MTRNet: An Efficient Hybrid Network Model for Surface Defect Detection of Sheet Metal

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

Abstract— **Addressing the challenges posed by the insufficient computational power of low-spec devices to achieve the inference efficiency of prevailing deep learning models, alongside the variability in type and shape within industrial metal sheet datasets, while also catering to the demand for high-precision detection, this paper proposes a novel single-stage detection architecture termed MTRNet. In this paper, we introduce three key enhancements: Firstly, to reduce the number of parameters, we propose a novel Partial Depth Convolution (PDC) structure. By minimizing redundant computations, we aim to enhance the efficiency of spatial feature extraction. Secondly, we propose a Convolution Transformer (CTR) structure that replaces the self-attention module with a convolutional module. This modification addresses the computational inefficiency inherent in the self-attention mechanism of existing Visual Transformers (ViT). Finally, we introduce a novel Attention Convolution Transformer (ACTR) structure to enhance the extraction of global and local feature information. This architecture seamlessly integrates the strengths of both attention and convolution mechanisms, working synergistically to improve performance. Our proposed MTRNet network demonstrates superior detection performance compared to existing industrial standards, achieving a detection accuracy of 81.5% on the NEU-DET dataset.**

Index Terms— **Defect Detection, Visual Transformer, Hybrid Structures, Convolutional Neural Networks**

I. Introduction

In the manufacturing process of sheet metal, the metal surface frequently encounters environmental factors such as extrusion, friction, oxidation, and chemical corrosion, leading to diverse surface defects that compromise the sheet metal quality. Consequently, machine vision inspection technology has witnessed rapid advancements in addressing the demands of industrial defect detection. Propelled by continuous breakthroughs in deep learning, numerous researchers have introduced a spectrum of deep learning algorithms [1]–[3]. In recent years, Transformer [4] has garnered significant attention within the industry owing to its remarkable capability in capturing long-range dependencies, thereby achieving notable success in domains such as target detection [5] and image classification [6].

Presently, the most prevalent detector models in the industry are one-stage detectors, owing to their capacity to deliver expedited detection speeds and consistent detection accuracy essential for real-time detection tasks. An illustrative example is the YOLO series [7]–[9]. Experimental evidence has demonstrated that while mainstream generalized models in the field of defect detection on metal plate surfaces have achieved competitive performance, they still exhibit challenges, including high false detection and leakage rates. These challenges are particularly pronounced in handling global and minute features such as stains and scratches. Furthermore, defect detection devices for sheet metal surfaces based on deep learning encounter issues such as large model sizes, slow recognition response times, and limitations in industrial deployment feasibility.

II. Method

The structure of the MTRNet model proposed in this paper is illustrated in Fig 1. Upon input of the image, standard convolution operations are employed to adjust the image size and channels, facilitating feature information extraction through the CTR and ACTR modules. Subsequently, the Partial Depth Convolution (PDC) module consolidates shallow and deep feature information, enabling the extraction of minute local defects. Ultimately, detection is executed via the Detect layer. The detailed design of the PDC, CTR, and ACTR modules is elaborated in the subsequent section.

A. Partial Depth Convolution (PDC)

Convolutional Neural Networks (CNNs) have demonstrated strong performance in various vision tasks, with models such as VGGNet [10], GoogleNet [11], ResNet [12], and MobileNet [13] being extensively utilized. However, CNNs entail significant computational costs when applied to defect detection on sheet metal surfaces. Hence, we propose a novel convolutional structure to enhance detection accuracy by reducing parameter overhead.

We propose the PDC convolution module structure, as shown in Fig 1. During feature extraction, redundant information can arise; therefore, we introduce a partial convolution (PConv) [14] operation that extracts feature information only for a part of the channels. This approach improves computational efficiency while maintaining detection accuracy. Additionally, we employ depth convolution to reduce parameter count while capturing spatial features. Standard convolution facilitates cross-channel information interaction, enabling effective spatial and channel-wise feature extraction. We integrate residual concatenation into the PDC structure to enhance performance and propagate gradients across layers, facilitating multi-scale feature capture. Experimental

[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

DOI reference number: 10.18293/SEKE2024-092

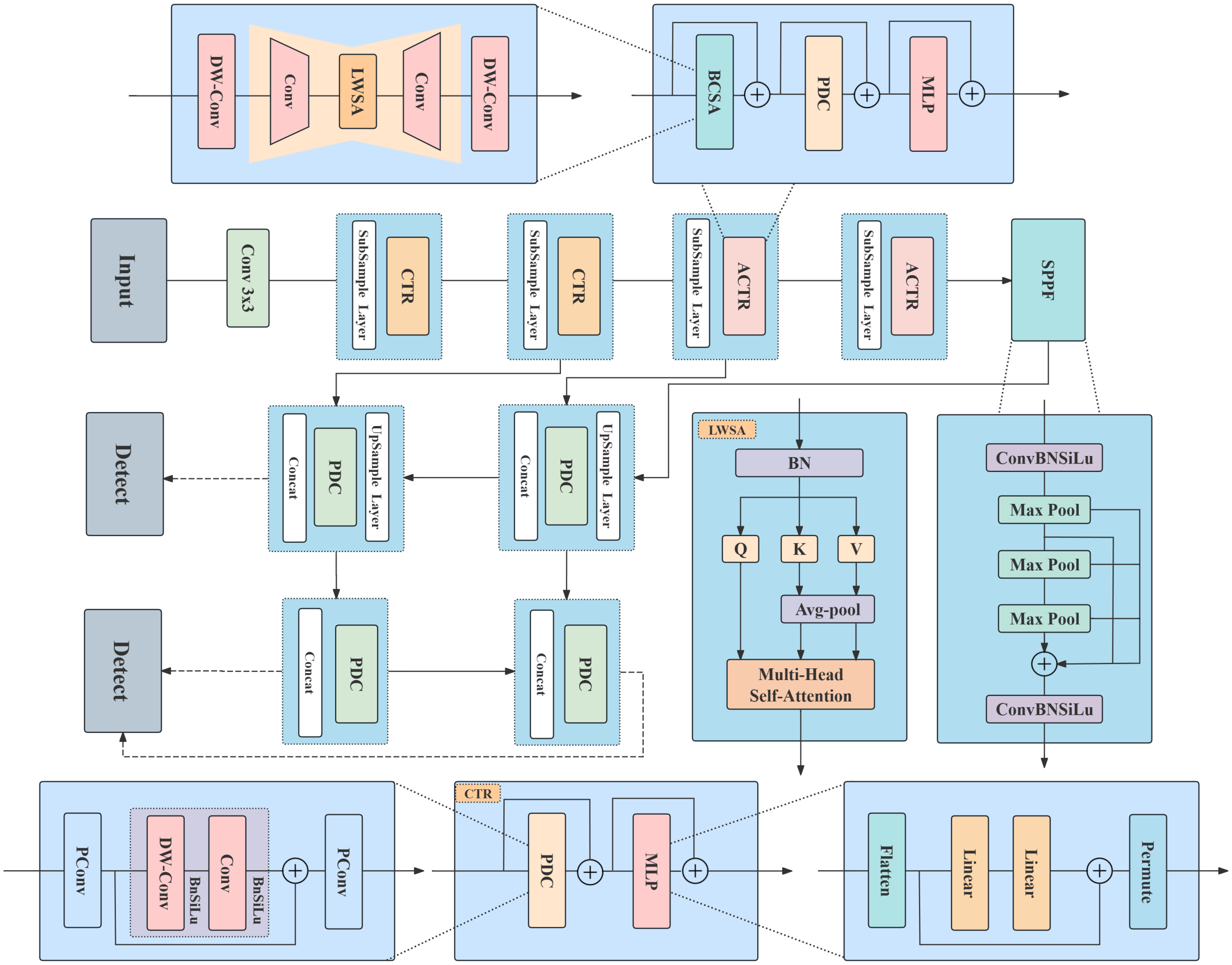

Fig. 1. Structure Diagram of the MTRNet Model.

results validate the efficacy of our approach. In the MTRNet model, the PDC module further integrates different feature types by fusing shallow features from the CTR module with deeper features from the ACTR module, thereby enhancing defect detection and optimizing small defect extraction in metal sheets. The relevant formulas are provided below.

$$\varphi\left(X\right)=pF^{3\times3}\left(F^{1\times1}\left(dF^{3\times3}\left(pF^{3\times3}\left(X\right)\right)\right)+pF^{3\times3}\left(X\right)\right) \tag{1}$$

$\varphi(X)$ represents the output of the PDC model structure, where $\mathrm{pF}^{3\times3}$ denotes the PConv convolution with a 3×3 kernel, $F^{1\times1}$ represents the ordinary convolution with a 1×1 kernel, $\mathrm{dF}^{3\times3}$ indicates the depthwise separable convolution with a 3×3 kernel, and X denotes the input feature map.

B. Convolution Transformer (CTR)

Transformers have made remarkable advancements in addressing vision tasks; however, their computational complexity remains a significant concern. Recent studies such as EdgeViTs [15] have demonstrated improved performance by simplifying the model structure. Nevertheless, traditional Transformers encounter limitations when applied to the sheet metal dataset due to the abundance of localized features.

In practical applications constrained by limited computational resources, designing more lightweight model architectures is necessary. Considering the prevalence of local features within most metal sheet datasets, we introduce the CTR model structure, depicted schematically in Fig 1. While vision transformer(ViT) [16] enhances the model's long-range dependencies through the self-attention mechanism, it overlooks the slices' local relationships and structural information. To address this, we incorporate convolutional neural networks to substitute the self-attention mechanism of ViT. Convolutional neural networks are proficient in extracting local features and entail fewer parameters than ViT's self-attention mechanism, thus maintaining the model's real-time performance. Moreover, the CTR model aligns with the MetaFormer [17] generic architecture. Consequently, this approach circumvents the drawbacks of extensive computation and impractical deployment associated with the self-attention mechanism and enhances the model's capability to extract local feature defects. The relevant formulas are provided below.

$$MLP\left(X\right)=Pe\left(Linear\left(Linear\left(Fl\left(X\right)\right)\right)+Fl\left(X\right)\right) \tag{2}$$

$$\omega=\left(X+\varphi\left(X\right)\right)+MLP\left(X+\varphi\left(X\right)\right) \tag{3}$$

Pe is the operation to restore the feature map to its original dimension, Fl flattens the feature map, and ω represents the output of the CTR model.

C. Attention Convolution Transformer (ACTR)

Although the CTR model has demonstrated exemplary performance in extracting local information, there is still room for improvement in acquiring both global and local information, particularly for global defects such as scratches and wrinkles on the surface of the metal sheet dataset. Therefore, leveraging the self-attention mechanism of ViT to extract global features remains necessary. Previous studies have indicated that Transformers may compromise the extraction of local information to some extent [18], potentially affecting details such as local texture. ViTAE [19] proposed a method that utilizes both convolutional and self-attention modules in parallel to address this issue, but its complexity is high.

In response to the abovementioned challenges, this paper introduces a novel hybrid model structure, ACTR, depicted in Fig 1. To mitigate redundant feature information and further reduce parameter count, we propose novel models of sparse self-attention, termed Bottleneck Convolutional Self-Attention (BCSA) structures. Initially, depthwise separable convolution and ordinary convolution are employed for spatial downsampling to decrease the dimensionality of the feature map upon entering the self-attention mechanism. Subsequently, a proportional sparsity operation is applied to K and V using average pooling. The global self-attention mechanism aggregates information in the feature maps captured by convolution. Finally, convolution is reapplied for upsampling to restore the dimensionality of the feature map. This sparse attention method has been experimentally validated to demonstrate competitive performance. However, reducing the feature map to such a small size results in losing significant local information, which ViT may exacerbate. Therefore, we propose a fusion approach that combines the sparse self-attention structure BCSA with the convolutional neural network PDC to integrate feature information and synergistically address the challenge of the model losing local information. Additionally, the MLP layer captures the nonlinear relationships between features to extract more essential and significant features. Our proposed ACTR model structure effectively extracts global and local information, enhancing the overall modelling capability. The relevant formulas are provided below.

$$Lw\left(X\right)=At\left(X\cdot W^{Q},Ap\left(X\cdot W^{K}\right),Ap\left(X\cdot W^{V}\right)\right) \tag{4}$$

$$\eta\left(X\right)=dF^{3\times3}\left(F^{1\times1}\left(Lw\left(F^{1\times1}\left(dF^{3\times3}\left(X\right)\right)\right)\right)\right) \tag{5}$$

$$\delta\left(X\right)=X+\eta\left(X\right)+\varphi\left(X+\eta\left(X\right)\right) \tag{6}$$

$$M=\delta\left(X\right)+MLP\left(\delta\left(X\right)\right) \tag{7}$$

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 $At\left(Q,K,V\right)=softmax\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V$. 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. Lw is the result obtained after sparsifying K and V. $\eta(X)$ represents the output of the sparse attention model structure BCSA, where $\mathrm{dF}^{3\times3}$ denotes depthwise separable convolution with a 3×3 kernel, and $F^{1\times1}$ represents standard convolution. $\delta(X)$ represents the output after the fusion of the convolution module PDC and the self-attention BCSA. M represents the output of the model structure ACTR, and MLP refers to our fully connected layer.

III. Experiments

A. DataSet

The publicly available datasets in this study comprise the Steel Surface Defect Dataset (NEU-DET) provided by Northeastern University. The NEU-DET dataset contains six typical surface defects found on hot-rolled strip steel: rolled-in scale, patches, crazing, pitted surface, inclusion, and scratches. It comprises 1,800 samples, each image weighing 200 × 200 pixels.

B. Device

The experiment was conducted on a Linux system with PyTorch 1.9.1, CUDA 11.4, and cuDNN 8.0. The NVIDIA RTX A6000 graphics card was utilized for all experiments. An SGD optimizer was employed with an initial learning rate of 0.01, 500 training epochs, and a zero initialization seed.

C. Comparison and ablation experiments

Table I compares experiments between our proposed MTRNet model and existing models on the NEU-DET dataset. Our model is evaluated against advanced industrial detection models, including PPYOLOE-s, PicoDet, and others. Experimental results demonstrate that the detection precision and recall rate of our proposed MTRNet model exceed 80%, with performance surpassing PPYOLOE-s by 3.6% and PicoDet by 3.8% on mAP@0.5. Additionally, on mAP@0.5:0.95, our model outperforms PPYOLOE-s by 2.4% and PicoDet by 2.3%. Notably, our model achieves superior detection performance compared to existing industrial detection models on the NEU-DET dataset.

The MTRNet, our designed single-stage detector, underwent ablation experiments on the innovative module using YOLOv5s as the baseline on the NEU-DET dataset. Results presented in Table II demonstrate significant improvements with each modified module: detection accuracy increased from 76.9% to 81.5%, recall rate improved from 74.6% to 80.4%, mAP@0.5 rose from 80.3% to 82.2%,

TABLE I

Comparative experiments of MTRNet and other models on the NEU-DET dataset are presented.

Detection	**Detection Result**			
Model	***Precision***	***Recall***	***mAP@.5***	***mAP@.5-.95***
YOLOv3	76.1%	75.5%	78.6%	41.2%
YOLOv4	78.1%	74.5%	79.1%	41.0%
YOLOv5s	76.9%	74.6%	80.3%	41.5%
YOLOv7-T	73.2%	70.1%	74.6%	38.8%
YOLOv8	78.4%	77.3%	78.7%	39.7%
YOLOR-P6	75.9%	74.7%	75.2%	39.3%
PPYOLOE-s	78.0%	73.9%	78.6%	42.1%
PicoDet	77.6%	78.2%	78.4%	42.2%
MTRNet	**81.5%**	**80.4%**	**82.2%**	**44.5%**

mAP@0.5:0.95 increased from 41.5% to 44.5%. At the same time, the parameter count decreased from 7.02M to 6.98M. The overall parameter count of our proposed model is lower than the Baseline while achieving improved accuracy.

TABLE II

Ablation experiments were conducted on the NEU-DET dataset to assess the improved module

Detection	**Detection Result**				
Model	***Precision***	***Recall***	***mAP@.5***	***mAP@.5-.95***	***Param***
BaseLine	76.9%	74.6%	80.3%	41.5%	7.02M
PDC	77.7%	76.5%	80.4%	43.1%	6.92M
CTR	77.5%	77.8%	80.2%	43.5%	**6.88M**
ACTR	77.3%	76.9%	80.8%	43.3%	6.91M
CTR+ACTR	78.2%	79.0%	81.6%	43.7%	6.95M
All	**81.5%**	**80.4%**	**82.2%**	**44.5%**	6.98M

IV. CONCLSION

This paper introduces MTRNet, featuring a Partial Depth Convolution (PDC) structure designed to reduce parameter count and optimize the standard convolutional architecture effectively. Given the prevalence of localized defects in metal sheets, we advocate for adopting convolutional neural networks as an alternative to self-attention mechanisms, aiming to streamline model complexity while bolstering its capacity to extract local features. Furthermore, we propose a hybrid architecture integrating sparse self-attention and CNNs to encapsulate global feature information. Experimental results validate the superior performance of our proposed model.

Acknowledgment

This work has been supported by several projects: the research on intelligent fault diagnosis methods for analyzing big data from mechanical equipment based on the industrial internet platform (ZR2022MF279); the research and application of critical technologies for intelligent control and information platforms for agricultural machinery (2023TSGC0587); and the integration technology and demonstration of networked collaborative manufacturing for textile and clothing enterprises (2021GXRC074). It also benefits from the research on the Classroom Teaching Evaluation System utilizing Intelligent Recognition of Human Body Posture and Gestures (2022TATSGC022); the development of High-Precision Industrial Vision Defect Detection Technology for Steel Materials (S202310431041); and the advancement of Key Technologies for Intelligent Control and Informatization Platforms for Agricultural Machinery (2023TATSGC043).

References

[1] Krizhevsky, A., Sutskever, I. & Hinton, G. ImageNet classification with deep convolutional neural networks. *Communications Of The ACM*. **60**, pp. 84-90 (2017)

[2] Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 779-788 (2016)

[3] Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 580-587 (2014)

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Advances In Neural Information Processing Systems*. **30** (2017)

[5] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. & Zagoruyko, S. End-to-end object detection with transformers. *European Conference On Computer Vision*. pp. 213-229 (2020)

[6] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. & Jégou, H. Training data-efficient image transformers & distillation through attention. *International Conference On Machine Learning*. pp. 10347-10357 (2021)

[7] Redmon, J. & Farhadi, A. Yolov3: An incremental improvement. *ArXiv Preprint ArXiv:1804.02767*. (2018)

[8] Ge, Z., Liu, S., Wang, F., Li, Z. & Sun, J. Yolox: Exceeding yolo series in 2021. *ArXiv Preprint ArXiv:2107.08430*. (2021)

[9] Wang, C., Bochkovskiy, A. & Liao, H. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 7464-7475 (2023)

[10] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*. (2014)

[11] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. Going deeper with convolutions. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 1-9 (2015)

[12] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 770-778 (2016)

[13] Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. & Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv Preprint ArXiv:1704.04861*. (2017)

[14] Chen, J., Kao, S., He, H., Zhuo, W., Wen, S., Lee, C. & Chan, S. Run, don't walk: chasing higher FLOPS for faster neural networks. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 12021-12031 (2023)

[15] Pan, J., Bulat, A., Tan, F., Zhu, X., Dudziak, L., Li, H., Tzimiropoulos, G. & Martinez, B. Edgevits: Competing light-weight cnns on mobile devices with vision transformers. *European Conference On Computer Vision*. pp. 294-311 (2022)

[16] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. & Others An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv Preprint ArXiv:2010.11929*. (2020)

[17] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J. & Yan, S. Metaformer is actually what you need for vision. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 10819-10829 (2022)

[18] Yang, R., Ma, H., Wu, J., Tang, Y., Xiao, X., Zheng, M. & Li, X. Scalablevit: Rethinking the context-oriented generalization of vision transformer. *European Conference On Computer Vision*. pp. 480-496 (2022)

[19] Xu, Y., Zhang, Q., Zhang, J. & Tao, D. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances In Neural Information Processing Systems*. **34** pp. 28522-28535 (2021)