Recognizing Wafer Map Patterns using Semi-Supervised Contrastive Learning with Optimized Latent Representation Learning and Data Augmentation

Zihu Wang
Electrical and Computer Engineering
University of California
Santa Barbara, CA 93106
zihu_wang@ucsb.edu

Hanbin Hu*
Electrical and Computer Engineering
University of California
Santa Barbara, CA 93106
hanbinhu@ucsb.edu

Chen He
Automotive Processing
NXP Semiconductors
Austin, TX 78735
chen.he@nxp.com

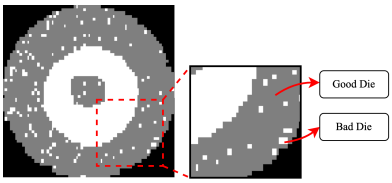
Peng Li
Electrical and Computer Engineering
University of California
Santa Barbara, CA 93106
lip@ucsb.edu

2023 IEEE International Test Conference (ITC) | 979-8-3503-4325-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ITC51656.2023.00030

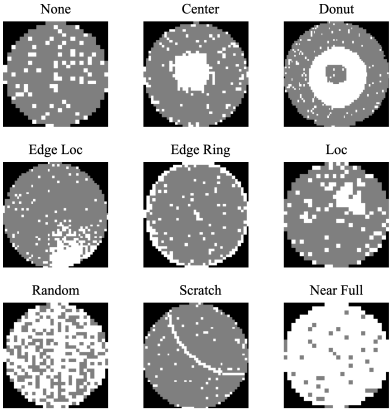
***Abstract*—Wafer map analysis is essential for process issue detection and yield improvement in semiconductor manufacturing. Accurate wafer map pattern recognition facilitates root-causing of abnormal chip fabrication conditions. However, manually annotating wafer map data is expensive and time-consuming, which drives up the demand for exploring label-efficient methods for wafer analysis. This paper proposes a novel contrastive learning framework for wafer map pattern feature extraction and classification. Under the semi-supervised learning setting, the proposed approach aims at learning from a large amount of unlabeled data while efficiently exploiting a small amount of expensive labeled data. Our method utilizes supervised contrastive learning on a small amount of labeled data to learn a better latent space representation with well-separated wafer pattern classes. Furthermore, a dual-encoder latent-space model is incorporated to best optimize the simultaneous use of labeled, unlabeled data, and varying types of data augmentations for representation learning. Finally, we enrich the semantics of the learned latent representation space by introducing a novel inter-wafer data augmentation to synthesize data which are not present in the given dataset. Experiments show that our method leads existing wafer pattern recognition techniques including recent contrastive learning based approaches by a large performance gain, and suggest that superior accuracy may be achieved simply by semi-supervised learning without resorting to labeling-intensive supervised learning.**

I. Introduction

In semiconductor manufacturing, wafer map analysis is a key to detect process issues and improve yield of integrated circuits (ICs). As a visualization of chip test failures on wafers, wafer maps help engineers identify the root cause of systematic chip fabrication issues when they show a certain pattern. For example, clustered failing dies on the wafer may indicate equipment and chemical stains, while a ring on the edge of the wafer may signify problems in etching [16]. Fig. 1 shows examples of wafer maps and some types of failure patterns. Thus, wafer map pattern recognition is one of the most essential tasks in semiconductor manufacturing.



(a) Passed and failed dies on a wafer



(b) Common wafer map patterns

Fig. 1: Chips on a wafer and different wafer patterns.

Traditionally, wafer map pattern recognition relies on expe-

*Hu is currently with Google. The work was performed while being at University of California, Santa Barbara

2378-2250/23/$31.00 ©2023 IEEE
DOI 10.1109/ITC51656.2023.00030

Regular Paper

rienced engineers for visual inspection and empirical judgment [6], which are subjective and time-consuming. Hence, the exploration of automatic and efficient approaches of wafer pattern identification is highly desired in semiconductor manufacturing.

Wafer map pattern recognition may benefit from the latest computer-aided methods for improving the efficiency and accuracy of wafer analysis. Various modern machine learning techniques have been adapted for wafer map pattern recognition. [23] and [4] propose to extract geometric and Randon-based features from wafer maps and utilize support vector machine (SVM) to classify wafer patterns. A decision tree ensemble-based method is introduced to recognize wafer map patterns [16]. Recently, deep convolutional neural networks (CNNs) have been used as the feature extractor for wafer pattern recognition. [14] is the first to propose a systematic approach to leverage CNNs for feature extraction by combining data mining and machine learning. Following CNN-based works [9], [18] improves supervised learning accuracy in wafer map pattern classification.

A major drawback of these previous methods is that all these supervised data-driven approaches require a significant amount of accurately labeled wafer map data, which is a time-consuming and expensive process. To respond to the above issue, unsupervised and semi-supervised methods of wafer map recognition have emerged. [8] proposes to train a variational autoencoder (VAE) on unlabeled data for wafer map pattern clustering. [12] and [17] also incorporate VAE in their framework to learn wafer pattern latent representations and classify them in a semi-supervised manner. However, these methods require a large proportion of the overall training data to be labeled in order to achieve satisfying performance in pattern classification. As a state-of-the-art deep learning framework, generative adversarial network (GAN) is adopted to give rise to unsupervised learning in wafer pattern recognition [15], but the proposed training process heavily relies on manual tuning and human intervention. [20] also incorporates GANs for pattern recognition; however, this approach also requires large amounts of labeled data to achieve high accuracy in wafer pattern classification.

In recent years, contrastive learning has emerged as a promising semi-supervised approach to learn powerful representations of visual and language data. Among contrastive learning approaches, SimCLR [2] is one of the most renowned frameworks for learning visual representations. At the core of SimCLR are augmentations that are applied to unlabeled images to get different views of them. By maximizing the agreement between features extracted from the same data's different views, SimCLR learns clustered data representations from unlabeled data. A classifier is then trained on labeled data upon the pre-trained feature extractor to learn the classification. [7] adapts SimCLR to wafer map pattern recognition by proposing domain-specific augmentations. Although this framework gives rise to label-efficient training where models trained on small amount of labels achieve comparable performance with fully-supervised training, it is unable to recognize all patterns accurately. In other words, the approach exhibits relatively low accuracy in identifying some specific wafer map patterns. Furthermore, we believe that more effective utilization of the small amount of labeled data can further improve the performance of contrastive learning in wafer pattern recognition.

To this end, we propose a novel contrastive learning framework of wafer map pattern recognition. Fig. 2 illustrates the architecture of our proposed framework. First, we propose to train the encoder by **S**upervised **C**ontrastive learning for **W**afer **M**aps (SCWM) on the labeled portion of the data. Leveraging labels, SCWM contributes to learning better clustered representations of wafer patterns. Alternating between supervised and conventional unsupervised contrastive training can effectively take advantage of both the unlabeled data and the small amount of expensive labeled data. Second, we propose a dual-encoder model consisting of a 'key' and 'query' encoder for learning wafer map representations. The moving-averaged 'key' encoder, which is updated faster in SCWM than in unsupervised contrastive learning, can help the 'query' encoder learn better-clustered representations by placing 'anchors' in the feature space. Finally, we introduce *Inter-Wafer Data Augmentation (IW-DA)* as a novel data augmentation method of synthesizing new wafer pattern classes that are not present in the given training dataset. Unlike conventional data augmentation techniques that create different views of a given input instance, IW-DA superimposes two existing instances to create realistic wafer which contains wafer patterns from both of its components.

Experiments show that SCWM greatly improves classification accuracy compared to existing supervised learning approaches. Our overall framework integrating supervised contrastive learning (SCWM), dual-encoder model architecture, and Inter-Wafer Data Augmentation (IW-DA) outperforms existing supervised and semi-supervised approaches for wafer pattern recognition.

II. Background

A. Problem Formulation

As shown in Fig. 1a, each wafer map can be represented by a matrix $\mathbf{x} \in \mathcal{X} = \{-1, 0, 1\}^{W \times H}$, where W and H are the wafer width and height, respectively, and each chip (die) is denoted by a pixel in the matrix. 0 is used to denote good dies, 1 is for bad dies, and -1 means that there is no dies at the location. For a labeled wafer map, there is a corresponding label $y \in \mathcal{Y}$ which specifies the pattern type of the wafer map, while no label is attached to an unlabeled wafer map.

In deep learning, to classify wafer maps, a deep neural network $f_\theta(\cdot)$ parameterized by θ is trained to extract key features from wafer maps and to predict their pattern type, such that $\hat{y} = f_\theta(\mathbf{x})$, where $\hat{y} \in \mathcal{Y}$ is the prediction.

B. Existing Contrastive Learning Methods in Wafer Pattern Recognition

Following [2], [7] proposes a semi-supervised contrastive learning framework for wafer pattern recognition. The frame-

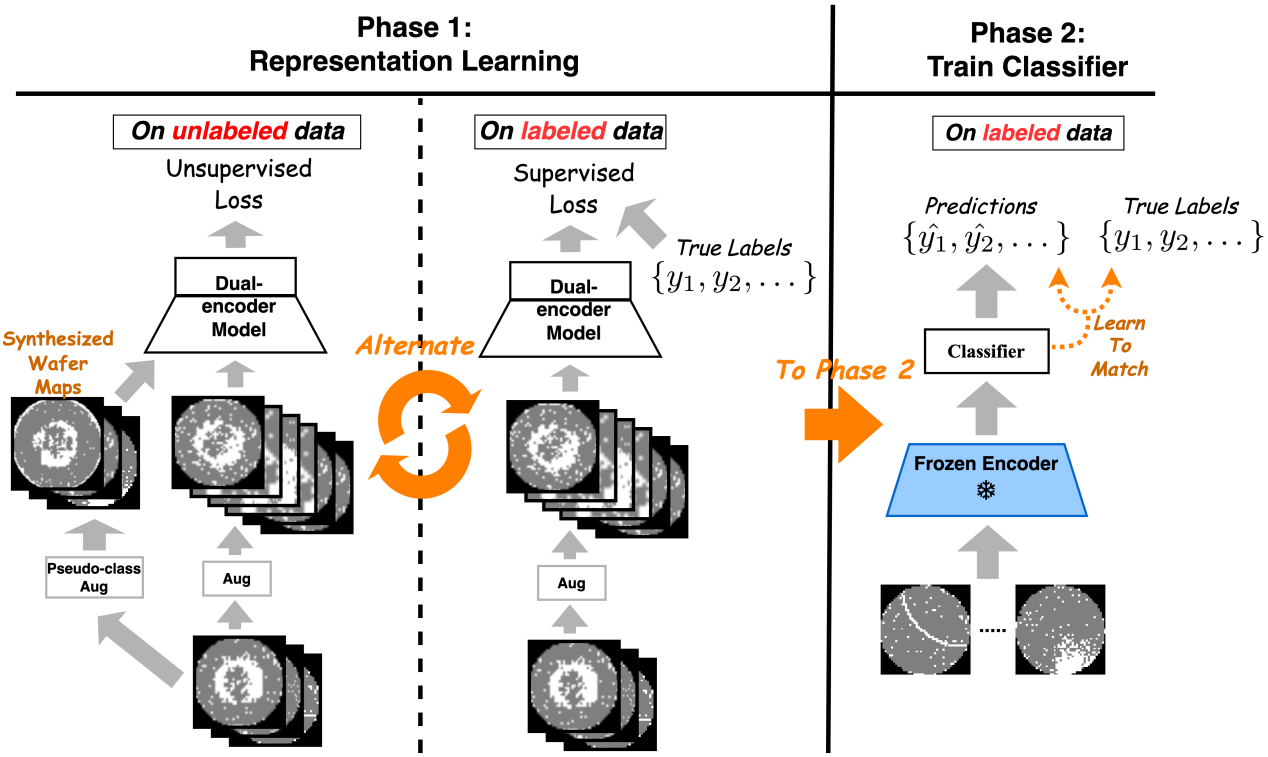


Fig. 2: An overview of the proposed framework.

work contains two steps, *unsupervised representations learning* and *classifier training*.

1) Unsupervised Representations Learning: The major goal of the first step is to learn an encoder $g_{\theta_g}(\cdot)$ that can extract the most distinctive features $\mathbf{v}$ for further classification, i.e., $\mathbf{v} = g_{\theta_g}(\mathbf{x})$, $\mathbf{v} \in \mathbb{R}^M$.

To achieve this goal on unlabeled data, two random data augmentations $Aug_1, Aug_2 : \mathcal{X} \to \mathcal{X}$ are sampled from an augmentation set $\mathcal{A}$ to provide different views of each wafer map, as shown at the top of Fig. 3:

$$\mathbf{x}_1 = Aug_1(\mathbf{x}), \quad \mathbf{x}_2 = Aug_2(\mathbf{x}) \tag{1}$$

The augmentations are designed to ensure the identity of the original wafer pattern is maintained in the augmented views. An encoder (CNN) $g_{\theta_g}(\cdot)$ is then used to encode the views. A projection head (MLP) $h_{\theta_h}(\cdot)$ projects the encoded views to a lower dimensional *feature space* $\mathcal{Z} = \mathbb{R}^N$, $N < M$.

$$\mathbf{z_1} = h_{\theta_h}(g_{\theta_g}(\mathbf{x}_1)), \quad \mathbf{z_2} = h_{\theta_h}(g_{\theta_g}(\mathbf{x}_2)) \tag{2}$$

In each training iteration, a batch of B wafer map data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_B\}$ is sampled and then go through the above process to form a batch of feature vectors $\{\mathbf{z}_{11}, \mathbf{z}_{12}, \ldots, \mathbf{z}_{B1}, \mathbf{z}_{B2}\}$ in the $\mathcal{Z}$ space where a contrastive loss ℓ_{CL} [2], [7] is applied:

$$\ell_{CL}(\mathbf{z}_{i1}, \mathbf{z}_{i2}) = -\log \frac{\exp(\text{sim}(\frac{\mathbf{z}_{i1}}{\|\mathbf{z}_{i1}\|}, \frac{\mathbf{z}_{i2}}{\|\mathbf{z}_{i2}\|})/\tau)}{\sum_{\mathbf{z}_j \in B'_{z_i}} \exp(\text{sim}(\frac{\mathbf{z}_{i1}}{\|\mathbf{z}_{i1}\|}, \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|})/\tau)} \tag{3}$$

where $B'_{z_i} = \{\mathbf{z}_{11}, \mathbf{z}_{12}, \ldots, \mathbf{z}_{B1}, \mathbf{z}_{B2}\} \backslash \{\mathbf{z}_{i1}, \mathbf{z}_{i2}\}$ is the set of feature vectors of the batch excluding the $\mathbf{z}_{i1}$ and $\mathbf{z}_{i2}$, $\text{sim}(\cdot)$ calculates the cosine similarity of two normalized vectors, and τ denotes a temperature parameter.

Note that in (3), the numerator is formed by a pair of views from the same raw instance, while the denominator is a summation over pairs formed by views from different instances. The pairs in the numerator and denominator are referred to as positive pairs and negative pairs, respectively. By minimizing $\ell_{CL}(\mathbf{z}_{i1}, \mathbf{z}_{i2})$, the similarity between positive pair instances is maximized, while the similarity between negative pair instances is minimized. The overall loss function of this step can be written as:

$$\mathcal{L}_{CL} = \sum_{i=1}^{B} (\ell_{CL}(\mathbf{z}_{i1}, \mathbf{z}_{i2}) + \ell_{CL}(\mathbf{z}_{i2}, \mathbf{z}_{i1})) \tag{4}$$

2) Classifier Training: After the unsupervised training of the encoder, the projection head is discarded. A linear classifier $s_{\theta_s}(\cdot)$ is then connected to the parameter-frozen encoder to learn the classification of representations on labeled data.

$$\hat{y} = s_{\theta_s}(g_{\theta_g}(\mathbf{x})) \tag{5}$$

The linear classifier is trained by minimizing a cross entropy loss to match the model's predictions $\hat{y}$ to true labels y. Fig. 3 illustrates the overall pipeline of this method.

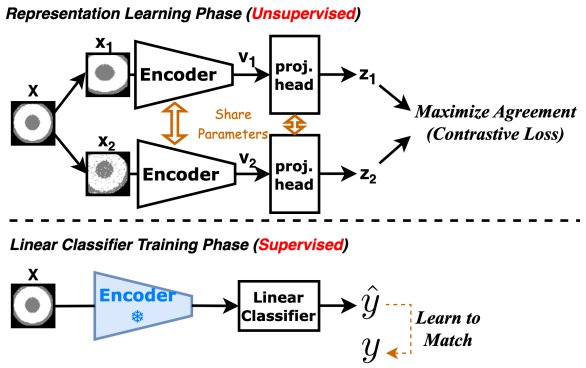


Fig. 3: Pipeline of the existing contrastive learning method [7] for wafer recognition.

C. Drawbacks of Existing Contrastive Learning Work [7]

Despite the highly label-efficient manner displayed by this semi-supervised method, the classification accuracy of the trained model is imbalanced among classes, i.e., this method has difficulties in identifying some specific wafer patterns. This model's failure in recognizing these wafer patterns may be attributed to the fact that plenty of instances from some classes exhibit similar geometrical features. For example, as it is shown in Fig. 4, the three wafer maps from three different classes share a similar appearance. Without human supervision, it is difficult for a black-box machine learning model to categorize them correctly.

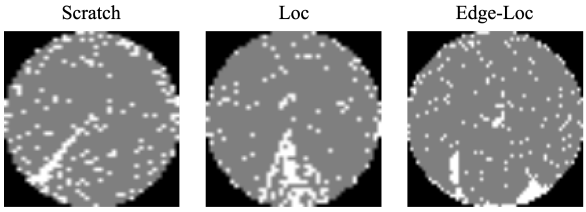


Fig. 4: Examples of wafer patterns from different classes that share similar appearance.

III. Methodology

To this end, we propose a novel contrastive learning framework for wafer recognition. We believe that, by more effectively leveraging the information provided by the small amount of labels in the dataset, contrastive learning has the potential to further enhance the performance of semi-supervised learning methods in wafer pattern recognition.

A. Supervised Contrastive Learning for Wafer Recognition

As it has been analyzed in Section II-C, contrastive learning may mis-classify instances from those classes which share similar geometric features. Addressing this issue on unlabeled data is challenging, which highlights the importance of human supervision.

To address the above challenge, here our key idea is to utilize labeled data for training the encoder in addition to the use of unlabeled data. This would allow for more reliable learning of key features that can distinguish similar but differently annotated wafer maps. Inspired by self-supervised contrastive learning [2], supervised contrastive learning (SupCon) [10] proposes to learn good visual representations of image data by forming positive pairs between all instances with the same label in computer vision.

In a similar spirit, we propose **S**upervised **C**ontrastive learning for **W**afer **M**aps (SCWM) to help the encoder better separate those classes with similar geometric features in the feature space. We form positive pairs between all wafer maps that share the same label. It is noteworthy that the denominator of loss (3) used in the more conventional unsupervised contrastive learning contains 'fake' negative pairs in which the two views have the same label. However, these views are not supposed to be pushed apart in the feature space because instances with the same label have similar semantics. Therefore, we also filter out those fake negative pairs in the denominator of (3). We define $A(i) = \{1, 2, \ldots, 2B\} \backslash \{i\}$ as the set of all but the i_{th} indices of feature vectors in a training data batch. The loss function $\mathcal{L}_{SCWM}$ to achieve these goals in SCWM is:

$$\mathcal{L}_{SCWM} = \sum_{i=1}^{2B} \frac{-1}{\|P(i)\|} \sum_{p \in P(i)} \log \frac{\exp(\mathrm{sim}(\frac{\mathbf{z}_i}{\|\mathbf{z}_i\|}, \frac{\mathbf{z}_p}{\|\mathbf{z}_p\|})/\tau)}{\sum_{a \in Q(i)} \exp(\mathrm{sim}(\frac{\mathbf{z}_i}{\|\mathbf{z}_i\|}, \frac{\mathbf{z}_a}{\|\mathbf{z}_a\|})/\tau)}, \quad (6)$$

where $P(i) = \{p \in A(i) : y_i = y_p\}$ is the set of feature vector indices that have the same label of $\mathbf{z}_i$, $Q(i) = A(i) \backslash P(i)$ is the set of indices whose labels are different from that of $\mathbf{z}_i$, and only the instances from different classes are used to form negative pairs.

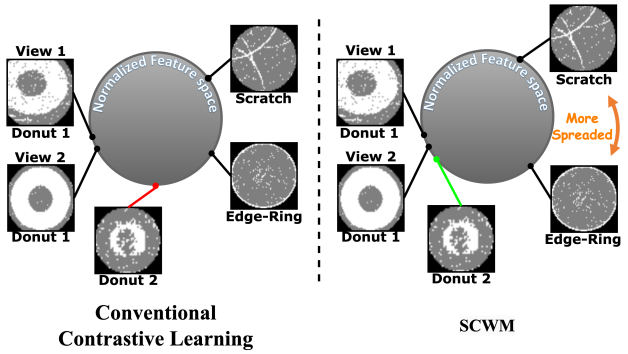


Fig. 5: The conventional contrastive learning is instance-based which maximizes the agreement between feature vectors of the same instance's views on the unit hypersphere in the feature space but misplace views without considering their class labels. By leveraging labeled data, SCWM maximizes the agreement among all instances of the same class.

As shown in Fig. 5, benefiting from the reformulated loss function, SCWM can help better cluster together feature vectors from a same class. Meanwhile, clusters of representations of distinct classes are more separated, which makes the downstream classification of them easier.

In the proposed overall semi-supervised learning framework, we alternate between the conventional unsupervised contrastive learning on the unlabeled portion of the dataset and SCWM on the labeled portion during the representation learning step. This allows us to best exploit the whole semi-supervised dataset.

B. Dual-encoder Model

It is important to note that the representation learning processes of conventional contrastive learning and SCWM are not in full concordance of each other. As discussed in Section III-A, SCWM tends to learn representations clustered according to labels. However, conventional unsupervised contrastive learning may blur the boundary between those clusters which have similar appearance but different labels. Therefore, when alternating between SCWM and conventional contrastive learning, these two processes may not be able to complement each other in learning a better representation distribution. To address this issue, we propose a dual-encoder model that optimizes the interactions between unsupervised and supervised contrastive learning (SCWM) during the representation learning phase as shown in Phase 1 of Fig. 2.

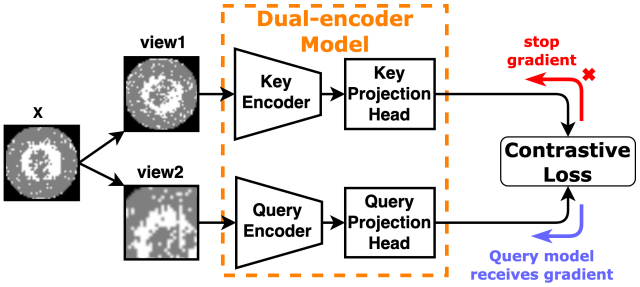


Fig. 6: The architecture of dual-encoder model.

While bearing some resemblance to Momentum Contrast (MoCo) [5], which enhances the vanilla contrastive learning for computer vision applications by introducing a momentum-updated encoder, the proposed dual-encoder model operates under a rather different training mechanism and for a very different goal.

In our proposed framework, we use the dual-encoder model as the feature extractor of wafer maps to best orchestrate SCWM and unsupervised contrastive learning.

The dual-encoder model comprises two encoders and projection heads with identical architecture as shown in Fig. 6. For each positive pair, the two views are fed to the two encoders respectively. The 'query' model is updated by the gradient it receives, while the 'key' model, which receives no gradient, evolves by using a weighted sum of its own parameters and the query model's parameters:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q, \tag{7}$$

where θ_k and θ_q denote the parameters of the key and query model, respectively, $m \in [0, 1)$ controls the key model's evolving rate. A larger m slows down the update of the key model.

The supervised (SCWM) and unsupervised contrastive learning phases are orchestrated by setting different values of m in the dual-encoder model. We use a relatively small m for the key model's update in SCWM and a large m in unsupervised contrastive learning. During SCWM, a small m enables the key model to quickly learn well-separated representations for different classes leveraging the available class labels. While processing unlabeled data in the unsupervised contrastive learning phase, a larger m slows down the update of the key model and preserves the representation distribution learned from SCWM. Since the key model receives no gradient from the contrastive loss, the query model always pulls query representations closer to key representations. Specifically, for the two views generated for each unlabeled input, the view generated by the key model places an 'anchor' in the feature space to which the view of the query encoder gets pulled. By pulling its view closer to the anchor provided by the key model, the query encoder learns the rich semantics presented in the unlabeled data while maintaining the separations between different classes. After the overall representation learning phase, the query encoder is used as the encoder for training the classifier.

C. Exploring Strong Data Augmentations for Wafer Maps

Data augmentations play a crucial role in contrastive learning as they guide the training process of the encoder by creating additional views of the input data while allowing to learn more robust and discriminative representations without requiring class labels. Therefore, carefully selecting and applying appropriate augmentations can greatly improve the utilization of large amounts of unlabeled data and hence the quality of the learned representations.

In [7], 5 random data augmentations are used to form the random augmentation strategy which includes visual augmentations and augmentations designed specifically for wafer maps. Empirical studies on wafer maps show that *random resized-crop* provides the most significant performance boost for wafer map representation learning. Similarly, in visual contrastive learning, empirical experimental results suggest that *random resized-crop* is the most crucial data augmentation [2], [5], [19].

For wafer pattern recognition, we believe that it is instrumental to explore *random resized-crop* as a strong data augmentation by adapting it to wafer data because of intrinsic geometrical features of wafer patterns. Unlike natural images, the location of bad die clusters on a wafer may determine the category to which the wafer belongs. For example, the major difference between 'Loc' and 'Edge-Loc' is where a cluster of

bad dies is located. A wafer map is classified as 'Edge-Loc' if the cluster is on the edge, otherwise it is 'Loc'.

Therefore, we propose to form positive pairs between a cropped view and an uncropped view to learn representations. Specifically, when applying the augmentations in (1), Aug_1 consists of a sequence of basic augmentations including *random resized-crop* while Aug_2 contains all basic augmentations of Aug_1 except *random resized-crop*.

Furthermore, *random resized-crop* is properly exploited under our dual-encoder architecture, where the key encoder is used to encode uncropped views. Since the key encoder is responsible for mapping anchor views, distinguishable wafer pattern features must be captured by it. Hence, the uncropped views of the essential bad die locations are fed to the key encoder during training.

D. Inter-wafer Data Augmentation

Recent works in visual representation learning demonstrate that including more views in the feature space can enhance the performance of the encoder in classification tasks [1], [21], [22]. Increasing the number of diverse views of the data presented to the encoder can enhance the semantics of the feature space, resulting in improved representational generalization capabilities.

Mixco [11] considers mix-up as a data augmentation for contrastive learning for computer vision applications, where a convex combination of two images is encoded and projected to the feature space. The similarity between the mixed image and the original images in the feature space is then constrained by a loss function.

Motivated by the success of mix-up for visual data, we propose *Inter-wafer Data Augmentation (IW-DA)* as a method to synthesize wafer map data. However, naive convex combination of wafer maps will result in unrealistic wafer maps since the values used to represent wafer dies are discrete. Thus, IW-DA is formulated as follows:

$$s_{ij}^{IW} = \begin{cases} w_{ij}^1, & w_{ij}^1 \geq w_{ij}^2 \\ w_{ij}^2, & w_{ij}^1 < w_{ij}^2 \end{cases} \tag{8}$$

where s_{ij}^{IW}, w_{ij}^1, and w_{ij}^2 are the die at the i_{th} row and j_{th} column of the synthesized new wafer map and those of the two original wafer maps used in synthesis, respectively. Since a value of 1 indicates a bad die and 0 represents a good die, IW-DA is equivalent to a 'logic or' operation that preserves the bad die patterns from both wafer maps. An example of the IW-DA operation is illustrated in Fig. 7.

Different from [13], which uses superimposed wafer maps to learn mutli-label classification under supervised learning setups, we propose to utilize synthesized wafer maps to enhance the quality of learned representations. To achieve this, we propose a loss function that constrains the mapping of the synthesized data. We synthesize a set of new wafer maps $\mathcal{X}_{IW} = \{\mathbf{x}_{IW_1}, \mathbf{x}_{IW_2}, \dots\}$ by randomly combining data from each training batch. Let the new map $\mathbf{x}_{IW_p}$ be generated from

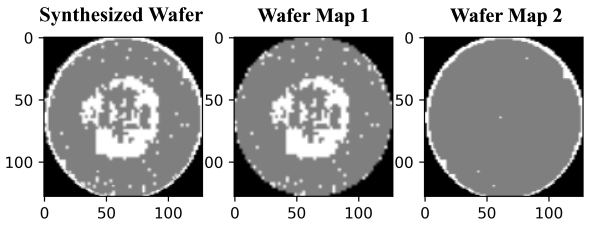


Fig. 7: An example of IW-DA operation on two wafer maps.

$\mathbf{x_i}$ and $\mathbf{x_j}$, the loss function to constrain the encoding of $\mathbf{x_{IW_p}}$ is:

$$\begin{aligned} \ell_{IW_p} = &|\text{sim}(\mathbf{z}_{IW_p}, \mathbf{z}_i) - \text{sim}(\mathbf{z}_{IW_p}, \mathbf{z}_j)| \\ &+ \sum_{k \neq i,j} \log(\exp(\text{sim}(\mathbf{z}_{IW_p}, \mathbf{z}_k)/\tau)) \end{aligned} \tag{9}$$

where the first term requires $\mathbf{z}_{IW_p}$ to be located at the middle point between $\mathbf{z_i}$ and $\mathbf{z_j}$ in the feature space, while the second term forms negative pairs between $\mathbf{z}_{IW_p}$ and all other instances in the batch. IW-DA introduces realistic wafer patterns, which are not present in the training dataset, to the feature space. The IW-DA loss ℓ_{IW} (9) constrains the representation of the synthesized wafer pattern to the middle point between the two original maps in the feature space as shown in Fig. 8.

The total loss on synthesized wafer map data in a batch is:

$$\mathcal{L}_{IW} = \sum_{p:\mathbf{x}_{IW_p} \in \mathcal{X}_{IW}} \ell_{IW_p}. \tag{10}$$

The overall loss to be minimized on unlabeled data is the sum of (4) and (10):

$$\mathcal{L}_u = \mathcal{L}_{CL} + \mathcal{L}_{IW}. \tag{11}$$

The overall representation learning phase proceeds by alternating between the minimization of the unsupervised contrastive loss (11) and SCWM loss (6).

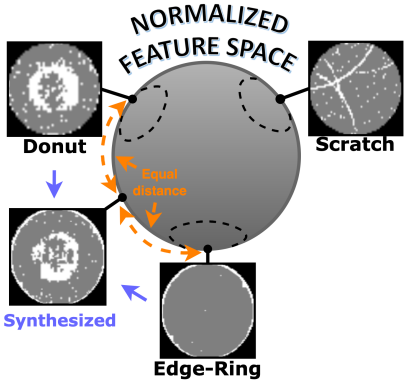


Fig. 8: The IW-DA loss ℓ_{IW} (9) constrains the representation of the synthesized wafer pattern to the middle point between the two original maps in the feature space.

TABLE I: Comparison of balanced accuracy on WM-811K between various methods.

		Labeled Data Percentage				
		5%	10%	20%	50%	100%
Type	Method	Balanced Accuracy				
Supervised	SVM	41.57%	45.88%	48.66%	46.99%	-
	Weighted SVM	46.99%	54.72%	55.40%	48.76%	-
	CNN(cross-entropy)	66.10%	66.74%	77.90%	74.20%	90.59%
Semi-supervised	SimCLR [2]	80.22%	80.19%	82.42%	-	
	5TCLWMPR [7]	77.41%	83.44%	82.35%	84.57%	-
	CLLD (ours)	**81.35%**	**87.51%**	**86.34%**	**90.56%**	-

TABLE II: Comparison of accuracies of CLLD and 5TCLWMPR [7] in recognizing each category of wafer pattern. 10% of labeled data is used for training.

	None	Edge-Ring	Center	Edge-Loc	Loc	Random	Scratch	Donut	Near-Full
5TCLWMPR [7]	92.94%	95.86%	94.44%	50.64%	92.02%	30.72%	82.33%	32.19%	65.66%
CLLD (ours)	80.97%	93.44%	95.32%	74.31%	95.88%	57.72%	88.87%	60.21%	63.33%

TABLE III: Dataset details.

Wafer patterns	Training	testing
None	33051	3679
Edge-Ring	7735	819
Center	3113	349
Edge-Loc	2150	267
Loc	1458	162
Random	546	63
Scratch	446	54
Donut	372	37
Near-Full	49	5
Total	48920	5435

IV. Experiments

A. Implementation Details

a) Details of the dataset: We adopt the most widely used public wafer map dataset, WM-811K [23], for training and testing our models. The dataset contains 811,457 wafer maps collected from real IC manufacturing processes. Following [7], we use 54,355 labeled wafer maps from the dataset to build our training and testing datasets. The numbers of wafer maps contained in the training and testing datasets are shown in Table III. To conduct experiments under semi-supervised setups, we use all 48,920 wafer maps to build the unlabeled training dataset, and take out a small portion $p_l\%$ of them and their corresponding labels to construct a labeled training dataset. We experiment with a wide range of $p_l\%$ for a comparison between our proposed method and other existing supervised methods including support vector machine (SVM), weighted SVM, and convolutional neural network (CNN), and semi-supervised contrastive learning approaches including SimCLR [2] and [7].

TABLE IV: Balanced accuracy comparison between cross-entropy loss based supervised learning and SCWM.

	Supervised (cross-entropy)	SCWM
Acc (Balanced)	90.59%	**92.89%**

It can be seen from Table III that the original WM-811K dataset [23] is highly imbalanced, e.g., the most dominant class 'None' has approximately 674 times more instances than the least dominant class 'Near-Full'. Thus, it is not fair to use the conventional accuracy metric to evaluate model performance, as a model's ability to recognize different classes has a highly imbalanced contribution to the final result. For example, a model that predicts all samples as 'None' can achieve approximately 67.6% accuracy. To deal with this issue, we use balanced accuracy (BAC) [7] to evaluate model performance:

$$BAC = \frac{\sum_{i=1}^{N} w_i \mathbb{1}[\hat{y}_i = y_i]}{\sum_{i=1}^{N} w_i}, \tag{12}$$

where y_i is the true label of a wafer map $\mathbf{x}_i$, and $\hat{y}_i$ is the prediction made by a model. $w_i = \frac{1}{N_{y_i}}$, N_{y_i} denotes the number of instances with the same label y_i, used to balance the accuracy. In (12), a method's capability to recognize each class in the testing dataset has equal contribution to BAC.

Since the wafer maps in WM-811K are of different sizes, which is intractable for CNNs, we resize all wafer maps to the size of $1 \times 128 \times 128$.

b) Details of model setups: For the SVM based models, we used Randon-based and geometric-based feature extractor [23] and a radial basis function kernel. Fig. 9 illustrates the neural network architectures adopted in the proposed method. The encoder contains three convolutional layers each of which is followed by a max pooling layer. A flatten layer and a fully connected layer project the output of the last convolutional

TABLE V: Ablation study results of the proposed IW-DA data augmentation.

	labeled Data Percentage			
	5%	10%	20%	50%
Configurations	Balanced Accuracy			
CLLD (IW-DA removed)	80.86%	86.58%	85.89%	90.48%
CLLD	**81.35%**	**87.51%**	**86.34%**	**90.56%**

layer to a 256 dimensional vector. The projection head is a two-layer MLP that projects the output of the encoder to a 64-dimensional $\mathcal{Z}$ space, and the linear classifier is constructed by a single fully-connected layer. The dual-encoder model consists of two encoders with identical architecture and the associated projection heads follow the architecture in Fig. 9. For the sake of a fair comparison with other CNN-based methods, we adopt the same CNN architecture for all other semi-supervised and supervised learning methods. Specifically, the supervised model trained with a cross-entropy loss consists of a CNN-based encoder followed by a linear classifier.

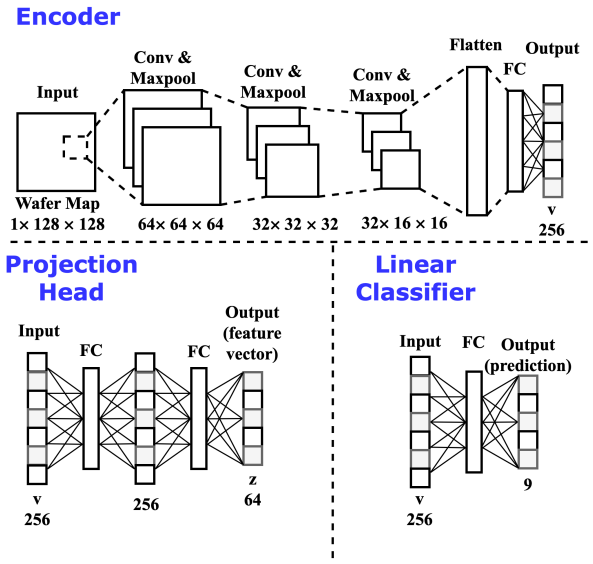


Fig. 9: Details of neural network architectures of our method.

c) Details of hyperparameters: As discussed in Section III-C, we use two differently composed data augmentations in our framework. One is formed by combining commonly used basic augmentations of *random resized-crop*, *random noise*, *random horizontal flip*, *random rotate*, and *random rotate-twist* [7]. On the other hand, the other one contains only *random noise*, *random horizontal flip*, *random rotate*, and *random rotate-twist*.

In the proposed framework, network parameters are updated by gradient descent [3], [24], [25]. An Adam optimizer with a cosine scheduler is used to update all neural networks. A starting learning rate of 10^{-3} is chosen for updating the encoder and the projection head, while the linear classifier's starting learning rate is 5×10^{-4}. The optimizer has a weight decay of 10^{-4}. The batch size of the unlabeled data is 256, and that of the labeled data is 64. The temperature hyperparameter is set to $\tau = 0.1$ for unsupervised training and $\tau = 0.7$ for supervised training. For the cross-entropy loss based supervised CNN training, similarly, an Adam optimizer with a learning rate of 10^{-3} and a weight decay of 10^{-4} is used to update the network. The batch size of the cross-entropy based supervised training is 64. All of the experiments are implemented on an NVIDIA GeForce RTX 3090 GPU.

For semi-supervised training methods, models are trained on unlabeled data for 100 epochs and are tuned on labeled data every 5 epochs. For our proposed method, where the dual-encoder model is used, $m = 0.999$ is used for updating the query model on unlabeled data and $m = 0$ is used for updating the query model on labeled data. It takes around 2.5 hours to complete the 100-epochs training of our proposed method.

B. Performance of the proposed supervised contrastive learning (SCWM)

First, we focus on evaluating only the performance of SCWM, the proposed supervised contrastive learning technique. For this, we make use of all labels in the WM-811K dataset for training a model using SCWM, and contrast it with a CNN-based model trained with the standard cross-entropy based supervised learning on the same dataset as a fair comparison. Note that the SCWM model is trained without any unsupervised contrastive learning phase and the proposed inter-wafer data augmentation (IW-DA).

Table IV presents a comparison of balanced accuracy between the two models which are all based on supervised learning. SCWM achieves a balanced accuracy of 92.89%, which is more than 2% higher than the performance of the cross-entropy based method. This demonstrates that by making use of the proposed SCWM loss in Equation 6, contrastive learning in the supervised manner can train the encoder to better capture the desired distinguishing features, enabling learning latent representations well clustered within the class boundaries.

C. Proposed Semi-supervised Learning Performance

The balanced classification accuracies of our method and other methods are shown in Table I. There are three pure supervised learning methods included in the comparison, which are SVM, weighted SVM, and a cross-entropy loss based CNN training method. It can be seen from the table that the CNN-based wafer map pattern recognition gains a large

improvement over the more conventional machine SVM based approach.

We integrate our proposed supervised contrastive learning (SCWM), dual-encoder model, and inter-wafer data augmentation (IW-DA) and refer to the resulting approach as Contrastive Learning with optimized Latent representation learning and Data augmentation (CLLD). CLLD can more effectively leverage a small amount of labeled data with the dual-encoder model and SCWM. Learning representations with the novel IW-DA based data augmentation can further improve classification accuracy by creating a more semantically meaningful feature space. As seen in Table I, CLLD significantly outperforms SimCLR and 5TCLWMPR in terms of accuracy by up to 8.14%, and 5.99%, respectively. Notably, when trained on a dataset containing only 50% labeled data, our approach achieves an accuracy comparable to that of conventional supervised CNN-based methods trained on the fully labeled dataset, demonstrating the great potential of semi-supervised learning.

Table II reports the models' performance of recognizing each pattern in the WM-811K dataset. Compared with [7], our method improves the accuracy in recognizing pattern classes such as 'Donut' and 'Edge-Loc', which makes the model's performance more balanced among classes.

D. Ablation Study on Inter-Wafer Data Augmentation (IW-DA)

Enriching semantics of the feature space is shown to be an effective way of enhancing representation learning performance. We run experiments on our proposed framework with and without the proposed inter-wafer data augmentation (IW-DA) to validate its effectiveness.

Table V shows results of the ablation study. The use of IW-DA leads to a performance boost of approximately 0.5% across datasets with varying percentages of labeled data. With different percentage of labeled data in the dataset, IW-DA provides a performance boost of up to 0.93%. The semi-supervised learning incorporating IW-DA outperforms the one without it on all experiments, showing the effectiveness of the proposed data augmentation in learning more powerful wafer patterns representations.

V. Conclusion

We have proposed a novel semi-supervised contrastive learning framework for wafer map pattern recognition. We introduce supervised contrastive learning (SCWM) to improve the quality of the learned latent representations by better exploiting a given small amount of labeled data. A dual-encoder model is employed to best manage the supervised and unsupervised contrastive learning phases and augmented data views used in our framework. By placing anchors in the latent feature space, the key encoder helps the query encoder learn clustered representations conforming to the boundaries between different wafer pattern classes. Finally, the inter-wafer data augmentation (IW-DA) is shown to be able to enrich semantics of the latent feature space during the learning process. Experiments show that our approach outperforms existing supervised and semi-supervised learning methods across a wide range of labeled data proportions. Ablation studies on SCWM and IW-DA indicate that these proposed techniques are effective and worth further exploration.

Acknowledgment

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 1956313. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[3] Haskell B Curry. The method of steepest descent for non-linear minimization problems. *Quarterly of Applied Mathematics*, 2(3):258–261, 1944.

[4] Mengying Fan, Qin Wang, and Ben van der Waal. Wafer defect patterns recognition based on optics and multi-label classification. In *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, pages 912–915. IEEE, 2016.

[5] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[6] Chia-Yu Hsu, Wei-Ju Chen, and Ju-Chien Chien. Similarity matching of wafer bin maps for manufacturing intelligence to empower industry 3.5 for semiconductor manufacturing. *Computers & Industrial Engineering*, 142:106358, 2020.

[7] Hanbin Hu, Chen He, and Peng Li. Semi-supervised wafer map pattern recognition using domain-specific data augmentation and contrastive learning. In *2021 IEEE International Test Conference (ITC)*, pages 113–122. IEEE, 2021.

[8] Jonghyun Hwang and Heeyoung Kim. Variational deep clustering of wafer map patterns. *IEEE Transactions on Semiconductor Manufacturing*, 33(3):466–475, 2020.

[9] Cheng Hao Jin, Hyun-Jin Kim, Yongjun Piao, Meijing Li, and Minghao Piao. Wafer map defect pattern classification based on convolutional neural network features and error-correcting output codes. *Journal of Intelligent Manufacturing*, 31(8):1861–1875, 2020.

[10] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[11] Sungnyun Kim, Gihun Lee, Sangmin Bae, and Se-Young Yun. Mixco: Mix-up contrastive learning for visual representation. *arXiv preprint arXiv:2010.06300*, 2020.

[12] Yuting Kong and Dong Ni. A semi-supervised and incremental modeling framework for wafer map classification. *IEEE Transactions on Semiconductor Manufacturing*, 33(1):62–71, 2020.

[13] Chenwei Liu and Qiaoyue Tang. Triplet convolutional networks for classifying mixed-type wbm patterns with noisy labels. In *2021 IEEE International Test Conference (ITC)*, pages 200–207. IEEE, 2021.

[14] Kouta Nakata, Ryohei Orihara, Yoshiaki Mizuoka, and Kentaro Takagi. A comprehensive big-data-based monitoring system for yield enhancement in semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 30(4):339–344, 2017.

[15] Matthew Nero, Chuanhe Shan, Li-C Wang, and Nik Sumikawa. Concept recognition in production yield data analytics. In *2018 IEEE International Test Conference (ITC)*, pages 1–10. IEEE, 2018.

[16] Minghao Piao, Cheng Hao Jin, Jong Yun Lee, and Jeong-Yong Byun. Decision tree ensemble-based wafer map failure pattern recognition based on radon transform-based features. *IEEE Transactions on Semiconductor Manufacturing*, 31(2):250–257, 2018.

[17] Ho Sun Shon, Erdenebileg Batbaatar, Wan-Sup Cho, and Seong Gon Choi. Unsupervised pre-training of imbalanced data for identification of wafer map defect patterns. *IEEE Access*, 9:52352–52363, 2021.

[18] Ghalia Tello, Omar Y Al-Jarrah, Paul D Yoo, Yousof Al-Hammadi, Sami Muhaidat, and Uihyoung Lee. Deep-structured machine learning model for the recognition of mixed-defect patterns in semiconductor fabrication processes. *IEEE Transactions on Semiconductor Manufacturing*, 31(2):315–322, 2018.

[19] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

[20] Junliang Wang, Zhengliang Yang, Jie Zhang, Qihua Zhang, and Wei-Ting Kary Chien. Adabalgan: An improved generative adversarial network with imbalanced learning for wafer defective pattern recognition. *IEEE Transactions on Semiconductor Manufacturing*, 32(3):310–319, 2019.

[21] Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[22] Zihu Wang, Yu Wang, Hanbin Hu, and Peng Li. Contrastive learning with consistent representations. *arXiv preprint arXiv:2302.01541*, 2023.

[23] Ming-Ju Wu, Jyh-Shing R Jang, and Jui-Long Chen. Wafer map failure pattern recognition and similarity ranking for large-scale data sets. *IEEE Transactions on Semiconductor Manufacturing*, 28(1):1–12, 2014.

[24] Yifan Yang, Alec Koppel, and Zheng Zhang. A gradient-based approach for online robust deep neural network training with noisy labels. *arXiv preprint arXiv:2306.05046*, 2023.

[25] Yequan Zhao, Xinling Yu, Zhixiong Chen, Ziyue Liu, Sijia Liu, and Zheng Zhang. Tensor-compressed back-propagation-free training for (physics-informed) neural networks. *arXiv preprint arXiv:2308.09858*, 2023.