2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
October 6-10, 2024. Sarawak, Malaysia

An Efficient Token Mixer Model for Sheet Metal Defect Detection

Huijuan Hao[1,2], Sijian Zhu[1,2,†], Changle Yi[1,2], Yu Chen[1,2], Hongge Zhao[1,2], Yue Feng[1,2] and Rong Yang[1,2]

***Abstract*— Defects such as scratches, patches, and cracks frequently occur during sheet metal production. However, the low detection accuracy and slow processing speed of industrial defect detection models significantly impede enterprise production efficiency. The aforementioned issues primarily manifest in three aspects. Firstly, the model complexity and computational overhead are substantial. Secondly, detecting small local defects poses a significant challenge. Thirdly, extracting global features, such as elongated scratches, proves to be difficult. To address these challenges, this paper introduces a novel network architecture called SATRNet. Firstly, within the model backbone, the STR module is devised. Through incorporation of the sparse self-attention method and the CNNs parallel vision Transformer model in the shallow layers, this module significantly enhances the model's capability to extract global features. Secondly, the SCATR module is designed in this paper. By substituting self-attention with the designed SCA soft attention as the token mixer, the module aims to enhance detection accuracy while reducing the number of parameters, thereby fundamentally addressing the problem of model complexity. Finally, this paper presents the GCD bottleneck convolution module. This module combines shallow and deep features, enabling the fusion of more information beneficial for detection, thereby achieving improved efficacy in capturing minute targets. Experiments demonstrate that SATRNet surpasses existing advanced models in detection accuracy on public datasets.**

***Index Terms*— Defect detection, Vision transformer, Token mixer, Soft attention mechanism**

I. Introduction

Surface defect detection on metal sheets is a critical aspect of quality control in industrial settings, essential for ensuring production safety. However, the inherent nature of manufacturing processes inevitably leads to the production of defective items. With the ongoing advancement of defect detection technologies, conventional machine vision techniques have found widespread application in detecting surface defects on metal sheets. Wang et al. [1] introduced a novel method for template creation, leveraging statistical features and sorting operations, achieving an average detection rate of 96.2%. Song et al. [2] employed a scattering convolutional network based on wavelet transform to detect surface defects on hot-rolled strip steel, achieving a detection accuracy of 98.60%. Despite significant improvements over manual inspection, traditional machine vision detection methods still exhibit certain limitations.

[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China.

[2]Shandong Provincial Key Laboratory of Computer Networks, Shandong Fundamental Research Center for Computer Science, Jinan, China. 10431220410@stu.qlu.edu.cn

† Corresponding author

With the rapid evolution of deep learning technology, the surface defect detection algorithm for sheet metal, based on convolutional neural networks [3], has made remarkable strides, consistently enhancing detection efficiency and accuracy. Presently, one-stage detectors, including YOLOv5 [4], YOLOv6 [5], YOLOR [6], PP-YOLOE [7], PicoDet [8], among others, are commonly employed for surface defect detection in metal sheets. Despite the significant achievements in the realm of metal sheet surface defect detection, several challenges persist, such as insufficient model lightweightness, a considerable number of missed small target defects, and low efficiency in extracting global features like slender shapes.

To summarize, the innovations of this paper are as follows:

- A STR module is proposed for extracting global features.The principle is to combine the sparse self-attention mechanism with CNNs, allowing the module to leverage the strengths of each and synergistically enhance performance. In terms of extracting global information, the module demonstrates competitive performance.
- A lightweight SCATR module is proposed, which is based on the principle that the designed soft attention [9] replaces the self-attention in the vision Transformer [10]. This method not only improves detection accuracy but also reduces the number of parameters.
- This paper proposes a GCD bottleneck convolution module for small and other local defects. The principle is to utilize a multi-path feature fusion method and a bottleneck structure design idea to extract shallow and deep local features, thus achieving improved efficacy in capturing small defects.

II. Related Work

A. Convolutional Neural Network

As an early deep learning model, the convolutional neural network has been widely employed in numerous scenarios. In 2016, the bottleneck block proposed by ResNet [11] facilitated the deployment of the model on most hardware platforms and addressed the issue of gradient disappearance caused by increasing depth in deep neural networks, thereby laying a foundation for the development of convolutional neural networks. Building upon this foundation, DenseNet [12] emerged, enhancing feature transmission within the network and reducing the number of parameters by multiplexing the feature maps of each layer. ConvNeXt [13] proposes a pure convolutional model that preserves the simplicity and efficiency of standard CNNs. In the field of sheet metal surface defect detection technology, convolutional neural

2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC) | 978-1-6654-1020-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/SMC54092.2024.10831039

networks are extensively utilized owing to their compact size and ease of deployment. However, for defects on the surface of metal sheets, convolutional networks still cannot meet the requirements of high-precision detection. Therefore, in this paper, we propose new convolution modules to further enhance the detection accuracy.

B. Transformer

Transformer [14] utilizes the self-attention mechanism to achieve parallel computing and global correlation, demonstrating excellent performance.In 2021, the introduction of ViT enabled Transformers to demonstrate competitive performance in vision. However, ViT faces challenges such as high computational complexity and a large number of parameters, prompting researchers to explore these issues. Swin Transformer [15] utilizes hierarchical design and the shifted window method to address the computational complexity problem of self-attention. PvT [16] introduced a progressive pyramid architecture that decreases the sequence length of the Transformer as the network depth increases, thereby significantly reducing the computational overhead. Next-ViT [17] downsamples the spatial dimension before the self-attention operation, which enhances detection accuracy while reducing the number of parameters.However, the issue of significant computational overhead persists, which fails to meet the requirements of metal sheet surface defect detection. Therefore, this paper proposes a sparse attention method to diminish the number of parameters and mitigate computational costs.

C. Hybrid Model

Several studies have demonstrated that integrating the hybrid architecture of other network models and transformers enables the amalgamation of their respective advantages, resulting in enhanced performance [18]. CMT [19] proposed a hybrid network that combines vision Transformer and CNN, leveraging the Transformer for capturing global feature information and the CNN for capturing local features. CvT [20] introduces depth convolution and point convolution before self-attention to enhance performance and efficiency. MetaFormer [21] achieves highly competitive performance by employing pooling as a token mixer for Vits. However, further improvement is needed for these models. In this paper, we propose a new ViT architecture to meet the real-time requirements of industrial inspection.

III. Method

The structure of the proposed SATRNet model is depicted in Fig 1. Firstly, after inputting the image, the convolutional layer is used to adjust the image size and channel, and then the STR module and the lightweight SCATR module are employed to extract feature information. Secondly, following the multi-scale design concept, the GCD module is employed to fuse feature information for extracting small defects. Finally, the feature information is fed into the Detect module for defect detection. Subsequently, detailed descriptions of the three modules, namely STR, SCATR, and GCD, will be provided.

A. STR

Convolutional neural networks effectively capture the local features of images by leveraging the characteristics of local connections and weight sharing. However, traditional convolutional models often struggle to effectively handle global information, such as slender scratches on the surface of metal sheets.Therefore, we introduce the self-attention mechanism of ViT to enhance the extraction of global information. For close-range images like those in the metal sheet dataset, the defect shapes undergo minimal changes and exhibit a single color, resulting in a large number of redundant features during model feature extraction. If the Transformer module is employed for feature extraction, it encounters the same issue, where a large amount of repetitive information is highly likely to interfere with the model's selection of a small amount of other useful information, thereby impacting detection accuracy.

Therefore, we propose the novel sparse self-attention method, and the model structure is depicted as BSA in Fig 1. This structure introduces a bottleneck Token into the multi-head self-attention mechanism of ViT. The bottleneck block [22] is integrated into the Token to extract the most representative and abstract features of the input data through dimensionality reduction.Building upon the aforementioned idea, this paper implements proportional sparse operations on the Q, K, V variables within the Transformer. The sparse factor is determined based on the module's location and the total number of layers in the backbone network, thus regulating the proportion of feature sparsity to reduce the interference of redundant features on the model. However, reducing the feature map to such a small size may lead to the loss of local information, while the Transformer may degrade high-frequency local information to some extent [23], such as local texture and patch details. Therefore, we propose another novel architecture that integrates CNNs and self-attention. The model structure is depicted in the BSA module in Fig 1. It consists of two parallel blocks: a convolution module for extracting local information from the feature map, and a sparse self-attention module for capturing global features. This dual-branch design enables the STR model structure to efficiently capture both high-frequency and low-frequency information, achieving a balance between optimization accuracy and efficiency, thus comprehensively enhancing the detection performance of the model. The formula for the STR module is as follows:

$$\varphi = \begin{cases} \sqrt{\left|\frac{1}{2} - \frac{X_i}{N}\right|} + \frac{1}{2} & 0 \le \left|\frac{1}{2} - \frac{X_i}{N}\right| \le 0.25 \\ 1 & 0.25 < \left|\frac{1}{2} - \frac{X_i}{N}\right| < 0.5 \end{cases} \tag{1}$$

$$Sa = Up\left(At\left(\varphi \cdot X \cdot W^Q, \varphi \cdot X \cdot W^K, \varphi \cdot X \cdot W^V\right)\right) \tag{2}$$

$$\psi = Sa + F^{1\times1}\left(F^{1\times1}\left(gF^{3\times3}\left(X\right)\right)\right) \tag{3}$$

$$\mu = X + \psi + MLP\left(X + \psi\right) \tag{4}$$

In the above formula, φ represents the sparsity factor, N is a constant derived from the total number of backbone layers, and X_i denotes the number of layers where the position to be

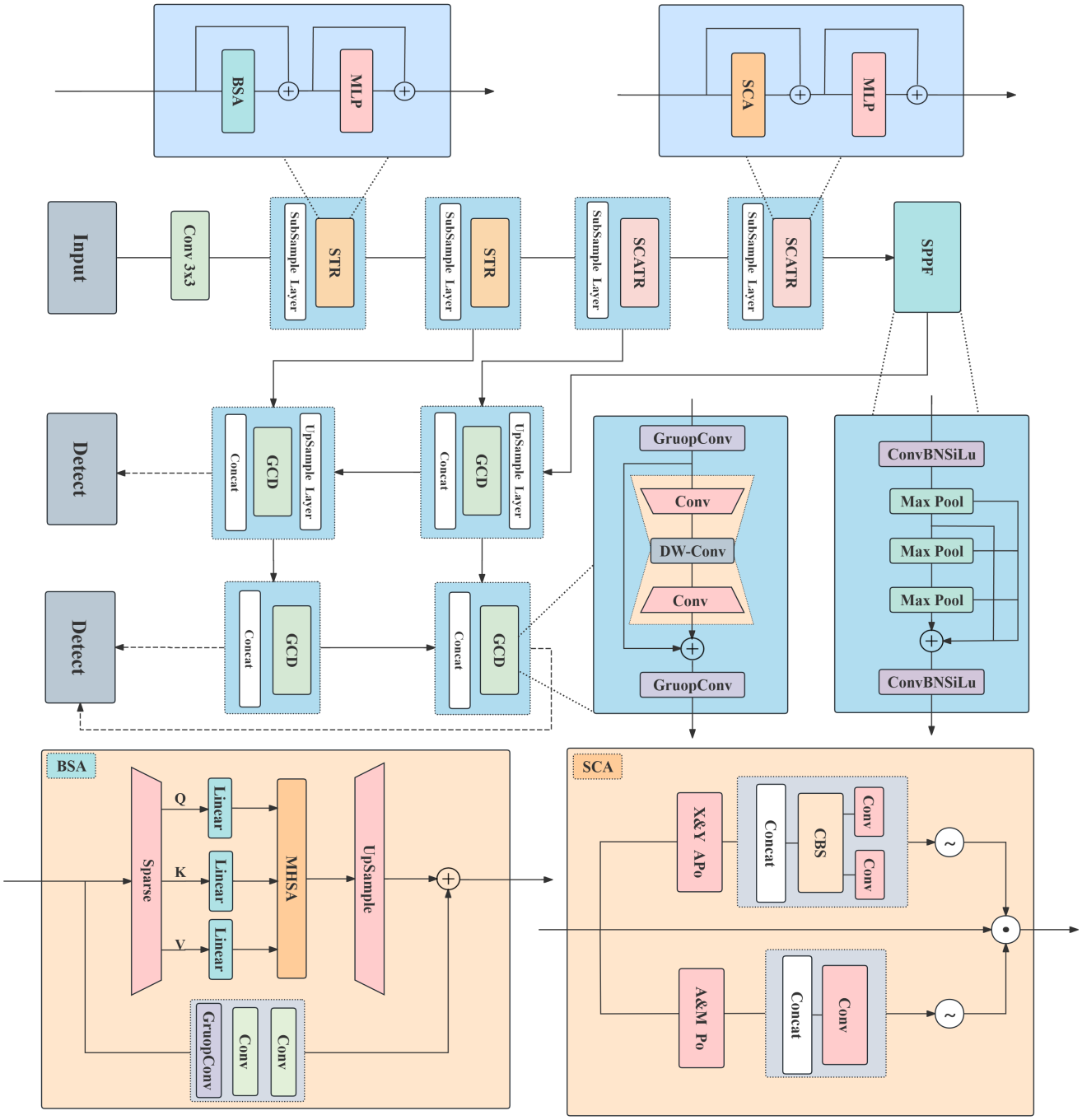

Fig. 1. Structure Diagram of the SATRNet Model.

sparsified is located. X represents the feature map representing the input, and W^Q, W^K, and W^V are the learned weight matrices. At represents the standard self-attention with the formula $At\,(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$. Where $\frac{QK^T}{\sqrt{d_k}}$ is the attention score obtained by dividing the dot product of the query vector and the key vector by the scaling factor $\sqrt{d_k}$, where d_k is the dimension of the key vector, and the $softmax$ function converts the attention score to be between 0 and 1. Up stands for the upsampling operation, and Sa represents the output result of sparse self-attention. $gF^{3\times3}$ denotes grouped convolution with 3×3 kernels, $F^{1\times1}$ represents standard convolution, ψ represents the output of the BSA model, and μ represents the output of the STR model.

B. SCATR

To showcase the superiority of the proposed SCATR, we review several classical architecture designs based on ViT enhancements, as depicted in Fig 2. PoolFormer replaces the multi-head self-attention mechanism in ViT with pooling,

and the model achieves competitive performance, demonstrating that ViT's performance is influenced by both its architecture and token mixer design. Next-ViT employs a convolutional neural network model to replace the multi-head self-attention mechanism, thereby enhancing the capability to extract local features while reducing the number of parameters. While the hybrid model demonstrates satisfactory performance currently, it falls short of meeting the real-time detection demands in practical industrial scenarios.

To further enhance the model's performance, we devised the SCATR model structure, depicted in Fig 1. The model structure introduces a novel soft attention mechanism, SCA, to replace the multi-head self-attention mechanism in ViT. Additionally, the model adheres to the general architecture of MetaFormer. The MLP layer maps original features into a new high-dimensional feature space through two fully connected layers, enhancing its capability to capture nonlinear relationships between features. Our design of SCA soft attention is flexible and lightweight, capable of capturing long-distance dependencies within a single channel while preserving precise location and spatial information. We utilize SCA soft attention as a token mixer to address the challenges posed by the relative complexity and heavy computational burden of self-attention. This approach significantly improves the accuracy of model detection while ensuring that the model maintains lightweight performance to meet the real-time requirements of devices with relatively low computing power. The formula for implementing the SCATR module is as follows:

$$F^h, F^w = S\left(CBS\left(T\left(GAP^h\left(X\right), GAP^w\left(X\right)\right)\right)\right) \quad (5)$$

$$S^h, S^w = \delta\left(Conv^{1\times1}\left(F^h, F^w\right)\right) \quad (6)$$

$$M = \delta\left(Conv^{7\times7}\left(T\left(F_{max}\left(X\right), F_{avg}\left(X\right)\right)\right)\right) \quad (7)$$

$$Y = X \times S^h \times S^w \times M \quad (8)$$

$$\eta = X + (X \times Y) + MLP\left(X + (X \times Y)\right) \quad (9)$$

X represents the input feature map, GAP^h and GAP^w denote the global average pooling of the feature maps in the width and height dimensions respectively, T represents the concatenation operation, CBS represents the convolution, normalization, and activation function δ, S represents the split operation, and F^h, F^w represent the feature maps obtained in both the width and height directions. $Conv^{1\times1}$ represents convolution with a 1×1 kernel, and S^h, S^w represent the attention weights of the feature map in the height and width dimensions, respectively. F_{max} and F_{avg} stand for average pooling and max pooling of input feature maps, respectively. $Conv^{7\times7}$ represents convolution with a 7×7 kernel, and M stands for the output spatial attention weights. Y represents the attention weight of the soft attention SCA output. MLP stands for a fully connected layer, and η represents the output of the SCATR model structure.

C. GCD

In the metal sheet dataset, numerous small target defects are prevalent, yet they are frequently overlooked during the detection process, leading to diminished detection accuracy. Strengthening the model's capability to detect small objects is imperative for enhancing both accuracy and efficiency. Therefore, we have devised a novel bottleneck convolutional structure named GCD, and the corresponding model architecture is depicted as the GCD module in Fig 1. Firstly, we process the input data through grouped convolution with a 3x3 convolution kernel, aiming to reduce the number of parameters and enhance the learning and generalization abilities of the model. Secondly, drawing inspiration from the bottleneck structure design, we utilize standard convolution for cross-channel information interaction, reducing and restoring the dimensionality of the input feature map. Additionally, we employ depthwise separable convolution with a 3x3 kernel to extract spatial feature information, enabling the model to better capture image features across both space and channels. Finally, we applied residual connections to GCD structures to enhance performance, improve gradient propagation across layers, and capture multi-scale features. Experimental results verify the effectiveness of this structural modification.

In the SATRNet model, the GCD module can also perform the fusion of different features, merging the shallow features extracted by the STR module and the deep features extracted by the SCATR module, thereby incorporating more information beneficial for detection. The GCD module effectively enhances the model's ability to detect small targets and further optimizes the detection results. The formula is provided below.

$$\rho = F^{1\times1}\left(dF^{3\times3}\left(F^{1\times1}\left(gF^{3\times3}\left(X\right)\right)\right)\right) \quad (10)$$

$$\omega = gF^{3\times3}\left(gF^{3\times3}\left(X\right) + \rho\right) \quad (11)$$

X is the input feature map, $gF^{3\times3}$ represents grouped convolution with a 3×3 kernel, $F^{1\times1}$ represents standard convolution with a 1x1 kernel, $dF^{3\times3}$ is depthwise separable convolution with a 3x3 kernel, and ω represents the output of the GCD module.

IV. Experiments

A. DataSet

We utilize the publicly available NEU-DET dataset for steel surface defects and an industrial-grade aluminum surface defect detection dataset as our research data sources. The steel dataset provided by Northeastern University comprises 1800 grayscale images, representing six typical defects, with 300 samples for each defect, such as patches and scratches. The industrial-grade aluminum surface defect detection dataset consists of a total of 400 grayscale images, featuring four typical defects: pinholes, scratches, wrinkles, and dirt.

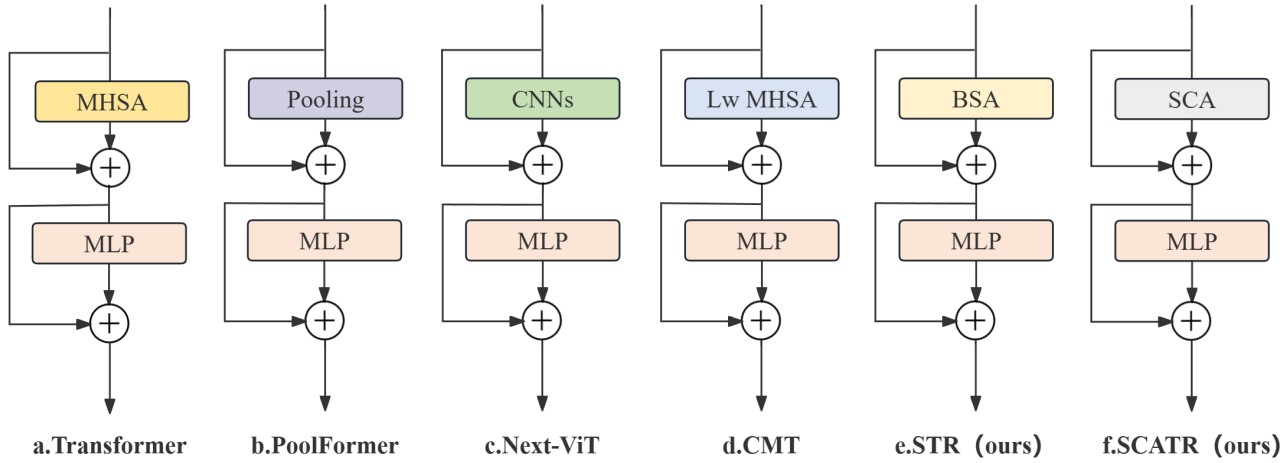

Fig. 2. Comparing Different Transformer Models.

B. Device

The experiments were conducted on a Windows system with PyTorch 2.2.2, CUDA 12.1, and cuDNN 8.8 configurations. The GPU utilized was an NVIDIA GeForce RTX 4060. We employed the SGD optimizer with an initial learning rate of 0.01.

C. Comparison and ablation experiments

Table I presents the comparative experiments of our proposed model, SATRNet, with existing models on the NEU-DET dataset. Currently, mainstream detection models such as YOLOv8, PPYOLO, and PicoDet have demonstrated high performance in the field of defect detection. Building upon this foundation, our model achieves further improvements. Experimental results indicate a respective increase of 4.6% and 4.5% in detection accuracy and recall compared to YOLOv8, and an increase of 3.2% and 4.4% in mAP@0.5 and mAP@0.5:0.95. Relative to mainstream industrial models like PPYOLO and PicoNet, our model exhibits significant enhancements in detection accuracy, recall, and mean average precision.

TABLE I

COMPARATIVE EXPERIMENTS OF SATRNET AND OTHER MODELS ON THE NEU-DET DATASET ARE PRESENTED.

Detection Model	Detection Result			
	Precision	*Recall*	*mAP@.5*	*mAP@.5-.95*
YOLOv3	76.2%	74.8%	77.4%	40.1%
YOLOv4	79.1%	75.6%	78.3%	40.5%
YOLOv5s	78.0%	76.2%	78.1%	41.7%
YOLOv7-T	74.4%	72.3%	73.3%	39.5%
YOLOv8	78.6%	78.1%	78.5%	40.2%
YOLOR-P6	75.7%	72.2%	71.0%	38.4%
PPYOLOE-s	78.5%	74.8%	77.6%	42.3%
PicoDet	78.3%	78.6%	78.1%	42.1%
SATRNet	**83.2%**	**82.6%**	**81.7%**	**44.6%**

Table II presents a comparative analysis of our proposed SATRNet model with existing methodologies on an industrial-grade aluminum surface defect detection dataset. Our experimental findings reveal the superior performance of our model over PPYOLO and PicoDet, demonstrating improvements of 4.7% and 5.0% in detection accuracy and recall, respectively, compared to PPYOLO, and 3.0% and 3.3%, respectively, compared to PicoDet. Additionally, our model achieves enhancements of 5.6% and 4.6% in mAP@0.5 and mAP@0.5:0.95, respectively, compared to PPYOLO, and 2.9% and 3.5%, respectively, compared to PicoDet. Notably, our model surpasses mainstream industrial detection models on publicly available aluminum datasets.

TABLE II

COMPARATIVE EXPERIMENTS BETWEEN SATRNET AND OTHER MODELS ON THE INDUSTRIAL-GRADE ALUMINUM DATASET ARE PRESENTED.

Detection Model	Detection Result			
	Precision	*Recall*	*mAP@.5*	*mAP@.5-.95*
YOLOv3	82.4%	85.3%	85.1%	46.1%
YOLOv4	77.2%	69.8%	71.3%	39.6%
YOLOv5s	84.2%	85.4%	83.0%	44.1%
YOLOv7-T	78.6%	77.3%	74.9%	40.8%
YOLOv8	85.9%	84.8%	85.5%	43.1%
YOLOR-P6	82.0%	81.6%	80.2%	42.3%
PPYOLOE-s	84.2%	84.5%	82.7%	45.0%
PicoDet	85.9%	86.2%	85.4%	46.1%
SATRNet	**88.9%**	**89.5%**	**88.3%**	**49.6%**

Table III presents ablation experiments evaluating various enhancement methods based on YOLOv5s using the NEU-DET dataset. The results demonstrate significant accuracy improvement achieved by the STR method, with a detection accuracy of 79.3%. The SCATR method reduces parameter count while maintaining detection accuracy. The GCD method exhibits outstanding performance in recall rate and mAP@0.5. Overall, the model's detection accuracy increases from 75.7% to 83.5%, the recall rate increases from 74.8% to 82.4%, and the mAP@0.5 and mAP@0.5:0.95 improve by 2.4% and 3.6% respectively compared to the baseline. Additionally, the parameter count decreases from 7.02M to

6.98M.

TABLE III

ABLATION EXPERIMENTS WERE CONDUCTED ON THE NEU-DET DATASET TO ASSESS THE IMPROVED MODULE

Detection Model	Detection Result				
	Precision	*Recall*	*mAP@.5*	*mAP@.5-.95*	*Param*
BaseLine	75.7%	74.8%	79.5%	41.1%	7.02M
STR	79.3%	77.1%	80.8%	42.3%	6.94M
SCATR	78.4%	78.2%	79.5%	42.9%	6.85M
GCD	76.8%	78.0%	80.4%	42.4%	6.93M
STR+SCATR	81.4%	79.6%	81.7%	43.3%	6.96M
All	**83.5%**	**82.4%**	**81.9%**	**44.7%**	**6.98M**

Based on experiments, our model has demonstrated competitive performance in the field of industrial inspection. To present the detection results on the steel and aluminum datasets more clearly, we provide a heatmap of the detection outcomes, as depicted in Fig 3.

V. CONCLSION

In this paper, we introduce the STR structure featuring parallel CNNs and sparse self-attention, alongside the lightweight SCATR structure, with the objective of notably enhancing the model's detection accuracy and efficiency. To address the challenge of detecting numerous small target defects effectively, we propose the GCD module, which refines the model's focus on these targets through network deepening and multi-feature fusion operations. Experimental results affirm the superior performance of our model.

ACKNOWLEDGMENT

This paper gratefully acknowledges the support received from six research projects: 'Exploration of Intelligent Fault Diagnosis Methods for Analyzing Big Data from Mechanical Equipment on Industrial Internet Platforms (ZR2022MF279),' 'Research and Application of Key Technologies for Intelligent Control and Information Platforms in Agricultural Machinery (2023TSGC0587),' 'Integration Technology and Demonstration of Networked Collaborative Manufacturing for Textile and Clothing Enterprises (2021GXRC074),' 'Classroom Teaching Evaluation System Based on Intelligent Recognition of Human Body Posture and Gestures (2022TATSGC022),' 'Research on Intelligent High-Precision Industrial Vision Defect Detection Technology for Steel Materials (S202310431041),' and 'Key Technologies Research and Application of Intelligent Control and Informatization Platform for Agricultural Machinery (2023TATSGC043).'

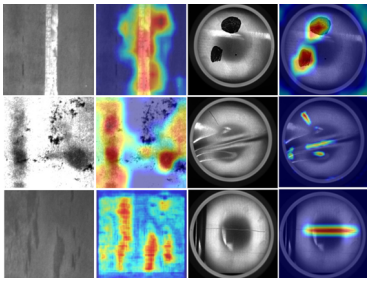

Fig. 3. The heat map illustrates the detection results.

REFERENCES

[1] H. Wang, J. Zhang, Y. Tian, et al., "A simple guidance template-based defect detection method for strip steel surfaces," IEEE Transactions on Industrial Informatics, vol. 15, no. 5, pp. 2798-2809, 2018.
[2] K. Song, S. Hu, and Y. Yan, "Automatic recognition of surface defects on hot-rolled steel strip using scattering convolution network," J. Comput. Inf. Syst., vol. 10, no. 7, pp. 3049-3055, 2014.
[3] R. Girshick, J. Donahue, T. Darrell, et al., "Region-based convolutional networks for accurate object detection and segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 1, pp. 142-158, 2015.
[4] G. Jocher, A. Chaurasia, A. Stoken, et al., "ultralytics/yolov5: v6.2-yolov5 classification models, apple m1, reproducibility, clearml and deci.ai integrations," Zenodo, 2022.
[5] C. Li, L. Li, H. Jiang, et al., "YOLOv6: A single-stage object detection framework for industrial applications," arXiv preprint arXiv:2209.02976, 2022.
[6] C. Y. Wang, I. H. Yeh, H. Y. M. Liao, "You only learn one representation: Unified network for multiple tasks," arXiv preprint arXiv:2105.04206, 2021.
[7] S. Xu, X. Wang, W. Lv, et al., "PP-YOLOE: An evolved version of YOLO," arXiv preprint arXiv:2203.16250, 2022.
[8] G. Yu, Q. Chang, W. Lv, et al., "PP-PicoDet: A better real-time object detector on mobile devices," arXiv preprint arXiv:2111.00902, 2021.
[9] S. Woo, J. Park, J. Y. Lee, et al., "CBAM: Convolutional block attention module," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-19, 2018.
[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[11] K. He, X. Zhang, S. Ren, et al., "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.
[12] G. Huang, Z. Liu, L. Van Der Maaten, et al., "Densely connected convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700-4708, 2017.
[13] Z. Liu, H. Mao, C. Y. Wu, et al., "A convnet for the 2020s," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976-11986, 2022.
[14] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, pp. 5998-6008, 2017.
[15] Z. Liu, Y. Lin, Y. Cao, et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012-10022, 2021.
[16] W. Wang, E. Xie, X. Li, et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568-578, 2021.
[17] J. Li, X. Xia, W. Li, et al., "Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios," arXiv preprint arXiv:2207.05501, 2022.
[18] Z. Dai, H. Liu, Q. V. Le, et al., "Coatnet: Marrying convolution and attention for all data sizes," Advances in Neural Information Processing Systems, vol. 34, pp. 3965-3977, 2021.
[19] J. Guo, K. Han, H. Wu, et al., "CMT: Convolutional neural networks meet vision transformers," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12175-12185, 2022.
[20] H. Wu, B. Xiao, N. Codella, et al., "CVT: Introducing convolutions to vision transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22-31, 2021.
[21] W. Yu, M. Luo, P. Zhou, et al., "Metaformer is actually what you need for vision," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10819-10829, 2022.
[22] A. Srinivas, T. Y. Lin, N. Parmar, et al., "Bottleneck transformers for visual recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16519-16529, 2021.
[23] R. Yang, H. Ma, J. Wu, et al., "Scalablevit: Rethinking the context-oriented generalization of vision transformer," in European Conference on Computer Vision, Springer Nature Switzerland, pp. 480-496, 2022.