Activation-Tunable Network for Surface Defect Detection

Xunkuai Zhou[1,2], Xi Chen[2*], Jie Chen[1,3], and Ben M. Chen[2]

***Abstract*— Vision-based defect detection effectively monitors the condition and quality of construction and industrial products. This work presents an accurate detection network augmented by an environmental interaction module and a flexible, tunable activation function. The environmental interaction module is designed to localize and detect defects more accurately, while the flexible activation improves accuracy without increasing parameters. To mitigate information loss after downsampling, we restructure features and introduce a simple deep-global fusion module that integrates deep and global cues to enhance detection performance. Extensive experiments demonstrate the superiority of the proposed network, and real-world tests highlight its portability and practicality. On an edge-computing device, the model achieves real-time inference at 15 FPS, underscoring its suitability for resource-constrained deployment. Furthermore, the proposed activation function enhances the nonlinear representational capacity of neural networks, outperforming 20 widely used activation functions in detection accuracy.**

I. INTRODUCTION

Detecting defects in buildings and industrial products is essential for ensuring safety and maintaining quality control. These structures can develop a variety of defects, such as cracks, corrosion, and stains, which, if left unchecked, can lead to significant damage and financial loss. Traditional inspection methods primarily rely on human visual assessments, which are challenging for inspecting tall structures due to the need for high-altitude operations or specialized imaging equipment [1]. These methods often compromise safety and are prone to inaccuracies resulting from human fatigue and equipment limitations. Therefore, automated defect detection is crucial for ensuring safety, upholding production and environmental standards, and enabling efficient operations and maintenance in large-scale construction and infrastructure management.

In recent years, there have been notable advancements in vision-based defect detection techniques. Yang *et al.* [2] proposed a convolutional neural network (CNN) for defect detection that improves efficiency while maintaining high inspection speeds by integrating EIoU and modification loss functions into YOLOv3. However, this method's inspection speed of 93.5ms per image on an NVIDIA GTX1050Ti GPU makes it unsuitable for edge-computing devices like the Nvidia Orin NX. Similarly, YOLO-M [3] modifies YOLOv3 with an acceleration algorithm and a median flow (MF) algorithm for crack counting, but it is limited by its low processing speed and the narrow scope of defect types it can detect, specifically pavement cracks. The Convolutional Recurrent Reconstructive Network (CRRN) [4] improves defect detection performance by incorporating convolutional spatiotemporal memory (CSTM), with its effectiveness validated on two public datasets. Despite these advancements, current algorithms still face several challenges:

- Using convolutional neural networks (CNNs) for downsampling and feature extraction often results in the loss of important features. Improving feature retention during the propagation process is expected to enhance detection performance.
- Predominantly focusing on target features often overlooks environmental context/cues. Integrating environmental features, as certain targets are intrinsically connected to specific contexts, can improve detection accuracy.
- Limited activation capabilities result in inadequate representation of complex defect data, which contributes to suboptimal detection accuracy.
- Improving detection accuracy often requires adding more parameters or computational cost, but balancing accuracy with memory efficiency is essential for practical applications on memory-constrained devices.

Moreover, Drones' high maneuverability allows them to access areas that are otherwise unreachable by humans. Drone-mounted defect detection systems can improve inspection efficiency and minimize accuracy loss due to human fatigue. Advances in accurate defect detection methods can enhance inspection effectiveness, while faster detection rates improve overall efficiency [5]–[9].

This work aims to develop a defect detection framework that optimizes the balance between parameters and accuracy, making it suitable for edge-computing devices. We introduce a novel detection network, the Environmental Interaction and Activation Representation Network (EARNet), which provides accurate and fast defect detection with minimal parameters and computational cost. To achieve this, we: 1) advocate for using space-to-depth downsampling instead of traditional convolutional layers to ensure complete feature propagation [11]; 2) propose an environmental interaction module to enhance detection performance; and 3) propose the Kernelized Input-Logarithmic Unit activation function

The work was supported in part by the Research Grants Council of Hong Kong SAR under Grant 14217922 (*Corresponding author: Xi Chen*)

[1]School of Electronics and Information Engineering, Tongji University, Jiading, Shanghai 201804, China (e-mail: 2010474@tongji.edu.cn; chenjie206@tongji.edu.cn).

[2]Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: {xunkuaizhou; xichen002; bmchen}@cuhk.edu.hk)

[3]Harbin Institute of Technology, Harbin, 150001, China

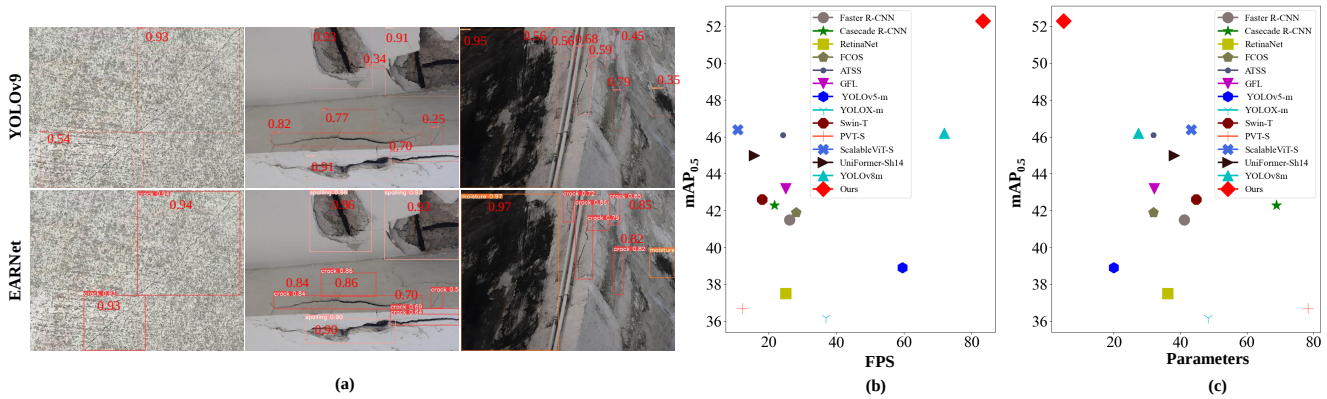


Fig. 1. These illustration demonstrate the superior performance of our method. (a) The first column shows the visualization results from accurate YOLOv9 [10], while the second column displays the results from our method, which detects defects with higher confidence scores. (b) The trade-off between inference speed and accuracy. (c) The trade-off between the number of parameters and accuracy. Please zoom in for a clearer view.

(Kilu), which offers flexible nonlinear representation capabilities through adjustable parameters, thus improving detection performance. As shown in Fig 1(a), EARNet achieves more accurate defect detection compared to YOLOv9 [10]. These visualizations highlight the practical effectiveness of our approach. Fig 1(b) and Fig 1(c) demonstrate that our method achieves higher accuracy with fewer parameters and the fastest speed. Specifically, our method achieves an accuracy of 51.1% with only 4.9M parameters, delivering the fastest inference speed, making it highly competitive for drone-based applications.

In summary, the main contributions of this work are:

1) Introducing environmental interactive information in defect detection research marks a novel development that enhances defect localization and assessment, thereby improving detection performance.
2) A novel, adaptable activation function is proposed to augment the nonlinear representation capabilities of neural networks, thereby enhancing detection accuracy without an increase in parameters.
3) Comprehensive ablation studies illustrate the effectiveness of the proposed strategies. The efficacy of EARNet is validated across three challenging datasets. Deploying EARNet on an edge computing device with 1920×1080 resolution videos confirms its real-time detection capabilities in UAV onboard applications.

II. Methodology

As depicted in Fig. 2, the Environmental Interaction and Activation Representation Network (EARNet) for accurate defect detection comprises two key components: the encoder and the decoder. The encoder incorporates five Convolutional Spatial-to-Depth (CSD) modules (black blocks) and four Contextual Residual Mapping (CRM) modules (light red blocks). The CSD modules perform downsampling while preserving essential convolutional features, whereas the CRM modules enhance environmental interaction. The encoder concludes with the Depth Global Module (DGM), which captures and integrates both depth and global feature dynamics by expanding the receptive field to improve detection accuracy.

The decoder comprises two pathways: bottom-up and top-down. The bottom-up branch includes two upsampling modules and one CRM module, designed to increase feature scale and refine residual features. Conversely, the top-down branch contains two CRM and two CSD modules, which focus on downsampling and feature extraction, with each CSD halving the feature map dimensions. This branch only integrates features from high-level CRM modules to reduce computational costs. The bottom-up pathway strengthens the localization accuracy in lower feature layers, reinforcing the hierarchical structure and minimizing information propagation distance. The variable "N" denotes the number of Contextual Action networks (CA) within each CRM module, and the CSD module is composed of convolutional layers, normalization, the Kilu activation function, and space-to-depth operations.

A. CSD Module

As illustrated in Fig. 3, the CSD module initially extracts convolutional features using a convolutional layer, ensuring that the input and output retain the same feature dimensions. Subsequently, the spatial features are converted into depth features through pixel reorganization. With the application of the CSD operation, the channels in the output increase to four times that of the input, while the spatial dimensions are reduced by half. This process allows for the retention of a larger quantity of feature information while the reduction in spatial dimensions.

B. Proposed CRM Module

Neural networks usually focus on directly extracting features from objects for detection tasks. However, many objects are linked to specific environments. For instance, penguins are native to Antarctica, and distinguishing between turtles and tortoises is easier when considering their habitats—turtles live in water, while tortoises live on land. Similarly, the defects studied here are tied to certain environmental conditions. Moisture is found in humid areas,

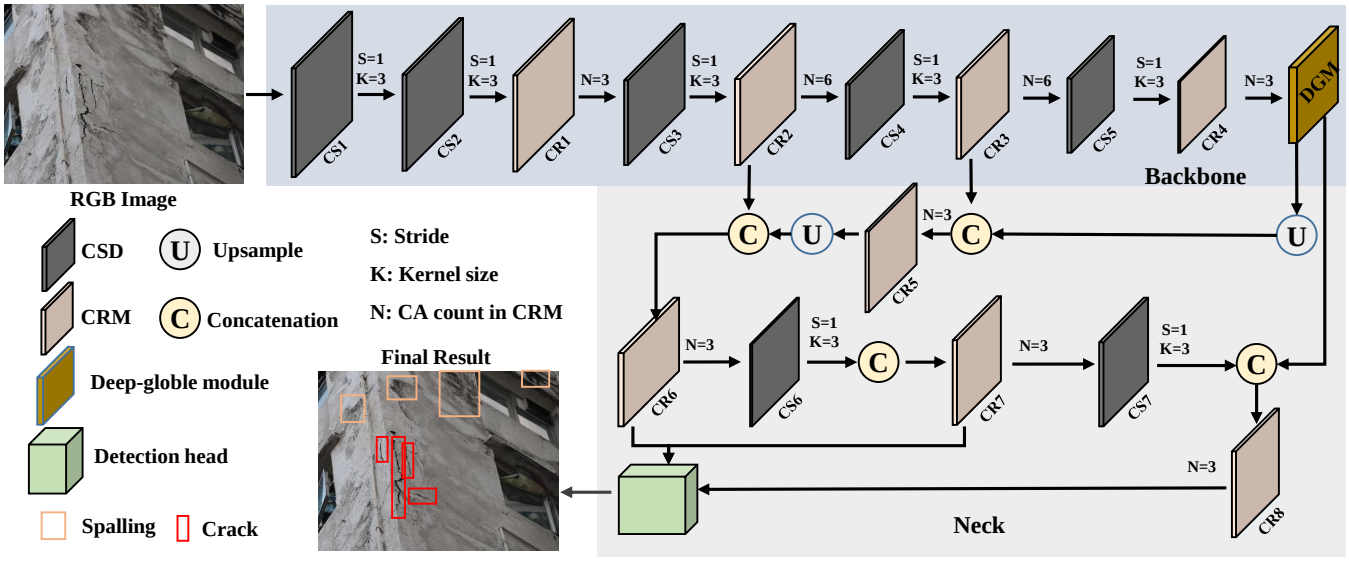


Fig. 2. **Defects detection framework**. First, we perform four rounds of downsampling on the original input image. Then, we utilize DGM to fuse the deep convolutional features with global features. We employ one top-down branch followed by one bottom-top branch and perform feature fusion. Finally, the fused features are passed to the detection head to output the detection results. Please zoom in for the best view.

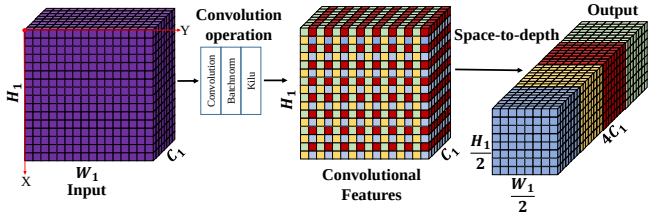


Fig. 3. **Illustration of CSD**. Following the convolution operation and subsequent downsampling, the output dimensions are reduced to half of the input dimensions, while the number of channels increases to four times that of the input.

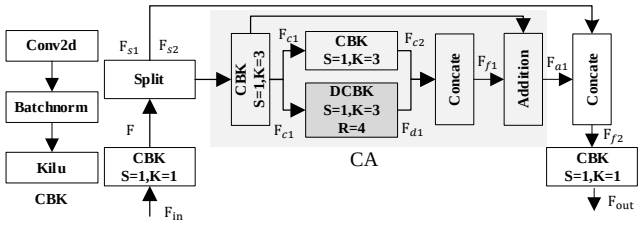


Fig. 4. **Illustration of CRM**. The CA module is capable of multiple serial connections.

and cracks often occur in places with vibrations. Thus, including environmental information helps improve defect classification accuracy.

To do this, we propose a module for integrating environmental information, as shown in Fig. 4. The process starts with feature extraction $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ through a convolutional layer. The features are then split into two parts, $\mathbf{F_{s1}}$, $\mathbf{F_{s2}} \in \mathbb{R}^{C/2\times H\times W}$, which go through the CA module for enhancement. A dilated convolution (gray DCBK block) captures contextual environmental features. After extraction, the features are fused, with $\mathbf{F_{s1}}$ being combined with the enhanced features via a residual connection, resulting in $\mathbf{F_{a1}} \in \mathbb{R}^{C/2\times H\times W}$. This is then merged with $\mathbf{F_{s2}}$ and

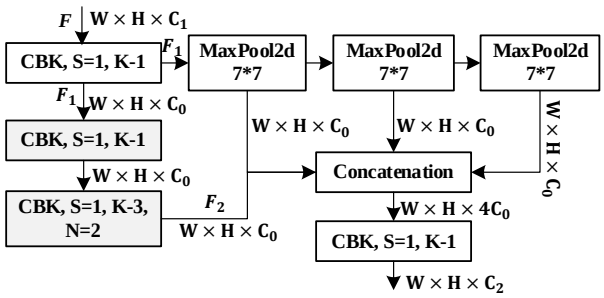


Fig. 5. **The structure of the proposed DGM**. The three-channel global features are fused with the deep convolutional features.

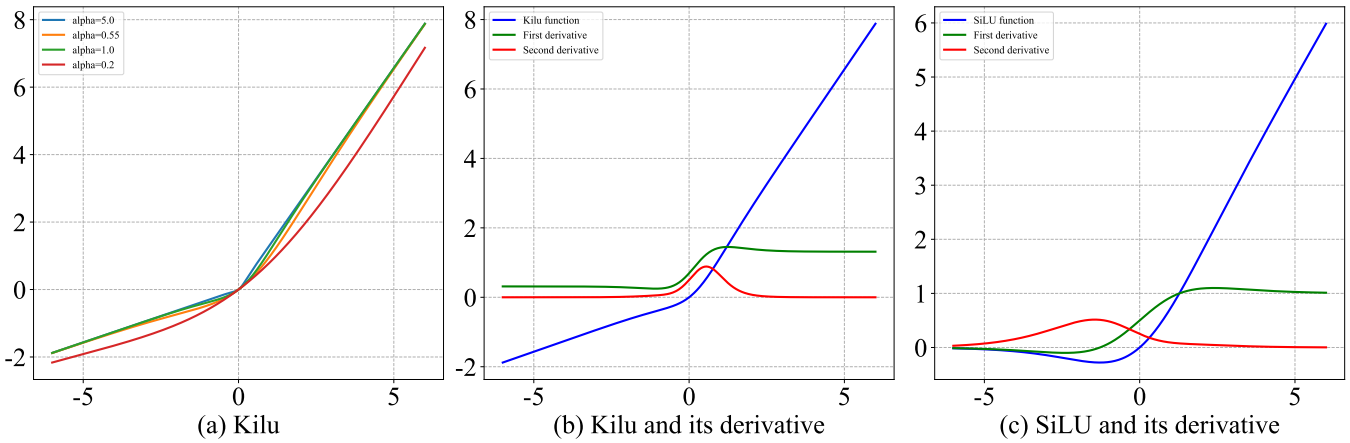
processed through another convolutional layer to produce the final features. When using multiple CA operations, a fusion step ensures the features are refined before the final output.

C. DGM Module

Fig. 5 depicts the Deep Global Module (DGM) architecture. The input feature $\mathbf{F} \in \mathbb{R}^{C_1\times H\times W}$ undergoes an initial convolutional layer, generating the feature $\mathbf{F_1} \in \mathbb{R}^{C_0\times H\times W}$. Subsequently, $\mathbf{F_1}$ is enhanced through three successive 7×7 global pooling layers, each refining the features further. Concurrently, $\mathbf{F_1}$ passes through a separable convolution (indicated by the gray block), resulting in a nuanced feature set $\mathbf{F_2} \in \mathbb{R}^{C_0\times H\times W}$. This feature set $\mathbf{F_2}$ is then merged with the globally pooled features to form a composite feature matrix, which is processed by another convolutional layer to produce the final output feature. This configuration leverages separable convolutions to delve deeper into the feature space efficiently, and the 7×7 pooling size enhances global contextual capture by extending the receptive field.

D. Proposed Kilu Activation Function.

Finding an effective and robust activation function for deep neural networks (DNNs) is challenging, mainly due to



(a) Kilu (b) Kilu and its derivative (c) SiLU and its derivative

Fig. 6. **The graph of the function.** (a) Kilu function's graph for different values of α. (b) The first derivative plot, and the second derivative plot of the Kilu function. (c) The first derivative plot, and second derivative plot of the SiLU function. Please zoom in for the best view.

the saturation properties of traditional functions. Saturation occurs when the derivative of an activation function, $\delta(x)$, approaches zero in both the positive and negative ranges, leading to vanishing gradients. Classic activation functions like *Sigmoid* and *Tanh* are particularly prone to this issue, which often results in poor gradient propagation during training, especially when input values are very large or very small. The introduction of the Rectified Linear Unit (ReLU), defined as $\delta(x) = \max(0, x)$, was a major step forward in activation functions, as it allowed for more efficient training dynamics by mitigating the vanishing gradient problem. However, ReLU is still has its drawbacks, such as the "dying neuron" problem, where neurons become inactive and output zero for any negative input, which hinders gradient flow through these neurons.

To address these challenges, we propose a flexible activation function called Kernelized Input-Logarithmic Unit (*Kilu*). As shown in Fig. 6(b) and Fig. 6(c), similar to the SiLU function used in YOLOv9 [10], *Kilu* exhibits an unbounded upper limit on the right side of the activation curve. The *Kilu* activation function is designed by multiplying the logarithm of the exponential function of *Tanh* with its input x and is defined as:

$$\delta(x) = xlog(1 + e^{tanh(\alpha x)}) \tag{1}$$

where α is the scaling parameter.

The first column in Fig. 6 depicts the graph of *Kilu* function for different values of α. It is observable that the value of α affects the amplitude of the circular arc in the middle. As α increases, the amplitude of the circular arc diminishes.

For substantial positive inputs, the Kilu function exhibits characteristics akin to SiLU, with the output approximating a linear relationship to the input. Distinctively, the Kilu function maintains a linear response even for negative inputs, unlike SiLU and other prevalent activation functions.

The second and third columns of Fig. 6 depict the graphs of the first and second derivatives of the Kilu and SiLU functions, respectively. Analyzing the first derivatives, it is evident that the gradients of our activation functions do not approach zero as they extend towards negative or positive infinity. Notably, the second-order derivative of the proposed Kilu function resembles the negative Laplacian operator, similar to the second-order derivative of the Gaussian operator. This resemblance is advantageous for function maximization.

REFERENCES

[1] K. Li, B. Wang, Y. Tian, and Z. Qi, "Fast and accurate road crack detection based on adaptive cost-sensitive loss function," *IEEE Transactions on Cybernetics*, vol. 53, no. 2, pp. 1051–1062, 2023.

[2] Z. Yang, Z. Xu, and Y. Wang, "Bidirection-Fusion-YOLOv3: An improved method for insulator defect detection using uav image," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–8, 2022.

[3] D. Ma, H. Fang, N. Wang, C. Zhang, J. Dong, and H. Hu, "Automatic detection and counting system for pavement cracks based on pcgan and yolo-mf," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 22 166–22 178, 2022.

[4] Y.-H. Yoo, U.-H. Kim, and J.-H. Kim, "Convolutional recurrent reconstructive network for spatiotemporal anomaly detection in solder paste inspection," *IEEE Transactions on Cybernetics*, vol. 52, no. 6, pp. 4688–4700, 2022.

[5] C. Gao, X. Wang, R. Wang, Z. Zhao, Y. Zhai, X. Chen, and B. M. Chen, "A uav-based explore-then-exploit system for autonomous indoor facility inspection and scene reconstruction," *Automation in Construction*, vol. 148, p. 104753, 2023.

[6] L. Hachemi, M. Guiatni, and A. Nemra, "Fault diagnosis and reconfiguration for mobile robot localization based on multi-sensors data fusion," *Unmanned Systems*, vol. 10, no. 01, pp. 69–91, 2022.

[7] X. Zhou, L. Li, and B. M. Chen, "Lenet: Lightweight and effective detector for aerial object," *Unmanned Systems*, vol. 12, no. 06, pp. 1105–1121, 2024.

[8] X. Zhou, X. Chen, J. Chen, and B. M. Chen, "A low-complexity and high-accuracy defect detection network," *Journal of Systems Science and Complexity*, vol. 38, no. 2, pp. 573–596, 2025.

[9] X. Zhou, G. Yang, Y. Chen, L. Li, and B. M. Chen, "Vdtnet: A high-performance visual network for detecting and tracking of intruding drones," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 8, pp. 9828–9839, 2024.

[10] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," *arXiv preprint arXiv:2402.13616*, 2024.

[11] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new cnn building block for low-resolution images and small objects," in *Machine Learning and Knowledge Discovery in Databases*. Cham: Springer Nature Switzerland, 2023, pp. 443–459.