IEICE Electronics Express, Vol.VV, No.NN, 1–6

LETTER

UnionMixup and Max-Min-Saliency Mixup for Mixed-type Defect Recognition of Wafer Bin Maps

Qingqing Yu [1a)]

Abstract Defect pattern detection of wafer bin maps (WBMs) is vital in wafer quality improvement owing to preventing further defects and resource waste. We proposed two Mixup approaches to train Vision Transformer under only single defect WBM samples for mixed-type defects recognition. We use UnionMixup and Token level Max-Min-Saliency Mixup to generate mixed-type defect WBMs to feed Vision Transformers. In the recognition of two-mixed defect types WBMs, our method improves 17.1 % compared to baseline (none mixup) and we have 1.7% accuracy gain compared with state-of-the-art mixup approaches. In the recognition mixed defect samples containing more than two-mixed defects (three-mixed and four-mixed), we gain at least 24.7% (compared with baseline) and 11.1% (compared with single SOTA mixup) respectively. The combination of Union Mixup and Token level Max-Min-Saliency Mixup become better than other SOTA mixup methods obviously in mixed patterns including more than three defects.

key words: Defect pattern classification, Mixed-type defects , WBM classification, Mixup, Vision Transformer

Classification: Electron devices, circuits and modules (silicon, compound semiconductor, organic and novel materials)

1. Introduction

The requirement and manufacture outputs of wafers have boosted in the latest decades which results in manual detection inadequate. Each wafer has thousands of chips which should be analyzed by several detection devices at different phases of manufacturing procedure for defect detection. A wafer bin map (WBM) is a two dimensional binary map illustrates the electrical detection outcomes. Correct dies are specified a value of zero, while error dies are specified a value of one. We classify these defect WBM patterns into 9 categories: Edge-Loc, Edge-Ring, Center, Scratch, Donut, Loc, Random, Near-Full, and Normal (with none defect) as shown in Fig.1. To achieve high outputs and high quality in wafer manufacture, we need better automatic apparatuses with high recognition accuracy of WBMs defect patterns [1, 2]. So far, most researches focus on designing data-oriented models for WBM defect patterns classification [3, 4]. A supervised model of convolutional neural networks trained by an end-to-end way with minimal feature [5] is one of the most efficient methods. Since obtaining the huge size datasets for training are time-costly and labour-intensively, various efforts have been made on synthesis of two labelled samples. Moreover, the work of label the samples of two or more type defect patterns is more difficult for mixed-type patterns grow exponentially according to the number of single-defect patterns. So the build of a labelled WBM dataset containing mixed-type defect labels is impractical. To solve this issue, we propose an architecture that training vision transformers only by single-defect samples and classify samples containing mixed-type defects. We use improved inputMixup [6] methods and Token-level mixup which makes significant improvements in recognizing mixed-type defect WBMs that have not appeared in the single-type dataset. We propose Max-Min saliency Token Mixup(MMSTM for short) to improve the recognition accuracy of three or four mixed-types. MMSTM surpasses state-of-the-art mixup approaches on five different vision transformer models, and it demonstrates that our approach is not restricted to specific framework. After we have done UnionMixup of two type defect patterns, when the mixed WBMs are further mixed in MMSTM, three-mixed defect patterns or four-mixed patterns will be generated to feed the transformer. Specially, compared with other SOTA methods, the architecture we proposed perform more efficiently in classification of mixed type containing three defect patterns or more than three type defect patterns. In short, our contributions are as follows: (1) We propose UnionMixup which is effective for mixed type defect patterns classification in WBMs in the early stage. (2) We propose MMSTM a Token level Mixup method based on Max-Min saliency optimal assignment which is efficient in augmenting the training of Vision Transformer. (3) The architecture of UnionMixup

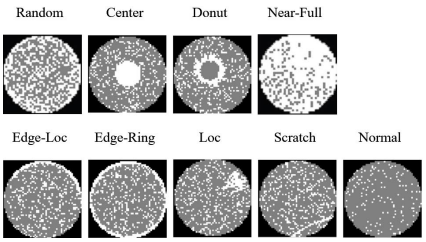

Fig. 1. Visualization of the nine single defect patterns in WM-811 K

[1]Dept. of Information Management, MinNan University of Science and Technology, Quanzhou 999001-362700, China
a) 476931017@qq.com

DOI: 10.1587/elex.21.20240054
Received January 26, 2024
Accepted March 18, 2024
Publicized April 01, 2024

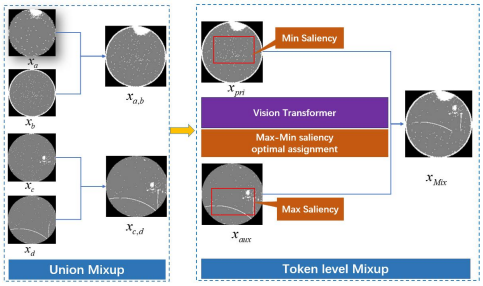

Fig. 2. Framework of the combination of Union Mixup and MMSTM. (Here, $\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c, \mathbf{x}_d \in \mathcal{D}_{low}$, $\mathbf{x}_{a,b}$,is the Union Mixup of $\mathbf{x}_a$ and $\mathbf{x}_b$,and $\mathbf{x}_{c,d}$ is the Union Mixup of $\mathbf{x}_c$ and $\mathbf{x}_d$. $\mathbf{x}_{Mix}$ is the Token level mixup with Max-Min Saliency optimal assignment of $\mathbf{x}_{a,b}$ and $\mathbf{x}_{c,d}$.)

combined with MMSTM improve the robustness of Vision Transformer's capability in WBM defect pattern classification, especially when three or more type defect patterns are mixed. The framework of this approach is illustrated in Fig.2.

2. Related work

So far, most works focused on recognition of single-defect pattern WBMs, our target is the classification of mixed-type defect patterns in this paper. Seldom approaches have been suggested for the classification of mixed-type WBM defects. [7] trained individual CNN models for single defect and then combined the result of single-type CNNs for predictions of the mixed-type defects classification. [8] adopted semi-supervised deep generative network in training unlabeled data for recognition of mixed-type wafer defect pattern. Memory-augmented CNN model containing a triplet loss function had been mentioned in [9] to obtain effective low-dimensional embedding perform efficiently in imbalanced WBM dataset. For the mixed-type WBM defects diversify vastly in angles and positions, correct detection of defect patterns is challenging. Due to standard training in CNNs and transformers can only learn the common features while neglecting the rare features, as a result it brings bad generalization performance. [10] have shown that Mixup can successfully mix the common and rare features so that the gradients along these two features are correlated. Therefore, learning of rare feature can be boosted by the fast learning of common features, and eventually reaches a rather high level to outweigh the influences of noise on test samples. In order to detect mixed-type defect patterns, we propose a unique framework to incorporate UnionMixup and MMSTM into training Vision Transformers based on only single defect dataset. Mixup firstly introduced for data augmentation by interpolate two instances linearly [6, 11, 12]. Nonlinear mixup of two different random areas is one replacement of masking square areas [13] in cutMix method [14], while some variants considering arbitrary regions [15, 16, 17]. To avoid selecting localities randomly, saliency is applied to detect objects from different instances and combine them into one sample [18, 19, 20, 21]. And we do Token level Mixup based on Max-Min saliency optimal assignment in transformers. We proposed a novel architecture to solve more challenging task like the detection of mixed-type patterns WBMs composed of four defects. The Transformer model will converges faster when the architecture has two phases of mixup algorithms: UnionMixup and MMSTM.

3. Methodology

Here, we introduce UnionMixup and MMSTM. The goal of our method is to augment intermediate tokens while maximizing the saliency level. The whole architecture can be accomplished in three major steps: 1) Calculate defect ratio, 2) The samples whose defect ratio lower than 0.5 can be used to do UnionMixup, 3) All samples no matter whether defect ratio is lower 0.5 participate in MMSTM. The following subsections provide detailed descriptions of the steps.

3.1 Problem formulation

Suppose that a single labelled defect dataset including m WBMs in which each image is labelled with one of the c classes. The i^{th} WBM, $\mathbf{x}^i \in \mathbb{R}^{H\times W}$ $(i = 1, 2, \ldots, m)$. Let $(\mathbf{x}, \mathbf{y})$ be an WBM image $\mathbf{x} \in \mathcal{X}$ with its one-hot encoded class label $\mathbf{y} \in \mathbf{Y}$, where $\mathcal{X}$ is the input image space, $\mathbf{Y} = [0, 1]^c$ and c is denoted as the defective pattern categories in the classification of WBMs. Suppose a trainable parameters θ , the object of WBM defect pattern recognition is to learn a function f_θ (e.g., Vision Transformer or CNN) based on a training dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})_{i=1}^m\}$ only containing single defect-labeled samples. So as to correctly map WBMs to their corresponding target labels.

3.2 Defect ratio of WBM

Defect ratio r_{defect} for wafer bin map can be defined as Equation (1). For normal dies and null dies (outside the circular distribution) are defined a value of zero. So, all dies in a normal WBM are zero while considering none noise in the image of WBM.

$$r^{(d)}_{defect} = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{x}^{(d)}_{i,j}}{H\times W}\mathbf{x}^{(d)}_{i,j} \in \{0, 1\}, d = 1, 2, \ldots, m \quad (1)$$

Here, m is the total amount of WBMs. $\mathbf{x}^{(d)}$ is the d^{th} WBM in the single defect-labeled training dataset $\mathcal{D}$, $r^{(d)}_{defect}$ is defined as the defect ratio of $\mathbf{x}^{(d)}$. If $\mathbf{x}^{(d)}_{defect} < 0.5$, $\mathbf{x}^{(d)} \in \mathcal{D}_{low}$ that is $\mathbf{x}^{(d)}$ will be included in set $\mathcal{D}_{low}$ which is the image space of WBM with low defect degree. And inputMixup and its variants can be adopted based on the dataset $\mathcal{D}_{low}$. Based on the defect ratio defined in Equation (1), the defect complexity is determined. Considering the damage severity degree is rather high when defect ratio is larger than 0.5, so the high degree damaged wafers does not need to be mixed in the inputMixup process. However, all wafers will be trained in the transformer.

3.3 Union-Mixup for WBMs

Let $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ be an WBM image $\mathbf{x}^{(i)} \in \mathcal{D}_{low}$ with its one-hot encoded class label $\mathbf{y}^{(i)} \in \mathbf{Y}$. Union Mixup virtually train instances are synthesized by taking the union of defect dies sets within two WBMs. Considering the binary characteristics of WBM data, we can only calculate the union of abnormal dies sets included in two source data instances $(\mathbf{x}^{(a)}, \mathbf{y}^{(a)})$ and $(\mathbf{x}^{(b)}, \mathbf{y}^{(b)})$. Union Mixup $U(\mathbf{x}^{(a)}, \mathbf{x}^{(b)})$ is denoted as the union of dies sets of two data instances $(\mathbf{x}^{(a)}, \mathbf{y}^{(a)})$ and $(\mathbf{x}^{(b)}, \mathbf{y}^{(b)})$. Here, $\mathbf{x}_{mix}^{(a,b)} = U(\mathbf{x}^{(a)}, \mathbf{x}^{(b)}), \mathbf{y}_{mix}^{(a,b)} = f_\theta(U(\mathbf{x}^{(a)}, \mathbf{x}^{(b)}))$. Because abnormal dies are defined a value of one, while normal dies and null dies (outside the circular distribution) are defined a value of zero. $U(x^{(a)}, x^{(b)})$ can be defined as $\mathbf{x}^{(a)} || \mathbf{x}^{(b)}$ to avoid unpractical values that exceed one.

$\mathbf{y}_{mix}^{(a,b)} = \lambda \mathbf{y}^{(a)} + (1-\lambda)\mathbf{y}^{(b)}$. Here, $\lambda = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} \mathbf{x}_{i,j}^{(a)}}{\sum_{i=1}^{H}\sum_{j=1}^{W} \mathbf{x}_{i,j}^{(a)} + \sum_{i=1}^{H}\sum_{j=1}^{W} \mathbf{x}_{i,j}^{(b)}}$.

$\mathbf{y}_{mix}^{(a,b)} = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} \mathbf{x}_{i,j}^{(a)}}{\sum_{i=1}^{H}\sum_{j=1}^{W} \mathbf{x}_{i,j}^{(a)} + \sum_{i=1}^{H}\sum_{j=1}^{W} \mathbf{x}_{i,j}^{(b)}} \mathbf{y}^{(a)} + \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} \mathbf{x}_{i,j}^{(b)}}{\sum_{i=1}^{H}\sum_{j=1}^{W} \mathbf{x}_{i,j}^{(a)} + \sum_{i=1}^{H}\sum_{j=1}^{W} \mathbf{x}_{i,j}^{(b)}} \mathbf{y}^{(b)}$.

Where $a = 1, 2, \ldots, m, b = 1, 2, \ldots, m$. Substitute equation(1), we can get the following equation to simplify the algorithm. $\mathbf{y}_{mix}^{(a,b)} = \frac{r_{defect}^{(a)}}{r_{defect}^{(a)} + r_{defect}^{(b)}} \mathbf{y}^{(a)} + \frac{r_{defect}^{(b)}}{r_{defect}^{(a)} + r_{defect}^{(b)}} \mathbf{y}^{(b)}$. The Union mixed up samples will be added to the original dataset of WBMs $\mathcal{D}$ to generate new dataset of WBM is defined as $\mathcal{D}_{new}$.

3.4 Max-Min Saliency optimal assignment

Given a pair of training samples $(\mathbf{x}^{(a)}, \mathbf{y}^{(a)})$ and $(\mathbf{x}^{(b)}, \mathbf{y}^{(b)})$, We first partition the input image $\mathbf{x}$ into non-overlapping patches $\mathbf{x}^p \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P} \times P^2}$, (P, P) is the resolution of each image patch. The flattened 2D patches is $\mathbf{x}^p \in \mathbb{R}^{n \times d}$, here $d = P^2$,$n = \frac{H}{P} \times \frac{W}{P}$ is the resulting number of patches. [22, 23, 24, 25] investigated the imbalanced information of tokens, and token level mixing randomly induces significant information loss and useless token exchanges. Rather than using gradient-based saliency detectors which requires substantial computation [18, 19], we exploit the benefit of self-attention map, an inherent saliency approximation in transformer modules. So we can deduce the saliency of the tokens from the attention imposed on the tokens as $\mathbf{A} = \frac{1}{N_h}\sum_{h=1}^{N_h} \mathbf{A}_h$, where N_h is the number of heads in the multi-head attention layer, and $\mathbf{A}_h$ is the h^{th} attention head. Then, the saliency score $\mathbf{S}_t$ can be computed as $\mathbf{S}_t = \frac{1}{n}\sum_{i=1}^{n} \mathbf{A}_{i,t}$, where $t = 1, 2, \ldots, n$ is the token index.

Based on the estimated saliency of tokens, we aim at maximizing the overall saliency level by optimally assigning a different mixup target for each instance. Max-Min saliency optimal assignment is carried out between two instances: primary WBM and auxiliary WBM. It replaces a minimum-scored rectangle area of the primary image with a maximum-scored rectangle area of the auxiliary image, so as to maximize saliency of mixed WBMs. The visualization is in Fig.2, we partition the primary image $\mathbf{x}_{pri}$ and the auxiliary image $\mathbf{x}_{aux}$ into non-overlapping patches of size $P \times P$. We define $n_{row} = \frac{H}{P}$ and $n_{col} = \frac{W}{P}$ to simplify the formulas. A total of $n = n_{row} \times n_{col}$ patches are obtained for each image. However, $\mathbf{x}_{pri}$ and $\mathbf{x}_{aux}$ are reorganized as $\mathbf{x}_{pri}, \mathbf{x}_{aux} \in \mathbb{R}^{n_{row} \times n_{col} \times P^2}$, one element of which is a token in vision transformer. Then, they are the input for a Vision Transformer to get the corresponding image saliency scores $\mathbf{S}_{pri} \in \mathbb{R}^n$ and $\mathbf{S}_{aux} \in \mathbb{R}^n$. Similarly, we rearrange the shape of their image saliency score vectors, $\mathbf{S}_{pri}$ and $\mathbf{S}_{aux}$, to matrices of $n_{row} \times n_{col}$. This method clips a rectangle area from the auxiliary WBM and pastes the rectangle area into the primary WBM to rebuild a mixed WBM. For optimal assignment, we use a selection ratio ρ, sampled from a uniform distribution (0.15, 0.85), to determine the total $\lfloor \rho n_{row} \rfloor \times \lfloor \rho n_{col} \rfloor$ patches within the selected rectangle area. The kernel of the optimal algorithm is to choose the most informative rectangle area in the auxiliary WBM, and the least informative rectangle area in the primary WBM. The start indices of these two rectangle areas are defined as the following equations.

$$i_{pri}^*, j_{pri}^* = \underset{i,j}{argmin} \sum_{p,q} \mathbf{S}_{pri}^{i+p, j+q}, \mathbf{S}_{pri} \in \mathbb{R}^{n_{row} \times n_{col}}$$

$$i_{pri}^*, j_{pri}^* = \underset{i,j}{argmin} \sum_{p,q} \mathbf{S}_{pri}^{(i-1+p) \times n_{col} + j+q}, \mathbf{S}_{pri} \in \mathbb{R}^n$$

$$i_{aux}^*, j_{aux}^* = \underset{i,j}{argmax} \sum_{p,q} \mathbf{S}_{aux}^{i+p, j+q}, \mathbf{S}_{aux} \in \mathbb{R}^{n_{row} \times n_{col}}$$

$$i_{aux}^*, j_{aux}^* = \underset{i,j}{argmax} \sum_{p,q} \mathbf{S}_{aux}^{(i-1+p) \times n_{col} + j+q}, \mathbf{S}_{aux} \in \mathbb{R}^n$$

Here,$1 \le i \le n_{row}, 1 \le i \le n_{col}$, $h_r = \lfloor \rho n_{row} \rfloor$, $w_r = \lfloor \rho n_{col} \rfloor$, $p \in \{0, 1, \ldots, h_r - 1\}$, and $q \in \{0, 1, \ldots, w_r - 1\}$.

3.5 Token level Mixup

Intuitively, the selected rectangle area contains patches with the maximum saliency score of the auxiliary WBM and the minimum saliency score of the primary WBM. Then, we obtain the new mixed training instance $\mathbf{x}_{Mix}$, $\mathbf{y}_{Mix}$ as follows: Firstly we set $\mathbf{x}_{Mix} = \mathbf{x}_{pri}$, then $\mathbf{x}_{Mix}^{i_{pri}^*+p, j_{pri}^*+q} = \mathbf{x}_{aux}^{i_{aux}^*+p, j_{aux}^*+q}$. $(p \in \{0, 1, \ldots, h_r - 1\}, q \in \{0, 1, \ldots, w_r - 1\})$. $\mathbf{y}_{Mix} = \lambda_{Mix}\mathbf{y}_{pri} + (1 - \lambda_{Mix})\mathbf{y}_{aux}$, where λ_{Mix} is defined in the following section. Similarly to CutMix, we then generate an appropriate binary mask $\mathbf{M} \in n_{row} \times n_{col}$ according to the selection ratio ρ. The mixed new training sample $(\mathbf{x}_{mix}, \mathbf{y}_{mix})$ is created as follows:

$$\mathbf{x}_{mix} = \mathbf{M} \odot \mathbf{x}_{pri} + (1 - \mathbf{M}) \odot \mathbf{x}_{aux},$$

$$\mathbf{y}_{Mix} = \frac{\sum_{t=1}^{n} \mathbf{M}_t \cdot \mathbf{S}_{pri}^t}{\sum_{t=1}^{n} (\mathbf{M}_t \cdot \mathbf{S}_{pri}^t + (1 - \mathbf{M}_t) \cdot \mathbf{S}_{aux}^t)} \cdot \mathbf{y}_{pri} + \frac{\sum_{t=1}^{n} (1 - \mathbf{M}_t) \cdot \mathbf{S}_{aux}^t}{\sum_{t=1}^{n} (\mathbf{M}_t \cdot \mathbf{S}_{pri}^t + (1 - \mathbf{M}_t) \cdot \mathbf{S}_{aux}^t)} \cdot \mathbf{y}_{aux}$$

So $\lambda_{Mix} = \frac{\sum_{t=1}^{n} \mathbf{M}_t \cdot \mathbf{S}_{pri}^t}{\sum_{t=1}^{n} (\mathbf{M}_t \cdot \mathbf{S}_{pri}^t + (1 - \mathbf{M}_t) \cdot \mathbf{S}_{aux}^t)}$. where $\mathbf{S}$ indicates the set of saliency scores of all tokens, $\mathbf{S}_{pri}^t$, $\mathbf{S}_{aux}^t$ indicates

Table I. Labeled samples distribution of 9 categories in WM-811k

Class label	Count	Proportion(%)	Class label	Count	Proportion(%)
Donut	555	0.32	Near-Full	149	0.09
Center	4294	2.48	Random	866	0.50
Edge-Loc	5189	3.00	Scratch	1193	0.69
Loc	3593	2.08	None	147431	85.25
Total	172950	100			

the $t^{th}(t = 1, 2, \ldots, n)$ saliency score of primary WBM and auxiliary WBM respectively. $\odot$ denotes element-wise multiplication, $\mathbf{M}_t$ denotes the t^{th} token of the mask $\mathbf{M}$.

4. Experiments

4.1 Dataset

We use WM-811 K dataset an open real-world dataset of 811,457 WBM samples to demonstrate the efficiency of our method. The dataset has a subset of 172,950 wafer images with single-defect labels. As shown in Fig.1, the labelled single-defect pattern has nine categories: Normal (none error die), Center ,Edge-Loc, Loc, Edge-Ring, Donut , Scratch, Random, and Near-Full. WBMs in the WM-811 K have different resolutions varying from 6 × 21 to 300 × 302, we change the size of every WBM to 64 × 64 resolution by nearest neighbor interpolation for experiment.

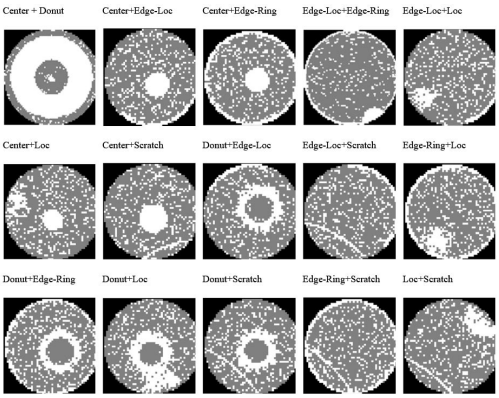

Fig. 3. Visualization of 15 possible mixed categories generated by two distinct defect patterns.

As illustrated in Table I, the Near-Full pattern occupies merely 0.09% and the Normal pattern has the highest proportion of 85% in WM-811 K dataset. The category distribution of labelled data is imbalance. So, we use the inverse class frequencies to calculate WBM sample weights and dynamically do sampling from multinomial distribution to acquire balanced mini-batches to feed Transformer. We do not need mixed-type defect data in transformer training, but some mixed-type data are needed for testing to demonstrate the final accuracy. To get the mixed-type test samples, we apply the following approach to build test datasets containing mixed-type defect pattern data. Firstly, among WM-811 K, we get samples of two different WBMs in six single-defect pattern categories including: Scratch, Loc, Donut,Center, Edge-Ring, and Edge-Loc. We do not inspect mix-type about

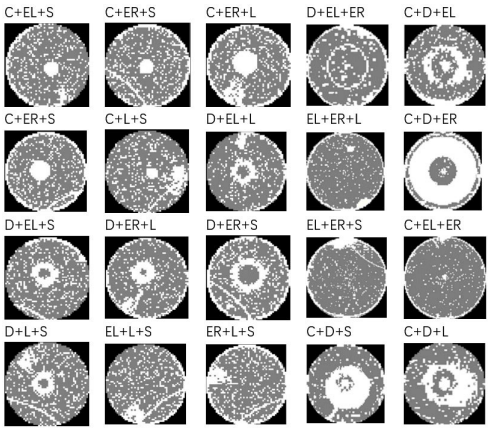

Fig. 4. Visualization of 20 categories of three-mixed different defect patterns (D: Donut, C: Center,ER: Edge-Ring, EL: Edge-Loc, L:Loc, S: Scratch.)

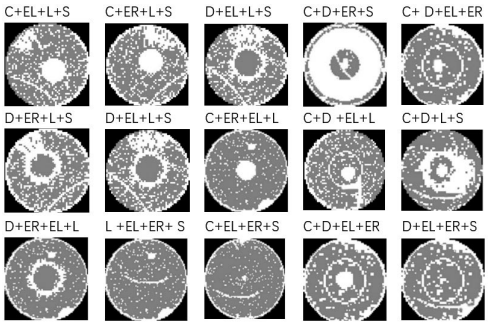

Fig. 5. Visualization of 15 categories of four-mixed different defect patterns

Table II. WBM classification Accuracy of test dataset including two-mixed defect pattern.

Accuracy(%)	ViTLite	CVT	CCT	Swin	LeViT
Baseline	74.6	79.1	81.1	81.3	83.2
InputMixup	85.4	86.3	88.2	88.1	89
Cutmix	84.9	86.3	87.9	88	88.6
PuzzleMix	89.4	89.6	91	90.5	91.4
ManifoldMixup	85.7	86.9	88.2	87.4	89
UnionMix	85.6	86.4	89.2	88.1	90.1
TokenMix	86.2	87.4	88.9	88.1	90.2
MMSTM	89.1	90.1	92.1	91.2	92.5
Input-CutMix	88.3	89.3	90.1	89.7	91.3
Input-TokenMix	89.5	90.3	92.4	91.9	92.8
Union-TokenMix	90.4	91.6	93.1	92.5	93.4
Input-MMSTM	90.8	91.7	93.4	93.1	93.8
Union-MMSTM	91.7	92.1	93.8	94.3	94.7

Random and Near-Full for most of the time they dominate the map while mixing with any other types.Moreover, we synthesize two WBMs to generate a mixed-type defect pattern WBM and then sort out and pick up the most realistic WBMs for test. So, we have 15 mixed-type defect patterns WBMs as illustrated in Fig.3. We chosen 200 wafer bin maps in each of 15 possible compositions of mixed-type defect patterns. At last, we examine our approach by test dataset consisting of 3000 samples. We also get 20 categories samples of three mixed type patterns visualized in Fig.4. Similarly 15 categories of four mixed type samples visualized in Fig.5.

Table III. WBM classification Accuracy of test dataset including two-mixed and three-mixed defect pattern.

Accuracy(%)	ViTLite	CVT	CCT	Swin	LeViT
Baseline	65.3	66.3	70.1	69.1	70.4
InputMixup	78.2	81.2	84.3	84.2	85.3
Cutmix	77.5	80.1	82.4	83.6	84.2
PuzzleMix	80.8	81.9	82.7	83.4	84.9
ManifoldMixup	78.6	80.1	82.2	82.9	83.5
UnionMix	80.2	82.3	84.2	84.7	85.3
TokenMix	82.2	84.1	85.2	85.8	86.3
MMSTM	83.1	84.9	86.6	86.2	87.4
Input-CutMix	82.5	84.1	86.1	85.6	87.2
Input-TokenMix	84.3	85.2	86.8	87.1	87.4
Union-TokenMix	85.3	85.7	87.5	87.8	88.3
Input-MMSTM	86.2	87.1	88.7	89.2	89.8
Union-MMSTM	89.1	90.1	91.9	92.4	93.6

Table IV. WBM classification accuracy of test dataset including two-mixed, three-mixed, four-mixed defect pattern.

Accuracy(%)	ViTLite	CVT	CCT	Swin	LeViT
Baseline	53.2	55.7	58.2	59.7	60.8
InputMixup	62.5	64.6	66.1	65.7	66.3
Cutmix	61.4	63.8	66.2	66.1	68.1
PuzzleMix	64.2	66.2	68.4	68.9	69.3
ManifoldMixup	62.5	64.3	66.8	67.4	69.1
UnionMix	62.8	64.7	66.3	67.1	68.6
TokenMix	64.1	65.8	67.8	66.9	68.9
MMSTM	66.8	68.3	70.4	69.8	71.2
Input-CutMix	69.3	70.2	72.5	72.1	73.6
Input-TokenMix	71.4	72.2	74.5	74.3	75.1
Union-TokenMix	73.5	75.2	77.8	77.5	78.9
Input-MMSTM	75.8	77.2	79.5	79.8	81.2
Union-MMSTM	77.9	82.8	86.5	87.3	89.1

4.2 Implementation Details

We evaluate our method on several recent vision transformer architectures, including ViT Lite, Compact Vision Transformer (CVT) [26], Compact Convolution Transformer (CCT) [26], Swin Transformer [27], LeViT [28]. The batch size is set to 1024. We adopt AdamW [29] as the optimizer and set the learning rate as 0.001 with 5 warm-up epochs. The learning rate is decayed following a cosine scheduler down to 10^{-6}. Without other specification, we train the models for 300 epochs. For training architectures with smaller model sizes: ViT-Lite-6/4 [26], CVT-6/4 [26], CCT-7/3×1 [26], Swin-T [27], LeViT-192 [28]. Considering the mixed images are more likely to contain multiple labels, we adopt binary cross-entropy (BCE) loss rather than the cross-entropy (CE) loss [30, 31]. In order to examine the performance of Union Mixup and MMSTM, we try to combine MMSTM with other mixup algorithms, also try the combination of UnionMixup with other mixup methods. In TableII, Input-CutMix is the combination of inputMixup and CutMix. Similarly, Input-TokenMix is inputMixup + TokenMixup, Union-TokenMix is UnionMixup + TokenMixup, and Union-MMSTM is UnionMixup + MMSTM.

4.3 Classification of mixed-defect pattern WBMs

We apply the official code designed by the authors to execute the experiment of the state-of-the-art (SOTA) approaches

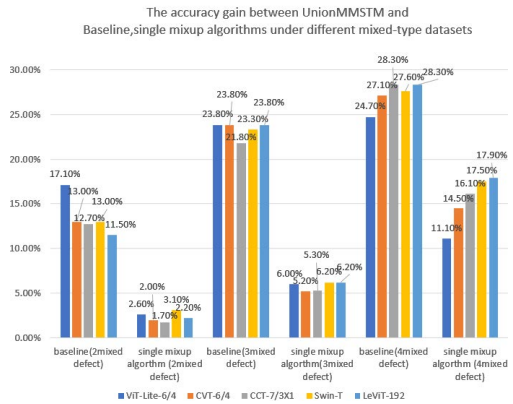

Fig. 6. Compared with Baseline and single mixup algorithm, the accuracy gain of UnionMMSTM under different mixed-type datasets.(e.g. Compared with baseline(no mixup), the accuracy gain of UnionMMSTM based on ViTLite-6/4 under dataset containing 2-mixed type defect patterns is 17.1%)

of mixup including: baseline network (none mixup), input mixup [6], CutMix [14], Manifold mixup [32], PuzzleMix [19], TokenMix[33] , our Union-Mixup and MMSTM. Using top-1 accuracy(%) for judgement, we examine the effectiveness of Union-Mixup and MMSTM in WBMs defect classification. WBMs defect classification of test dataset including two-mixed type defect samples are illustrated in TableII, UnionMixup performs better than the inputmixup and CutMix. When only consider single mixup algorithm, MMSTM achieves the highest recognition accuracy. While using two mixup algorithms, Union-MMSTM get the best performance in the classification of WBMs. Based on ViT-Lite-6/4, Union-MMSTM performs better than baseline by 17.1%. However, Union-MMSTM performs better than the method of only using sing mixup at least 1.7%. In addition, as shown in TableIII, when considering three-mixed type samples, Compared with none mixup and single mxiup, Union-MMSTM gain at least 21.8% and 5.2% respectively. Furthermore, in the classification including four-mixed type patterns, compared with none mixup and single mxiup, Union-MMSTM gain at least 24.7% and 11.1% respectively. As illustrated in Fig.6, when the mixed pattern become more complicated, our method make greater increment of classification accuracy. It proves that Union-MMSTM works efficient in the defect classification of WBMs, especially when the defect pattern mixed.

5. Conclusion

In this paper, we have proposed two mixup methods for training transformers based on single-defect pattern data to classify mixed-type defect WBMs. In the early stage, Union-Mixup can be used to do the mixup of WBMs with low defect ratio. Later, MMSTM is applied during the training of Vision Transformer. Our approaches solve the difficulty of obtaining mixed-type defect training dataset of WBMs in the recognition of WBM defect patterns. In addition, we have proved that the combination of different mixup algorithms is

a very effective data augmentation method. Especially when the test data become more complicated with three or more mixed type patterns, the framework composed of Union-Mixup and MMSTM performs much more better than other approaches. Though the classification accuracy of WBMs has been improved by our methods, the time complexity of algorithm still need to be improved in the following research. So, the issue of how to guarantee the accuracy of classification of mixed-type WBMs without compromising speed and simplicity still need more efforts in the next step.

Acknowledgments

This work was supported by Fujian Province Program of Youth and Middle-aged Education under Grant(JAT210509).

References

[1] C.W. Liu and C.F. Chien, "An intelligent system for wafer bin map defect diagnosis: An empirical study for semiconductor manufacturing," Engineering Applications of Artificial Intelligence, vol.26, no.5-6, pp.1479–1486, 2013.

[2] C.F. Chien, C.Y. Hsu, and K.H. Chang, "Overall wafer effectiveness (owe): A novel industry standard for semiconductor ecosystem as a whole," Computers and Industrial Engineering, vol.65, no.1, pp.117–127, 2013.

[3] J.C.H. Pan and D.H. Tai, "A new strategy for defect inspection by the virtual inspection in semiconductor wafer fabrication," Computers and Industrial Engineering, vol.60, no.1, pp.16–24, 2011.

[4] I. Tirkel, "The efficiency of inspection based on out of control detection in wafer fabrication," Computers and Industrial Engineering, vol.99, pp.458–464, 2016.

[5] Z. Kang, C. Catal, and B. Tekinerdogan, "Machine learning applications in production lines: A systematic literature review," Computers and Industrial Engineering, vol.149, p.106773, 2020.

[6] H. Zhang, M. Cisse, Y.N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," Int. Conf. Learning Representations, 2018.

[7] K. Kyeong and H. Kim, "Classification of mixed-type defect patterns in wafer bin maps using convolutional neural networks," IEEE Trans. on Semiconductor Manufacturing, vol.31, no.3, pp.395–402, 2018.

[8] H. Lee and H. Kim, "Semi-supervised multi-label learning for classification of wafer bin maps with mixed-type defect patterns," IEEE Trans. on Semiconductor Manufacturing, vol.33, no.4, pp.653–662, 2020.

[9] Y. Hyun and H. Kim, "Memory-augmented convolutional neural networks with triplet loss for imbalanced wafer defect pattern classification," IEEE Trans. on Semiconductor Manufacturing, vol.33, no.4, pp.622–634, 2020.

[10] D. Zou, Y. Cao, Y. Li, and Q. Gu, "The benefits of mixup for feature learning," 2023.

[11] H. Inoue, "Data augmentation by pairing samples for images classification," arXiv preprint arXiv:1801.02929, 2018.

[12] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," vol.abs/1711.10282, 2017.

[13] T. Devries and G.W. Taylor, "Improved regularization of convolutional neural networks with cutout," vol.abs/1708.04552, 2017.

[14] S. Yun, D. Han, S. Chun, S.J. Oh, Y. Yoo, and J. Choe, "Cutmix: Regularization strategy to train strong classifiers with localizable features," Int. Conf. Computer Vision), pp.6022–6031, 2019.

[15] E. Harris, A. Marcu, M. Painter, M. Niranjan, A. Prügel-Bennett, and J.S. Hare, "Understanding and enhancing mixed sample data augmentation," vol.abs/2002.12047, 2020.

[16] C. Summers and M.J. Dinneen, "Improved mixed-example data augmentation," vol.abs/1805.11272, 2018.

[17] R. Takahashi, T. Matsubara, and K. Uehara, "Ricap: Random image cropping and patching data augmentation for deep cnns," Proc.10th Asian Conf. Machine Learning, ed. J. Zhu and I. Takeuchi, pp.786–798, PMLR, 14-16 Nov 2018.

[18] J. Kim, W. Choo, H. Jeong, and H.O. Song, "Co-mixup: Saliency guided joint mixup with supermodular diversity," vol.abs/2102.03065, 2021.

[19] J. Kim, W. Choo, and H.O. Song, "Puzzle mix: Exploiting saliency and local statistics for optimal mixup," Proc. 37th Int. Conf. Machine Learning, 2020.

[20] J. Qin, J. Fang, Q. Zhang, W. Liu, X. Wang, and X. Wang, "Resizemix: Mixing data with preserved object information and true labels," vol.abs/2012.11101, 2020.

[21] A.F.M.S. Uddin, M.S. Monira, W. Shin, T. Chung, and S. Bae, "Saliencymix: A saliency guided data augmentation strategy for better regularization," vol.abs/2006.01791, 2020.

[22] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S.J. Oh, "Rethinking spatial dimensions of vision transformers," Int. Conf. Computer Vision, pp.11916–11925, 2021.

[23] T. Wang, L. Yuan, Y. Chen, J. Feng, and S. Yan, "Pnp-detr: Towards efficient visual analysis with transformers," Proc. Int. Conf. Computer Vision, 2021.

[24] B. Roh, J. Shin, W. Shin, and S. Kim, "Sparse DETR: efficient end-to-end object detection with learnable sparsity," vol.abs/2111.14330, 2021.

[25] Y. Rao, W. Zhao, B. Liu, J. Lu, and C.J. Hsieh, "Dynamicvit: Efficient vision transformers with dynamic token sparsification," Proc. 35th Int. Conf. Neural Information Processing Systems.

[26] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," vol.abs/2104.05704, 2021.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," Int. Conf. Computer Vision, pp.9992–10002, 2021.

[28] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze, "Levit: a vision transformer in convnet's clothing for faster inference," Int. Conf. Computer Vision, pp.12239–12249, 2021.

[29] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," Proc. 15th Int. Conf. Learning Representations.

[30] L. Beyer, O.J. Hénaff, A. Kolesnikov, X. Zhai, and A. van den Oord, "Are we done with imagenet?," vol.abs/2006.07159, 2020.

[31] R. Wightman, H. Touvron, and H. Jégou, "Resnet strikes back: An improved training procedure in timm," vol.abs/2110.00476, 2021.

[32] V. Verma, A. Lamb, C. Beckham, A. Najafi, A. Courville, I. Mitliagkas, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," 2019.

[33] J.C. Hyeong Kyu Choi and H.J. Kim, "Tokenmixup: Efficient attention-guided token-level data augmentation for transformers," Proc. 36th Int. Conf. Neural Information Processing Systems, 2022.