A Lightweight Convolutional Network for Few-Shot and Multi-Class Detection of Tiny Aluminum Defects

Jianbin Zhong*, Daitao Wang†, Bo Xie‡, Wenjing Yu§, Hongjun Qiu¶
*†‡§Software Engineering Institute of Guangzhou
Guangzhou, Guangdong, China
510900
*netgonight@gmail.com, †wangdaitao7@gmail.com, ‡1732209494@qq.com, §ywj@mail.seig.edu.cn
¶Guangzhou Civil Aviation College
Guangzhou, Guangdong, China
510403
qiuhongjun@gcac.edu.cn

Abstract*—For the task of detecting tiny defects on industrial aluminum surfaces, existing methods face two major challenges: 1) Limited training data makes it difficult for models to learn small defect features; 2) Industrial deployment requires high efficiency and lightweight structures. To address these issues, a novel solution based on an attention mechanism and a lightweight architecture is proposed in this paper. An efficient encoder-decoder network has been designed, where the encoder adopts lightweight convolutions to reduce computational cost; spatial and channel attention modules are embedded within the skip connections between the encoder and decoder modules to enhance the recognition capability for subtle defects. Experimental results demonstrate that this method achieves high accuracy (73.54% mIoU) on a small-sample aluminum defect dataset, while realizing extremely fast inference speed (314.55 FPS) and small model size (0.114M parameters). Compared to classic models such as DeepLabV3+ and U-Net, the proposed method exhibits significant superiority in key metrics, especially in identifying tiny defect categories like pinholes and abrasions. This work provides an efficient, lightweight, and reliable solution for industrial vision inspection tasks. The code and models will be made publicly available at *https://github.com/NETgonight/Aluminum-defect

***Index Terms*—Defect detection, lightweight network, tiny defect recognition, attention mechanism, few-shot learning.**

I. Introduction

In the modern aluminum processing industry, precise surface quality inspection is crucial for ensuring product quality. As a material widely used in aviation, automotive, construction, and other fields, the detection of defects on aluminum sheets directly impacts the safety and durability of products. Traditional manual visual inspection methods have become insufficient in terms of efficiency and accuracy to meet the demands of large-scale production. With the advancement of Industry 4.0, the integration of computer vision and deep learning techniques has provided a new perspective for aluminum defect detection. These technologies can automatically analyze image data and identify various defects on aluminum surfaces, enabling rapid and accurate quality assessment.

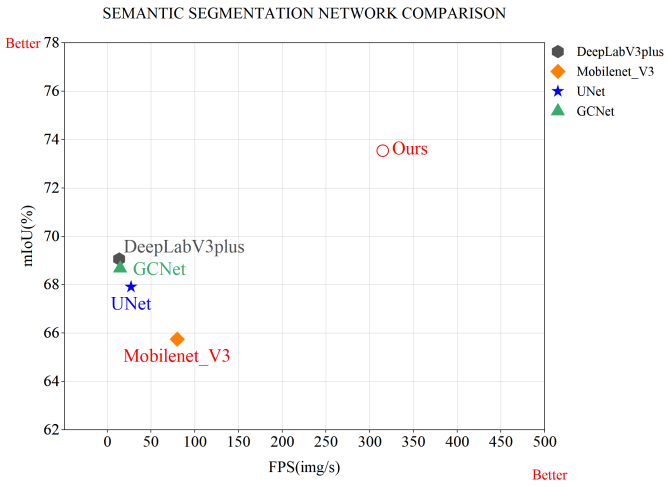

Fig. 1: Comparison of our Net with other semantic segmentation networks.

However, despite the significant progress made by deep learning methods in defect detection, several challenges remain. Firstly, real-time performance constraints have limited the application of these methods in production environments requiring fast response times. Secondly, the precise detection of small-sized defects remains a challenge, as these defects are often difficult to accurately recognize in many lightweight networks, especially in resource-constrained scenarios. Furthermore, most research has focused on foreground-background anomaly detection, while fine-grained segmentation of multiple defect categories, particularly in cases where various defect types coexist within the same image, has received relatively less attention.

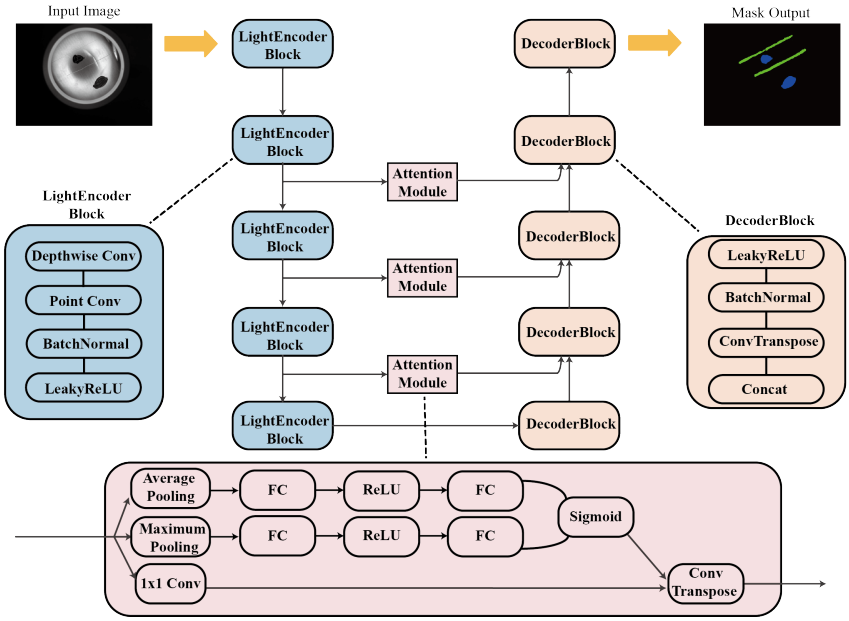

Fig. 2: Framework of the proposed defect segmentation network

To address the aforementioned issues, a novel highly efficient real-time detection network that combines lightweight convolutions and an attention mechanism has been proposed. This network is not only capable of effectively recognizing small-sized defects but also enhances the model's focusing ability on key features through the introduction of the attention mechanism, enabling precise identification of small defects in multi-class semantic segmentation tasks. Experimental results demonstrate that the proposed method excels in real-time performance, rapidly processing image data while maintaining high accuracy. On a single NVIDIA V100 GPU, it achieves 314.55 FPS with only 0.591 GFLOPs and 0.114M parameters, attaining 73.54% mIoU. Figure 1 illustrates its advantages over other semantic segmentation networks.

II. Related Work

Semantic Segmentation: Semantic segmentation is a fundamental task in the field of computer vision, aiming to classify each pixel in an image into a specific category. With the advancement of deep learning, convolutional neural networks (CNNs) have become the mainstream method for achieving high-accuracy semantic segmentation. Networks such as SegNet [1], the DeepLab series [2], and HRNet [3], which effectively captures high-resolution information through multi-scale feature fusion, have significantly improved segmentation accuracy. Despite this progress, existing methods often overlook the trade-off between segmentation quality and computational efficiency, which is crucial for real-time applications and algorithm deployment in industrial settings.

Lightweight Networks: The demand for real-time and resource-efficient models has driven the development of lightweight networks tailored for industrial inspection tasks. Networks like MobileNet [4], EfficientNet [5], and GhostNet [6] adopt depthwise separable convolutions and channel pruning to reduce computational complexity. While maintaining acceptable accuracy, these networks have been successfully applied in defect detection for industrial manufacturing. However, their application in complex defect detection scenarios, particularly involving small defects with low contrast or microscopic regions, remains challenging.

Attention Mechanisms: Attention mechanisms are increasingly integrated into encoder-decoder architectures, such as Swin-UNet [7], to enhance fine-grained defect segmentation. Wang et al. [8] demonstrated the effectiveness of attention modules in improving segmentation performance for various material defects. Yang et al. [9] combined attention mechanisms with residual connections, but accurate detection of small defects within the U-Net architecture remains challenging, especially for tiny defects or few-shot learning scenarios.

Few-Shot Multi-Class Defect Detection: Multi-class defect detection poses unique challenges, including class imbalance, difficulty in recognizing small defects, and real-time performance requirements. Class imbalance may cause the model to be biased towards the majority class, neglecting minority classes. Recent studies by Huang et al. [10] and Dong et al. [11] have attempted to address these challenges, but a gap remains in developing a model that can effectively balance all these factors in the few-shot learning scenario.

Despite progress in semantic segmentation, lightweight network design, and attention mechanisms, existing models still require improvement for small-sample and tiny defect detection scenarios. For multi-class defect detection in industrial environments, there is an urgent need to develop efficient and precise models that meet real-time requirements.

TABLE I: Network Architecture Table

Layer	Type	Kernel	Stride	Padding	Output Size
Input		-			640x480x3
Encoder1	LightEncoder	3x3	2	1	320x240x6
Encoder2	LightEncoder	3x3	2	1	160x120x12
Encoder3	LightEncoder	3x3	2	1	80x60x24
Encoder4	LightEncoder	3x3	2	1	40x30x48
Bottleneck	LightEncoder	3x3	2	1	20x15x96
Attention1	AttentionModule	1x1	1	0	20x15x96
Decoder1	Decoder	3x3	2	1	40x30x48
Attention2	AttentionModule	1x1	1	0	40x30x48
Decoder2	Decoder	3x3	2	1	80x60x24
Attention3	AttentionModule	1x1	1	0	80x60x24
Decoder3	Decoder	3x3	2	1	160x120x12
Attention4	AttentionModule	1x1	1	0	160x120x12
Decoder4	Decoder	3x3	2	1	320x240x6
Output	Decoder	3x3	2	1	640x480x1

III. METHOD

A. LightBlock-Encoder-Decoder

The backbone network constructed in this study follows an encoder-decoder architecture similar to U-Net [12], as illustrated in Figure 2. The architecture consists of 9 modules in total, as shown in Table I. To ensure industrial applicability, a lightweight convolutional structure has been adopted to reduce computational cost while maintaining segmentation accuracy.

The encoder comprises 4 LightEncoder modules, each aiming to reduce the spatial resolution of the input feature maps while increasing the number of channels. All four encoders utilize 3×3 convolutional kernels with a stride of 2×2 and padding of 1×1 to implement depthwise separable convolutions, effectively extracting spatial features. Subsequently, 1×1 pointwise convolutions are applied to double the number of channels, enhancing feature representation capacity. Batch normalization and leaky ReLU activation functions are applied after each convolutional layer. In the final step of the encoder, a special LightEncoder module, referred to as the bottleneck, is introduced, further increasing the number of channels to 96 to provide richer feature information for the decoder. The decoder section contains 4 convolutional transpose layers (ConvTranspose layers) corresponding to the encoder, progressively restoring the spatial dimensions of the feature maps. Each decoder module includes a 3×3 convolutional transpose layer with stride and padding consistent with its corresponding encoder module to ensure precise feature map alignment. To improve segmentation accuracy, skip connections are employed in the decoder at each step, concatenating the feature maps from the encoder with the decoder's output.

B. Skip Connection

Skip Connections [13], as a key connectivity strategy, are integrated into the encoder and decoder parts of the network to enhance the model's context information fusion ability and segmentation accuracy. Within the encoder and decoder blocks, skip connections allow low-level features from the encoder to be directly passed to the corresponding layers in the decoder. Specifically, the output from encoder layer 1 is concatenated with the output from decoder layer 4, encoder layer 2 with decoder layer 3, encoder layer 3 with decoder layer 2, and encoder layer 4 with decoder layer 1. This parallel connection approach enables the model to learn both local and global feature representations simultaneously.

C. Attention Module

Attention mechanisms enhance deep learning models by allowing them to focus on the most relevant parts of the input data, mimicking the human ability to prioritize sensory information. These mechanisms have proven to be highly effective in improving the performance of models in tasks requiring detailed feature analysis, such as image segmentation. To further enhance the network's feature extraction capability and segmentation accuracy, while addressing the trade-off between small defect recognition and computational efficiency, an attention module integrating channel attention and spatial attention has been designed.

The attention mechanism is incorporated into the key layers of the encoder and decoder, following the LightEncoder modules in the encoder and preceding the Decoder modules in the decoder. This allows the model to more effectively extract and utilize critical information when processing small defects, thereby improving segmentation accuracy.

Channel Attention: The channel attention module learns the importance of different channels through global pooling operations and fully connected layers. As shown in Figure 2, an adaptive average pooling and an adaptive max pooling are utilized to obtain global statistics from the feature maps. The outputs of these pooling operations are processed separately through two 1×1 convolutional layers, with the second layer downsampling the channels to 1/4 to save computational cost and generate channel weights. This enables the learning of differences between various channels in the feature maps.

Spatial Attention: The spatial attention module captures spatial information from the feature maps through convolutional operations. 1×1 convolutional kernels are employed to learn spatial weights, allowing the model to adjust spatial information without altering the feature map dimensions.

TABLE II: Ablation Study Results

Class	**Background**		**Wrinkling**		**Abrasion**		**Contamination**		**Pinhole**	
Method	IOU	ACC	IOU	ACC	IOU	ACC	IOU	ACC	IOU	ACC
BaseLine	98.21	99.38	53.76	65.33	39.70	42.37	93.36	97.11	46.48	53.85
BaseLine+LightBlock	98.12	98.95	59.07	76.69	33.91	51.20	**94.92**	**98.45**	53.72	75.11
BaseLine+LightBlock+CA	98.16	98.71	61.45	**84.62**	45.65	**71.82**	93.79	98.12	53.45	80.64
BaseLine+LightBlock+SA	98.21	99.18	57.84	72.63	33.54	43.05	94.33	98.01	**67.02**	**85.52**
BaseLine+LightBlock+CA+SA	**98.47**	**99.35**	**62.29**	74.09	**46.10**	59.81	94.78	96.54	66.04	82.81

The outputs of channel attention and spatial attention are concatenated and then upsampled through a convolutional transpose (ConvTranspose2d) layer to restore the spatial dimensions of the feature maps. This combined strategy enables the model to simultaneously focus on channel and spatial information, achieving more refined feature extraction.

IV. Experiments

A. Datasets

The dataset used in this study, provided by the Intel FPGA China Innovation Center, comprises 400 original images of aluminum surface defects. The images encompass four defect types: wrinkling, abrasion, contamination, and pinholes, with a total of 1,062 annotated defects. Figure 3 presents examples from the dataset. Precise manual annotations were performed by an expert team using the LabelMe tool to generate segmentation masks. To ensure model generalization, the dataset was divided into a training set of 280 images and a validation set of 120 images.

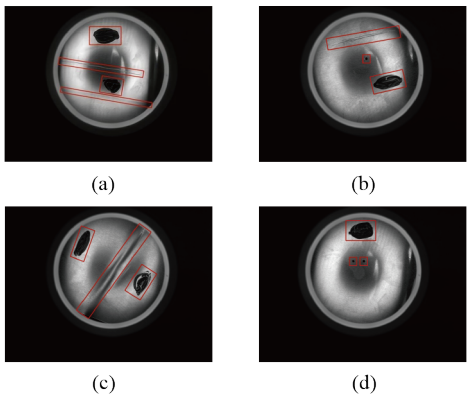

Fig. 3: Dataset sample

B. Experimental Setup

The experiments were conducted using the PyTorch framework, with a hardware configuration consisting of an NVIDIA GPU-V100-SXM2-32GB (32GB) and 12 virtual CPU cores of an Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz. On this hardware platform, ablation studies and model comparisons were carried out on the validation set to validate the effectiveness of the proposed method. The loss function employed is CrossEntropyLoss, with Adam as the optimizer, 2000 epochs, an initial learning rate of 0.01, and a MultiStepLR learning rate policy that decreased the learning rate to 1/10 of its original value at the 500th and 1400th iterations. Model performance was evaluated using the mean Intersection over Union (mIoU) and mean Accuracy (mAcc) as the primary metrics. The formulas for these metrics are given as follows:

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c} \tag{1}$$

$$\text{mAcc} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \tag{2}$$

where C represents the number of classes, and for each class c, TP_c, FP_c, and FN_c denote the true positives, false positives, and false negatives, respectively. Additionally, the number of floating-point operations (FLOPs) and frames per second (fps) were calculated to assess the model's lightweight nature and efficiency for practical applications.

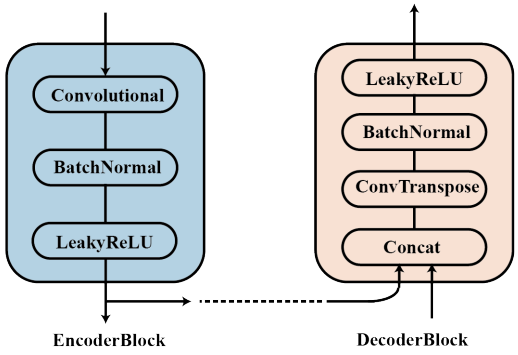

Fig. 4: Basic building blocks of Baseline

C. Ablation Study

To validate the effectiveness of the proposed method, a series of ablation experiments were designed: lightweight convolutional blocks were added to the baseline model, and attention modules were incorporated into the skip connections. Furthermore, the effects of adding spatial attention, channel attention, and their combination to the skip connections were compared. The ablation study results are presented in Table II. The baseline model consists of 4 encoder modules with 3x3 standard convolutional layers, 1 bottleneck, and 4 decoder modules, as shown in Figure 4 and Table I.

TABLE III: Comparison Study Results

Model	mIoU	mAcc	mDice	FLOPs (G)	Params (M)	fps (img/s)
DeepLabV3plus [2]	69.07	73.56	73.84	298.00	60.210	13.55
MobilenetV3 [4]	65.75	69.66	71.62	10.21	3.283	79.87
UNet [12]	67.91	74.10	73.05	254.00	28.987	27.31
GCNet [14]	68.68	72.11	73.53	323.00	66.250	14.75
Ours	**73.54**	**82.52**	**83.20**	**0.591**	**0.114**	**314.55**

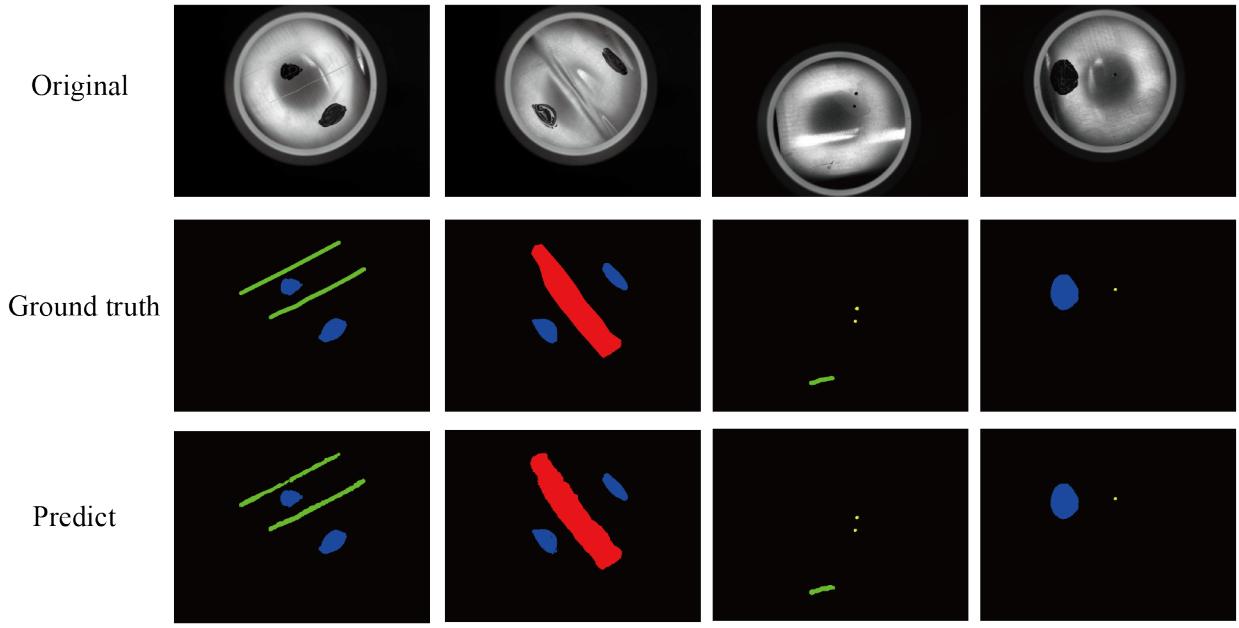

Fig. 5: Segmentation samples from the validation set.

Adding Lightweight Convolution Blocks: By introducing lightweight convolutional blocks into the encoder of the baseline model, replacing the original standard convolutions, significant performance improvements were observed, as shown in Table II. For the pinhole category, the IoU increased from 46.48% to 53.72%, while for the wrinkling category, the IoU rose from 53.76% to 59.07%. This demonstrates the crucial role of lightweight convolutional blocks in extracting effective features and detecting small defect targets.

Adding Attention Modules and Combinations: To evaluate the impact of different attention mechanisms on defect detection, the effects of using spatial attention, channel attention, and their combination within the skip connections were compared. As shown in Table II, when using channel attention (CA) alone, the model achieved ACC values of 84.62% and 71.82% for the wrinkling and abrasion categories, respectively, compared to 42.37% for the baseline, representing an approximately 70% improvement in ACC. The results indicate that the model performed best in recognizing subtle and easily confused wrinkling and abrasion categories, validating the ability of channel attention to focus on critical semantic information.

When using spatial attention (SA) alone, the IoU for the pinhole category was 67.02%, demonstrating a high IoU score in extracting localized small targets like pinholes. Spatial attention increased the IoU of the pinhole category from 46.48% in the baseline to 67.02%, an improvement of 44.2%, which is consistent with the design intent of spatial attention. When combining spatial attention and channel attention (CA+SA), the model achieved an IoU of 98.47% for the background category, and the IoU for the wrinkling category significantly improved from 53.76% in the baseline to 62.29%. This evidence highlights the complementary nature of spatial attention and channel attention in target recognition, allowing the model to balance global and local information when processing complex surface defects while maintaining computational efficiency.

Through the ablation study, we can conclude: 1) Lightweight convolutional blocks and attention modules are crucial for improving the model's segmentation accuracy and computational efficiency. 2) The combined use of spatial attention and channel attention enhances the model's ability to recognize small-area and easily confused defect categories while maintaining high computational efficiency.

D. Comparison Experiments

In this section, the proposed model is evaluated through comparative experiments against multiple open-source models, all of which are based on the baseline implementations provided by mmlab [15]. As shown in Table III, the experimental results demonstrate the superior performance of our model across multiple evaluation metrics. In terms of accuracy, the model achieves 73.54% mIoU, an improvement of 7.24 percentage points compared to the baseline model, and significantly outperforms other typical models such as DeepLabV3+ [2] and UNet [12]. This evidence substantiates that the combination of lightweight convolutions and dual attention mechanisms can enhance detailed recognition capabilities without relying on excessively large model sizes.

Moreover, the lightweight design of the model results in a parameter count of only 0.114M, significantly lower than existing models such as DeepLabV3plus, MobilenetV3 [4], UNet, and GCNet [14]. This supports the premise that the attention mechanism can effectively extract features from regions of interest, reducing unnecessary computational overhead. Furthermore, the model achieves an inference speed of 314.55 img/s, ensuring efficient operation in resource-constrained environments and meeting real-time application requirements. Figure 5 showcases the segmentation and detection results of the best-performing model on the validation set. Figure 1 provides a more intuitive observation of the model's superiority, highlighting its comprehensive advantages in accuracy, lightweight design, and speed.

V. Conclusion and Future Work

To address the challenges of surface defect detection in industrial production, this research proposes an encoder-decoder network that combines lightweight convolutions and attention mechanisms. Through a carefully designed lightweight structure and attention modules, the network effectively improves the recognition accuracy for small-sized defects and achieves real-time performance in multi-class defect detection tasks. Experimental results demonstrate the model's outstanding performance in addressing lightweight requirements and improving small defect detection accuracy, significantly outperforming other typical models.

Despite these achievements, our work still faces challenges in practical applications. The model's robustness in handling complex backgrounds requires further improvement. In future work, we plan to incorporate Transformer networks into the defect detection task. Through self-attention mechanisms, Transformers can capture long-range dependencies, enabling the model to better understand complex patterns in images, thereby further enhancing the recognition capability for small defects.

Acknowledgment

The research is partially supported by the Guangdong Provincial Science and Technology Innovation Strategic Fund Climbing Program, 2023.(No:PDJH2023a0714), and in part by the College Innovative Scientific Research Project of Guangdong(under Grant No. 2022KTSCX224).

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[2] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[3] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019.

[4] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for mobilenetv3," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019, pp. 1314–1324.

[5] B. Koonce and B. Koonce, "Efficientnet," *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, pp. 109–123, 2021.

[6] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1580–1589.

[7] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[8] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 534–11 542.

[9] L. Yang, S. Song, J. Fan, B. Huo, E. Li, and Y. Liu, "An automatic deep segmentation network for pixel-level welding defect detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–10, 2021.

[10] Y. Huang, C. Qiu, and K. Yuan, "Surface defect saliency of magnetic tile," *The Visual Computer*, vol. 36, pp. 85–96, 2020.

[11] H. Dong, K. Song, Y. He, J. Xu, Y. Yan, and Q. Meng, "Pga-net: Pyramid feature fusion and global context attention network for automated surface defect detection," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 12, pp. 7448–7458, 2019.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[15] M. Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," https://github.com/open-mmlab/mmsegmentation, 2020.