

# CMTSiMBA-light: Industrial Defect Classification and Detection Based on Deep Learning

Wencheng Ding<sup>1</sup> and Shan Chang<sup>1</sup> and Tianhui Cai<sup>1</sup> and Hongya Wang<sup>1</sup>

<sup>1</sup>School of Information and Intelligent Science, Donghua University, Shanghai 201620, China  
18817323219@163.com

**Abstract.** Surface quality inspection of industrial products is a crucial step in ensuring production line efficiency and product quality. In response to the problem of insufficient fusion of local and global features in image classification algorithms in complex industrial scenarios and the difficulty of a single architecture to balance texture details and long-range dependencies, this paper proposes and optimizes the relevant algorithm CMTSiMBA-light. Experiments show that the model improves accuracy by 0.7% and 1.4% respectively on the NEU-CLS and FSC-20 datasets compared to CMT, reduces parameters by 10.8%, lowers FLOPs by 5%, and increases throughput by 5.2%.

**Keywords:** Industrial Classification, Industrial Detection, Deep Learning.

## 1 Introduction

In the industrial field, image classification methods have undergone a qualitative change from traditional pattern recognition to deep learning, eventually forming a technical system [错误!未找到引用源。](#)

The core advantages of pure CNNs stem from their local perception mechanisms and hierarchical feature extraction capabilities. However, the local receptive field and fixed weight mechanism of CNNs limit their global modeling capabilities. Transformer network architectures exhibit strong global modeling capabilities and efficient parallelism, but they also incur high computational complexity due to their global modeling paradigm.

However, in terms of industrial defect classification, current image classification algorithms have deficiencies in the fusion of local and global features. First, in industrial scenarios, there are numerous interfering factors, which makes it difficult to effectively capture and fuse the features of small-sized defects

Secondly, traditional Transformer [错误!不能识别的开关参数。](#) models have exposed many problems in the application of industrial defect classification. The computational complexity of its self-attention mechanism increases in squares with the sequence length, which greatly reduces the efficiency of model training and inference, making it difficult to meet the demand for fast processing of large amounts of data in actual industrial production. Moreover, when long sequences are input, the model

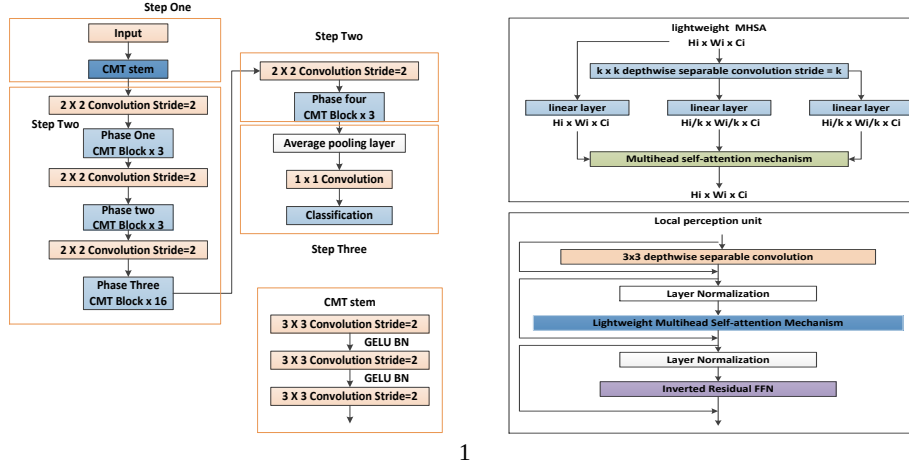
throughput is low, failing to meet the high concurrency and low latency real-time standards required by industrial equipment, which restricts the efficient operation of the production line.

The primary contributions are as follows (1)We enhance feature extraction capabilities by incorporating the SENet convolution module into the CMT Stem , and we apply low-rank decomposition to the linear layer of the lightweight MHSA to decrease parameters and computational costs. (2)We replace the traditional convolution with the PConv convolution module in the SiMBA stem to lower computational costs while preserving detailed information .

## 2 Related Work

### 2.1 CMT Model

In the CMT block structure, as shown in **Fig. 1**, it can be seen that feature extraction is achieved by three layers of convolution. However, since the channel modeling of traditional CNNs ~~错误!不能识别的开关参数。~~ is channel feature homogenization, when the input channels contain high-frequency edge information and low-frequency texture information respectively, the model cannot perform differentiated feature fusion on different channels, resulting in reduced accuracy of defect classification. And although lightweight MHSA models have improved efficiency to some extent, with the increasing complexity of industrial scenarios and the sharp increase in data volume, their linear layers face issue of large parameters and significant computational demands, requiring further reduction of both parameters and computational load.



1

**Fig. 1.** Structure Diagram of the CMT Model and the CMT Block2

## 2.2 SiMBA Model

Due to the computational redundancy and high memory consumption caused by the three layers of convolution in SiMBA, and the multiple layers of convolution stacking resulting in the smoothing of high-frequency details (such as edges, textures), especially in the middle layer (the second layer), abstract features gradually take over, and the original input information is severely lost.

The conventional Transformer model encounters two main challenges. The one is that the computational complexity of its self - attention mechanism grows quadratically with the sequence length, leading to decreased efficiency during model training and inference. The other is that the model ' s throughput is limited with long sequence inputs, making it challenging to satisfy the high concurrency and low-latency real-time requirements of industrial devices.

## 2.3 Mamba Network Architecture

Mamba achieves linear complexity through SSM ' s recursive structure ; Flash Attention optimizes parallel computing and memory access; Dynamic weighting mechanism adjusts parameters based on input to enhance time series modeling. Combining the long-term dependency capabilities of recursion with parallel training, taking into account both efficiency and long context modeling, to break through the computational bottleneck of the Transformer. Among them, the SSM state space model is the foundation of Mamba.

Combining the base block of most SSM architectures such as H3 with the gated MLP that is common in modern neural networks such as transformer to form a new Mamba block, repeating this block and combining it with normalization and residual connections, the Mamba architecture is formed Fig. 2 shows that.

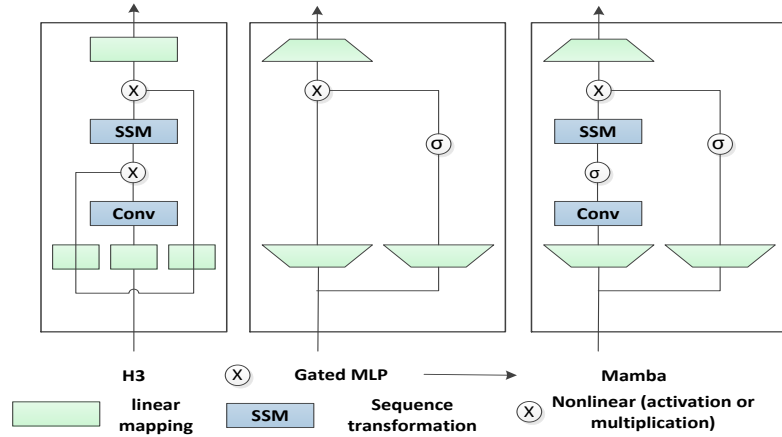


Fig. 2. Mamba Network Architecture

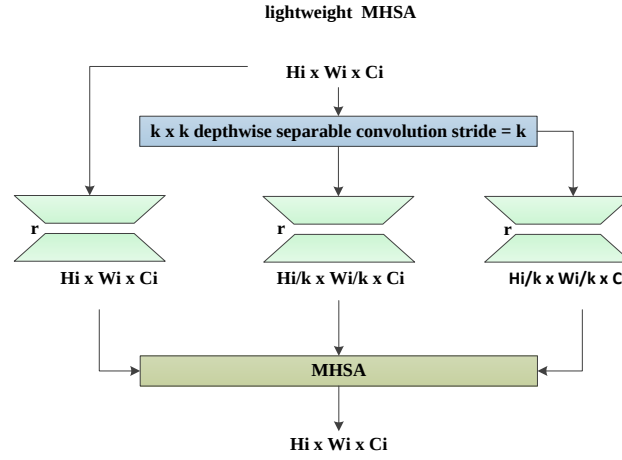
Among them, the H3 architecture is mainly composed of SSM and convolutional layers, and has the characteristics of efficient processing of long sequences. Gated MLP is another core component of Mamba, used to enhance the network's nonlinear expressibility, mainly composed of MLP structure and gating mechanism.

### 3 Methods

#### 3.1 Hybrid Model CMT-light based on Channel Attention Mechanism and Low-Rank Decomposition

The CMT - SEmodule is largely made of two  $3 \times 3$  convolutional layers and one SE-Module attention module. Each convolutional layer has a step size of 2 and is used to reduce the spatial dimension of the feature map.

As shown in **Fig. 3**, we use  $k \times k$  deep convolution and  $\text{step} = k$  to reduce the space size of K and V before MHSA. 错误!未找到引用源。 The input feature map first passes through a  $k \times k$  DWConv layer, an operation that takes a  $\text{stride} = k$  step size and reduces the spatial resolution of the feature map to  $1/k$  (i.e.  $H_i/k \times W_i/k$ ) of the original size.  $q, k, v$  also passes through a layer of convolution, with the aim of lightening the model while maintaining the effectiveness of the attention mechanism.



#### 3.2 Add the SENet Convolution Enhancement Module to CMT Stem

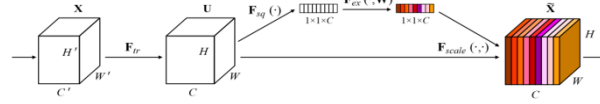
Because the channel modeling flaws of conventional CNNs are mainly manifested in channel feature homogenization processing. The mathematical expression of the conventional convolution operation can be defined as (1) :

$$Y_c = \sum_{i=1}^{C_{in}} X_i * K_c, i + bc(i \leq c \leq C_{out}) \quad (1)$$

$$Y_c = \sum_{i=1}^{C_{in}} X_i * K_c, i + bc(i \leq c \leq C_{out}) \quad (1)$$

Where:  $i$  is the channel of the input feature map  $X_i \in R^{H \times W}$ ,  $C_{in}, C_{out}$  is the number of input and output channels.

To address the channel flaws of traditional CNNs, this chapter introduces the SE-Module convolution enhancement module in CMT Stem, mainly to strengthen the responsiveness of key channels during the initial feature extraction stage of the network, thereby enhancing the model's sensitivity and expressiveness to industrial defect features. SEmodule dynamically adjusts the channel weights 错误!不能识别的开关参数。错误!未找到引用源。



**Fig. 4.** SEmodule Structure

Three steps make up the core operation:

**Squeeze operation**, which is a global average pooling of the input feature layer, compressing the feature of each channel into a global feature value to obtain a channel descriptor.

**Excitation operation**: In order to take advantage of the aggregated information, the SEmodule learns the weights of each channel through two layers of fully connected networks.

**Feature reweighting**: Once the weight values for each channel are obtained, the SE-Module multiplies these weights by the corresponding channels of the original input feature layer, effectively reweighting the features.

### 3.3 Lightweight MHSA based on Low-Rank Decomposition

The self-attention module in the original MHSA can be represented as (2) 错误!不能识别的开关参数。错误!未找到引用源。

$$Attn(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

$Q \in R^{n \times d_k}, K \in R^{n \times d_k}, V \in R^{n \times d_v}$  is obtained by the linear variation  $n = H \times W$  of the input  $X \in R^{n \times d}$ . In the formula,  $Q$  and  $K$  require a dot product operation  $QK^T$ . Before the

attention operation, CMT uses  $k \times k$  depth-separable convolution with a step size of  $k$  to reduce the spatial dimensions of the key ( $\kappa$ ) and value ( $v$ ), and the formula can be

$$\text{LightAttn}(Q, K, V) = \text{Softmax}\left(\frac{QK'^T}{\sqrt{d_k}}\right)V' \quad (3)$$

expressed as (3)

找到引用源。

$$\text{LightAttn}(Q, K, V) = \text{Softmax}\left(\frac{QK'^T}{\sqrt{d_k}}\right)V' \quad (3)$$

$Q \in R^{n \times d_k}, K' = \text{DWConv}(K) \in R^{\frac{n}{k^2} \times d_k}, V' = \text{DWConv}(V) \in R^{\frac{n}{k^2} \times d_v}$ . In the formula, the dot product operation has a computational complexity of  $O(\frac{n^2}{k^2} \times d_k)$ , which reduces the computational complexity. In the original linear layer, take the Q matrix as an example:

$$Q' = Q \cdot W_Q + b_Q \quad (4)$$

$W_Q \in R^{d_k \times qk\_dim}, b_Q \in R^{qk\_dim}$ . The number of parameters is  $d_k \times qk\_dim$ . Because the low-rank decomposition decomposes the original weight matrix into two low-rank matrices  $U \in R^{d_k \times r}$  and  $V \in R^{r \times qk\_dim}$ , satisfies  $W_Q \approx UV$ , the number of parameters is  $d_k \times r + r \times qk\_dim$ . The model performance is at its best when  $r = 80$  the experimental results on the basis of SEModule are demonstrated, as shown in Table 1.

Comparison of rank results

**Table 1.** Comparison of rank results

| rank | Params/ $10^6$ | FLOPs/ $10^9$ | Acc@1/% |
|------|----------------|---------------|---------|
| 32   | 20.5           | 3.5           | 97.5%   |
| 48   | 21.1           | 3.6           | 99.2%   |
| 64   | 21.7           | 3.7           | 99.3%   |
| 80   | 22.3           | 3.8           | 100%    |
| 96   | 22.9           | 3.9           | 99.2%   |

At that time  $r = 80$ , the model reached its highest accuracy (100%), the parameter count decreased by 10.8%, and the FLOPs decreased by 5%.

### 3.4 Research on Image Classification Algorithms Based on Linear Attention Mechanism

As shown in **Fig. 5**, firstly, the model input data passes through the stem-PC module; Then through the Mamba module: responsible for deep modeling of global features; Through EinFFT module: Complementary to Mamba, for the first time, bi-directional

feature fusion in the spatio-spatial and frequency-domain in the model. And the regularization effect was enhanced by three Dropout layers (embedded in Mamba and EinFFT modules respectively) to significantly suppress the risk of overfitting. The overall design takes into account the efficient interaction of local and global while reducing computational complexity.

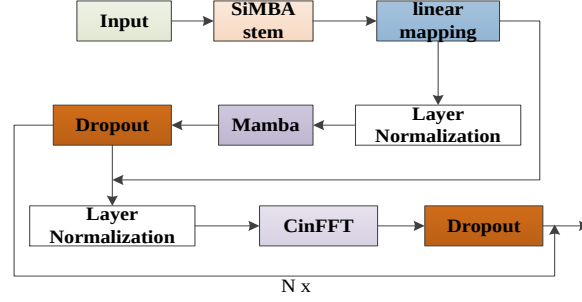


Fig. 5. SiMBA-PC Model Structure

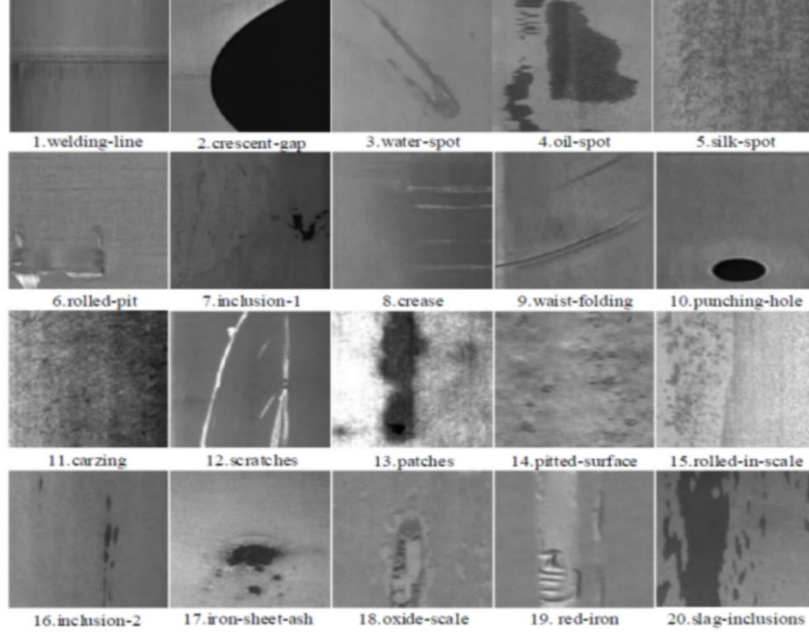
Due to the computational redundancy and high memory consumption caused by the three-layer convolution in SiMBA, and the multiple layers of convolution stacking resulting in the smoothing of high-frequency details (such as edges, textures), especially in the middle layer (the second layer). Therefore, replacing the second layer in stem with PConv reduces the computational load while alleviating feature smoothing and retaining detail information. 错误!未找到引用源。

While maintaining the remaining portion of the channel, PConv is a partial convolution technique that only performs a portion of the input channel convolution. This method alleviates feature smoothing while reducing computational load, retains detailed information, and improves model performance 错误!未找到引用源。 1

## 4 Experiment

### 4.1 Experiment Setup and Dataset

The datasets include the NEU-CLS dataset (1,800 grayscale images) and the FSC-20 dataset (small sample classification dataset specifically for classifying surface defects of steel strips). The dataset integrates datasets from three different sources: NEU-CLS, X-SDD, and GC10-DET, which contain a total of 20 categories of steel strip surface defects, as shown in Fig. 6. 错误!未找到引用源。



**Fig. 6.** categories of FSC-20 dataset

The training set to test set ratio was set at 8:2, and set the resolution of the input images uniformly to  $256 \text{ pixel} \times 256 \text{ pixel}$ , center cropping  $224 \text{ pixel} \times 224 \text{ pixel}$ , and normalizing the distribution. GPU uses the NVIDIA GeForce RTX 4090D.

#### 4.2 Hybrid Model CMT-light based on Channel Attention Mechanism and Low-Rank Decomposition

In the actual deployment of deep learning models, the number of parameters (Params), computational power (FLOPs), and accuracy (Acc@1) are the core metrics for measuring model performance. This study was tested on the industrial defect detection datasets NEU-CLS and FSC-20, and benchmarked against seven mainstream classification networks, including EfficientNet, VanillaNet, VMamba-T.

The results of the NEU-CLS comparison are shown in Table 2. Cmt-light has 10.8% fewer parameters and 5% fewer FLOPs compared to the original CMT model, while Acc@1 has increased by 0.7%. Compared with other models, Acc@1 increased by up to 2.7%.



**Table 2.** NEU-CLS Comparison Experiment

| Methods         | Params/ $10^6$ | FLOPs/ $10^9$ | Acc@1/% |
|-----------------|----------------|---------------|---------|
| CMT-S           | 25.0           | 4.0           | 99.3    |
| CMT-light       | 22.3           | 3.8           | 100     |
| EfficientNet-B4 | 19             | 4.2           | 99.2    |
| VanillaNet-10   | 45.7           | 9.4           | 99.1    |
| VMamba-T        | 22             | 5.7           | 98.3    |
| Swin-T          | 29             | 4.5           | 97.6    |
| CPVT-S-GAP      | 23             | 4.6           | 97.3    |
| RegNetY-8GF     | 32.9           | 8             | 97.6    |
| CeiT-S          | 24.2           | 4.5           | 98.2    |

As shown in Table 3, CMT-light improved by 1.4% compared to the original CMT model Acc@1. 错误!未找到引用源。 Compared with other models, Acc@1 increased by up to 3.5%.

**Table 3.** FSC-20 Comparative Experiment

| Methods         | Params/ $10^6$ | FLOPs/ $10^9$ | Acc/% |
|-----------------|----------------|---------------|-------|
| CMT-S           | 25.0           | 4.0           | 89.3  |
| CMT-light       | 22.3           | 3.8           | 90.7  |
| EfficientNet-B4 | 19             | 4.2           | 88.9  |
| VanillaNet-10   | 45.7           | 9.4           | 89.0  |
| VMamba-T        | 22             | 5.7           | 88.6  |
| Swin-T          | 29             | 4.5           | 87.4  |
| CPVT-S-GAP      | 23             | 4.6           | 87.2  |
| RegNetY-8GF     | 32.9           | 8             | 87.9  |
| CeiT-S          | 24.2           | 4.5           | 88.2  |

Based on the above comprehensive analysis, it can be concluded that the method proposed in this chapter performs well on the public datasets NEU-CLS and FSC-20. It outperforms all other mainstream classification networks in terms of classification accuracy. This indicates that the method is more effective when applied to image classification problems in real industrial scenarios.

错误!未找到引用源。错误!未找到引用源。 CMT+ low-rank decomposition mainly achieved significant reductions on Params.

**Table 4.** NEU-CLS ablation experiment

| Methods                   | Params/ $10^6$ | FLOPs/ $10^9$ | Acc@1/% |
|---------------------------|----------------|---------------|---------|
| CMT                       | 25.0           | 4.0           | 99.3    |
| CMT+SENet                 | 25.0           | 3.8           | 100     |
| CMT+ low-rank composition | 22.3           | 4.0           | 99.3    |
| CMT-light                 | 22.3           | 3.8           | 100     |

**Table 5.** FSC-20 ablation experiment

| Methods                   | Params/ $10^6$ | FLOPs/c | Acc@1/% |
|---------------------------|----------------|---------|---------|
| CMT                       | 25.0           | 4.0     | 89.3    |
| CMT+ SENet                | 25.0           | 3.8     | 90.7    |
| CMT+ low-rank composition | 22.3           | 4.0     | 90.2    |
| CMT-light                 | 22.3           | 3.8     | 90.7    |

With the introduction of the SENet, low-rank factorization module, the model significantly improves performance while maintaining low Params and FLOPs by enhancing the responsiveness of the key channels and reducing the parameter size by multiplying the factorization matrices. This shows the importance of designing effective modules for improving model performance. Additionally, it shows how these modules may be used to improve the underlying model.

#### 4.3 Research on Image Classification Algorithms Based on Linear Attention Mechanism

Comparative experiments were conducted on the datasets NEU-CLS and FSC-20 as shown in Table 6 and 7. The CMT model in Chapter 2, the CMT-Light model and the vit model were selected as the basic comparative models.

**Table 6.** NEU-CLS Comparison Experiment

| Methods     | Params/ $10^6$ | FLOPs/ $10^9$ | Sample/s(test) |
|-------------|----------------|---------------|----------------|
| SiMBA-S     | 25.8           | 3.2           | 623.12         |
| SiMBA-light | 25.7           | <b>3.0</b>    | <b>655.47</b>  |
| Vit-B       | 85.8           | 17.5          | 421.55         |
| CMT-S       | 25.0           | 4.0           | 570.69         |
| CMT-light   | <b>22.3</b>    | 3.8           | 583.35         |

错误!未找到引用源。According to the data in Table 4-1, the SiMBA-light method showed significant advantages in FLOPs and Sample/s (test) as well as accuracy (Acc/%).

**Table 6.** NEU-CLS Comparison Experiment

| Methods     | Params/ $10^6$ | FLOPs/ $10^9$ | Sample/s(test) |
|-------------|----------------|---------------|----------------|
| SiMBA-S     | 25.8           | 3.2           | 623.12         |
| SiMBA-light | 25.7           | <b>3.0</b>    | <b>655.47</b>  |
| Vit-B       | 85.8           | 17.5          | 421.55         |
| CMT-S       | 25.0           | 4.0           | 570.69         |
| CMT-light   | <b>22.3</b>    | 3.8           | 583.35         |

Specifically, the FLOPs was 3.0G, which was 6.2 percent lower than the original SiMBA-S model, and SiMBA-light processed 655.47 samples in 1 second in inference, which was 5.2 percent higher than the original SiMBA-S model. SiMBA-light achieved 98.9% accuracy, which is 0.8% higher than that of SiMBA-S.

Compared with the CMT model, params improved by 2.8%, decreased accuracy by 0.4%, but reduced FLOPs by 25% and increased throughput by 14.8%. Compared with

the CMT-light model, params improved by 15.2 percent, with a 1.1 percent drop in accuracy but a 21 percent drop in FLOPs and a 12.3 percent increase in throughput. Compared with the vit model, params has been significantly optimized across all three metrics, with a 70% reduction in FLOPs and an 88.5% increase in throughput.

**Table 6.** NEU-CLS Comparison Experiment

| Methods     | Params/ $10^6$ | FLOPs/ $10^9$ | Sample/s(test) |
|-------------|----------------|---------------|----------------|
| SiMBA-S     | 25.8           | 3.2           | 623.12         |
| SiMBA-light | 25.7           | <b>3.0</b>    | <b>655.47</b>  |
| Vit-B       | 85.8           | 17.5          | 421.55         |
| CMT-S       | 25.0           | 4.0           | 570.69         |
| CMT-light   | <b>22.3</b>    | 3.8           | 583.35         |

**Table 7.** FSC-20 Comparison Experiment

| Methods     | Params/ $10^6$ | FLOPs/ $10^9$ | Sample/s(test) |
|-------------|----------------|---------------|----------------|
| SiMBA-S     | 25.8           | 3.2           | 601.12         |
| SiMBA-light | 25.6           | <b>3.0</b>    | <b>625.44</b>  |
| Vit-B       | 85.8           | 17.5          | 417.83         |
| CMT-S       | 25.1           | 4.0           | 566.73         |
| CMT-light   | <b>22.3</b>    | 3.8           | 571.79         |

In the FSC-20 dataset, SiMBA-light reduced FLOPs by 6.2% compared to the original SiMBA-S model and improved accuracy by 1.3%. In inference, SiMBA-light processed 625.44 samples in 1 second and improved by 4% compared to the original SiMBA-S model.

## 5 Conclusion

This study focuses on the research of efficient and reliable surface defect classification and detection, and proposes the comprehensive hybrid model CMTSiMBA-light. Thus the precision of identifying tiny defects in complex industrial scenarios has been improved. The next step could be to extend the task to high-resolution images, taking advantage of SSM's ability to reduce complexity to the linear level  $O(n)$ .

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Wang W, Xie E, Li X, et al. Pvt v2: Improved baselines with pyramid vision transformer[J]. Computational visual media, 2022, 8(3): 415-424.
2. Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

3. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. *Advances in neural information processing systems*, 2017, 30.
4. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. *IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, 2016: 770–778.
5. CHEN H, WANG Y, GUO J, et al. VanillaNet: The Power of Minimalism in Deep Learning[C]// *NeurIPS Foundation. 2023 Advances in Neural Information Processing Systems (NeurIPS)*. New Orleans: Curran Associates, 2023, 36: 7050-7064.
6. Liu Y, Tian Y, Zhao Y, et al. Vmamba: Visual state space model[J]. *Advances in neural information processing systems*, 2024, 37: 103031-103063.
7. Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. *arXiv preprint arXiv:1704.04861*, 2017.
8. Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//*Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017: 2117-2125.
9. KOONCE B. EfficientNet[M/OL]//KOONCE B. *Convolutional Neural Networks with Swift for Tensorflow*. Berkeley, CA: Apress, 2021: 109-123. [http://link.springer.com/10.1007/978-1-4842-6168-2\\_10](http://link.springer.com/10.1007/978-1-4842-6168-2_10). DOI:10.1007/978-1-4842-6168-2\_10.
10. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//*Proceedings of the IEEE/CVF international conference on computer vision*. 2021: 10012-10022.
11. Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//*European conference on computer vision*. Cham: Springer International Publishing, 2020: 213-229.
12. Graham B, El-Nouby A, Touvron H, et al. Levit: a vision transformer in convnet's clothing for faster inference[C]//*Proceedings of the IEEE/CVF international conference on computer vision*. 2021: 12259-12269.
13. Dai Z, Liu H, Le Q V, et al. Coatnet: Marrying convolution and attention for all data sizes[J]. *Advances in neural information processing systems*, 2021, 34: 3965-3977.
14. Srinivas A, Lin T Y, Parmar N, et al. Bottleneck transformers for visual recognition[C]//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021: 16519-16529.
15. Wu H, Xiao B, Codella N, et al. Cvt: Introducing convolutions to vision transformers[C]//*Proceedings of the IEEE/CVF international conference on computer vision*. 2021: 22-31.
16. Guo J, Han K, Wu H, et al. Cmt: Convolutional neural networks meet vision transformers[C]//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022: 12175-12185.
17. Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces[J]. *arXiv preprint arXiv:2312.00752*, 2023.
18. Redmon, Joseph, Santosh Kumar Divvala, Ross B. Girshick and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 779-788.

19. Li S, Singh H, Grover A. Mamba-nd: Selective state space modeling for multi-dimensional data[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 75-92.
20. Yang C, Chen Z, Espinosa M, et al. Plainmamba: Improving non-hierarchical mamba in visual recognition[J]. arXiv preprint arXiv:2403.17695, 2024.
21. Huang T, Pei X, You S, et al. Localmamba: Visual state space model with windowed selective scan[J]. arXiv preprint arXiv:2403.09338, 2024.
22. Pei X, Huang T, Xu C. Efficientvmamba: Atrous selective scan for light weight visual mamba[J]. arXiv preprint arXiv:2403.09977, 2024.
23. Patro B N, Agneeswaran V S. Simba: Simplified mamba-based architecture for vision and multivariate time series[J]. arXiv preprint arXiv:2403.15360, 2024.