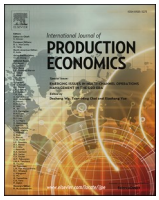
Int. J. Production Economics 273 (2024) 109275

Contents lists available at ScienceDirect

International Journal of Production Economics

journal homepage: www.elsevier.com/locate/ijpe



A new ViT-Based augmentation framework for wafer map defect classification to enhance the resilience of semiconductor supply chains

Shu-Kai S. Fan [a,b,*], Shang-Hao Chiu [a]

[a] Department of Industrial Engineering and Management, National Taipei University of Technology, Taipei, 106344, Taiwan, ROC
[b] Department of Industrial Engineering and Management, Yuan Ze University, Taoyuan, 32003, Taiwan, ROC

ARTICLE INFO

Keywords:
Semiconductor supply chain
Vision transformer (ViT)
Data augmentation
Wafer defect pattern classification
Deep learning

ABSTRACT

Wafer map defect classification plays a crucial role in sustaining the semiconductor supply chain during industrial disruptions by ensuring continuity, resilience, and efficiency while aligning with sustainability principles. Up to 30 percent of production costs is lost to chip testing and yield losses for semiconductor manufacturing supply chain. In the semiconductor practice, wafer map defects recognition plays a linchpin role in the front-end-of-line stage. Defect pattern recognition can directly pinpoint the assignable causes and provide the domain experts with actionable insights. However, various wafer defect types occur differently from each other, which makes the collected wafer map dataset to be highly imbalanced in defect classes. In the face of data imbalance, the classification model usually cannot provide satisfactory classification performance. In this aspect, this paper intends to investigate the wafer defect map classification problem by using Vision Transformer (ViT) as an alternative data augmentation approach. The primary purpose is to alleviate the class-imbalance issue and then the performance of a deep learning-based convolutional neural network for wafer defect classification can be effectively improved. The experimental results demonstrate that the proposed data augmentation method by using ViT proves to be a potential generative model for improving wafer map defect classification, particularly robust on the individual minority class. In a word, the proposed augmentation framework for wafer map defect classification facilitates a more targeted and efficient approach to quality control, resource utilization, and production optimization in maintaining a sustainable semiconductor supply chain, particularly in times of industrial disruption.

1. Introduction

The global electronics industry is currently being under rapid development and its technological innovations are accelerating with full speed. Integrated circuits (ICs) play an important role of electronic devices. With the advancement of science and technology, the requirements for chip production are also getting increasingly stricter due to Moore's Law. The demand for high-speed computing chips such as state-of-the-art applications of 5G, Internet of Things (IoT), smart edge, and artificial intelligence (AI) continues to inject growth momentum into the semiconductor industry. To collectively improve products' performance and cost-effectiveness throughout the semiconductor supply network, the coopetition among the different-tier suppliers may take place for orders and contracts between each other (Katsaliaki et al., 2024). The semiconductor manufacturing process is, in essence, highly complex as it involves a series of sophisticated inter-layer fabrication steps to create the electrical properties on the semiconductor materials within supply chain networks. One of the key challenges to the semiconductor supply chain is chip production processing time, i.e., the cycle time. The time between initial processing and the final product takes weeks and even more than one month. In the meantime, a job processing tool is often scheduled to go through a maintenance operation with an objective of minimizing risks that could affect wafer quality (Yu and Han, 2021). During the period, up to 30 percent of production costs is lost to testing and yield losses (Fan et al., 2023c). Therefore, ensuring production resource management is not only crucial for coping with the potential risks posed by geopolitical constraints or barriers, regulatory changes, and environmental differentiation, but also for the long-term viability and resilience of industries. Making the typical resource management practice to be sustainable aims to mitigate the aforementioned risks by adopting strategies that promote responsible usage, conservation, and replenishment of production resources. By prioritizing

* Corresponding author. Department of Industrial Engineering and Management, National Taipei University of Technology, Taipei, 106344, Taiwan, ROC.
E-mail address: morrisfan@ntut.edu.tw (S.-K.S. Fan).

https://doi.org/10.1016/j.ijpe.2024.109275
Received 2 January 2024; Received in revised form 22 April 2024; Accepted 8 May 2024
Available online 13 May 2024

sustainable resource management, the semiconductor industry can enhance its resilience, mitigate risks related to resource scarcity or domestic tensions, and contribute to a more sustainable and responsible future.

By supporting continuity, resource optimization, remote monitoring, adaptive practices, waste reduction, supplier relationships and innovation, wafer map defect classification significantly contributes to sustaining the semiconductor supply chain during industrial disruptions. Its role is of primary essence in ensuring that semiconductor manufacturing remains efficient, resilient, and aligned with sustainability objectives during challenging times. In order to enhance productivity and increase production yields, the establishment of an effective wafer defect recognition system is of primary importance in the advanced process control (APC) practice. Wafer maps are thus created, after the wafer acceptance test (WAT) process, which provides visual, insightful information about potential production irregularities in the processing steps. Therefore, the central element of wafer defect recognition is to pinpoint the root causes of defects and deploy corrective actions to prevent further wafer nonconformities. The main goal of this research is to develop a novel image augmentation framework of wafer maps to enhance the performance of defect classification, given that the numbers of different wafer map patterns are extremely imbalanced.

This paper re-visits the well-known WM-811K dataset (Wu et al., 2015), which contains totally 811,457 defect maps and has 9 different classes of wafer defect patterns. Specifically, there are *Center*, *Donut*, *Edge-Loc*, *Edge-Ring*, *Loc*, *Near-full*, *Random*, *Scratch*, and *None*), which are illustrated from left to right, top to bottom, respectively, in Fig. 1. However, there are only 172,950 maps in the original dataset that had been manually annotated by domain experts. Particularly, the numbers of sample maps between defect classes are extremely imbalanced. Only 25,519 wafers had been identified as distinguishable failure patterns; while 147,431 wafers could be at best only annotated as pattern *None*. The number of *Center* is 4,294, *Donut* is 555, *Edge-Loc* is 5,188, *Edge-Ring* is 9,680, *Loc* is 3,593, *Near-full* is 148, *Random* is 867, *Scratch* is 1193 and *None* is 147,432. The remaining 638,507 samples are all unlabeled wafer maps. Class-imbalance issue in the WM-811K dataset affects the classification performance significantly (Wang et al., 2019, 2021; Tsai and Lee, 2020; Saqlain et al., 2020; Yu et al., 2021; Yu and Liu, 2021; Fan et al., 2023b). Towards that end, many researches took advantage of data augmentation to mitigate the undue difficulty of data imbalance and showed that data augmentation improved the performance of defect classification. Note that even using data augmentation techniques for alleviating data imbalance, class *None* will serve as the main influential factor since it accounts for 85.2% of the original labeled data. The classification accuracy obtained could therefore be misleading since it dominates six-seventh percentage of the labeled data. A related research of a hybridized data mining method for investigating the patterns of wafer bin maps (WBMs) was addressed in Hsu and Chien (2007). For a comprehensive understanding of wafer map defect classification using WM-811K, readers are encouraged to explore the seminal works in this field (Wu et al., 2015; Tello et al., 2018; Kyeong and Kim, 2018; Wang et al., 2019; Wang and Chen, 2019; Yu et al., 2019; Yu, 2019; Kong and Ni, 2020; Tsai and Lee, 2020; Saqlain et al., 2020; Shon et al., 2021).

Recently, deep learning (DL) has become a state-of-the-art (SOTA) methodology in many practical applications. Particularly in computer vision, convolutional architectures are really best suited to image classification and recognition. More recently, in another field, natural language processing (NLP) tasks, the transformer presented a great breakthrough. The transformer was proposed by Vaswani et al. (2017). The architecture of the transformer includes an encoder and a decoder. A critical mechanism inside is called self-attention. It collects the correlation information between the input sequences via the calculations of query, key and value, as will be introduced shortly in a later section. The transformer performs very well in the NLP field; however, its application in the field of computer vision is quite limited.

Later, a new image-based framework of the transformer was proposed, which can be directly applied to sequences of image patches and perform very well without recourse to convolutional architectures on image classification assignments (Dosovitskiy et al., 2021). This proposal was given a unique and innovative name—the Vision Transformer (ViT) model. Upon the advent of the ViT model, potential research works by applying the ViT model in various fields were evaluated. In the medical field, ViT was applied to X-ray detection to determine whether it can detect the presence of COVID-19. By means of fine-tuning the ViT model, an overall classification accuracy rate of 97.61% was attained, showing the promise of applying the ViT model to the medical field

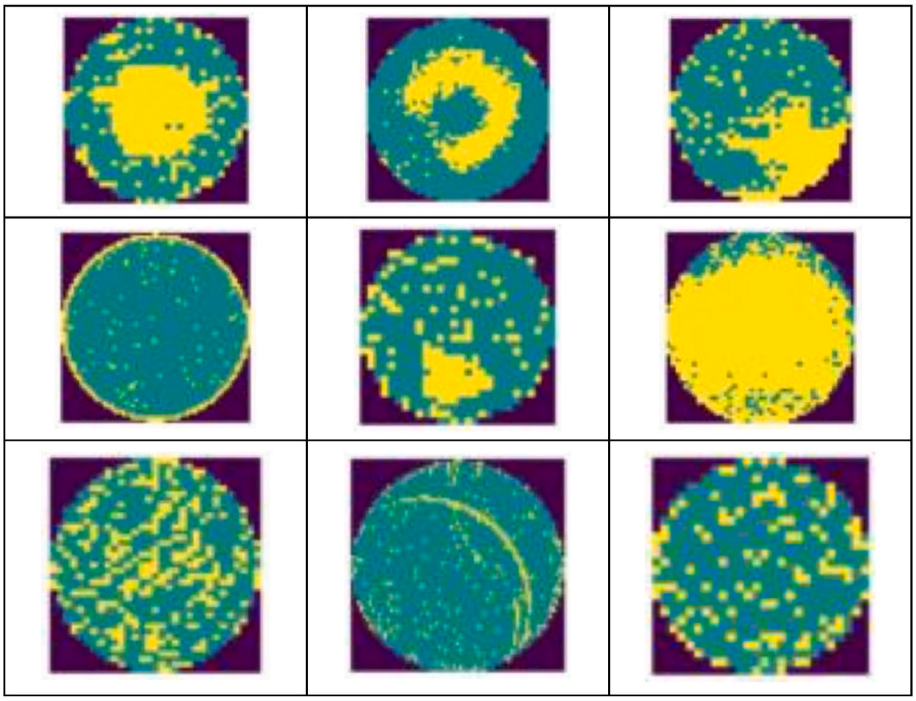


Fig. 1. Nine different defect patterns of wafer maps.

(Krishnan and Krishnan, 2021). In the research field of traffic scene analysis, ViT was applied to the prediction of the pedestrian crossing intention. The experimental results showed that the proposed Action-ViT outperformed the other competing models (Zhao et al., 2022). In medical image segmentation, a multi-scale efficient transformer attention (META) mechanism for fast and high-accuracy polyp segmentation was proposed (Wu et al., 2023). The efficient transformer blocks were adopted to generate multi-scale element-wise attentions for adaptive feature fusion in the existing U-shape encoder-decoder architecture.

Data imbalance can frequently take place in the real-world circumstances, especially in the semiconductor practice. In semiconductor manufacturing, attributed to stable production and high yields, the proportion of fraction nonconforming of each defect type is extremely low. Therefore, from a data scientist's perspective, the purpose of data augmentation is to create sufficient data of minority classes from the existing data. The created data can be added to the original training data required by the DL model; meanwhile, the data can be so diversified as to avoid over-fitting. Generative Adversarial Network (GAN) is a very representative data augmentation method among various solutions for data imbalance. GAN was proposed by Goodfellow et al. (2014). The GAN model architecture consists of two sub-neural-networks, generator and discriminator. A generator for generating new examples and the discriminator is used to judge whether the generated sample is true or fake. The generator neural network competes with the discriminator neural network like interacting in a zero-sum game. In the semiconductor manufacturing literature, Fan et al. (2022) introduced a novel approach for fault diagnosis in wafer acceptance test (WAT) and chip probing (CP) utilizing machine learning techniques. Leveraging the process flow of wafers alongside their corresponding process data, the study employed a sampling method known as Synthetic Minority Oversampling Technique (SMOTE) to address imbalances within the process dataset. Through experimentation with four robust classifiers, the study aimed to determine the optimal SMOTE ratio for enhancing classification models in this context. To address the challenge of class imbalance, Fan et al. (2023a) introduced a novel approach utilizing a variational autoencoder (VAE) as a data augmentation technique for rebalancing temporal raw trace data within the semiconductor manufacturing process. By extracting latent variables that encapsulate the distribution characteristics of defective samples, the study leveraged the statistical randomness inherent in these variables to synthesize defective samples through the VAE's decoder scheme, trained specifically for this purpose. However, many existing GAN-based and VAE-based augmentation methods may encounter challenges such as limited diversity in the generated data, inconsistent quality of generated samples, training instability, and sensitivity to hyperparameters.

As of today, a definitive "most robust" generative model for image augmentation in deep learning remains unanswered. Nevertheless, numerous models have gained widespread use and demonstrated effectiveness across various tasks, primarily stemming from generative-adversarial-network- and variational-autoencoder-based approaches. The CycleGAN (Zhu et al., 2020) approach is a competent deep-learning architecture facilitating image-to-image translation without the requirement for paired training data. Leveraging two generative adversarial networks (GANs), it learns the mapping between two domains of images. The model is capable of capturing the characteristics of the target domain, enabling the generation of new images from the source domain that exhibit corresponding traits. In this paper, two domains represent the different types of wafer defect maps. Aside from this, the major advantages of CycleGAN include (i) the constraint of cycle consistency helps improve the quality and coherence of augmented images, reducing artifacts and inconsistencies, (ii) semantic preservation to preserve the semantic content of input images during translation, resulting in visually coherent and meaningful transformations, particularly for the WM-811K dataset, (iii) bidirectional translation between two domains, allowing for versatile transformations without recourse to separate models or additional training, (iv) robustness to domain shifts and variations in input data distribution. The aforementioned adaptability makes CycleGAN suitable for applications where the characteristics of input data may vary across different environments or datasets and the overfitting can be more likely circumvented. Given the aforementioned evaluation, we decide to compare the proposed ViT augmentation methodology to the CycleGAN model.

As mentioned previously, imbalance data might have affected model classification performance. To remedy this situation, a novel data augmentation strategy in terms of ViT is proposed in this article, and the wafer defect map identification problem is addressed while being classified by using the convolutional neural network (CNN) models. The remaining of this paper is organized as follows. The proposed ViT augmentation framework is elaborated in Section II. In Section III, the augmentation performances between the proposed ViT model and the CycleGAN model while a CNN model supporting 16 layers proposed by Visual Geometry Group (VGG) at Oxford are assessed for wafer pattern classification. The CNN model using the VGG algorithm in this paper will be justified as well. The conclusion and the directions of future research are remarked in Section IV.

2. Proposed ViT augmentation framework

In this section, how the ViT model is designed to generate affined images is described. The evaluation procedure of the proposed data augmentation is briefly stated as follows.

(i) Given a set of training samples that involve imbalanced class samples;
(ii) Train the ViT model to anchor the best-practice hyper-parameter setting;
(iii) Apply the trained ViT model as an image generator;
(iv) Add generated data to the original training data to form a new training set so that the samples of minor pattern classes can be effectively increased.
(v) Evaluate classification performances before and after ViT augmentation to verify the efficacy.

2.1. Architecture of the vision transformer model

In order to apply the ViT model to the image classification task, first, the input image of a wafer map will be split into several patches via linear projection of flattened patches, and the position information of each image is attached to the patches to facilitate the spatial identification of each individual patch. Eventually, the output of the transformer encoder will be sent to the classification module in terms of multilayer perceptron (MLP) for classification, as illustrated in Fig. 2.

In the first step, the input image $x \in \mathbb{R}^{H\times W\times C}$ is given and the ViT model divides the image into N patches with $x_{patch} \in \mathbb{R}^{N\times P^2\times C}$, where H and W are the height and width of the input image, C is the number of channels, and (P, P) is the resolution of each image patch. The total number of patches is $N = HW/P^2$. The original wafer map is the size of $H \times W \times C$. The ViT model splits the wafer map into N patches; P is the side length of each patch, as is shown in Fig. 3, where the number of patches N is 25. The input image of size $H \times W$ will be divided into $n \times n$ grids, where $N = n^2$, and used in a later process. Considering both computational cost and classification accuracy, an image size of 75×75 pixels was selected for subsequent experiments conducted in this paper. The original WM-811K dataset comprises 632 different image sizes, ranging from 6×21 to 300×202 pixels. To standardize images across the dataset, nearest-neighbor interpolation, also referred to as proximal interpolation, was employed in the study.

Next, by flattening each patch to a one-dimensional (1D) vector, the size of those 1D vectors becomes $P^2 \times C$. All the vectors are projected to

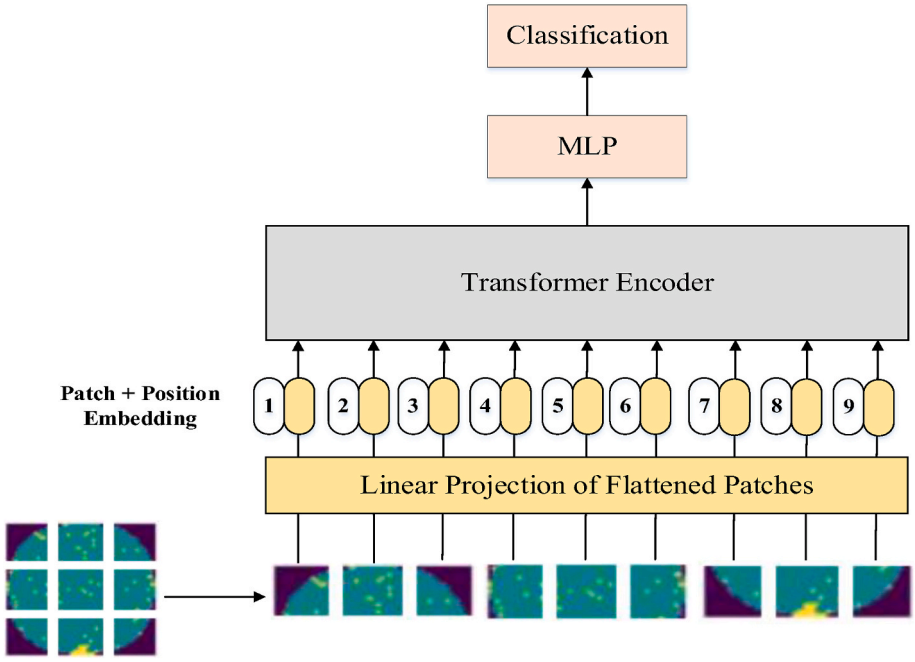


Fig. 2. Overview of the ViT model.

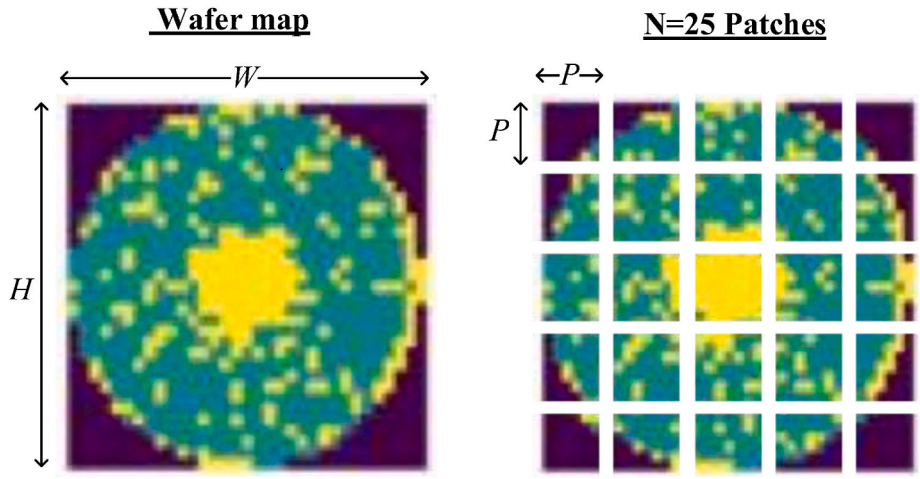


Fig. 3. Wafer map and patch size.

the same length D through a trainable linear projection with spatial information added to each vector to preserve the original positional (i.e., spatial) information of the image, which is thus the so-called position embedding. These embedding vectors are used as the input to Transformer Encoder, as illustrated in Fig. 2.

Transformer Encoder primarily consists of multi-head self-attention (MSA) and multilayer perceptron (MLP) blocks. Layer normalization (LN) (Ba et al., 2016) is utilized prior to every MSA/MLP blocks, and residual connection (He et al., 2015) is used after every MSA/MLP blocks. The MSA mechanism allows the model to attend to different positions of the input sequence to compute a representation of each position. The MLP block applies a fully connected feed-forward network to each position separately. The LN layer normalizes the input to have zero mean and unit variance before applying the MSA and MLP blocks (see Fig. 4). The LN layer helps to reduce the effect of the input distribution on the network's performance and improves the stability of the training process. The residual connection allows the gradient to flow directly through the block without passing through non-linear activation functions, which helps in training deep models. In some circumstances, the non-linear activation function can cause the gradient to vanish or explode in training models with deep layers.

From a high-level perspective, the MSA block operates by applying multiple self-attention mechanisms in parallel, with each mechanism focusing on a different aspect of the input sequence. This approach allows the model to capture long-range dependencies and patterns in the data more effectively than a single self-attention mechanism. The MSA block is a key component in the Transformer encoder, having a form of operating many self-attentions at the same time as shown in Eqs. (1) and (2) in the following:

$$head_i = Attention(Q_i, K_i, V_i), i = 1, \ldots, k \tag{1}$$

$$MSA(Q, K, V) = Concact(head_i, \ldots, head_k)W_o, i = 1, \ldots, k \tag{2}$$

where W_o denotes the trained output matrices. In Fig. 4(b), the detailed operations of the MSA block with 4 input vectors are demonstrated. In the MSA block of the Transformer architecture, there have three weight

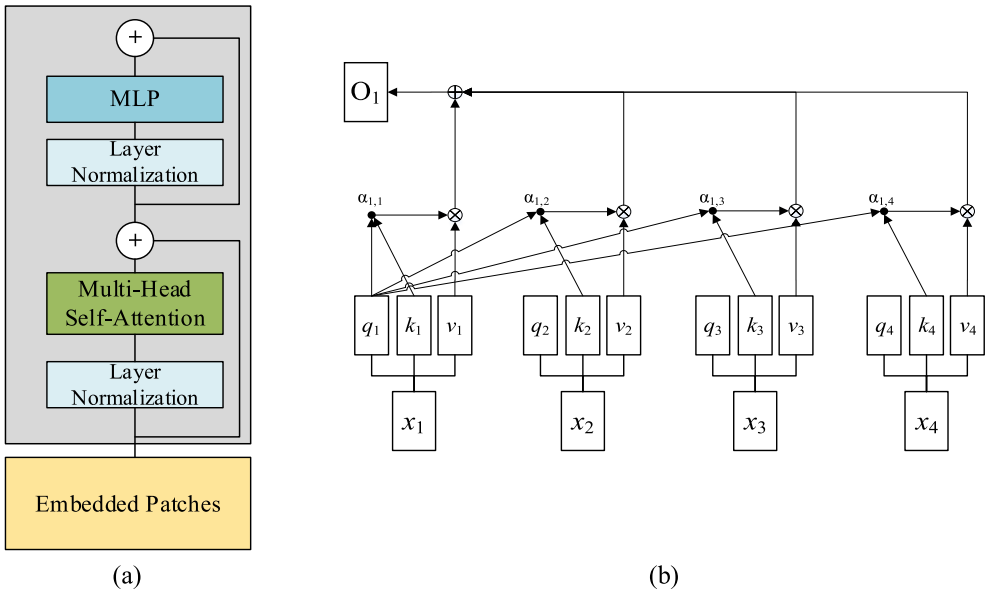


Fig. 4. The architecture of the ViT model: (a) the architecture of Transformer encoder, (b) an illustration of self-attention using 4 input vectors.

matrices that need to be trained: one for Query matrix (Q), one for Key matrix (K), and one for Value matrix (V). These three matrices are used to project the input sequence (accompanied by the added position embeddings) into the query, key, and value spaces, respectively. Through these three learned weight matrices $\mathbf{W}_q$, $\mathbf{W}_k$ and $\mathbf{W}_v$, the input with position embedding matrix $\mathbf{X}$ is projected to calculate the query, key and value components that can be mathematically expressed by Eqs. (3)–(5):

$$Query(\mathbf{Q}) = \mathbf{W}_q \cdot \mathbf{X} \tag{3}$$

$$Key(\mathbf{K}) = \mathbf{W}_k \cdot \mathbf{X} \tag{4}$$

$$Value(\mathbf{V}) = \mathbf{W}_v \cdot \mathbf{X} \tag{5}$$

An attention function is operated through a combination of a pair of query and key. The attention function is used to represent the degree of relationship between the provided query position and the provided key position to reflect relevance or similarity according to

$$AttentionMatrix(\mathbf{A}) = \mathbf{Q} \cdot \mathbf{K}^{\mathrm{T}} \tag{6}$$

where Q indicates Query matrix and K indicates Key matrix. The output O of self-attention in matrix form is calculated according to Eq. (7) as follows:

$$Output(\mathbf{O}) = \mathbf{A} \cdot \mathbf{V} \tag{7}$$

Attention score is thus utilized to compute a weighted sum of the values associated with the keys, where the weights are determined by the softmax function applied to the scores. The resulting weighted sum is referred to the attention output **O**, which represents a contextual representation of the query based on the information contained in the key-value pairs. As shown in Fig. 4(b), attention score $\alpha_{i,j}$ in **A** represents the attention score between input vectors of positions i and j, respectively. For clarification purposes, the pseudo code of the ViT model is provided in Appendix A for interested readers seeking further information.

2.2. Comparisons of different restructured ViT models in classification performance

To select a classification model, ViT and VGG16 (Simonyan and Zisserman, 2014) are preliminarily tested for an initial experimental comparison, and the decision will be made on which model will better serve as the benchmark model for classification. Here, totally, 25,519 labeled wafer maps in the WM-811K dataset are utilized, where sample maps in 8 different classes (excluding the None-type pattern) were randomly selected according to three partitions: 60% for training, 15% for validation and 25% for testing, respectively. To conduct a fair comparison, ten experiments were run independently, and the computational results are shown in Table 1. The best-practice average accuracy of ViT is about 91.51% as 6 × 6 patches and 4 HEADs are adopted; the average accuracy of VGG16 is about 93.56%. Note that full comparison results between ViT and VGG16 under the WM-811K scenario are available upon request. On the whole, the average classification performance of VGG16 is achieved by approximately 2% higher than ViT. Note that the training dataset contains imbalanced numbers of wafer patterns. All the experimental studies of image augmentation, henceforward, will be performed to assess the classification performances of the VGG16 model. The usage of the VGG16 model in this paper will be justified in a later section.

2.3. Modified structure of ViT using different heads

By means of multiple heads, the original ViT model is able to capture different types of information and attend to different features. The outputs from the different heads are "concatenated" and passed through a linear projection to produce the final output of the self-attention layer. This output is then passed through a feedforward neural network, including an activation function and two fully connected layers, to generate the final features used for classification.

For further assessment purposes, different modifications to the MLP design in ViT are investigated. In Table 2, there are the modified ViT model using 2 heads (with each connected to its own independent MLP for classification), the modified ViT model only with an individual Head 1, and the modified ViT model only with an individual Head 2. See the corresponding results in column from left to right, respectively. In the same way, the corresponding classification results of using 3 and 4 heads in a modified ViT version are tabulated in Tables 3 and 4, respectively. The instance of modified ViT with four independent heads is illustrated in Fig. 5.

It can be conjectured from the preceding experimental comparisons in Tables 2–4 that the modified ViT models using an individual head compete very well with the ViT model with multiple heads in classification performance. It evidently indicates that each head carries its own distinctive correlation information between different positions in the

Table 1
Comparison results between ViT and VGG16.

Class	ViT (12×12)	ViT (9×9)	ViT (6×6)	ViT (4×4)	ViT (2×2)	VGG16
						
Center	97.73%	97.90%	97.87%	96.87%	97.17%	97.31%
Donut	81.15%	84.89%	81.51%	87.19%	81.73%	83.45%
Edge-Loc	90.39%	90.26%	90.61%	88.77%	87.06%	92.30%
Edge-Ring	98.14%	98.17%	97.98%	98.10%	98.01%	98.10%
Loc	80.53%	79.07%	79.88%	76.95%	70.71%	85.49%
Near-full	89.19%	85.95%	87.57%	84.87%	88.11%	82.43%
Random	86.30%	86.20%	87.87%	89.91%	89.44%	91.99%
Scratch	60.34%	64.63%	62.79%	56.31%	48.05%	80.33%
Accuracy	91.43%	91.49%	91.51%	90.48%	88.76%	93.56%

Table 2
Classification results of ViT with 2 heads.

Class	Modified ViT with 2 HEADs	Modified ViT Head 1 classifier	Modified ViT Head 2 classifier
Center	95.72%	96.55%	96.93%
Donut	92.09%	85.61%	83.45%
Edge-Loc	85.04%	85.89%	89.98%
Edge-Ring	98.51%	98.84%	97.52%
Loc	68.82%	64.59%	65.26%
Near-full	78.38%	75.68%	83.78%
Random	93.98%	86.57%	88.43%
Scratch	51.68%	50.00%	58.72%
Accuracy	88.52%	87.88%	88.84%

Table 3
Comparison results of ViT with 3 heads.

Class	Modified ViT with 3 HEADs	Modified ViT Head 1 classifier	Modified ViT Head 2 classifier	Modified ViT Head 3 classifier
Center	97.11%	97.21%	97.11%	98.7%
Donut	82.73%	85.61%	79.14%	75.54%
Edge-Loc	91.06%	87.82%	89.21%	87.12%
Edge-Ring	97.77%	98.55%	98.02%	97.73%
Loc	71.27%	67.82%	74.28%	70.82%
Near-full	78.38%	81.08%	72.97%	78.38%
Random	86.11%	85.65%	87.96%	87.04%
Scratch	59.4%	50.67%	44.3%	51.34%
Accuracy	89.94%	88.76%	89.32%	88.82%

Table 4
Comparison results of ViT with different heads.

Class	Modified ViT with 4 heads	Modified ViT Head 1 classifier	Modified ViT Head 2 classifier	Modified ViT Head 3 classifier	Modified ViT Head 4 classifier
Center	97.39%	97.77%	97.58%	97.39%	97.39%
Donut	90.65%	87.77%	87.77%	87.05%	85.61%
Edge-Loc	87.43%	87.9%	86.43%	86.89%	85.89%
Edge-Ring	98.1%	98.35%	98.88%	98.43%	98.43%
Loc	69.38%	66.82%	69.93%	67.71%	66.82%
Near-full	83.78%	81.08%	94.59%	89.19%	86.49%
Random	87.96%	86.57%	87.96%	89.35%	89.35%
Scratch	48.99%	51.68%	48.32%	55.37%	56.38%
Accuracy	88.89%	88.78%	89.06%	88.96%	88.63%

score matrix. The attention matrices of the first 4 heads in ViT can be the building block of generating new wafer defect maps for WM-811K. In comparison to the original ViT shown in Table 1, the original ViT model with 4 heads and 6 × 6 patches performs best and will provide a basis for the proposed image augmentation framework to be presented in a later section.

2.4. Modification of attention matrix in ViT

In the original ViT architecture, the attention matrix of different heads is multiplied by the value vector to obtain a weighted vector in the multi-head mechanism. The idea of the modification shown here is to directly concatenate all the attention matrices, and then to be multiplied with weighted matrix to produce the output vector for classification. Thus, Value matrix (V) is not used in this modified ViT as illustrated in Fig. 6.

The idea for the modified version is to enhance the ViT model with a different topology of spatial information, hoping to improve the classification performance. The classification results of the original ViT in 6 × 6 patches versus the modified version of the attention matrix are tabulated in Table 5. Evidently from the table, these two versions of the ViT model perform nearly equally well and the difference in accuracy is merely about 0.8%. The insight gained from the foregoing comparison implicitly indicates that the concatenation of the attention matrix plays a more pivotal role in classification than the subsequent MLP structure. Towards this end, the obtained attention matrix can be deemed informative sources to the generation of augmented image data, which will be elaborated in the next section as the primary contribution of this research. Anyway, potential modifications to the MLP network become trivial in the wafer map application addressed here.

2.5. Vision transformer model to generate new data

In this section, we take advantage of Vision Transformer to propose a new data augmentation framework for the purposes of classification improvement. The new idea arises from the attention matrix of ViT to calculate the row averages of attention scores as indicated in Fig. 7, therefore, the heatmap can be particularly created for augmentation, where the row averages of the attention scores are calculated and rearranged to form the heatmap.

As illustrated in Fig. 7, consider using the number of $n \times n$ gridded patches (i.e., $n^2 = N$ positions in total), and then the attention matrix of size $n^2 \times n^2$ will be created according to the ViT model. It is worth noting that the average attention score is computed on a row-basis, indicating all queries are taken into account. The attention matrix can be obtained from the ViT model, prior to the MLP mechanism for classification. The attention matrix represents the important relationship between different positions and different heads that focus on different areas.

Using the wafer map as an example, a ViT model is trained with four HEADs, dividing the map into 36 patches as illustrated in Fig. 8 ($n = 6$). All remaining hyper-parameters were set to their default values, with the exception of the recommended four HEADs, as detailed in Appendix

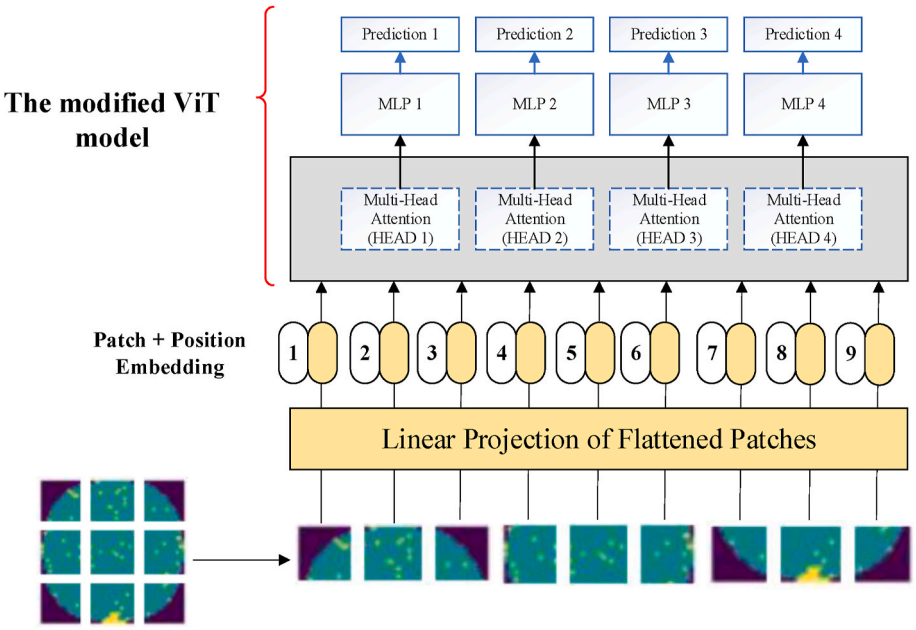


Fig. 5. Modified ViT architecture where 4 independent heads are utilized.

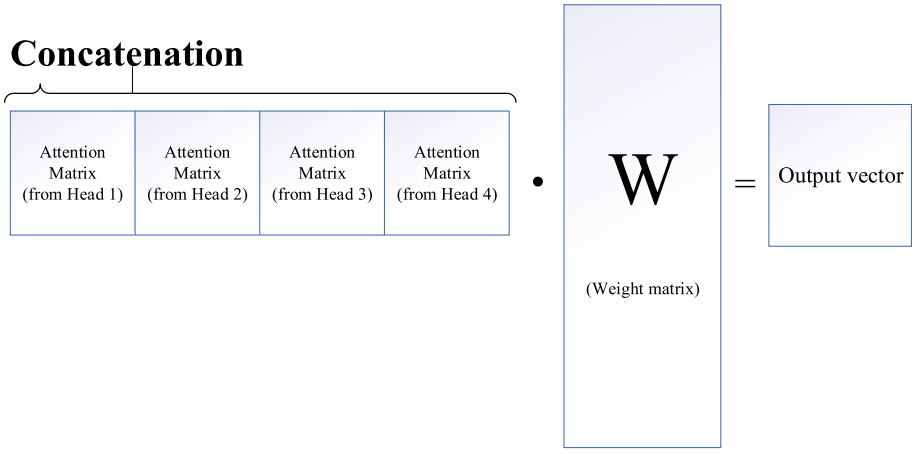


Fig. 6. Modification of the attention matrix.

Table 5
Comparison results between Original ViT and the Modified ViT by concatenating attention matrix.

Class	Original ViT (6 × 6)	Modified ViT
Center	97.87%	97.38%
Donut	81.51%	85.04%
Edge-Loc	90.61%	91.98%
Edge-Ring	97.98%	97.70%
Loc	79.88%	75.14%
Near-full	87.57%	85.14%
Random	87.87%	86.99%
Scratch	62.79%	57.72%
Accuracy	91.51%	90.73%

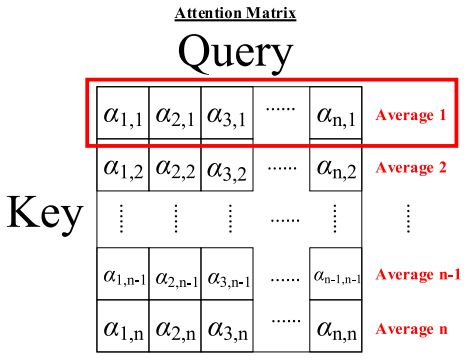


Fig. 7. Calculation of attention matrix of an $n \times n$ gridded patch.

B. For instance, 36 patches are generated if $n = 6$ is used to split the wafer map of Center type as shown in Fig. 8. The created attention per HEAD is a matrix of size 36 × 36. Visualizations of attention matrices for HEADS 1–4 can be seen from left to right, top to bottom, respectively, in Fig. 9. Next, the averages over the entire attention matrix are computed by row, so there are 36 averages indicated in Fig. 10. In essence, taking the row averages reflects a natural heat-like representation of the attention matrix.

To establish the heatmap from Fig. 11, the 36 row averages are sequentially placed from left to right, top to bottom to form a 6 × 6 square map as shown in the right-hand side of Fig. 11. From the visualization perspective, the proposed heatmap created from the self-

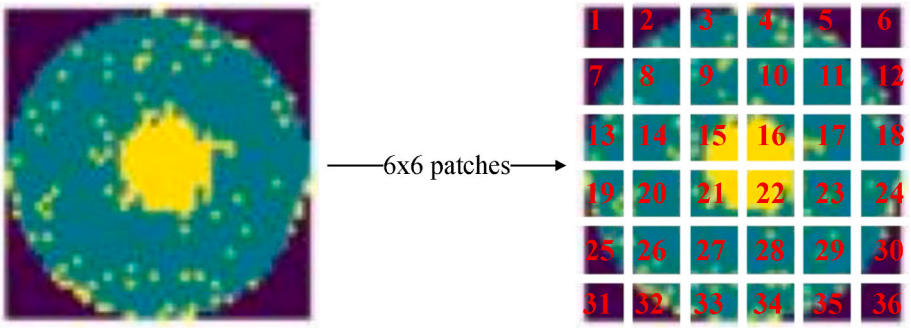


Fig. 8. The wafer map (of *Center* type) split into 36 patches.

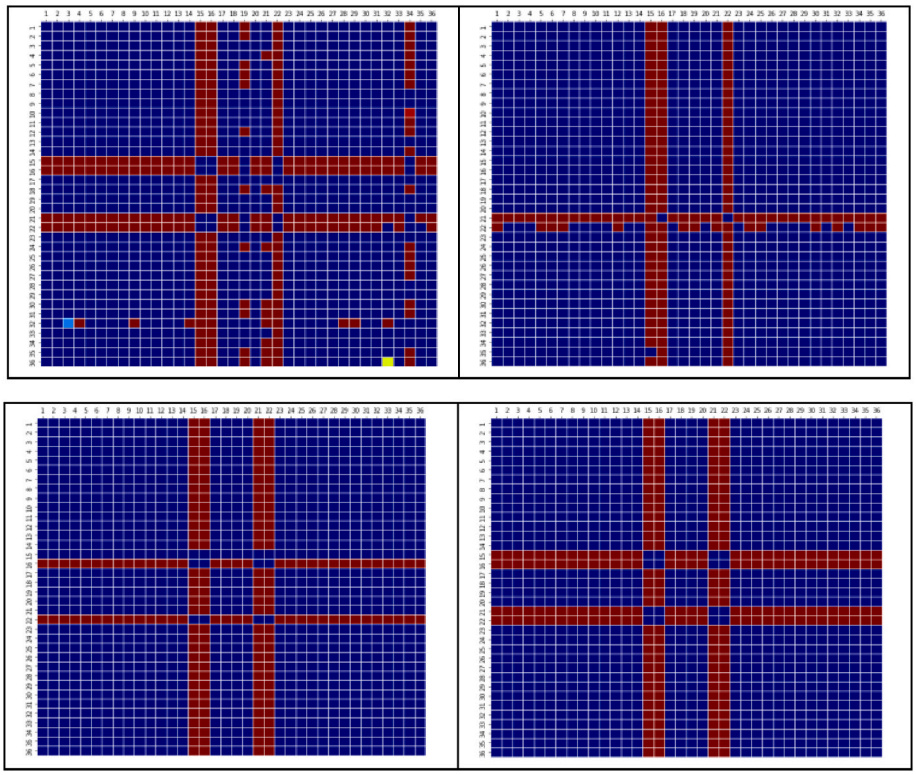


Fig. 9. Visualization of attention matrix of 4 different heads for the defect map of Center type.

attention computation through the ViT model serves an alternative to image augmentation.

In the aforementioned case of Center type shown in Fig. 8, wafer defects are mostly distributed around the patches numbered 15, 16, 21 and 22. Accordingly, the attention matrix shows that the previous 4 patches are of relational importance (indicated with color in red) as shown in Fig. 9.

2.6. Vision transformer image generation procedure

The procedure of the proposed vision transformer image generation is shown in Fig. 12. Firstly, a ViT model is trained to learn a variety of wafer defect maps. In the meanwhile, each wafer map is partitioned in patches. By means of the attention score of multi-head mechanism embedded in the ViT model, the attention matrix is calculated and the newly generated heatmap is created for the purposes of augmentation.

The generative model presented in sub-section E is adopted to generate additional defective wafer maps. Next, the traditional image processing method is opt for resizing the output of the proposed image generator when necessary. Finally, the generated image data is augmented to the original image dataset, and then the deep learning CNN model serves as the classifier to verify if the image data generation mechanism is helpful in improving defect map classification.

Therefore, the proposed data augmentation procedure using the ViT model can be formalized in the following algorithm.

ALGORITHM ViT Data Augmentation

1. Given the training image dataset, number of heads, number of patches to be used;
2. The training data are fed to the ViT model;
3. Apply the multi-headed self-attention mechanism of the ViT model to acquire the attention matrix for each head as shown in Fig. 7;
4. Apply row average calculation to the attention matrix in (3) per head and generate the corresponding heatmap as shown in Fig. 11;
5. Resize the heatmap if necessary;
6. Combine the original training dataset with the augmented dataset to train the CNN classifier;
7. Evaluate whether the data size of image augmentation is sufficient. If not, go to (1) to increase the number of heads and retrain the ViT model;

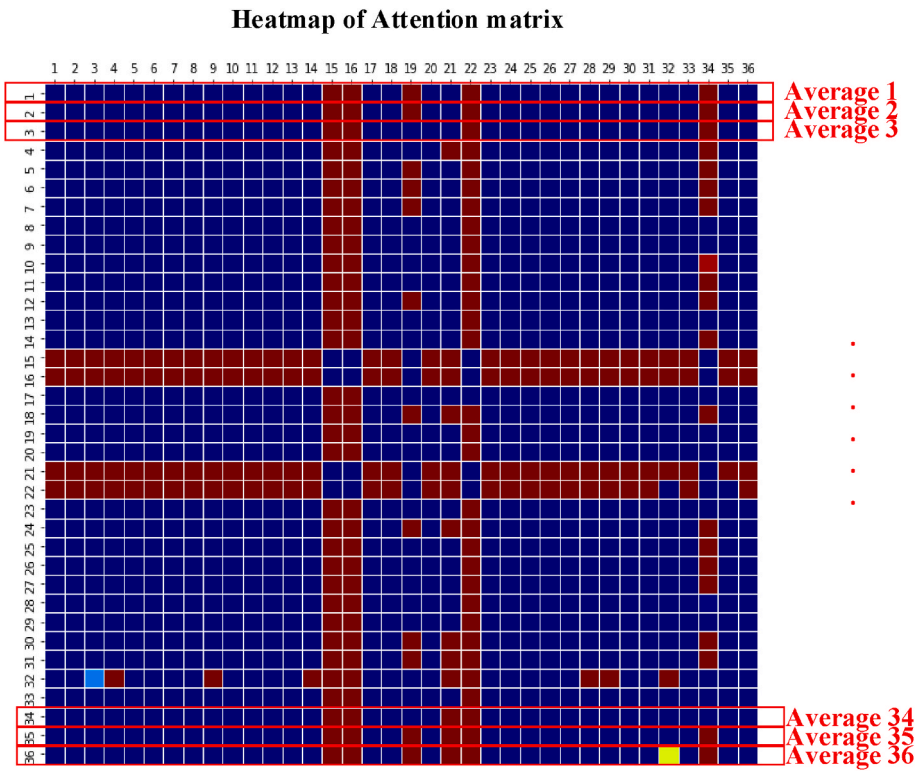


Fig. 10. Row average calculation of attention matrix in the wafer map of *Center* type.

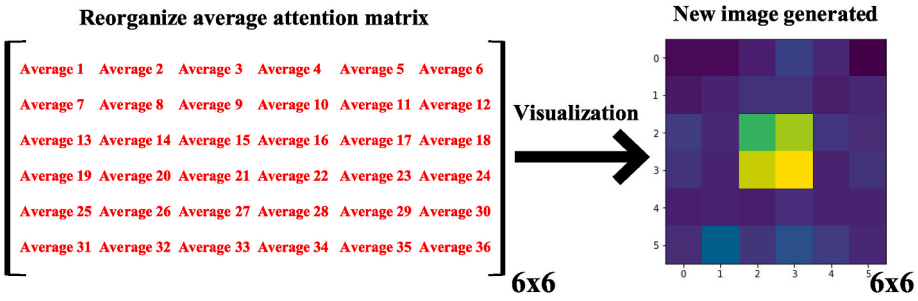


Fig. 11. New generated data in *Center* type wafer map.

3. Computational Experiences of proposed ViT augmentation framework

3.1. Data configuration of wafer defect patterns

WM-811K is a real industrial dataset, collected from 46,393 wafer lots and consisting of 811,457 original wafer defect maps. In the dataset, there are only 172,950 images that were labeled with particular defect patterns by domain experts. Among the 172,950 annotated images, 9 types of wafer map patterns are present, inclusive of 8 types of particular defect patterns (i.e. Center, Donut, Edge-Loc, Edge-Ring, Loc, Near-full, Random, Scratch) and a kind of wafer map without defect pattern (named None). The dataset is partitioned according to annotation and failure types as shown in Fig. 13. Note again that the ViT model will be used in the proposed augmentation framework, and then the VGG16 model will serve as the classifier for augmentation performance assessment investigated in Section III. All the comparison results demonstrated in this section are computed based on 10 independent runs.

In an earlier experimental study, four different CNN-based models, ResNet50 (He et al., 2016), VGG16 (Simonyan and Zisserman, 2014), EfficientNet (Tan and Le, 2019) and ViT (Dosovitskiy et al., 2021), are comprehensively assessed based on the WM-811K dataset. The average accuracies returned by using ResNet50, VGG16, EfficientNet and ViT are 97.12%, 97.45%, 97.49% and 96.73%, respectively. Among the compared CNN-based models, both VGG16 and EfficientNet generate comparably competitive performances on the wafer defect map dataset. Although EfficientNet was claimed state-of-the-art (SOTA) in 2019, however, VGG16 is chosen as the classifier in this paper due to its much wider popularity with 114,299 citations thus far in comparison to EfficientNet gaining 16,450 citations. It is also of primary importance to note that the major goal of our paper is to propose a new augmentation framework instead of a classifier for wafer defect pattern classification. The detailed comparison results are available upon request.

Class *None* accommodates 85.2% and the remaining 8 wafer defect patterns are distributed less than 15%. Due to the sheer number that might cause severe data imbalance, Class *None* will not be considered in this research. It worth noting particularly that class Random is hardly distinguishable from class *None* by human visual inspection. Meanwhile, these two wafer-pattern types seem to co-exist in a single wafer map frequently. After due consideration, the quality assessment of data annotation already done by domain experts in the WM-811K dataset rests with future research and is beyond the scope of this paper. All

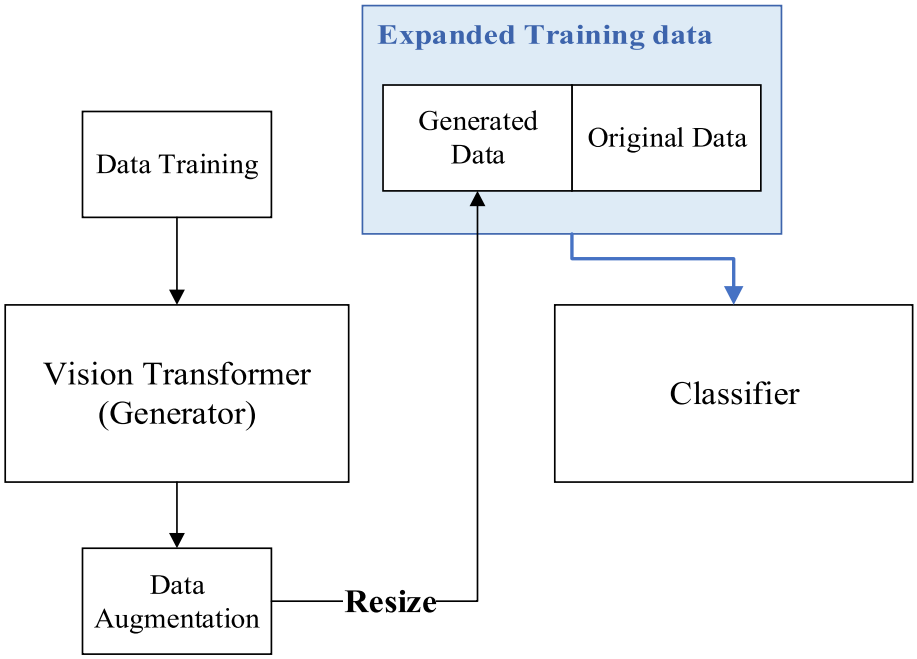


Fig. 12. Procedure of Vision Transformer image generation.

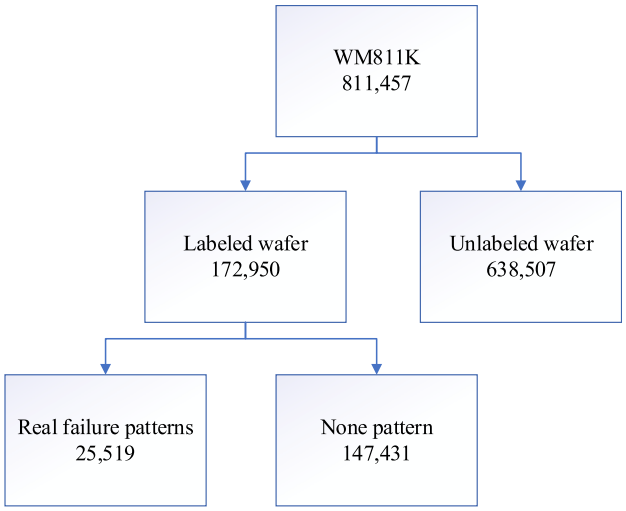


Fig. 13. The distribution of the WM-811K dataset.

experiments are focused on 8 types of wafer map patterns that have clearly-defined failure patterns. The percentages of 8 wafer defect patterns are tabulated in Table 6. The wafer maps in the WM-811K dataset are composed of 3-color bands: 0, 1 and 2, where 0 means no wafer grain in purple, 1 means non-defect grains in green, and 2 means defect grains in yellow.Note that all the experiments conducted in this section were performed on a computing machinery with INTEL I7-8700 CPU and NVIDIA RTX2080 GPU.

3.2. Image generation of wafer defect patterns via ViT

There are a total of 632 different sizes of wafer maps in the WM-811K dataset. To ensure a fair comparison, all the wafer maps used in experimentation were resized to 75 × 75 (in pixels). The dataset in Table 6 is partitioned into 60%, 15% and 25% of each class for training, validation and testing, respectively, as illustrated in Table 7.

First of all, making use of 2, 4 and 8 HEADs is examined for the comparison purposes. In order to extract maximum detailed information, splitting the original image size of 75 × 75 into 25 × 25 patches is adopted, each with a patch size of 3 × 3. The primary goal of this research is to alleviate the data imbalance difficulty encountered in WM-811K. Towards that end, additional image data will be generated in multiples of the number of HEADs in this paper. Some wafer defect maps generated via ViT by means of 2 and 4 HEADs are exhibited in Table 8.

As can be seen from Fig. 14, defect patterns can be clearly shown in the generated images as 8 HEADs are used in ViT; however, patterns of which in terms of some HEADs are inverted, as demonstrated in Fig. 14. Image reverse is a required pre-processing step prior to the model training. Therefore, it can be logically alleged that using 8 HEADs for image augmentation can be redundant for some instances in the WM-811K dataset.

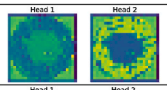
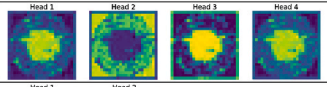
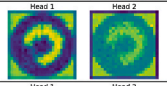
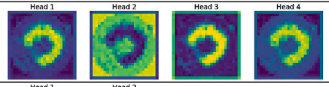
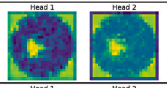
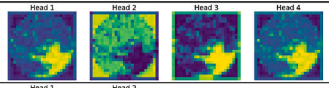
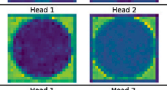
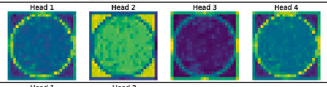
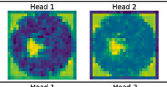
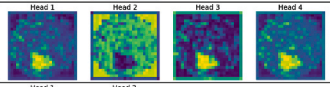
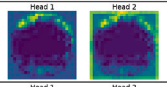
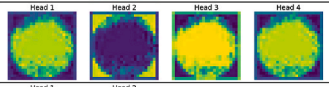
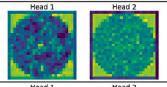
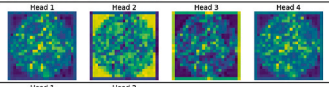
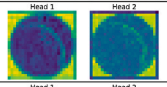
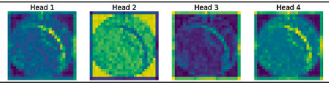
Table 6
8 Class distribution of the WM-811K dataset.

Defect class	Number	Percentage
Center	4294	16.83%
Donut	555	2.17%
Edge-Loc	5189	20.33%
Edge-Ring	9680	37.93%
Loc	3593	14.08%
Near-full	149	0.58%
Random	866	3.39%
Scratch	1193	4.67%
Total	25,519	100%

Table 7
Data arrangement for training, validation and testing.

Class	Training (60%)	Validation (15%)	Testing (25%)
Center	2576	644	1074
Donut	333	83	139
Edge-Loc	3113	778	1297
Edge-Ring	5808	1452	2420
Loc	2156	539	898
Near-full	89	22	37
Random	520	130	217
Scratch	716	179	298
Total	15,313	3827	6379

Table 8
Generated images using 2 and 4 HEADs.

Class	HEAD	generated data
Center	2	
	4	
Donut	2	
	4	
Edge-Loc	2	
	4	
Edge-Ring	2	
	4	
Loc	2	
	4	
Near-full	2	
	4	
Random	2	
	4	
Scratch	2	
	4	

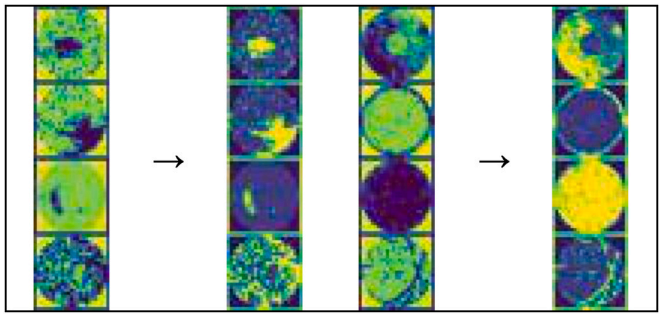


Fig. 14. Inverted defect patterns generated by using 8 HEADs.

3.3. Performance evaluation of the proposed ViT generative model

In this section, different augmentation strategies are presented in an attempt to mitigate data imbalance. First, on account of the benchmark results shown in Table 1, Donut, Loc, Near-full, and Scratch are considered more difficult defect patterns than the others for classification in the studied dataset. It is shown in Table 7 that these four "minority" classes contain a training size all less than 1000. These four minority classes will be used in Strategy A. Furthermore, class Loc is extremely confounded with class Edge-Loc; class Near-full is fairly confounded with class Random, as can be apparently seen from Table 8. On this account, only class Donut and Scratch will be used "individually" in Strategies B and C, respectively. The strategies are summarized as

follows:

Strategy A augments the 4 minority classes simultaneously using 4 heads;

Strategies B1–B3 augment class Donut individually using 2, 4 and 8 heads, respectively;

Strategies C1–C3 augment class Scratch individually using 2, 4 and 8 heads, respectively.

Certainly, augmented images from Strategies A, B and C can be used jointly or separately to train the CNN. To begin with Strategy A shown in Table 9, the wafer maps of 4 minority classes are synthesized by using the proposed augmentation framework where "4 HEADs" are utilized. That is, the training size after augmentation becomes 5 times the original one. Taking the minority class of Donut as an example in the table, the original training data of 333 maps are fed into ViT via 4 HEADs to generate additional training data of 1332 maps, amounting to the training data after augmentation of 1665 maps. In the cases of using 2 and 8 heads, additional training data are augmented in the same manner. In the table, the recall and the average are averaged over 10 independent runs for every defect class and its corresponding standard deviation is indicated in parenthesis.

As compared to the VGG16 accuracy of 93.56% as the benchmark (see Table 1), the accuracy of Strategy A improves to 93.77% in Table 9. Specifically, the accuracy of Donut improves to 87.70% from 83.45%, class Near-full improves to 89.73% from 82.43%, class Random declines minutely to 90.93% from 91.99%, and class Scratch improves to 83.46% from 80.33%. Overall, Strategy A quite succeeds in classification improvement; however, its performance appears to be mostly veiled by the training sizes of the minority classes.

To improve the recall of individual classes, the weight of generating individual class images (using 2, 4 to 8 HEADs) is placed on minority classes Donut and Scratch separately. Regarding Strategies B1–B3 shown in Tables 10–12, the augmentation of class Donut really boosts recall by increasing the number of HEADs with 86.69%, 89.35% and 86.33%, respectively. However, using 8 HEADs for augmentation of class Donut seems to be overkill. Individual augmentation in terms of 4 HEADs on class Donut raises its best recall up to 89.35%, leading to an overall accuracy of 93.94%.

With respect to Strategies C1–C3 shown in Tables 13–15, the individual augmentation of class Scratch, with the expanded training size by using 2, 4 and 8 HEADs, improves its recall to 81.14%, 83.89% and 86.21%, respectively. The overall accuracy improves to 93.74%, 93.78% and 93.80 accordingly as compared to the benchmark accuracy of 93.56%. As mentioned previously, recall improvement on individual classes is noticeable but the contribution of the proposed augmentation framework cannot be fully justified in merely looking at overall accuracy rates.

Table 9
Strategy A: Generate minority classes using 4 HEADs (Donut, Near-full, Random and Scratch).

Class	Original training data	Augmented training data	Total training data	Recall	Acc.
Center	2576	0	2576	97.40% (0.44%)	93.77% (0.2063%)
Donut	333	1332	1665	87.70% (0.48%)	
Edge-Loc	3114	0	3114	92.77% (0.62%)	
Edge-Ring	5808	0	5808	98.03% (0.41%)	
Loc	2156	0	2156	84.63% (0.62%)	
Near-full	90	360	450	89.73% (0.69%)	
Random	520	2080	2600	90.93% (0.64%)	
Scratch	716	2864	3580	83.46% (0.47%)	

Table 10
Strategy B-1: Expand training data of Donut to 999 using 2 HEADs.

Class	Original training data	Training data after augmentation	Recall	Accuracy
Center	2576	2576	97.92% (0.35%)	93.81% (0.1717%)
Donut	333	999 (333 + 666)	86.69% (0.55%)	
Edge-Loc	3114	3114	92.16% (0.59%)	
Edge-Ring	5808	5808	98.23% (0.56%)	
Loc	2156	2156	85.29% (0.70%)	
Near-full	90	90	87.03% (0.60%)	
Random	520	520	91.20% (0.52%)	
Scratch	716	716	81.98% (0.56%)	

Table 11
Strategy B-2: Expand training data of Donut to 1665 using 4 HEADs.

Class	Original training data	Training data after augmentation	Recall	Accuracy
Center	2576	2576	97.71% (0.67%)	93.94% (0.2412%)
Donut	333	1665 (333 + 1332)	**89.35%** (0.68%)	
Edge-Loc	3114	3114	92.27% (0.56%)	
Edge-Ring	5808	5808	98.24% (0.66%)	
Loc	2156	2156	85.95% (0.61%)	
Near-full	90	90	81.89% (0.50%)	
Random	520	520	91.53% (0.70%)	
Scratch	716	716	82.08% (0.55%)	

Table 12
Strategy B-3: Expand training data of Donut to 2997 using 8 HEADs.

Class	Original training data	Training data after augmentation	Recall	Accuracy
Center	2576	2576	97.84% (0.46%)	93.64% (0.1026%)
Donut	333	2997 (333 + 2664)	86.33% (0.23%)	
Edge-Loc	3114	3114	91.88% (0.11%)	
Edge-Ring	5808	5808	98.36% (0.16%)	
Loc	2156	2156	84.78% (0.56%)	
Near-full	90	90	84.87% (0.79%)	
Random	520	520	91.25% (0.28%)	
Scratch	716	716	80.81% (0.22%)	

From Tables 9–13, the augmented samples for different defect classes are still imbalanced. The purpose of the experimental study using Strategies A-C is only to pinpoint exactly what augmentation capabilities the proposed ViT framework can provide. There are two immediate solutions for the proposed framework to class imbalance: (i) increase the number of HEADs used in ViT and (ii) reallocate the training, validation

Table 13
Strategy c-1: Expanded scratch training data to 2148 using 2 HEADs.

Class	Original training data	Training data after augmentation	Recall	Accuracy
Center	2576	2576	97.50% (0.14%)	93.74% (0.2316%)
Donut	333	333	86.33% (0.71%)	
Edge-Loc	3114	3114	91.64% (0.39%)	
Edge-Ring	5808	5808	98.44% (0.43%)	
Loc	2156	2156	85.84% (0.43%)	
Near-full	90	90	85.14% (0.57%)	
Random	520	520	91.34% (0.33%)	
Scratch	716	2148 (716 + 1432)	**81.14%** (0.36%)	

Table 14
Strategy c-2: Expanded scratch training data to 3580 using 4 HEADs.

Class	Original training data	Training data after augmentation	Recall	Accuracy
Center	2576	2576	97.97% (0.58%)	93.78% (0.2632%)
Donut	333	333	84.96% (0.25%)	
Edge-Loc	3114	3114	92.57% (0.15%)	
Edge-Ring	5808	5808	98.06% (0.66%)	
Loc	2156	2156	84.61% (0.11%)	
Near-full	90	90	87.30% (0.34%)	
Random	520	520	90.93% (0.21%)	
Scratch	716	3580 (716 + 2864)	**83.89%** (0.85%)	

Table 15
Strategy c-3: Expanded scratch training data to 6444 using 8 HEADs.

Class	Original training data	Training data after augmentation	Recall	Accuracy
Center	2576	2576	97.93% (0.15%)	93.80% (0.1330%)
Donut	333	333	84.68% (0.54%)	
Edge-Loc	3114	3114	91.66% (0.28%)	
Edge-Ring	5808	5808	98.34% (0.20%)	
Loc	2156	2156	84.64% (0.14%)	
Near-full	90	90	86.49% (0.32%)	
Random	520	520	90.97% (0.27%)	
Scratch	716	6444 (716 + 5728)	**86.21%** (0.54%)	

and testing partitions of 60%, 15% and 25% randomly and the ViT model is retrained to create more heatmaps for augmentation. To our knowledge, the latter is recommended rather than the former in practice.

Lastly, the proposed augmentation method is compared to one of the most practically noted generative models for image augmentation, CycleGAN (Zhu et al., 2020). To make a fair comparison with the best-practice generative model, CycleGAN, the training size of each class is adjusted to an equal training size. It can be achieved by under- and over-sampling procedures. For example, if an equal training size of 1000 defect maps is needed, then 1000 defect maps will be randomly under-sampled out of the original training data (2576 maps, see Table 7) for class Center. As for class Donut (with the original training size of 333 maps), the proposed ViT framework and CycleGAN ought to synthesize 667 maps to meet a total of 1000 defect maps.

For the comparison purposes, the equal training size of each class is set up to 1,000, 1,500, 3000 and 5000 defect maps randomly selected by undersampling or oversampling via the two compared generators, respectively. The VGG16 classification results are shown in Tables 16–19, respectively, as Strategies D1-D4. Evidently seen from these tables, the proposed ViT framework outperforms CycleGAN for every equal training size, with improvement on overall accuracy of 1.25%, 1.14%, 2.17%, and 0.86%, respectively. From the viewpoint of image augmentation, the proposed ViT framework is deemed a viable alternative and/or add-on to the present image generative models in the literature.

A summary of experimental results is presented in Table 20, comparing the proposed ViT augmentation framework with the benchmarks, VGG16 (with and without augmentation) and CycleGAN. The proposed augmentation method exhibits a notable advantage over the benchmark comparisons. Despite a marginal improvement in percentage, this increase in accuracy and recall translates to substantial cost savings in in-process quality control (IPQC) and on-site inspection/appraisal within semiconductor manufacturing. The role of defect map classification quality in the front-end-of-line (FEOL) processes is multifaceted, significantly impacting the semiconductor industry's endeavors to enhance energy efficiency and mitigate environmental and economic risks. This impact unfolds through various avenues, including: (i) streamlining manufacturing processes for optimal efficiency, (ii) minimizing rework and scrap, thus reducing resource wastage, (iii) ehancing overall yield, ensuring maximum output from production, (iv) bosting equipment efficiency through targeted defect identification and resolution, (v) facilitating process innovation and intelligentization, driving continuous improvement, and (vi) ensuring compliance with regulatory standards and mandates, fostering industry sustainability. Overall, the inspection quality of defect map classification in the front-end-of-line (FEOL) is essential for semiconductor manufacturers to reduce energy consumption, improve energy efficiency, and mitigate environmental and economic risks in their supply chain management. By leveraging advanced inspection technologies and data analytics, companies can optimize their processes, enhance product quality, and contribute to a more sustainable semiconductor industry.

Before concluding this section, as recommended by an anonymous reviewer, we revisited two additional public datasets, CIFAR-10 (Krizhevsky and Hinton, 2009) and MixedWM38 (Wang et al., 2020), to validate the proposed ViT augmentation model. The first dataset, CIFAR-10, is a widely recognized benchmark dataset in the fields of machine learning and computer vision. It comprises 60,000 color images

Table 16
strategy d-1: Adjust all training data to 1000.

Class	Training data	*Recall (ViT)	*Accuracy (ViT)	*Recall (GAN)	*Accuracy (GAN)
Center	1000	95.38%	89.53%	75.94%	84.28%
Donut	1000	89.79%		90.94%	
Edge-Loc	1000	76.70%		70.35%	
Edge-Ring	1000	98.32%		95.30%	
Loc	1000	75.77%		79.69%	
Near-full	1000	90.81%		88.65%	
Random	1000	95.65%		94.08%	
Scratch	1000	89.73%		88.86%	

Table 17
STRATEGY D-2: ADJUST ALL TRAINING DATA TO 1500.

Class	Training data	*Recall (ViT)	*Accuracy (ViT)	*Recall (GAN)	*Accuracy (GAN)
Center	1500	93.82%	91.70%	95.96%	90.56%
Donut	1500	89.64%		86.62%	
Edge-Loc	1500	84.61%		81.81%	
Edge-Ring	1500	98.30%		97.89%	
Loc	1500	84.79%		78.73%	
Near-full	1500	69.19%		84.33%	
Random	1500	93.61%		93.24%	
Scratch	1500	84.63%		86.44%	

Table 18
STRATEGY D-3: ADJUST ALL TRAINING DATA TO 3000.

Class	Training data	*Recall (ViT)	*Accuracy (ViT)	*Recall (GAN)	*Accuracy (GAN)
Center	3000	97.66%	93.91%	96.17%	91.74%
Donut	3000	89.79%		88.49%	
Edge-Loc	3000	92.31%		85.19%	
Edge-Ring	3000	98.29%		97.47%	
Loc	3000	84.52%		83.05%	
Near-full	3000	91.89%		86.49%	
Random	3000	91.30%		93.05%	
Scratch	3000	84.16%		85.64%	

Table 19
STRATEGY D-4: ADJUST ALL TRAINING DATA TO 5000.

Class	Training data	*Recall (ViT)	*Accuracy (ViT)	*Recall (GAN)	*Accuracy (GAN)
Center	5000	97.78%	93.88%	96.83%	93.02%
Donut	5000	87.34%		86.47%	
Edge-Loc	5000	92.60%		90.08%	
Edge-Ring	5000	98.12%		97.70%	
Loc	5000	83.70%		86.66%	
Near-full	5000	91.35%		89.19%	
Random	5000	91.39%		84.72%	
Scratch	5000	86.85%		83.29%	

Table 20
SUMMARY OF EXPERIMENTAL RESULTS USING DIFFERENT STRATEGIES.

Strategy/Measure		Recall/Accuracy Best Achieved	Benchmark (VGG/GAN)	Improvement
A	ACCURACY	93.77%	93.56%	0.21%
B1	RECALL	86.69%	83.45%	3.24%
B2	(DONUT)	89.35%	83.45%	5.9%
B3		86.33%	83.45%	2.88%
C1	RECALL	81.14%	80.33%	0.81%
C2	(SCRATCH)	83.89%	80.33%	3.56%
C3		86.21%	80.33%	5.88%
D1	ACCURACY	89.53%	84.28%	5.25%
D2		91.70%	90.56%	1.14%
D3		93.91%	91.74%	2.17%
D4		93.88%	93.02%	0.86%

of dimensions 32 by 32 pixels, distributed across 10 classes, with 6000 images per class. These classes encompass everyday objects such as airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks. The dataset is segmented into 50,000 training images and 10,000 testing images. We allocated 60%, 15%, and 25% of the data for training, validation, and testing, respectively. The results of the augmentation results before and after augmentation are presented in Appendix C, and comprehensive computational reports are available upon request.

The second dataset, MixedWM38 (WaferMap), comprises over 38,000 wafer maps, including one normal pattern, eight single defect patterns, and 29 mixed defect patterns, totaling 38 defect patterns. The eight single defect patterns consist of Center, Donut, Edge-loc, Edge-ring, Local, Near-full, Scratch, and Random. For this experiment, we consider only seven single defect patterns (by excluding the random pattern, which has only 149 samples) with 1000 images per class. It is important to note that there are only 866 images for the near-full class in the dataset. The images in the dataset are of the size 52 by 52 pixels in grey level. Consistent with previous experiments, we allocated 60% of the data for training, 15% for validation, and 25% for testing. Classification results, both before and after augmentation, are demonstrated in Appendix D, and detailed computational reports are available upon request.

4. Conclusion

In the context of sustainable development, semiconductor manufacturers strive to balance economic growth with environmental stewardship and social responsibility. In particular, wafer map defect classification addressed in the paper contributes to this by supporting the principles of sustainable manufacturing through waste reduction, resource efficiency, and quality improvement. The integration of accurate defect classification methods into semiconductor manufacturing processes aligns with sustainable development goals by promoting eco-friendly and responsible production practices.

During semiconductor fabrication, an enormous amount of image data can be collected through in-line inspection. However, the proportion of data categories is often imbalanced, posing a great challenge to training an effective classification model. Under such circumstances, data augmentation is warranted. This paper is to propose a new augmentation framework by means of the ViT model with the purpose of relieving the class-imbalance issue and enhancing the accuracy of wafer defect classification.

This research rigorously investigates the impact of different ViT structures, in terms of modified MSA and MLP mechanisms, on the classification performance. Self-attention insights are gained to create a newly generated heatmap per HEAD in MSA, serving as the proposed augmentation method. The ingenious use of the multi-head outputs of ViT for data augmentation creates a new path of ViT applications in computer vision. The proposed ViT-based augmentation framework is evaluated through a number of augmentation strategies, i.e. based on a group of minority classes and individual classes. Both strategies can be surely applied at the same time. The experimental results reveal that the proposed augmentation framework is effective in increasing the classification accuracy of individual minority classes. Lastly, the well-known and most-popular CycleGAN is introduced as the benchmark augmentation model to assess the proposed ViT augmentation framework. The comparison results do legitimate justice to the merit of the proposed framework over the widely recognized generative model.

Continued research in these areas could lead to more efficient, sustainable, and environmentally friendly semiconductor manufacturing processes by leveraging insights obtained from wafer map defect classification. Technically speaking, an immediate direction of future research could be how the ViT model can be appropriately restructured in multi-headed self-attention for classification performance improvement without recourse to a gigantically large dataset. How the yield of the front-end-of-line (FEOL) and back-end-of-line (BEOL) processes can be safeguarded arising from the proposed augmentation framework bears a further scrutiny. Understanding defect patterns and their correlation with process parameters can lead to improved quality control

and enhanced yields. Future research might focus on utilizing wafer map defect classification to identify root causes of defects, enabling proactive measures to enhance product quality while minimizing resource consumption and waste.

CRediT authorship contribution statement

Shu-Kai S. Fan: Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Shang-Hao Chiu:** Data curation.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported in part by National Science and Technology Council, Taiwan, under Grant No. 111-2221-E-027-070-MY3.

Appendix A. Algorithm 1: Vision Transformer

Pseudo Code of ViT Model.

Step 1. Split an image into N patches, $x_{\text{Patch}} \in \mathbb{R}^{N\times(P\times P\times C)}$, where (P, P, C) is the height and width of each image patch and C is the number of channels;
Step 2. Flatten each patch to sequence $P^2 \times C$ and add positional embeddings;
Step 3. Feed the sequence as an input to a standard **Transformer Encoder (see Algorithm 2)**;
Step 4. Connect the classification module using Multi-layer Perceptron (MLP) to classify input images.

Algorithm2. Pseudo code of Standard transformer encoder Model

Input: the sequence x_{Patch}
1: Perform the layer normalization first;
2: Train three weight matrices **Query**, **Key** and **Value:** W_q, W_k,and W_v.
$Query(Q) = W_q \cdot X$
$Key(K) = W_k \cdot X$
$Value(V) = W_v \cdot X$
3: Obtain **attention function** by combination of a pair of query(Q) and key(K), where $Attention(\alpha) = Q \cdot K^T$;
Output: The process involves the multiplication of attention scores assigned to each input with their corresponding values, followed by the aggregation of weighted values to generate the output.

Appendix B. Parameter Settings of Vision Transformer

Interested readers can find detailed information by referring to the provided link: https://github.com/google-research/vision_transformer.

Appendix C. Validation Results of the Proposed ViT Augmentation Model Using CIFAR-10

Based on the data partitions, 60%, 15% and 25%, for training, validation and testing, respectively. The VGG16 model serves as the classifier and the classification results are presented in this appendix. In the balanced scenario, where 3,600, 900, and 1500 images are allocated for training, validation, and testing for each class respectively, the classification results from five independent runs using the VGG16 classifier are presented in Table R1. An average accuracy of 80.983% was achieved.

Table R1
Accuracy attained by using the VGG16 classifier for the balanced scenario

Class\Run	1	2	3	4	5	Recall (%)
airplane	86.87	82.04	85.33	83.87	88.41	85.30
automobile	92.20	84.21	92.53	87.04	89.23	89.04
bird	76.12	79.47	81.47	76.82	78.21	78.42
cat	66.07	63.67	57.82	63.93	61.27	62.55
deer	74.80	73.95	79.47	81.47	77.07	77.35
dog	74.93	72.73	70.53	71.67	71.73	72.32
frog	87.53	79.01	90.41	86.6	85.93	85.90
horse	83.33	85.13	84.27	82.93	82.73	83.68
ship	90.67	91.07	83.67	93.27	87.61	89.26
truck	85.93	88.13	83.93	89.93	82.13	86.01
Accuracy (%)	81.84	79.84	80.84	81.75	80.43	**80.983**

To simulate an imbalanced scenario similar to the WM-811K dataset, we restructured the original CIFAR-10 dataset into training, validation, and testing subsets, as illustrated in Table R2. The numbers in parentheses represent the additional synthesized images required, to be generated using our proposed ViT augmentation model.

Table R2
Re-organization of CIFAR-10 into an imbalanced scenario.

Class\Partition	Training	Validation	Testing	Proportion
Airplane	1347 (2253)	337 (563)	1500	10.56%
Automobile	530 (3070)	132 (768)	1500	4.15%
Bird	3600 (0)	900 (0)	1500	28.21%
Cat	1798 (1802)	449 (451)	1500	14.09%
Deer	866 (2734)	216 (684)	1500	6.79%
Dog	1531 (2069)	383 (517)	1500	12.00%
Frog	586 (3014)	146 (754)	1500	4.59%
Horse	105 (3495)	26 (874)	1500	0.82%
Ship	1466 (2134)	367 (533)	1500	11.49%
Truck	931 (2669)	233 (667)	1500	7.30%

The classification results for CIFAR-10 using the VGG16 classifier under the imbalanced scenario, without augmentation, are presented in Table R3, achieving an average accuracy of 70.724%. In contrast, the results after applying the proposed ViT augmentation model are displayed in Table R4, with the average accuracy increasing to 79.488%. A comparison of Tables R1-4 reveals that the accuracy rate has significantly improved from 70.724% in the imbalanced scenario to 79.488% in the balanced scenario, largely due to the introduction of the ViT augmentation model. The deviation in accuracy resulting from augmentation is merely within 1.5% from the original accuracy of 80.98%.

Table R3
Accuracy attained by using the VGG16 classifier for the imbalanced scenario without augmentation

Class\Run	1	2	3	4	5	Recall (%)
airplane	77.07	78.33	83.33	79.73	78.87	79.47
automobile	79.47	70.20	74.00	82.07	78.80	76.90
bird	86.47	85.73	84.73	87.67	84.27	85.77
cat	63.13	66.93	60.73	66.00	63.87	64.13
deer	59.60	65.73	63.87	66.60	63.73	63.91
dog	72.07	65.13	70.40	59.73	66.40	66.75
frog	62.00	67.47	70.93	63.07	72.27	67.15
horse	32.73	32.07	28.60	39.73	34.13	33.45
ship	86.60	90.00	87.13	85.33	89.93	87.79
truck	81.33	79.93	81.67	82.13	84.53	81.92
Accuracy (%)	70.05	70.15	70.54	71.21	71.68	**70.724**

Table R4
Accuracy attained by using the VGG16 classifier for the imbalanced scenario with augmentation

Class\Run	1	2	3	4	5	Recall (%)
airplane	82.97	84.44	80.24	80.57	81.71	81.986
automobile	86.09	83.15	87.09	86.95	85.35	85.726
bird	75.09	74.49	80.49	78.49	79.62	77.636
cat	61.61	54.94	53.48	56.61	58.61	57.05
deer	76.76	83.02	74.29	75.82	76.69	77.316
dog	77.47	73.87	70	73.94	75.2	74.096
frog	84.72	81.58	82.78	81.12	85.92	83.224
horse	76.06	79.86	85.46	79.26	76.53	79.434
ship	90.68	91.61	92.95	90.41	90.68	91.266
truck	87.52	88.12	82.25	88.59	89.25	87.146
Accuracy (%)	79.89	79.51	78.90	79.18	79.96	**79.488**

Appendix D. Validation Results of the Proposed ViT Augmentation Model Using MixedWM38

Based on the specified data partitions of 60% for training, 15% for validation, and 25% for testing, the VGG16 model was employed as the classifier, and the classification results are detailed in this appendix. In a balanced scenario, where 600 images are allocated for training, 150 for validation, and 250 for testing for each class (except for the Near-full class), the classification results from five independent runs using the VGG16 classifier are presented in Table R5, achieving an average accuracy of 97.00%.

Table R5
Accuracy attained by using the VGG16 classifier for the balanced scenario

Class\Run	1	2	3	4	5	Recall (%)
Center	98	99	100	99	100	99
Donut	100	99	99	99	100	99
Edge-Loc	97	97	93	95	89	94
Edge-Ring	92	91	93	92	92	92

(*continued on next page*)

Table R5 (*continued*)

Class\Run	1	2	3	4	5	Recall (%)
Loc	99	98	96	98	97	98
Near-full	92	90	96	96	95	94
Scratch	100	100	100	100	100	100
Accuracy (%)	96.92	96.47	96.70	97.04	96.28	**97**

To replicate an imbalanced scenario akin to the WM-811K dataset, we reorganized the original MixedWM38 dataset into training, validation, and testing sets, as detailed in Table R6. The figures in parentheses denote the extra synthetic images needed, which will be created using our proposed ViT augmentation model.

Table R6
Re-organization of MixedWM38 into an imbalanced scenario

Class\Partition	Training	Validation	Testing	Proportion
Center	88 (512)	22 (128)	250	4.26%
Donut	107 (493)	27 (123)	250	5.18%
Edge-Loc	586 (14)	147 (3)	250	28.35%
Edge-Ring	381 (219)	95 (55)	250	18.43%
Loc	495 (105)	124 (26)	250	23.95%
Near-full	57 (463)	14 (11)	216	2.76%
Scratch	353 (247)	88 (62)	250	17.07%

The classification results for MixedWM38, obtained using the VGG16 classifier in an imbalanced scenario without augmentation, are detailed in Table R7, generating an average accuracy of 92.32%. By contrast, the classification results after implementing the proposed ViT augmentation model are presented in Table R8, revealing an enhanced average accuracy of 95.86%. A comparison across Tables R5-8 underscores a significant enhancement in accuracy, rising from 92.32% under the imbalanced scenario to 95.86% in the balanced scenario, primarily attributed to the incorporation of the ViT augmentation model. The deviation in accuracy due to augmentation is only 1.14% from the initial accuracy of 97.00%.

Table R7
Accuracy attained by using the VGG16 classifier for the imbalanced scenario without augmentation

Class\Run	1	2	3	4	5	Recall (%)
Center	90.40	93.20	92.40	91.60	90.80	91.68
Donut	97.60	98.00	98.80	100.00	98.40	98.56
Edge-Loc	86.40	86.80	86.40	80.80	80.80	84.24
Edge-Ring	88.40	86.80	92.40	88.00	94.80	90.08
Loc	94.00	90.80	90.40	94.00	93.20	92.48
Near-full	90.40	91.20	89.20	92.40	90.80	90.80
Scratch	98.40	97.20	99.20	98.40	98.80	98.40
Accuracy (%)	92.23	92.00	92.69	92.17	92.51	**92.32**

Table R8
Accuracy attained by using the VGG16 classifier for the imbalanced scenario with augmentation

Class\Run	1	2	3	4	5	Recall (%)
Center	97.20	96.80	97.60	98.80	97.60	97.60
Donut	99.20	98.80	98.80	98.80	99.60	99.04
Edge-Loc	96.00	96.40	94.0	93.2	92.20	94.36
Edge-Ring	91.20	90.80	91.20	90.80	91.20	91.04
Loc	98.80	98.40	95.20	96.40	95.60	96.88
Near-full	90.30	89.35	96.30	96.30	92.60	92.97
Scratch	99.20	97.60	100	99.20	99.60	99.12
Accuracy (%)	95.99	95.45	96.16	96.21	95.49	**95.86**

References

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer Normalization arXivpreprint arXiv:arXiv: 1607.06450.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale arXiv: 2010.11929.

Fan, S.-K.S., Cheng, C.-W., Tsai, D.-M., 2022. Fault diagnosis of wafer acceptance test and chip probing between front-end-of-line and back-end-of-line processes. IEEE Trans. Autom. Sci. Eng. 19 (4), 3068–3082.

Fan, S.-K.S., Tsai, D.-M., Yeh, P.-C., 2023a. Effective variational-autoencoder-based generative models for highly imbalanced fault detection data in semiconductor manufacturing. IEEE Trans. Semicond. Manuf. 36 (2), 205–214.

Fan, S.-K.S., Tsai, D.-M., Shih, Y.-F., 2023b. Self-assured deep learning with Minimum pre-labeled data for wafer pattern classification. IEEE Trans. Semicond. Manuf. 36 (3), 404–415.

Fan, S.-K.S., Chen, M.-S., Hsu, C.-Y., Park, Y.-J., 2023c. An artificial intelligence transformation model—pod redesign of photomasks in semiconductor manufacturing. J. Ind. Prod. Eng. 41 (3), 201–216.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems 2, 2672–2680.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition arXiv preprint arXiv: arXiv:1512.03385.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
Hsu, S.-C., Chien, C.-F., 2007. Hybrid data mining approach for pattern extraction from wafer bin map to improve yield in semiconductor manufacturing. Int. J. Prod. Econ. 107, 88–103.
Katsaliaki, K., Kumar, S., Loulos, V., 2024. Supply chain coopetition: a review of structures, mechanisms and dynamics. Int. J. Prod. Econ. 267, 109057.
Kong, Y., Ni, D., 2020. A semi-supervised and incremental modeling framework for wafer map classification. IEEE Trans. Semicond. Manuf. 33 (1), 62–71.
Krishnan, K.S., Krishnan, K.S., 2021. Vision transformer based COVID-19 detection using chest X-rays. ISPCC 644–648.
Krizhevsky, A., Hinton, G., 2009. CIFAR-10 (Canadian Institute for Advanced Research). Retrieved from. https://www.cs.toronto.edu/~kriz/cifar.html.
Kyeong, K., Kim, H., 2018. Classification of mixed-type defect patterns in wafer bin maps using convolutional neural networks. IEEE Trans. Semicond. Manuf. 31 (3), 395–402.
Saqlain, M., Abbas, Q., Lee, J.Y., 2020. A deep convolutional neural network for wafer defect identification on an imbalanced dataset in semiconductor manufacturing processes. IEEE Trans. Semicond. Manuf. 33 (3), 436–444.
Shon, H.S., Batbaatar, E., Cho, W.-S., Choi, S.G., 2021. Unsupervised pre-training of imbalanced data for identification of wafer map defect patterns. IEEE Access 9, 52352–52363.
Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition arXiv preprint arXiv:1409.1556.
Tan, M., Le, Q.V., 2019. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, ICML 2019. Long Beach, pp. 6105–6114.
Tello, G., Al-Jarrah, O.Y., Yoo, P.D., Al-Hammadi, Y., Muhaidat, S., Lee, U., 2018. Deep-structured machine learning model for the recognition of mixed-defect patterns in semiconductor fabrication processes. IEEE Trans. Semicond. Manuf. 31 (2), 315–322.
Tsai, T.-H., Lee, Y.-C., 2020. A light-weight neural network for wafer map classification based on data augmentation. IEEE Trans. Semicond. Manuf. 33 (4), 663–672.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention Is All You Need arXiv:1706.03762.
Wang, J., Yang, Z., Zhang, J., Zhang, Q., Chien, W.K., 2019. AdaBalGAN: an improved generative adversarial network with imbalanced learning for wafer defective pattern recognition. IEEE Trans. Semicond. Manuf. 32 (3), 310–319.
Wang, J., Xu, C., Yang, Z., Zhang, J., Li, X., 2020. Deformable convolutional networks for efficient mixed-type wafer defect pattern recognition. IEEE Trans. Semicond. Manuf. 33 (4), 587–596.
Wang, R., Chen, N., 2019. Wafer map defect pattern recognition using rotation-invariant features. IEEE Trans. Semicond. Manuf. 32 (4), 596–604.
Wang, S., Zhong, Z., Zhao, Y., Zuo, L., 2021. A variational autoencoder enhanced deep learning model for wafer defect imbalanced classification. IEEE Trans. Compon. Packaging Manuf. Technol 11, 2055–2060.
Wu, M., Jang, J.R., Chen, J., 2015. Wafer map failure pattern recognition and similarity ranking for large-scale data sets. IEEE Trans. Semicond. Manuf. 28 (1), 1–12.
Wu, H., Zhao, Z., Wang, Z., 2023. META-unet: multi-scale efficient transformer attention unet for fast and high-accuracy polyp segmentation. IEEE Trans. Auto. Sci. Eng. https://doi.org/10.1109/TASE.2023.3292373 (in press).
Yu, N., Xu, Q., Wang, H., 2019. Wafer defect pattern recognition and analysis based on convolutional neural network. IEEE Trans. Semicond. Manuf. 32 (4), 566–573.
Yu, J., 2019. Enhanced stacked denoising autoencoder-based feature learning for recognition of wafer map defects. IEEE Trans. Semicond. Manuf. 32 (4), 613–624.
Yu, T.-S., Han, J.-H., 2021. Scheduling proportionate flow shops with preventive machine maintenance. Int. J. Prod. Econ. 231, 107874.
Yu, J., Liu, J., 2021. Multiple granularities generative adversarial network for recognition of wafer map defects. IEEE Trans. Ind. Electron. 18 (3), 1674–1683.
Yu, J., Shen, Z., Zheng, X., 2021. Joint feature and label adversarial network for wafer map defect recognition. IEEE Trans. Autom. Sci. Eng. 18 (3), 1341–1353.
Zhao, S., Li, H., Ke, Q., Liu, L., Zhang, R., 2022. Action-ViT: pedestrian intent prediction in traffic scenes. IEEE Signal Process. Lett. 29, 324–328.
Zhu, Y.J., Park, T., Isola, P., Efros, A.A., 2020. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks arXiv:1703.10593vol. 7.