A Convolutional Neural Network with Equal-Resolution Enhancement and Gradual Attention of Features for Tiny Target Detection

Mingyang Cheng, Junliang Wang, Yaqin Zhou, Chuqiao Xu, Ying Liu, and Jie Zhang

***Abstract*— The detection of tiny targets on the surface with high efficiency and accuracy is significant for the current intelligent manufacturing. Visual inspection methods based on deep learning are widely utilized to detect tiny objects. However, the tiny objects appear less distinct, less wide, and less area occupied in the image. At the same time, there is a lot of object-like noise, which further increases the difficulty of detecting tiny objects. In response to the challenges brought by the complexity of the detection environment, this paper proposes a network model that combines the enhancement of pixel-level features at equal resolution and the introduction of full-scale features based on attention. The model utilizes the subtle differences between the tiny target and the background and the semantic information of the tiny target outline to enhance the features of the tiny target while significantly reducing its loss in the equal-resolution feature layer. Additionally, a gradual attention mechanism is proposed to guide the network model to pay attention to tiny objects features on the full-scale feature layer. The performance of this network model is validated on a real dataset. Experiments show that the model exhibits superior performance and outperforms existing resNet50, DenseNet, Racki-Net, and SegDecNet in detecting tiny objects.**

***Index Terms*— detection of tiny targets, Visual inspection, enhancement of pixel-level features, gradual attention mechanism**

I. Introduction

Cracks are typical tiny targets and the ones on the surface of objects bring significant insecurity to equipment products in various fields, such as industry[1] and aerospace[2]. Thus, accurate and efficient techniques for detecting cracks are of great significance to the current manufacturing industry[3]. In the early days, most crack defects were detected by traditional manual methods, which showed low accuracy and low efficiency. With the progress of machine learning, the application of crack detection technology based on machine vision is gradually widespread and has become a focus of attention.

In the past two decades, traditional image processing techniques have been applied to simple detection tasks, such as thresholding methods[4], region segmentation[5], and morphological features[6]. However, these methods exhibit low accuracy and poor stability for detecting cracks under high noise. Subsequently, machine learning techniques have further allowed for a more sophisticated environment for detecting cracks. In [7], a method based on the random forest was proposed to generate descriptors of cracks. In addition, a crack detection method based on the support vector machine(SVM) was proposed in [8]. Although these methods show good performance in a specific complex detection environment, they mainly rely on manual or shallow feature extraction, resulting in significant limitations and poor robustness.

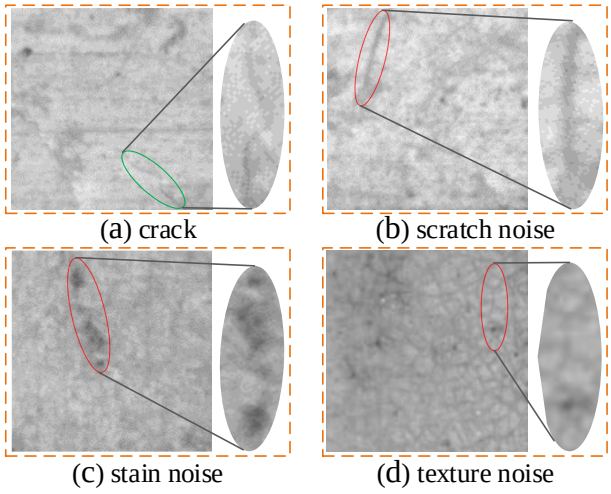

Figure 1. Cracks and various noises.

In recent years, the convolutional neural network(CNN) in the deep learning method[9] has been widely utilized in diverse scenarios such as industrial surveying[10] and object detection[11] because of its superiority, and it has also become the mainstream of crack detection[12](e.g., Qu et al.[13], Nguyen et al.[14], Han et al.[15]). In object detection, although the position of the crack can be located[16][17], its

*Research sponsored by Chenguang Program (Grant No.20CG41) which is supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission, and The Fundamental Research Funds for the Central Universities (2232021A-08).

Mingyang Cheng, College of Mechanical Engineering, Donghua University, Shanghai 201620, China (e-mail: 2200951@mail.dhu.edu.cn).

Junliang Wang, Engineering Research Center of Digitalized Textile and Fashion Technology, Ministry of Education, Shanghai 201620, China (corresponding e-mail: junliangwang@dhu.edu.cn).

Yaqin Zhou, College of Mechanical Engineering, Donghua University, Shanghai 201620, China (e-mail: zhouyaqin@dhu.edu.cn).

Chuqiao Xu, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China (e-mail: xuchuqiao@sjtu.edu.cn).

Ying Liu, School of Engineering Cardiff University, UK (e-mail: liuy81@cardiff.ac.uk).

Jie Zhang, Institute of Artificial Intelligence, Donghua University, Shanghai 201620, China (e-mail: mezhangjie@dhu.edu.cn.).

accuracy and efficiency have certain limitations. To quantitatively analyze cracks[18][19], the method based on pixel segmentation was applied to crack detection[20][21]. Although pixel segmentation shows superior performance in detecting cracks, its effect is still limited in the case of high noise, and some scenes need to determine the existence of cracks qualitatively. To detect cracks accurately and qualitatively, in [22] and [23], a network structure combining segmentation and classification is proposed to detect cracks. Experiments show that the combined network structure is beneficial to crack detection. However, complex application scenarios present new challenges to detection techniques. First, the cracks have low contrast in the image and are tiny, as shown in Fig. 1(a). The narrow width of the cracks on the surface also results in almost no difference in pixel values between cracked and non-cracked. Low-contrast and subtle cracks require the network model to have a more prominent ability to preserve features. Second, there is much crack-like noise in the non-cracked region, which is likely to lead to false detections by the model, as shown in Fig. 1(b)(c)(d). Third, the space ratio of cracks is small. The randomness of crack generation and its small width make the pixel points of the crack occupy a small proportion of the entire image, which increases the difficulty of the model to capture the features of the crack in the feature layer of a larger space.

To cope with the above difficulties, a new network structure based on CNN is proposed for detecting cracks in tiny targets. First, Equal Resolution Feature Enhancement Network(ERFE-Net) is designed to extract and enhance features of tiny targets through subtle differences between tiny targets and backgrounds and semantic information of tiny target contours in equal-resolution feature layers, which fully guarantees the ability of the network to retain useful features. Second, the fusion of multi-scale features can improve the detective performance of models[24][25]. Thus, the full-scale feature layers from ERFE-Net are extracted to obtain more adequate features of tiny targets. Additionally, the proposed gradual squeeze and excitation Network(GSE-Net) improves the ability of the network model to discover features of tiny targets, which further strengthens the effective features. Thus, both ERFE-Net and GSE-Net highly enhance and preserve features of tiny targets under high noise.

The rest of the paper is organized as follows: Section 2 describes the proposed tiny target detection network. Section 3 presents the experimental details and results analysis. Finally, the paper and future work are summarized.

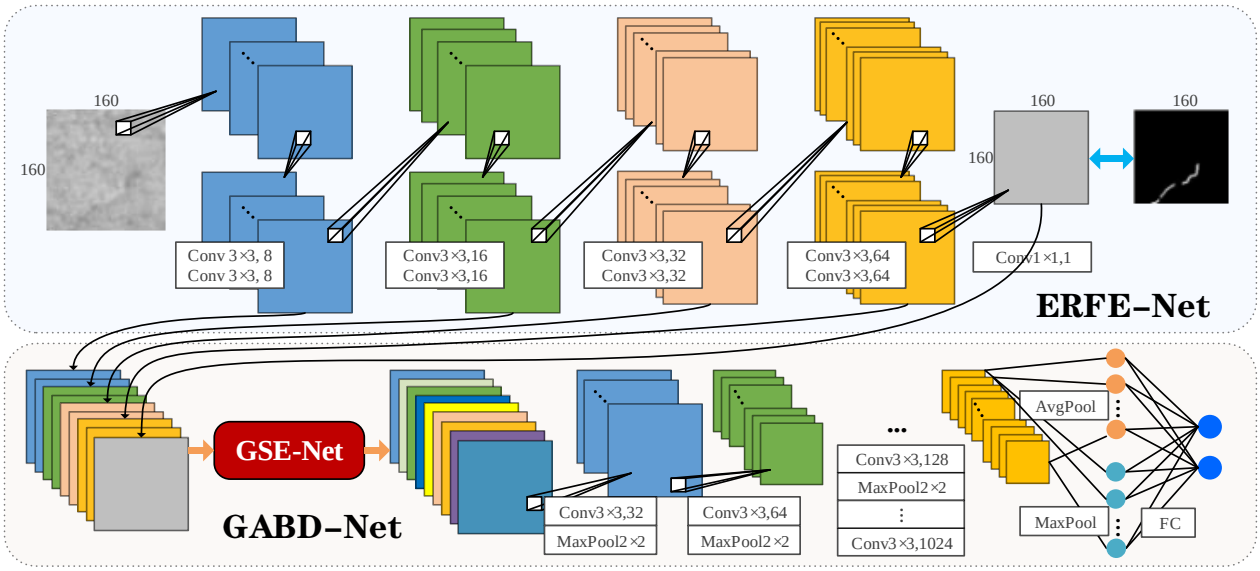

Figure 2. The structure for detecting tiny targets.

II. METHOD

The structure of the network proposed in this paper consists of two parts. In the first part, ERFE-Net is designed to extract and strengthen weak features of tiny targets. Then full-scale features, extracted from ERFE-Net and passed through an attention network with gradual squeezing and excitation, are used as input to the decision network to predict the probability of tiny targets' existence. Fig. 2 shows the structure of the entire network model.

A. Equal Resolution Feature Enhancement Network

To amplify the feature difference between tiny targets and non-target, ERFE-Net is proposed to enhance and extract features of tiny targets in the case of weak features and noise. To achieve this, the design of the network should satisfy the following requirements: (a) the requirement of effectively enhancing weak features and ignoring noise; (b) the requirement that effective features are highly preserved. In order to realize the above requirements, the structure of the network module is embodied in the following description. First, ERFE-Net achieves pixel-level segmentation by exploiting the subtle differences between tiny target and non-target pixel values and the semantic information of tiny target contours, enabling the convolution kernel to focus on tiny targets and ignore disturbances in a targeted manner. Second, the resolutions of all feature layers remain unchanged during extracting features, as shown in Fig. 3. ERFE-Net does not perform any downsampling operation because downsampling will further lose effective features to a certain extent.

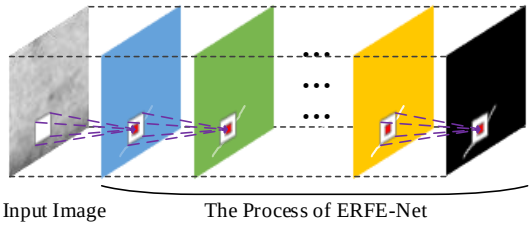

Figure 3. ERFE-Net process of strengthening features.

ERFE-Net is a fully convolutional network with 9 layers, as shown in Fig. 2 To avoid overfitting, batch normalization is performed after convolution operations except the last layer. Then, ReLU is utilized as an activation function. In order to extract features meticulously, the convolution kernels of the first 8 layers of the network are all 3×3, and the last layer is 1×1. The spatial size of each sample remains unchanged, and the output channel becomes 1 after passing through ERFE-Net. At this point, the output and the label image are consistent in space and channel. The loss is obtained by computing the point-to-point value between the output and the label image. The label value at each point on the label image is shown in the following formula.

$$y_{(i,j)} = \begin{cases} \rightarrow 1, \ (i, j) \in \textit{tiny target} \\ \rightarrow 0, \ (i, j) \in \textit{background} \end{cases} \tag{1}$$

where $y_{(i,j)}$ is the label value of the position (i, j) in the sample. If position (i, j) is the pixel of the tiny target, the corresponding true value on the label image is 1. Otherwise, the true value of the position is 0.

B. Gradual Attention Based Decision Network

1) Decision Network

In neural networks, shallow and deep features layers contain more details and semantic information, respectively. To efficiently utilize useful features, the rich features extracted from ERFE-Net are taken as the input of the data of the second part of the detective network. Specifically, the last feature layer and the feature layer of channel dimension transformation from ERFE-Net are taken out, as shown in Fig. 2. Then, these feature layers are concatenated along the channel direction without any operation since all feature layers have the same spatial size. At this point, the second part's input contains rich detailed and semantic information. Additionally, the spliced feature layer implements an attention mechanism for tiny targets through GSE-Net. The remaining structure of the second part is a binary classification network, and each image corresponding to a value of the label is shown in the following formula.

$$Y_{(i,j)} = \begin{cases} \rightarrow 1, \ \textit{image} \in \textit{tiny target images} \\ \rightarrow 0, \ \textit{image} \in \textit{without tiny target images} \end{cases} \tag{2}$$

where $Y_{(i,j)}$ is the label value of the image. If the sample contains tiny targets, the value of the corresponding label is 1. Otherwise, the label value of this sample is 0. The remaining network structure consists of 6 convolutional layers and 1 fully connected layer, as shown in Fig. 2. All convolutional layers undergo Batch normalization and ReLU activation after convolutional operations. In addition, the first five layers all perform a 2×2 max-pooling operation to realize the downsampling of the feature layer after the convolutional operation. To improve the confidence of network predictions, global max pooling and global average pooling are utilized in parallel after the sixth convolutional layer, resulting in 2048 neurons. Then, 2 neurons are obtained through the fully connected layer and the softmax function. That is the predicted probability of whether the image contains tiny targets is obtained.

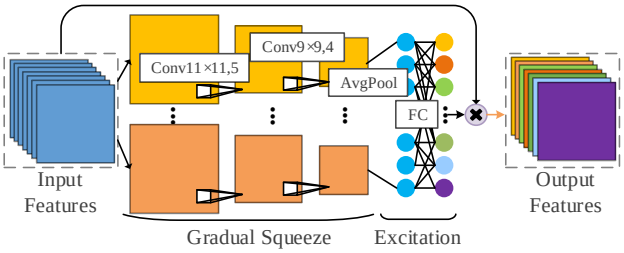

Figure 4. The structure of GSE-Net.

2) Gradual Squeeze and Excitation Network

The tiny targets appear inconspicuous, and the area ratio is also minimal in the image. Meanwhile, the content of useful information of each feature layer is also different. To further enhance the features of tiny targets, GSE-Net is proposed to pay attention to the region of tiny targets in the feature layer. Compared with SE-Net[26], the gradual squeeze can avoid abruptly compressing feature layers abruptly and ignoring effective features, and it enhances the expressive power of effective features. Thus, the gradual squeeze can gradually induce the network to discover features of tiny targets more easily.

GSE-Net consists of two parts: gradual squeeze and excitation, as shown in Fig. 4. First, gradual squeeze acts on each feature layer, including two convolutional operations and one average pooling operation, and then obtains the representative value of each feature layer after squeezing, as shown in the following formula.

$$\boldsymbol{F}_i^1 = \delta\left(Conv_5^{11\times11}\left(\boldsymbol{F}_i\right)\right) \tag{3}$$

$$\boldsymbol{F}_i^2 = \delta\left(Conv_4^{9\times9}\left(\boldsymbol{F}_i^1\right)\right) \tag{4}$$

$$f_i = GAP\left(\boldsymbol{F}_i^2\right) \tag{5}$$

where $\boldsymbol{F}_i \in \mathbb{R}^{160\times160\times1}$ is the feature of the i^{th} layer in the feature layer, with dimensions [160, 160, 1]. $\boldsymbol{F}^1 \in \mathbb{R}^{32\times32\times1}$ and $\boldsymbol{F}^2 \in \mathbb{R}^{8\times8\times1}$ are the features sequentially produced by the Gradual Squeeze process. $Conv_5^{11\times11}(\cdot)$ is a convolution kernel and a stride of 5 convolutional operations. $\delta(\cdot)$ is the function of rectified linear units. $GAP(\cdot)$ is global average pooling and f_i is the representative value of the features of the i^{th} layer after gradually extruding. The features of each layer in the feature layer are obtained sequentially after passing through the above stepwise extrusion operations. Then, all the representative values are activated through the fully connected layer and the sigmoid function, as shown in the following equation.

$$\boldsymbol{w} = \sigma\left(FC\left(f_1, f_2, \cdots, f_N\right)\right) \tag{6}$$

where N is the number of channels in the feature layer. $FC(\cdot)$ is the fully connected layer and $\sigma(\cdot)$ is the sigmoid function. $\boldsymbol{w} \in \mathbb{R}^{121}$ is the representative value after activation, and the size is 121. Finally, a gradual attention mechanism is implemented through element-wise multiplication between the activated representative values and the initial feature layer, as shown in the following equation.

$$\tilde{\boldsymbol{F}} = \boldsymbol{F} \otimes \boldsymbol{w} \tag{7}$$

where $\otimes$ is element-wise multiplication and $\tilde{\boldsymbol{F}}$ is the feature layer that has passed through GSE-Net.

C. Training

In this paper, the training of the proposed network for detecting tiny target consists of two parts: the training of ERFE-Net and the training of gradual attention based decision network(GABD-Net). There is a sequential relationship between the two parts of the training.

ERFE-Net is trained first, and GABD-Net is frozen during the training process. When the loss value of ERFE-Net reaches the specified value, the training is stopped. The loss function of ERFE-Net is shown below.

$$Loss_{ERFE} = -\frac{1}{b \times h \times w} \sum_{k=1}^{b} \sum_{i=1}^{h} \sum_{j=1}^{w} \left(y_{(k,i,j)} - p_{(k,i,j)}\right)^2 \tag{8}$$

where b is the number of samples in a batch. w and h are the width and height of the image, respectively. $y_{(k,i,j)}$ and $p_{(k,i,j)}$ are the label value and predicted value of position (i, j) in the k^{th} image, respectively. Then GABD-Net is trained and ERFE-Net is frozen and its loss function is shown below.

$$Loss_{GABD} = -\frac{1}{b} \sum_{k=1}^{b} \left[Y_k \cdot log\left(P_k\right) + \left(1 - Y_k\right) \cdot log\left(1 - P_k\right) \right] \tag{9}$$

where Y_k is the label value of the k^{th} image and P_k is the corresponding predicted value. When the number of training rounds reaches the specified value, the training of the entire network model is completed.

III. Experiments

To demonstrate the performance of the proposed detective network, related experiments are performed on a dataset derived from crack images of tantalum capacitors. Specifically, the experiment includes 3 parts: experimental settings, ablation experiment, and comparative experiment. Experimental results show that the network model exhibits superior performance than existing network models.

A. Experimental Settings

1) Dataset

The dataset of cracks was obtained by using an industrial camera to capture tantalum capacitors in a visual experiment platform, and the result was a total of 2400 images. In this dataset, there are 153 images containing cracks. Images containing cracks were taken as positive samples. 153 images without cracks are regarded as negative samples to balance positive and negative samples. Thus, the dataset has 306 images with a resolution of 160 × 160 pixels. Examples with and without cracks are shown in Fig. 5, and each image is provided with a manual pixel-level region mask. 266 images are randomly selected as an independent validation set, and the rest are training sets before experiments. To ensure the experimental validity, the images in the validation set of each experiment are the same. The training set and the validation set of each experiment contain half of the positive and negative samples.

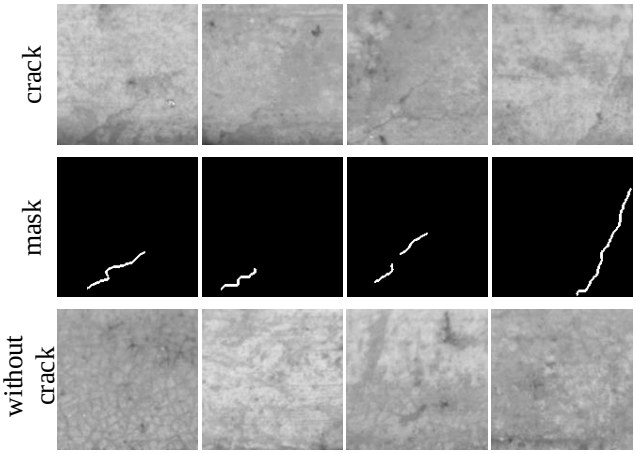

Figure 5. Examples of samples in the dataset: positive samples(crack), pixel-level region masks(mask), and negative samples(without crack).

2) Implementation Details

The dataset was acquired using a Conway 20x telecentric lens, model MVL-MY-2-110-MP. All experiments were run on CPU x Intel(R) Xeon(R) CPU E5-2698 v4 and NVIDIA Tesla V100. The proposed network employs the Adam optimizer during training with a learning rate set to 0.001, an average coefficient of gradient computation of 0.5, and a sum-squared coefficient of 0.999. The size of each batch is 16, and each batch is shuffled randomly. The loss threshold for training ERFE-Net is set to 0.0005, and the number of training epochs for GABD-Net is 100 epochs.

3) Evaluation Metrics

In order to evaluate the performance of the proposed detective model from multiple aspects, multiple evaluation metrics are introduced, including accuracy, precision, recall, and F1-score, as shown in the following equation.

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

$$Pr = \frac{TP}{TP + FN} \tag{11}$$

$$Re = \frac{TP}{TP + FP} \tag{12}$$

$$F1 = \frac{2 \times Pr \times Re}{Pr + Re} \tag{13}$$

where TP, FP, TN, FN is True Positive, False Positive, True Negative, and False Negative, respectively. In addition, to further demonstrate the stability of the proposed model, the receiver operating characteristic curve(ROC) and area under

the curve(AUC) also serve as evaluation means.

B. *Ablation experiments*

This subsection demonstrates the effectiveness of different innovative parts in the network model through ablation experiments. The effectiveness of ERFE-Net and GSE-Net are sequentially evaluated separately. Specific experimental details will be described below.

TABLE I. COMPARATIVE EXPERIMENTAL RESULTS ON THE PERFORMANCE OF ERFE-NET

Training	Ac	Pr	Re	F1	AUC
None ERFE	62.41%	59.76%	75.94%	66.89%	65.89%
Non-train ERFE	70.30%	71.77%	66.92%	69.26%	75.38%
Reduce FLR	96.24%	96.95%	95.49%	96.21%	98.33%
Ours	**98.50%**	**100.0%**	**96.99%**	**98.47%**	**99.96%**

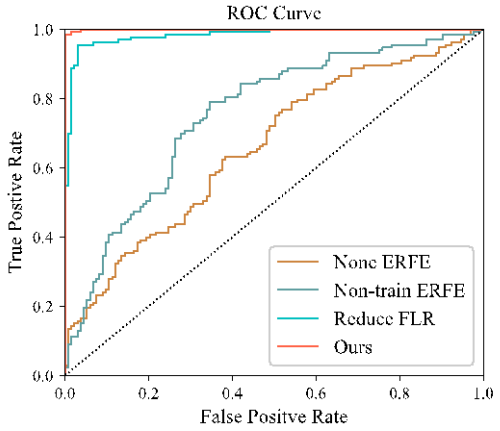

Figure 6. The ROC curve about the performance comparison experiment of ERFE-NET.

1) *Ablation of Equal Resolution Feature Enhancement Network*

To verify the superiority of ERFE-Net, the ways of participating in the comparison include None ERFE, Non-train ERFE, and Reduce FLR. Specifically, None ERFE means that the entire detective network only contains GABD-Net. That is, the original image is directly utilized as the input of GABD-Net. Non-train ERFE means that the training process is not divided into two steps. That is, the network training process is end-to-end. Reduce FLR means that the max-pooling layer is inserted after each convolutional operation of ERFE-Net. At this time, the feature layer of ERFE-Net decreases at twice the speed layer by layer.

2) *Ablation of Gradual Squeeze and Excitation Network*

To evaluate the superiority of GSE-Net, two terms involved in the comparison are introduced, including None Attention and +SE-Net. Among them, None Attention means that the model does not introduce an attention mechanism. +SE-Net refers to replacing GSE-Net with SE-Net.

From the experimental data in Table II, the introduction of GSE-Net improves the overall performance of the network model once again. In particular, the precision achieves a score of 100%, and the F1-score and AUC almost reach the score of a perfect classifier. However, SE-Net did not improve the performance of the network model. It is not difficult to analyze that GSE-Net is more suitable for feature layers with a larger resolution because it adopts a step-by-step extrusion method. Thus, this demonstrates the superiority of the proposed GSE-Net when applied to feature layers with larger spatial dimensions. In conclusion, GSE-Net can notice the features of cracks on feature layers with a larger resolution, thereby improving the model's overall performance.

TABLE II. COMPARATIVE EXPERIMENTAL RESULTS ON THE PERFORMANCE OF GSE-NET

Training	Ac	Pr	Re	F1	AUC
None Attention	97.74%	98.47%	96.99%	97.72%	99.79%
+ SE-Net	97.74%	98.47%	96.99%	97.72%	99.66%
+ GSE-Net	**98.50%**	**100.0%**	**96.99%**	**98.47%**	**99.96%**

C. *Comparative Experiments*

Given that deep learning has achieved good results in the task of detecting cracks, we compare several mainstream classification networks and more advanced methods with the network model proposed in this paper, including mainstream classification networks: resNet50[27] and DenseNet[28], and more advanced methods: Racki-Net[22] and SegDecNet[23]. For fairness, all participating networks were trained with the same choice of settings.

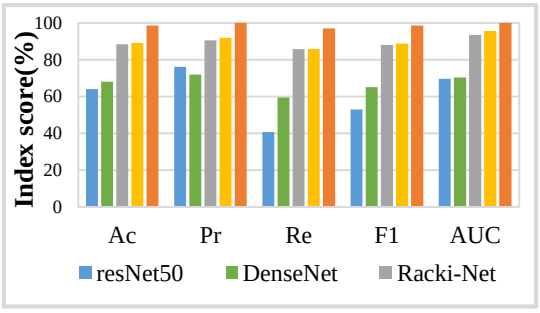

Figure 7. Experimental results comparing other models with ours.

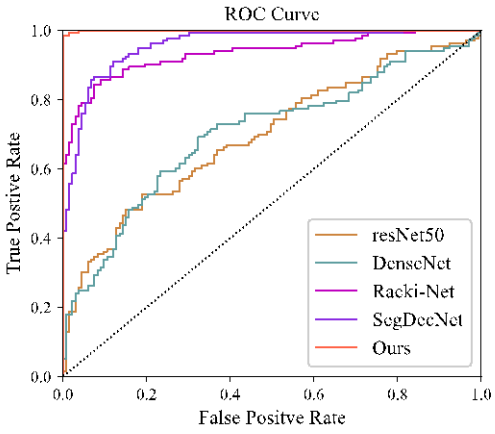

Figure 8. The ROC curve comparing the performance of other models with ours.

The experimental results of the performance comparison between the proposed and other network models are shown in

Fig. 7 and Fig. 8. Among other networks involved in the comparison, Racki-Net and SegDecNet did not reach 90% on the F1-score. The F1-score of the method proposed in this paper is almost 98.5%, about 10% higher than the above two methods. In terms of robustness, the score of 99.96% obtained in the experiments shows that the proposed network model has strong stability. From other performance evaluation indicators, the performance of the proposed network model is also extremely superior. It is easy to analyze that Racki-Net and SegDecNet reduce the resolution of feature layers in feature extraction and enhancement, which loses the features of cracks to a certain extent, while the proposed ERFE-Net reduces the loss of features. Additionally, introducing the gradual attention mechanism further strengthens the features of cracks. Furthermore, resNet50 and DenseNet clearly show poor performance. To sum up, the network model proposed in this paper has outstanding advantages in detecting cracks.

IV. Conclusion

In this paper, a network model combining ERFE-Net and GSE-Net is proposed for tiny target detection problems with weak features accompanied by high noise and a small area ratio, verifying its performance on a real crack dataset. Specifically, ERFE-Net is designed to overcome the problem of extracting and enhancing weak features under high noise because it efficiently extracts features while avoiding the loss of features to a certain extent. Additionally, the full-scale feature layers from ERFE-Net are exploited to improve the utilization of useful information. Meanwhile, GSE-Net is proposed to guide the network model to efficiently search for features of tiny targets in large-resolution feature layers with high noise. The role of GSE-Net in the full-scale feature layer is to strengthen the features of tiny targets further. The experimental results show that the method for detecting tiny targets proposed in this paper has significantly superior performance.

Future work will focus on investigating the generality of the detective method. The detection model is improved on the existing basis, and the applicability of the model is expanded to more tiny targets recognition, providing a powerful detection method for detecting tiny targets on the surface, such as precision instruments and electronic appliances.

Acknowledgment

Research sponsored by Chenguang Program (Grant No.20CG41) which is supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission, and The Fundamental Research Funds for the Central Universities (2232021A-08).

References

[1] H. Zhang, Z. Chen, C. Zhang, J. Xi, and X. Le, "Weld Defect Detection Based on Deep Learning Method," in *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, Aug. 2019, pp. 1574–1579, doi: 10.1109/COASE.2019.8842998.

[2] N. O. Larrosa, R. Akid, and R. A. Ainsworth, "Corrosion-fatigue: a review of damage tolerance models," *Int. Mater. Rev.*, vol. 63, no. 5, pp. 283–308, Jul. 2018, doi: 10.1080/09506608.2017.1375644.

[3] Y. Wang, L. Gao, Y. Gao, X. Li, and L. Gao, "Knowledge Graph-guided Convolutional Neural Network for Surface Defect Recognition," in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, Aug. 2020, pp. 594–599, doi: 10.1109/CASE48305.2020.9216752.

[4] B. Wu, J. Zhou, X. Ji, Y. Yin, and X. Shen, "An ameliorated teaching–learning-based optimization algorithm based study of image segmentation for multilevel thresholding using Kapur's entropy and Otsu's between class variance," *Inf. Sci. (Ny).*, vol. 533, pp. 72–107, Sep. 2020, doi: 10.1016/j.ins.2020.05.033.

[5] C.-M. Chen, S.-W. Zhang, and C.-Y. Hsu, "A sonography image processing system for tumour segmentation," *Enterp. Inf. Syst.*, vol. 14, no. 2, pp. 159–177, Feb. 2020, doi: 10.1080/17517575.2019.1575985.

[6] X. Ni, H. Liu, Z. Ma, C. Wang, and J. Liu, "Detection for Rail Surface Defects via Partitioned Edge Feature," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–17, 2021, doi: 10.1109/TITS.2021.3058635.

[7] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic Road Crack Detection Using Random Structured Forests," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 12, pp. 3434–3445, Dec. 2016, doi: 10.1109/TITS.2016.2552248.

[8] X. Zuo, B. Dai, Y. Shan, J. Shen, C. Hu, and S. Huang, "Classifying cracks at sub-class level in closed circuit television sewer inspection videos," *Autom. Constr.*, vol. 118, p. 103289, Oct. 2020, doi: 10.1016/j.autcon.2020.103289.

[9] W. J. Zhang, G. Yang, Y. Lin, C. Ji, and M. M. Gupta, "On Definition of Deep Learning," in *World Automation Congress Proceedings*, 2018, vol. 2018-June, doi: 10.23919/WAC.2018.8430387.

[10] L. Sun, H. Shi, and M. Bai, "Intelligent oil well identification modelling based on deep learning and neural network," *Enterp. Inf. Syst.*, vol. 16, no. 2, pp. 249–263, Feb. 2022, doi: 10.1080/17517575.2020.1722252.

[11] T. Li, W. Xu, W. Wang, and X. Zhang, "Obstacle detection in a field environment based on a convolutional neural network security," *Enterp. Inf. Syst.*, vol. 16, no. 3, pp. 472–493, Mar. 2022, doi: 10.1080/17517575.2020.1797180.

[12] L. Zhang, F. Yang, Y. Daniel Zhang, and Y. J. Zhu, "Road crack detection using deep convolutional neural network," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sep. 2016, pp. 3708–3712, doi: 10.1109/ICIP.2016.7533052.

[13] Z. Qu, C. Cao, L. Liu, and D.-Y. Zhou, "A Deeply Supervised Convolutional Neural Network for Pavement Crack Detection With Multiscale Feature Fusion," *IEEE Trans. Neural Networks Learn. Syst.*, pp. 1–10, 2021, doi: 10.1109/TNNLS.2021.3062070.

[14] N. H. T. Nguyen, S. Perry, D. Bone, H. T. Le, and T. T. Nguyen, "Two-stage convolutional neural network for road crack detection and segmentation," *Expert Syst. Appl.*, vol. 186, p. 115718, Dec. 2021, doi: 10.1016/j.eswa.2021.115718.

[15] C. Han, T. Ma, J. Huyan, X. Huang, and Y. Zhang, "CrackW-Net: A Novel Pavement Crack Image Segmentation Convolutional Neural Network," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–10, 2021, doi: 10.1109/TITS.2021.3095507.

[16] J. C. P. Cheng and M. Wang, "Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques," *Autom. Constr.*, vol. 95, pp. 155–171, Nov. 2018, doi: 10.1016/j.autcon.2018.08.006.

[17] Y. Tan, R. Cai, J. Li, P. Chen, and M. Wang, "Automatic detection of sewer defects based on improved you only look once algorithm," *Autom. Constr.*, vol. 131, p. 103912, Nov. 2021, doi: 10.1016/j.autcon.2021.103912.

[18] G. Li, X. Ren, W. Qiao, B. Ma, and Y. Li, “Automatic bridge crack identification from concrete surface using ResNeXt with postprocessing,” *Struct. Control Heal. Monit.*, vol. 27, no. 11, p. e2620, Nov. 2020, doi: 10.1002/stc.2620.

[19] F. Wei, G. Yao, Y. Yang, and Y. Sun, “Instance-level recognition and quantification for concrete surface bughole based on deep learning,” *Autom. Constr.*, vol. 107, p. 102920, Nov. 2019, doi: 10.1016/j.autcon.2019.102920.

[20] D. Kang, S. S. Benipal, D. L. Gopal, and Y.-J. Cha, “Hybrid pixel-level concrete crack segmentation and quantification across complex backgrounds using deep learning,” *Autom. Constr.*, vol. 118, p. 103291, Oct. 2020, doi: 10.1016/j.autcon.2020.103291.

[21] H. Chen and H. Lin, “An Effective Hybrid Atrous Convolutional Network for Pixel-Level Crack Detection,” *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021, doi: 10.1109/TIM.2021.3075022.

[22] D. Racki, D. Tomazevic, and D. Skocaj, “A Compact Convolutional Neural Network for Textured Surface Anomaly Detection,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2018, pp. 1331–1339, doi: 10.1109/WACV.2018.00150.

[23] D. Tabernik, S. Šela, J. Skvarč, and D. Skočaj, “Segmentation-based deep-learning approach for surface-defect detection,” *J. Intell. Manuf.*, vol. 31, no. 3, pp. 759–776, Mar. 2020, doi: 10.1007/s10845-019-01476-x.

[24] Y. Xu, Y. Bao, J. Chen, W. Zuo, and H. Li, “Surface fatigue crack identification in steel box girder of bridges by a deep fusion convolutional neural network based on consumer-grade camera images,” *Struct. Heal. Monit.*, vol. 18, no. 3, pp. 653–674, May 2019, doi: 10.1177/1475921718764873.

[25] W. Ma, C. Gong, S. Xu, and X. Zhang, “Multi-scale spatial context-based semantic edge detection,” *Inf. Fusion*, vol. 64, pp. 238–251, Dec. 2020, doi: 10.1016/j.inffus.2020.08.014.

[26] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-Excitation Networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.

[27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.