future:

1) Defects are not automatically classified. In fact, some defects do not require particular treatment and, if not differentiated, can increase the workload in post maintenance.
2) Since the focus of this paper is on defect segmentation, the work on track classification is done manually. However, this is clearly infeasible when applied on a wide scale and it is therefore necessary to design specialized CNN classifiers for track classification.

REFERENCES

[1] L. Xiao, B. Wu, and Y. Hu, "Surface defect detection using image pyramid," *IEEE Sensors J.*, vol. 20, no. 13, pp. 7181–7188, Jul. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9019620/

[2] J. Tao, Y. Zhu, W. Liu, F. Jiang, and H. Liu, "Smooth surface defect detection by deep learning based on wrapped phase map," *IEEE Sensors J.*, vol. 21, no. 14, pp. 16236–16244, Jul. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9419060/

[3] S. Hajizadeh, A. Núñez, and D. M. J. Tax, "Semi-supervised rail defect detection from imbalanced image data," *IFAC-PapersOnLine*, vol. 49, no. 3, pp. 78–83, 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2405896316302099

[4] A. R. Rizvi, P. R. Khan, and S. Ahmad, "Crack detection in railway track using image processing," *Int. J. Adv. Res., Ideas Innov. Technol.*, vol. 3, no. 4, pp. 489–496, 2017.

[5] Q. Li and S. Ren, "A real-time visual inspection system for discrete surface defects of rail heads," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 8, pp. 2189–2199, Aug. 2012. [Online]. Available: http://ieeexplore.ieee.org/document/6151140/

[6] Q. Li and S. Ren, "A visual detection system for rail surface defects," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 1531–1542, Nov. 2012. [Online]. Available: http://ieeexplore.ieee.org/document/6208896/

[7] Z. He, Y. Wang, F. Yin, and J. Liu, "Surface defect detection for high-speed rails using an inverse P-M diffusion model," *Sensor Rev.*, vol. 36, no. 1, pp. 86–97, Jan. 2016. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/SR-03-2015-0039/full/html

[8] J. Gan, Q. Li, J. Wang, and H. Yu, "A hierarchical extractor-based visual rail surface inspection system," *IEEE Sensors J.*, vol. 17, no. 23, pp. 7935–7944, Dec. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/8063875/

[9] S. Fekri-Ershad and F. Tajeripour, "Multi-resolution and noise-resistant surface defect detection approach using new version of local binary patterns," *Appl. Artif. Intell.*, vol. 31, nos. 5–6, pp. 395–410, 2017. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/08839514.2017.1378012

[10] S. Fekri-Ershad, "Bark texture classification using improved local ternary patterns and multilayer neural network," *Expert Syst. Appl.*, vol. 158, Nov. 2020, Art. no. 113509. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S095741742030333X

[11] Y. He, K. Song, H. Dong, and Y. Yan, "Semi-supervised defect classification of steel surface based on multi-training and generative adversarial network," *Opt. Lasers Eng.*, vol. 122, pp. 294–302, Nov. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0143816619306499

[12] V. Natarajan, T.-Y. Hung, S. Vaikundam, and L.-T. Chia, "Convolutional networks for voting-based anomaly classification in metal surface inspection," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Mar. 2017, pp. 986–991. [Online]. Available: http://ieeexplore.ieee.org/document/7915495/

[13] S. Wang, Y. Yin, G. Cao, B. Wei, Y. Zheng, and G. Yang, "Hierarchical retinal blood vessel segmentation based on feature and ensemble learning," *Neurocomputing*, vol. 149, pp. 708–717, Feb. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231214010169

[14] J. Masci, U. Meier, D. Ciresan, J. Schmidhuber, and G. Fricout, "Steel defect classification with max-pooling convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, Jun. 2012, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/6252468/

[15] D. Soukup and R. Huber-Mörk, *Convolutional Neural Networks for Steel Surface Defect Detection From Photometric Stereo Images*, vol. 8887. Cham, Switzerland: Springer, 2014, pp. 668–677. [Online]. Available: http://link.springer.com/10.1007/978-3-319-14249-4_64

[16] J. Masci, U. Meier, G. Fricout, and J. Schmidhuber, "Multi-scale pyramidal pooling network for generic steel defect classification," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Aug. 2013, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/document/6706920/

[17] Y. He, K. Song, Q. Meng, and Y. Yan, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1493–1504, Apr. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8709818/

[18] H. Zhang, X. Jin, Q. M. J. Wu, Y. Wang, Z. He, and Y. Yang, "Automatic visual detection system of railway surface defects with curvature filter and improved Gaussian mixture model," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 7, pp. 1593–1608, Jul. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8304596/

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440. [Online]. Available: http://ieeexplore.ieee.org/document/7298965/

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241. [Online]. Available: http://link.springer.com/10.1007/978-3-319-24574-4_28

[21] H. Yang, Y. Chen, K. Song, and Z. Yin, "Multiscale feature-clustering-based fully convolutional autoencoder for fast accurate visual inspection of texture surface defects," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 3, pp. 1450–1467, Jul. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8598794/

[22] H. Dong, K. Song, Y. He, J. Xu, Y. Yan, and Q. Meng, "PGA-Net: Pyramid feature fusion and global context attention network for automated surface defect detection," *IEEE Trans. Ind. Informat.*, vol. 16, no. 12, pp. 7448–7458, Dec. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8930292/

[23] X. Gibert, V. M. Patel, and R. Chellappa, "Material classification and semantic segmentation of railway track images with deep convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 621–625. [Online]. Available: http://ieeexplore.ieee.org/document/7350873/

[24] X. Gibert, V. M. Patel, and R. Chellappa, "Deep multitask learning for railway track inspection," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 153–164, Jan. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7506117/

[25] Z. Liu, L. Wang, C. Li, and Z. Han, "A high-precision loose strands diagnosis approach for isoelectric line in high-speed railway," *IEEE Trans. Ind. Informat.*, vol. 14, no. 3, pp. 1067–1077, Mar. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8114270/

[26] J. Zhong, Z. Liu, Z. Han, Y. Han, and W. Zhang, "A CNN-based defect inspection method for catenary split pins in high-speed railway," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 8, pp. 2849–2860, Aug. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8482333/

[27] J. Chen, Z. Liu, H. Wang, A. Nunez, and Z. Han, "Automatic defect detection of fasteners on the catenary support device using deep convolutional neural network," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 2, pp. 257–269, Feb. 2018. [Online]. Available: http://ieeexplore.ieee.org/document/8126877/

[28] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with convolutional neural networks," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, vol. 1, 2014, pp. 766–774, doi: 10.5555/2968826.2968912.

[29] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2, 2012, pp. 2843–2851, doi: 10.5555/2999325.2999452.
[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Sep. 2014, pp. 1–14.
[32] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. CVPR*, Jun. 2015, pp. 1–9. [Online]. Available: http://ieeexplore.ieee.org/document/7298594/
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778. [Online]. Available: http://ieeexplore.ieee.org/document/7780459/
[34] 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. [Online]. Available: http://ieeexplore.ieee.org/document/8100143/
[35] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7803544/
[36] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *Appl. Statist.*, vol. 28, no. 1, p. 100, 1979. [Online]. Available: https://www.jstor.org/stable/10.2307/2346830?origin=crossref
[37] B. L. Price, B. Morse, and S. Cohen, "Geodesic graph cut for interactive image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3161–3168. [Online]. Available: http://ieeexplore.ieee.org/document/5540079/
[38] J. Gan, J. Wang, H. Yu, Q. Li, and Z. Shi, "Online rail surface inspection utilizing spatial consistency and continuity," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 7, pp. 2741–2751, Jul. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8354938/
[39] X. Jin *et al.*, "DM-RIS: Deep multimodel rail inspection system with improved MRF-GMM and CNN," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1051–1065, Apr. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8692707/
[40] S. B. Block, R. D. da Silva, L. B. Dorini, and R. Minetto, "Inspection of imprint defects in stamped metal surfaces using deep learning and tracking," *IEEE Trans. Ind. Electron.*, vol. 68, no. 5, pp. 4498–4507, May 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9062515/
[41] X. Wei, D. Wei, D. Suo, L. Jia, and Y. Li, "Multi-target defect identification for railway track line based on image processing and improved YOLOv3 model," *IEEE Access*, vol. 8, pp. 61973–61988, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9050731/
[42] Z. Tu, S. Wu, G. Kang, and J. Lin, "Real-time defect detection of track components: Considering class imbalance and subtle difference between classes," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9559954/

Hongfei Yang received the master's degree in mechanical engineering from Jilin University, Changchun, China, in 2020. He is currently pursuing the Ph.D. degree in test and measurement technology & instrumentation. His research interests include industrial defect detection, environmental identification of engineering vehicles, and intelligent sensing instrumentation.

Yanzhang Wang received the B.S., M.S., and Ph.D. degrees from Jilin University in 2002, 2005, and 2010, respectively. He is now a Professor with the School of Instrumentation & Electronic Engineering, Jilin University. His research interests are infrared spectral detection, intelligent sensing, intelligent signal processing, weak magnetic signal processing, and quantum sensors.

Jiyong Hu received the B.Sc. and M.Sc. degrees from Jilin University in 2009 and 2012, respectively. Now, he is a Planning Engineer in FAW-VW and his main work focuses on conveyor system of automobile assembly shop.

Jiatang He received the B.Sc. degree in electrical engineering and automation from Wismar University Germany in 2007. Now, he is a Supervisor Planning Engineer in FAW-VW and his main work focuses on conveyor system of automobile assembly shop.

Zongwei Yao received the B.Sc. and Ph.D. degrees in engineering from Jilin University, Changchun, China, in 2008 and 2013, respectively. He is currently an Associate Professor with Jilin University, where he is heading a group working on intelligent of engineering vehicles. He has been focusing on research areas of 3D construction of environment, autonomous operation of construction vehicles, and mechanical system dynamics.

Qiushi Bi received the B.Sc. and Ph.D. degrees in engineering from Jilin University, Changchun, China, in 2014 and 2019, respectively. He is currently a Lecturer with Jilin University. He has been focusing on research areas of intelligent of engineering vehicles, autonomous operation of mining equipment, optimization, and advanced design method.