Axial-iFormer: Robust and Efficient Wafer Map Defect Classification via Modulated Axial Attention

Jaeho Song
Dept. of Intelligent Electronics and Computer Engineering
Chonnam National University
GwangJu, Korea
sjh990110@jnu.ac.kr

Juhyeon Noh
Dept. of Intelligent Electronics and Computer Engineering
Chonnam National University
GwangJu, Korea
uoahynz@jnu.ac.kr

Jihoon Lee
Dept. of Intelligent Electronics and Computer Engineering
Chonnam National University
GwangJu, Korea
suda34@jnu.ac.kr

Yonggwan Won
Dept. of Intelligent Electronics and Computer Engineering
Chonnam National University
GwangJu, Korea
ykwon@jnu.ac.kr

Jaehyung Park
Dept. of Intelligent Electronics and Computer Engineering
Chonnam National University
GwangJu, Korea
hyeoung@jnu.ac.kr

Jinsul Kim
Dept. of Intelligent Electronics and Computer Engineering
Chonnam National University
GwangJu, Korea
jsworld@jnu.ac.kr

***Abstract*— Reliable classification of wafer map defect patterns helps process engineers identify yield-limiting issues in semiconductor fabrication plants. Hybrid CNN–Transformer models such as iFormer, which combine local convolutions with global self-attention, have recently achieved strong results on this task. However, iFormer's Single-Head Modulation Attention (SHMA) still performs expensive global attention over all spatial tokens and provides little inductive bias for thin, line-shaped defects such as scratches. We introduce Axial-iFormer and a lightweight companion model that replace SHMA with a Modulated Axial Attention block. The new block factors 2D attention into row-wise and column-wise operations and modulates the resulting context through a gating branch, preserving long-range dependencies while lowering computational cost. On the WM-811K wafer map benchmark, the performance-oriented Axial-iFormer-S increases Scratch Recall to 82.85%, a gain of 3.35 percentage points over the iFormer-S baseline. The efficiency-oriented Axial-iFormer-S-Lite reduces the parameter count by 29% yet still surpasses the baseline on overall defect detection, indicating that Axial-iFormer is well suited to yield monitoring scenarios with tight computational budgets.**

Keywords— Axial Attention, Computer Vision, Defect Classification, iFormer, Wafer Map

I. INTRODUCTION

Modern semiconductor fabrication at nanometer technology nodes is highly sensitive to small process variations, so even subtle drifts can translate into substantial yield loss. To analyze such issues, engineers inspect wafer bin maps, two-dimensional images that encode the pass/fail status of individual dies. Characteristic spatial patterns on these maps are studied under Wafer Map Pattern Recognition (WMPR) and are strongly associated with specific tools or process steps, which makes automatic pattern classification an important component of yield analysis. Early WMPR systems relied on visual inspection or simple statistical rules, but the growing scale and diversity of wafer data have led researchers to adopt deep learning models instead.

Deep neural networks for wafer maps increasingly adopt hybrid designs that mix Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs). iFormer is a representative model in this category: it augments convolutional features with a global context branch based on Single-Head Modulation Attention (SHMA). SHMA keeps the attention module compact, but it still performs global context modeling over all spatial tokens at once. This full-image processing is costly for edge devices and, more importantly, treats the feature map as a single sequence without explicitly encoding axis-aligned structures. As a result, thin linear defects such as scratch patterns may be under-represented, causing false negatives in yield-critical situations.

To address these challenges, we propose Axial-iFormer, a specialized architecture designed for robust and efficient wafer map classification. Our approach replaces the standard global attention in iFormer with Modulated Axial Attention. This modification introduces two key contributions:

1. Efficient Axial Decomposition: We decompose the computationally expensive global attention into consecutive 1D axis-wise (row and column) operations. This structural change significantly reduces the computational overhead while enhancing the model's sensitivity to horizontal and vertical defect patterns, such as scratches.
2. Stabilized Learning via Modulation: A naive application of Axial Attention often leads to information loss and training instability. We resolve this by incorporating the gating mechanism of iFormer. This integration ensures that the efficient axial features are robustly modulated, preventing performance degradation. Experimental results on the WM-811K dataset show that our approach improves Scratch Recall by 3.35 percentage points compared to the baseline, while our lightweight variant (S-Lite) reduces parameters by 29%, demonstrating a superior trade-off between accuracy and efficiency.

The remainder of this paper is organized as follows. Section II reviews related works on wafer map pattern recognition and efficient vision architectures. Section III details the proposed Axial-iFormer architecture, focusing on the Modulated Axial Attention mechanism. Section IV presents the experimental setup, comparative results, and efficiency analysis. Finally, Section V concludes the paper with directions for future research.

II. Related Work

A. Wafer Map Pattern Recognition

Early approaches to WMPR primarily relied on manual feature extraction combined with traditional machine learning algorithms, such as clustering and density-based methods [1]. However, these methods often struggled to generalize to the increasing diversity and complexity of defect patterns in large-scale fabrication data. With the advent of deep learning, CNNs became the standard for wafer map classification due to their strong capability in extracting local spatial features [2],[3]. Despite their success, standard CNNs have a limited receptive field, which makes it challenging to capture global dependencies required for distinguishing complex or spatially spread-out defect patterns across the entire wafer.

B. Efficient Hybrid Vision Architectures

To overcome the locality limitations of CNNs, recent studies have introduced hybrid architectures that incorporate ViTs [4] and hierarchical designs like Swin Transformer [5]. A notable state-of-the-art model in this line of work is iFormer [6], which effectively combines the local representation power of CNNs with the global modeling ability of self-attention. iFormer utilizes a SHMA mechanism to capture global context while mitigating the memory overhead of standard Multi-Head Attention. However, its global attention mechanism still processes all spatial tokens simultaneously, which remains computationally intensive for resource-constrained edge applications [7]. Furthermore, treating the feature map as a unified sequence lacks the specific inductive bias needed to detect fine-grained linear defect patterns, such as scratches, which often appear along specific axes.

C. Axial Attention Mechanism

Axial Attention was proposed to improve the efficiency of self-attention on multidimensional data [8]. Instead of applying attention to the entire flattened feature map, this mechanism decomposes the operation into two separate 1D attention steps along the height (row) and width (column) axes. This factorization allows the model to capture long-range dependencies with significantly reduced computational redundancy compared to global attention. In this work, we adopt this mechanism to efficiently capture the horizontal and vertical correlations typical of scratch defects. Crucially, we integrate it with the gating mechanism of iFormer to resolve the training instability that can occur when applying raw Axial Attention to sparse wafer data.

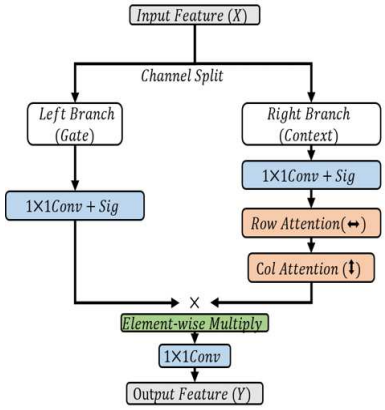

Fig. 1. Structure of the Proposed Modulated Axial Attention Block

III. Methodology

This section describes the proposed Axial-iFormer. We adopt the hierarchical pyramidal structure of the original iFormer to effectively process multi-scale defect patterns. The core innovation lies in the later stages, where we replace the original SHMA blocks with our proposed Modulated Axial Attention blocks.

A. Proposed Modulated Axial Attention Block

The Modulated Axial Attention module, illustrated in Fig. 1, is designed to replace the global attention in the original SHMA block and to address its limitations in detecting fine-grained linear defects. Distinct from the original design, the input features are first processed by a 1×1 convolution and Sigmoid activation, then split into a Gate branch and a Context branch. Instead of processing all spatial tokens simultaneously, this module decomposes the context modeling into two consecutive 1D operations: Row Attention followed by Column Attention.

- Axial Decomposition: This factorization allows the model to explicitly capture long-range dependencies along the horizontal and vertical axes, which directly corresponds to the geometric characteristics of scratch defects. Furthermore, this structural change significantly reduces the computational redundancy compared to standard all-to-all attention mechanisms.
- Gated Modulation: This efficient axial context is then integrated using the existing gating (modulation) mechanism of the SHMA block. The feature branch (Gate) and the context branch (Axial Attention output) are fused via element-wise multiplication, ensuring that the axial context is robustly incorporated into the final features, thereby stabilizing the training process. Empirically, removing this gating mechanism and using Axial Attention alone reduces Scratch Recall from 82.85% to 70.29%, confirming that modulation is essential for stable training on sparse wafer maps.

B. Model Variants

We propose two variants of Axial-iFormer to address different deployment scenarios:

- Axial-iFormer-S (Standard): This performance-oriented variant keeps the same channel configuration as the baseline iFormer-S, preserving feature expressiveness with a particular focus on recall for linear scratch patterns.
- Axial-iFormer-S-Lite (Lightweight): Designed for real-time edge monitoring, this variant strategically reduces the channel width in the deeper stages. As shown in our experiments, this reduction lowers the parameter count by 29% while preserving the structural advantages of the axial mechanism.

IV. Experiments

A. Experimental Setup

To ensure a fair and rigorous evaluation, we conducted experiments on the WM-811K dataset, which consists of real-world wafer maps. We removed unlabeled data and applied a stratified split, using 80% of the data for training and 20% for testing. Crucially, to guarantee that performance gains are derived solely from architectural improvements rather than data preprocessing, we applied an identical set of 7 custom

augmentations (including random rotation, random erasing, and Gaussian blur) to both the baseline and our proposed models.

To quantitatively evaluate the model performance, we utilize standard classification metrics: Precision, Recall, and F1-score. These metrics are defined as follows :

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN} \quad (1)$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2)$$

Here, TP, FP , and FN denote the number of true positives, false positives, and false negatives, respectively. In our experiments, considering the class imbalance, we report the Macro F1-score to treat all defect types equally.

B. Main Results: Robustness in Defect Classification

Table I summarizes the overall performance on the 9-class classification task. Our primary objective is to improve the detection rate (Recall) of 'Scratch' defects, which are typically fine-grained linear patterns that global attention mechanisms often fail to capture.

As shown in Table I, our performance-oriented model, Axial-iFormer-S, achieved a Scratch Recall of 82.85%, outperforming the baseline iFormer-S (79.50%) by 3.35 percentage points. In addition to scratches, Axial-iFormer-S also improved recall for other geometric defect patterns, such as Edge-Loc (+0.67 percentage points) and Donut (+0.90 percentage points). Taken together, these gains suggest that decomposing global attention into axial components strengthens the inductive bias for linear features and helps reduce false negatives in yield management.

We further investigate the role of the gating mechanism by training an Axial-only variant that replaces SHMA with pure Axial Attention while removing the modulation gate. Under the same training setup on WM-811K, this variant achieved an overall accuracy of 96.76%, a macro F1-score of 0.8365, and a Scratch Recall of 70.29%, all lower than both iFormer-S (97.80% / 0.9044 / 79.50%) and our Axial-iFormer-S (97.80% / 0.9067 / 82.85%). This result indicates that naive Axial Attention alone can even degrade performance on sparse wafer maps, and that the gating mechanism is crucial for stabilizing Axial Attention and achieving high recall on scratch defects.

TABLE I. OVERALL PERFORMANCE COMPARISON

Model	*Overall Acc (%)*	*Macro F1*	*Scratch Recall (%)*
iFormer-S	97.80	0.9044	79.50
Axial-iFormer-S	97.80	0.9067	82.85
Axial-iFormer-S-Lite	97.77	0.9058	82.43

C. Efficiency Analysis

For deployment in resource-constrained edge devices, model efficiency is as critical as performance. Our lightweight variant, Axial-iFormer-S-Lite, was designed by reducing channel dimensions in deeper stages. To validate its efficiency, we measured the number of parameters, FLOPs (using fvcore library), and inference latency on an NVIDIA H100 Tensor Core GPU with an input resolution of 224×224. As shown in Table II, the S-Lite model has only 4.43M parameters and 0.79G FLOPs. This represents a 29.1% reduction in parameters and a 28.2% reduction in computational cost compared to the baseline iFormer-S (6.25M parameters, 1.10G FLOPs). Furthermore, it achieves an inference latency of 4.69 ms, which is faster than the baseline (4.95 ms). This confirms that the proposed architecture offers a superior trade-off, maintaining computational efficiency suitable for real-time edge monitoring.

TABLE II. MODEL COMPLEXITY AND INFERENCE

Model	*Params (M)*	*FLOPs (G)*	*Latency (ms)*
iFormer-S	6.25	1.10	4.95
Axial-iFormer-S	8.44	1.28	6.76
Axial-iFormer-S-Lite	4.43	0.79	4.69

D. Robustness Verification

To further verify the robustness of our approach, we evaluated the models on a binary classification task (Scratch vs. Non-scratch). In this scenario, Axial-iFormer-S achieved a Scratch Precision of 93.97%, an improvement of 3.36 percentage points over the baseline (90.61%). This indicates that our model not only detects more defects (high recall) in complex multi-class scenarios but also distinguishes them more accurately (high precision) in targeted tasks, indicating its robustness against false positives.

V. CONCLUSION

In this paper, we proposed Axial-iFormer, an iFormer-based architecture tailored to wafer map defect classification. Motivated by the difficulty of capturing thin, axis-aligned defects with standard global attention, we replaced the SHMA block with a Modulated Axial Attention module that factorizes 2D context modeling into row-wise and column-wise attention and injects this axial context through the original gating branch.

On the WM-811K benchmark, the performance-oriented Axial-iFormer-S improves Scratch Recall by 3.35 percentage points over the iFormer-S baseline, while keeping overall accuracy and Macro F1 at a comparable level. In addition, Axial-iFormer-S increases recall for other geometric defect patterns such as Edge-Loc and Donut, indicating that axial decomposition strengthens the inductive bias for linear and edge-aligned structures. The lightweight Axial-iFormer-S-Lite reduces the number of parameters by about 29.1% and lowers FLOPs and latency, yet still achieves slightly higher Macro F1 and Scratch Recall than the baseline, making it more suitable for deployment on resource-constrained tools.

Despite these gains, not all defect types benefit equally from the proposed design, and improvements for several classes with more irregular spatial patterns remain modest compared to scratches. As future work, we plan to investigate a hybrid design that combines window-based attention, which is effective for local feature extraction, with axial attention for long-range axis-wise context. We also aim to evaluate Axial-iFormer and its variants in real in-line monitoring environments to study their robustness under process drifts and unseen defect patterns.

ACKNOWLEDGMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local

Intellectualization program grant funded by the Korea government(MSIT)(IITP-2025-RS-2022-00156287)

REFERENCES

[1] M. J. Wu, J. S. R. Jang, and J. L. Chen, "Wafer map failure pattern recognition and similarity ranking for large-scale data sets," *IEEE Transactions on Semiconductor Manufacturing*, vol. 28, no. 1, pp. 1–12, Feb. 2015.

[2] E. Shin and C. D. Yoo, "Efficient Convolutional Neural Networks for Semiconductor Wafer Bin Map Classification," *Sensors*, vol. 23, no. 4, p. 1926, Feb. 2023.

[3] H. Zheng, S. W. A. Sherazi, S. H. Son, and J. Y. Lee, "A Deep Convolutional Neural Network-Based Multi-Class Image Classification for Automatic Wafer Map Failure Recognition in Semiconductor Manufacturing," *Applied Sciences*, vol. 11, no. 20, p. 9769, Oct. 2021.

[4] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *In International Conference on Learning Representations (ICLR)*, 2021.

[5] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," *in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10012-10022.

[6] D. Qin et al., "MobileNetV4 -- Universal Models for the Mobile Ecosystem", *In European Conference on Computer Vision*, 2024, pp.78-96.

[7] C. Zheng, "iFormer: Integrating ConvNet and Transformer for Mobile Application," *In International Conference on Learning Representations (ICLR)*, 2025.

[8] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," *arXiv preprint arXiv:1912.12180,* 2019.