Proceedings of Machine Learning Research vol 278:1–10, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

An improved YOLOv11 algorithm for rice diseases

Tao Wang 202430310305@STU.SHMTU.EDU.CN
School of information engineering, Shanghai Maritime University, Shanghai, China

Changming Zhu* CMZHU@SHMTU.EDU.CN
School of information engineering, Shanghai Maritime University, Shanghai, China
**Corresponding author*

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

The timely identification of rice diseases is of vital importance to national food security. This paper proposes an improved model based on YOLOv11, and uses three key innovations to enhance the detection performance of the model. Firstly, DynamicConv enables the network to increase the number of parameters while maintaining a low number of floating-point operations (FLOPs), allowing these networks to benefit from large-scale visual pre-training. Secondly, the iterative attention feature fusion (iAFF) improves the detection accuracy by enhancing the feature fusion process. In addition, the Synergistic Cross-Scale Attention module (SCSA) is designed to effectively combine the advantages of channel and spatial attention, making full use of multi-semantic information, thus improving the performance of visual tasks. The experimental results show that the innovated model can effectively improve the detection efficiency of rice diseases, providing a reliable solution for agricultural security.

Keywords: Rice diseases; YOLOv11; DynamicConv; iAFF; SCSA

1. Introduction

Rice, as the staple food crop for more than half of the world's population, its production safety is directly related to national food security and social stability. However, during the growth cycle, rice is vulnerable to a variety of diseases such as rice blast, sheath blight, and bacterial leaf blight. These diseases are characterized by rapid spread and extensive harm. If not identified and controlled in a timely manner, they will lead to large-scale yield reduction or even complete crop failure. Traditional disease detection mainly relies on manual field inspections and experience-based judgments, which have problems such as low efficiency, high cost, and strong subjectivity, and it is difficult to meet the requirements of precision and real-time performance in modern agriculture. In recent years, computer vision technology based on deep learning has provided new ideas for the intelligent detection of agricultural diseases.

To break through the above bottlenecks, this paper proposes an improved rice disease detection model based on the YOLOv11 architecture, and achieves the coordinated optimization of detection accuracy and efficiency through three core technological innovations. Firstly, DynamicConv (Li et al., 2024) is introduced to dynamically adjust the parameters of the convolution kernel while maintaining low computational complexity (FLOPs), enhancing the model's adaptability to multi-scale disease features. At the same time, the generalization ability is improved through large-scale visual pre-training. Secondly, an iterative attention feature fusion module (iAFF (Goswami et al., 2023)) is designed to optimize the feature fusion process through multi-level feature interaction and

attention weighting mechanism, solving the problems of semantic information loss and noise interference in traditional methods. In addition, a Synergistic Cross-Scale Attention module (SCSA (Li et al., 2021)) is proposed to combine the advantages of channel attention and spatial attention, accurately locate the disease area and suppress background interference, and fully explore the representation potential of multi-semantic (Lu et al., 2021) features. The experimental results show that the improved model significantly surpasses existing benchmark methods on both the public rice disease dataset and the self-built field dataset, providing an efficient and reliable solution for real-time disease monitoring in complex farmland environments. This study not only provides technical support for the intelligent prevention and control of rice diseases, but also provides a transferable algorithm framework for the detection tasks of other crop disasters, which has important practical significance for promoting the intelligent transformation of agriculture and ensuring food security.

2. Methods and Principles

2.1. YOLOv11

YOLOv11 is the latest iterative version of the YOLO (You Only Look Once) series of object detection algorithms, which was released in 2024. Building on the advantages of its predecessors in real-time detection, it achieves a balanced optimization of accuracy and speed through multiple technological innovations (Riftiarrasyid et al., 2024), and it performs particularly well in small object detection and multi-task expansion. The structure is shown in Figure 1.

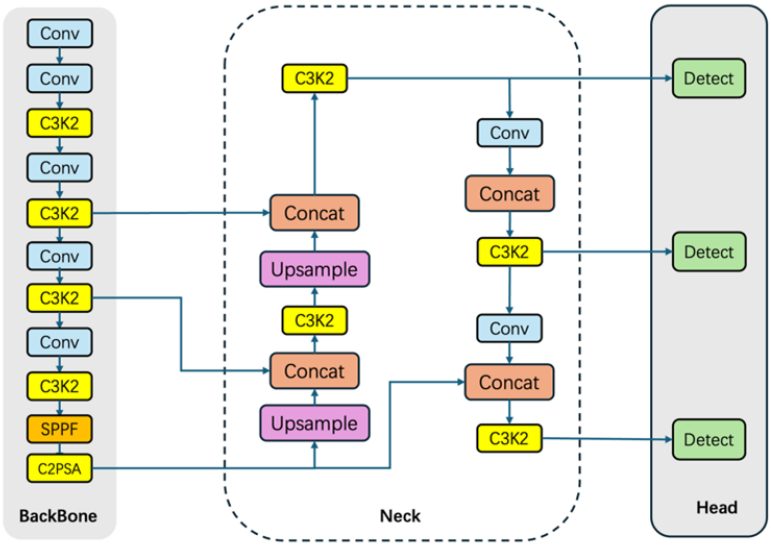

Figure 1: The original structure of YOLOv11

2.1.1. Innovative Design of the YOLOv11 Model Architecture

- **Improvement of the Backbone Network:** The C3k2 module is adopted to replace the traditional convolutional block. By stacking residual units, the feature extraction ability is enhanced. At the same time, the C2PSA module (Channel and Spatial Synergistic Attention)

is introduced to dynamically adjust the feature weights to focus on key areas. In addition, the backbone network supports the integration of the Transformer. Through the self-attention mechanism, it can capture long-range dependency relationships and improve the ability to understand the global context.

- **Optimization of the Neck Network:** The SPFF (Spatial Pyramid Fast Pooling) structure is used (Torres et al., 2022). By combining the multi-scale feature pyramid and cross-layer connections, the fusion efficiency of shallow detailed features and deep semantic features is strengthened, and the problem of missed detection of small objects is solved. Some variant models also introduce BiFPN (Weighted Bidirectional Feature Pyramid Network). Through the weighted fusion of multi-resolution features in a bidirectional path (Wu et al., 2024), the detection accuracy is further improved.
- **Multi-task Expansion of the Detection Head:** It supports five major tasks: detection, instance segmentation, pose estimation, oriented bounding box (OBB) detection, and classification. Among them, the segmentation head can generate high-precision masks through pixel-level prediction, and the pose detection head integrates a key-point regression network, which is suitable for scenarios such as human motion analysis.

2.2. Improvement of YOLOv11

2.2.1. THE STRUCTURE OF THE IMPROVED YOLOV11 MODEL

DynamicConv is adopted at the convolutional end, combined with large-scale visual pre-training; C2PSA is combined with Spatial and Channel Synergistic Attention (SCSA) to optimize the feature extraction process and improve the detection ability for small objects and diseases in complex backgrounds; Iterative Attention Feature Fusion (iAFF) is used to optimize the neck network, enhance the fusion effect of multi-scale features, and improve the detection accuracy. The structure diagram of the improved version is shown in Figure 2.

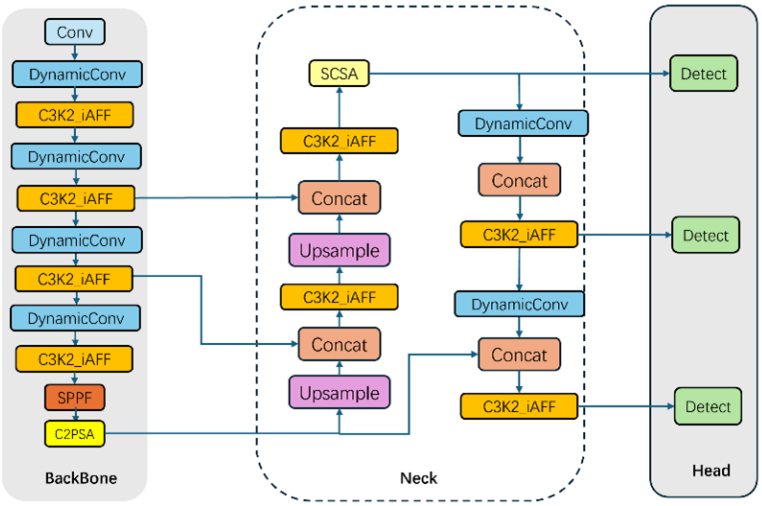

Figure 2: The improved structure

2.2.2. DYNAMICCONV (DYNAMIC CONVOLUTION)

DynamicConv solves the "low FLOPs trap" where low FLOPs models cannot benefit from large-scale pre-training by increasing the number of model parameters without significantly increasing the computational load (FLOPs) (Han et al., 2023). Dynamic convolution processes the input data by dynamically selecting or combining different convolution kernels for each input sample. This approach can be regarded as an extension of traditional convolution operations, which allows the network to adaptively adjust its parameters according to different inputs.

1.Discovery of the Low FLOPs Trap

Traditional low FLOPs models (such as lightweight CNNs/Transformers) perform poorly in large-scale pre-training (such as ImageNet-22K), while high FLOPs models can significantly benefit. This indicates that the insufficient number of parameters limits the ability of low computational models to utilize large amounts of data.

2.The Design Principle of Prioritizing Parameters

It is proposed that parameters are more important than FLOPs: Model performance is positively correlated with the number of parameters, because more parameters can capture more complex features. ParameterNet breaks the positive correlation between parameters and FLOPs by dynamically increasing the number of parameters (rather than the computational load).

3.Parameter Enhancement Achieved by Dynamic Convolution

Dynamic Convolution Layer: Replace the traditional convolution weights with a dynamically weighted combination of multiple "expert" weights ($X \in \mathbb{R}^{C_{\text{in}} \times H \times W}, W \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times K \times K}$ and the formula is as follows).

$$Y = X * W', \tag{1}$$

$$W' = \sum_{i=1}^{M} \alpha_i W_i$$

Dynamic Weight Generation: Weight coefficients are generated according to the input features through a lightweight Multi-Layer Perceptron (MLP) (Formula 2), which hardly increases the FLOPs.

$$\alpha = softmax(MLP\left(Pool\left(X\right)\right)) \tag{2}$$

Effect: The number of parameters is increased by M times (where M is the number of experts), while the FLOPs only increase slightly (for example, when $M = 4$, the FLOPs increase by about 1%).

2.2.3. IAFF (ITERATIVE ATTENTIONAL FEATURE FUSION)

The main idea is to improve the detection accuracy by enhancing the feature fusion process. A unified and general scheme - attentional feature fusion - is applicable to most common scenarios, including feature fusions triggered by short skip connections, long skip connections, and within the Inception layer.

To better fuse features with inconsistent semantics and mismatched scales, a multiscale channel attention module is proposed to specifically address the problems arising from multiscale feature fusion (Dai et al., 2020). The initial integration of feature maps can be a bottleneck, and this problem can be effectively alleviated by introducing the multiscale channel attention module (referred to as iterative attentional feature fusion). (The illustration of iAFF is shown in Figure 3)

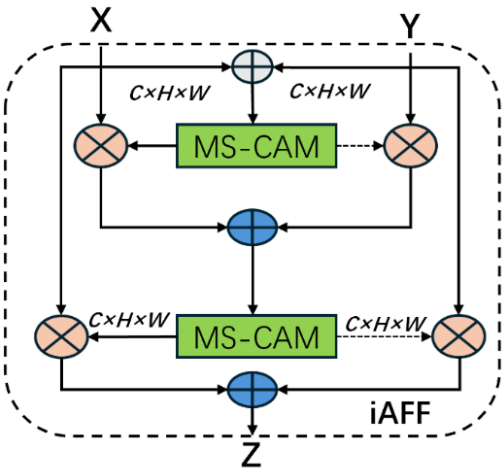

Figure 3: The illustration of iAFF

2.2.4. SCSA SYNERGISTIC ATTENTION MODULE

Traditional hybrid attention mechanisms (e.g., CBAM, CPCA) struggle to fully exploit multi-semantic guidance and overlook feature semantic differences, limiting their effectiveness in fine-grained tasks. To address this, the novel Spatial and Channel Synergistic Attention (SCSA) module introduces key innovations: Shared Multi-Semantic Spatial Attention (SMSA), which decomposes input features into four sub-features processed by multi-scale depthwise separable convolutions (kernels: 3,5,7,9) to capture hierarchical semantics, followed by Group Normalization to preserve semantic independence and suppress redundancy (Si et al., 2024); and (The structure is shown in Figure 4.)

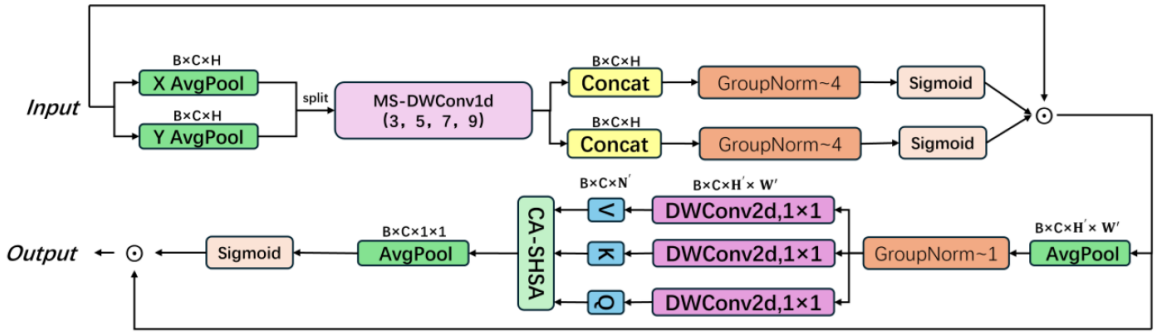

Figure 4: The structure of SCSA

3. Experiments and Analysis

3.1. Experimental Environment Configuration

The operating system used in this experiment is Windows 11. The GPU is NVIDIA GeForce RTX 5070ti, the CPU is 14th Gen Intel(R) Core(TM) i5-14600KF with a main frequency of 3.50GHz,

and the system is equipped with 64GB of memory. In the software environment, PyTorch 2.8.0, Python 3.10.15, and CUDA 12.8 are used.

3.2. Datasets

The dataset used in this experiment is Corp diseases Detection. This dataset contains a total of 8,334 images, which are divided into a training set, a validation set, and a test set according to a ratio of 8:1:1. There are four categories, namely bacterial blight, blast, brown spot, and tungro. Figure 5 shows some of the photos in the dataset.

Figure 5: Some of the photos in the dataset

3.3. Model Evaluation Indicators

The mean average precision (mAP50) is used as the evaluation indicator for the accuracy of the model. The mAP is obtained by calculating the average precision (AP) of each category and then calculating the average value of the average precisions of all categories. Parameters (Params) and FLOPS are used as the evaluation indicators for the complexity of the model. The calculation

process of mAPA50 is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{5}$$

The model parameters (Params) refer to the total number of parameters during the model training process, including weights, bias values, etc. This indicator measures the spatial complexity and size of the model. The model's computational power (GFLOPs) is the number of floating-point operations performed by the model during a single forward propagation process, usually measured in billions of floating-point operations per second.

3.4. Experimental Results of the Improved Model

During the training process, it is shown that the bounding box regression loss, classification loss, and distribution focal loss (presumably) all decrease with the number of training epochs. This indicates that the model is continuously learning and optimizing on the training set, and its capabilities in predicting the target position, judging the category, and predicting relevant attributes are gradually improving. The corresponding losses on the validation set also show a downward trend, and the trend is similar to that of the training set losses. This shows that the performance of the model on new data is also getting better, and there is no serious overfitting. (The training results are shown in Figure 6)

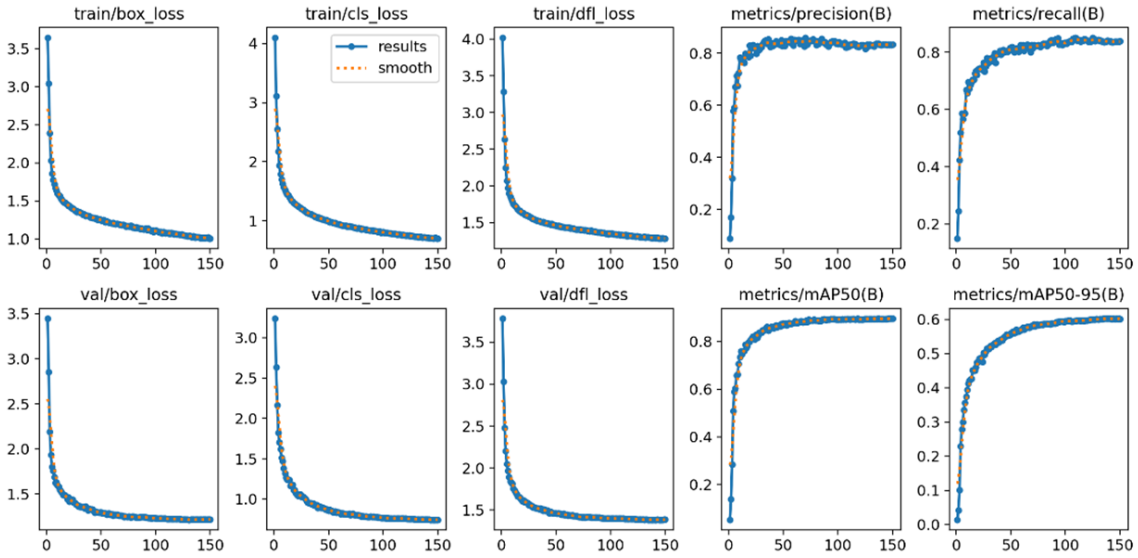

Figure 6: Experimental result figures

"F1 - Confidence Curve" demonstrates the changes in the F1 values of categories such as bacterial blight, brown spot, tungro, and all categories as the confidence level varies. The F1 value first increases and then decreases, with different trends for different categories.

Precision-recall curve, where the horizontal axis represents the recall rate and the vertical axis represents the precision rate. Curves of different colors correspond to categories such as bacterial blight, brown spot, tungro, and all categories, each having its own mean average precision (mAP). The curve shows that as the recall rate increases, the precision rate decreases, reflecting the differences in the trade-off between precision and recall for different categories and also reflecting the detection performance of the model. (The F1 - Confidence Curve and Precision-recall curve are shown in Figure 7)

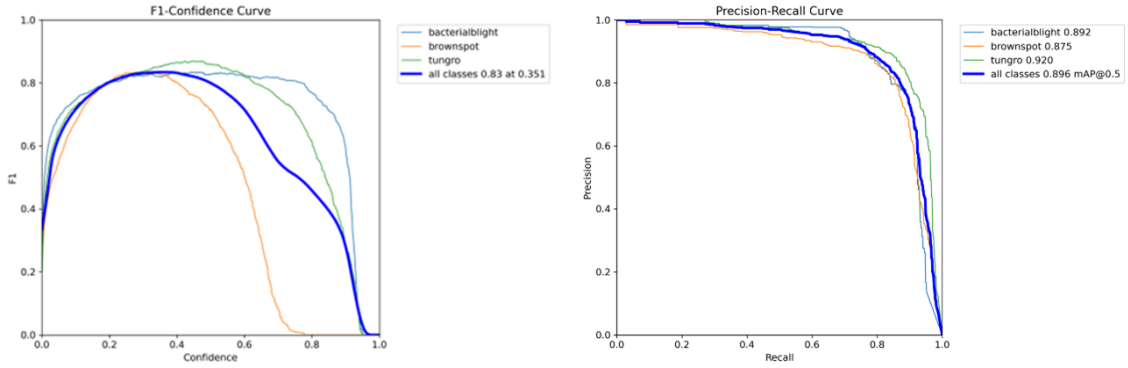

Figure 7: F1 - Confidence Curve and Precision-recall curve

3.5. Ablation Experiments

Ablation experiments systematically evaluate each component's contribution in our improved model by removing DynamicConv, iAFF, and SCSA modules. Results show: (1) Removing DynamicConv reduces multi-scale adaptability, lowering small-target detection accuracy; (2) Without iAFF, feature fusion weakens, increasing semantic loss and noise; (3) Omitting SCSA hampers multi-semantic utilization, degrading localization precision. By comparing mAP50, Params, and GFLOPs, these experiments validate each module's necessity and guide model optimization for rice disease detection.

From Table 1, it can be seen that after integrating DynamicConv into the YOLOv11 model, the mAP50 increases by 0.2%, and the mAP50-90 increases by 0.4%. After integrating iAFF, the mAP50 increases by 0.1%, and after integrating SCSA, the mAP50 increases by 0.1%. When integrating DynamicConv, iAFF, and SCSA simultaneously, the mAP50 increases by 0.4%, indicating a significant improvement in the detection performance of rice diseases. However, it also increases the number of model parameters and requires more computing resources.

Table 1: Experimental result table

Yolov11	DynamicConv	iAFF	SCSA	Precision	Recall	mAP50	mAP50-90	Params	GFLOP
✓				0.84	0.833	0.892	0.601	2582932	6.3
✓	✓			0.823	0.855	0.894	0.605	4588876	4.7
✓		✓		0.827	0.843	0.893	0.6	2631836	6.5
✓			✓	0.828	0.849	0.893	0.598	2583956	6.3
✓	✓	✓		0.84	0.845	0.895	0.6	4637780	4.9
✓	✓	✓	✓	0.832	0.837	0.896	0.602	4638804	4.9

4. Conclusion

This study takes the detection of rice diseases as the application scenario and makes three innovative improvements to the YOLOv11 model: Adopting DynamicConv at the convolutional end and combining it with large-scale visual pre-training to enhance the model's adaptability to different disease features; integrating the attention mechanisms of C2PSA and SCSA: Combining C2PSA with Spatial and Channel Synergistic Attention (SCSA) to optimize the feature extraction process and improve the detection ability for small targets and diseases in complex backgrounds; improving the feature fusion strategy: Using Iterative Attention Feature Fusion (iAFF) to optimize the neck network, enhancing the fusion effect of multi-scale features, and improving the detection accuracy. The experimental results show that the improved model has a 0.4% improvement in the mAP50 index compared with the original YOLOv11, significantly improving the accuracy of rice disease detection. At the same time, it provides new optimization ideas for the research in the field of object detection.

Acknowledgments

This work is supported by National Natural Science Foundation of China (CN) [62276164, 61602296], 'Science and technology innovation action plan' Natural Science Foundation of Shanghai [22ZR1427000], and Shanghai Oriental Talent Program-Youth Program. The authors would like to thank their supports.

References

Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard. Attentional feature fusion. *arXiv preprint arXiv:2009.14082*, 2020.

S. Goswami, K. Ashwini, and R. Dash. Grading of diabetic retinopathy using iterative attentional feature fusion (iaff). In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5, 2023. doi: 10.1109/ICCCNT56998.2023.10307892.

K. Han, Y. Wang, J. Guo, and E. Wu. Parameternet: Parameters are all you need for large-scale visual pretraining of mobile networks. *arXiv preprint arXiv:2306.14525v2*, 2023.

P. Li, E. Piliouras, V. Poghosyan, M. AlHameed, and T.-M. Laleg-Kirati. Automatic detection of epileptiform eeg discharges based on the semi-classical signal analysis (scsa) method. In *2021*

43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pages 928–931, 2021. doi: 10.1109/EMBC46164.2021.9631028.

S. Li, R. Jia, and J. Dai. Dgl-yolov8 a lightweight yarn detection algorithm. In *2024 China Automation Congress (CAC)*, pages 6410–6414, 2024. doi: 10.1109/CAC63892.2024.10865747.

D. Lu, X. Liao, F. Xu, and J. Bai. Anomaly detection method for substation equipment based on feature matching and multi-semantic classification. In *2021 6th Asia Conference on Power and Electrical Engineering (ACPEE)*, pages 109–113, 2021. doi: 10.1109/ACPEE51499.2021.9437096.

M. F. Riftiarrasyid, F. L. Gaol, H. Soeparno, and Y. Arifin. Suitability of latest version of yolov11 in drone development studies. In *2024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 504–509, 2024. doi: 10.1109/ISRITI64779.2024.10963542.

Y. Si, H. Xu, X. Zhu, W. Zhang, Y. Dong, Y. Chen, and H. Li. Scsa: Exploring the synergistic effects between spatial and channel attention. *arXiv preprint arXiv:2407.0512857*, 2024.

C. Torres, J. M. Blanes, A. Garrigós, D. Marroquí, and J. A. Carrasco. Single point failure free interleaved synchronous buck converter for microsatellite electrolysis propulsion. *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 10(5):5371–5380, 2022. doi: 10.1109/JESTPE.2022.3174358.

M. Wu, H. Dai, K. Yao, T. Tuytelaars, and J. Yu. Bg-triangle: Bézier gaussian triangle for 3d vectorization and rendering. *arXiv preprint arXiv:2407.05128*, 2024.