Semiconductor Image Analysis using Data-Efficient Machine Learning

Songhyun Yu*, Sangho Yoon, Sungjin Lim, Hoonseok Park, Subong Shon, Suil Son, Chulmoo Kang, Yoonsung Bae, Hansaem Park†, Sungsik Jo, Seungwoo Oh, Dongryul Lee, Seonghoon Jeong, Sungho Lee, and Myungjun Lee
Semiconductor Research & Development, Samsung Electronics Co., Ltd., 1-1 Samsungjeonja-ro, Hwaseong-si, Gyeonggi-do, 18448, Republic of Korea

ABSTRACT

Recent advancements in AI have intensified interest in automating Transmission/Scanning Electron Microscopy (TEM/SEM) image analysis for semiconductor metrology and inspection (MI). However, the scarcity of annotated datasets in the industry poses a critical challenge, as model performance heavily relies on the quantity of precisely labeled training data. To address this limitation, we propose a data-efficient machine learning framework optimized for structural and categorical semiconductor image analysis. Our approach integrates three key innovations: (1) Transfer learning of a foundation model for semiconductor image segmentation, which outperforms generic counterparts by 0.1920 in mIoU; (2) Semi-supervised learning with TEM-specialized augmentation strategy with limited labeled data, achieving competitive mIoU of 0.877 even with a few labeled examples; and (3) Self-supervised contrastive pre-training for SEM defect image classification, leveraging magnification-invariant consistency to enhance feature robustness. Evaluated under limited-data regimes, our framework demonstrates superior performance over conventional methods while significantly reducing annotation dependency. This solution directly tackles the semiconductor industry's persistent challenge of high labeling costs and rapidly expanding unlabeled data pools, offering a practical pathway for AI adoption in semiconductor developments.

Keywords: Semiconductor image analysis, metrology and inspection (MI), transmission electron microscope (TEM), scanning electron microscope (SEM), data-efficient learning, image segmentation, defect classification.

1. INTRODUCTION

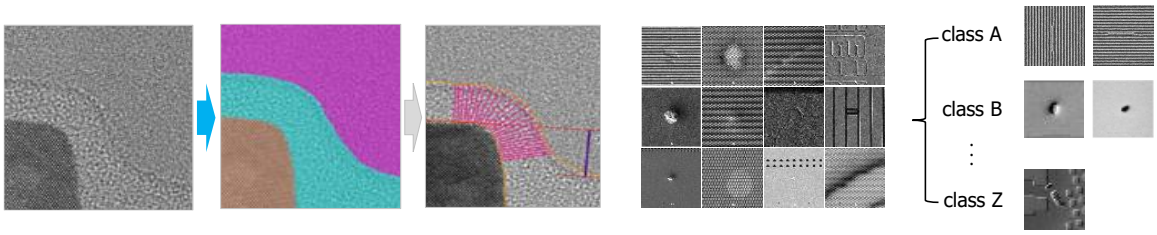
With the recent miniaturization of semiconductor devices, both product complexity and process difficulty have increased, raising the need for metrology and inspection (MI) technology to enable faster and more accurate process verification [1]. Among various MI methods, image-based analysis is crucial as it provides direct visualization and precise measurement of structures and defects.

Transmission electron microscopy (TEM) and scanning electron microscopy (SEM) are representative equipment for precise image analysis: TEM provides atomic-level structural visualization but requires destructive sample preparation, whereas SEM does not reach the atomic-level but enables nanometer-level structural measurements, without the need to destroy the sample. Although these imaging modalities are essential for validating semiconductor manufacturing processes, the manual interpretation of the resulting images is a labor-intensive and error-prone task that can lead to inconsistencies and reduced productivity.

Recent advancements in artificial intelligence (AI), particularly in deep learning (DL), have unlocked transformative potential for automating image-based tasks such as image denoising and super-resolution [2, 3, 4], segmentation-driven automated measurement [5], and defect inspection and classification [6, 7]. However, the practical deployment of DL in semiconductor manufacturing faces a critical bottleneck: the scarcity of labeled training data. For TEM analysis, obtaining annotated datasets is prohibitively expensive because it needs precise segmentation labels (often requiring expert

* sh07.yu@samsung.com
† hansaem.park@samsung.com

Metrology, Inspection, and Process Control XXXIX, edited by Matthew J. Sendelbach, Nivea G. Schuch, Proc. of SPIE Vol. 13426, 134261L · © 2025 SPIE
0277-786X · doi: 10.1117/12.3050624



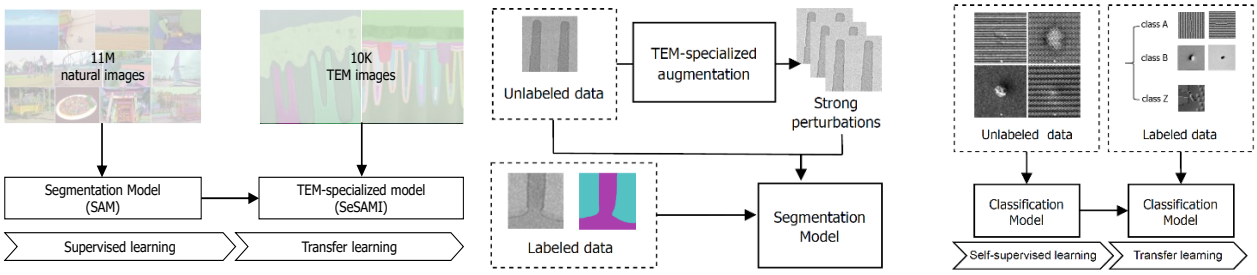
(a) TEM image segmentation (b) SEM defect classification

Figure 1. Semiconductor image analysis problems addressed in this paper. (a) Segmenting TEM images by materials. TEM imaging requires sample destruction, and material boundaries are often ambiguous, thus acquiring labeled TEM images for segmentation is time-consuming and costly. (b) Classifying SEM defect images to identify the root cause of defects. Collecting labeled datasets for new devices or defect types is challenging and resource-intensive. Data-efficient learning is essential to mitigate these challenges.

validation), and it requires sample destruction for imaging, making repeated analysis impossible. SEM imaging is primarily used for defect inspection and classification. Compared to segmentation, it allows for a larger number of samples and is easier to annotate. However, frequent data changes and the difficulty of obtaining sufficient labeled examples for early-stage R&D devices remain significant challenges.

These constraints have spurred interest in data-efficient learning which aims to maximize model performance with limited labeled data [8]. Previous research on data-efficient learning has explored various strategies, including developing foundation models [9], semi-supervised learning [10], and self-supervised learning [11, 12] to reduce the reliance on large labeled datasets. Kirillov et al. [9] introduced Segment Anything Model (SAM), a foundation model for segmentation that leverages a large-scale dataset and prompt-based segmentation, allowing it to generalize well with minimal fine-tuning. The ability of SAM to segment objects from a single prompt significantly reduces the need for extensive labeled data, making it an effective approach for data-efficient learning. Sohn et al. [10] developed FixMatch, a semi-supervised learning method that relies on a combination of pseudo-labeling and consistency regularization. By leveraging strongly augmented versions of unlabeled images and enforcing consistency with weakly augmented versions, FixMatch achieves high accuracy while requiring a minimal amount of labeled data, demonstrating the effectiveness of leveraging unlabeled data in a data-efficient manner. He et al. [11] proposed a self-supervised learning framework where a portion of the input image is masked to encourage the model to reconstruct missing components. This approach enables the model to learn strong representations with reduced labeled data dependency, significantly enhancing data efficiency in vision tasks. Chen et al. [12] introduced SimCLR, a contrastive learning framework which maximizes agreement between different augmented versions of the same image. By leveraging large-scale unlabeled datasets and simple architectural modifications, it achieves competitive results with fewer labeled samples, making it a key advancement in self-supervised learning for data-efficient learning. However, these methods often underperform when applied to semiconductor image analysis due to domain gap and domain-specific challenges. Therefore, there is a need for developing data-efficient learning methods tailored to semiconductor image analysis.

As shown in Figure 1, we address the two most challenging and important tasks in semiconductor image analysis: TEM image segmentation (Figure 1 (a)) and SEM image defect classification (Figure 1 (b)). Recent studies have explored domain-specific adaptations of data-efficient learning for semiconductor image analysis. Lee et al. [13] developed a weakly supervised image segmentation framework to address high cost and inefficiency of manual defect analysis in semiconductor manufacturing. While validated for single pattern bridge defects in a specific production process, its broader utility for diverse defect types remains unproven. Cai et al. [14] explored active learning to segment X-ray microscopy images with contrastive learning and rare sample selection. However, it does not account for the unique characteristics of semiconductor images when augmenting data for contrastive learning. Using augmentation specialized for semiconductor imagery could potentially enhance performance, as tailored strategies for semiconductor data may better preserve critical structural details or address domain-specific noise patterns. Geng et al. [15] developed a wafer defect classifier which combines few-shot learning to address class scarcity and self-supervised learning to utilize unlabeled wafer maps and augmentations. Kwak et al. [16] introduced SWaCo, which tackles out-of-distribution (OOD) unlabeled data in wafer bin maps via contrastive learning and demonstrated improved accuracy on small labeled datasets. Despite their success in constrained settings, these methods are designed for wafer maps which have simple patterns and limited defect classes,



(a) (b) (c)

Figure 2. The schematic diagram of the proposed methods. (a) Transfer learning of a foundation model to TEM image segmentation. It highly reduces labeling costs by providing robust initial annotation masks. (b) Semi-supervised learning with TEM-specialized perturbation. It leverages unlabeled TEM images to improve segmentation performances with limited labels. (c) Self-supervised learning for defect classification to adapt classification models to new devices with limited labeled data.

and they cannot be directly applied to the SEM-based defect classification tasks, which involve relatively large datasets, diverse class labels, and include multi-magnification images.

To overcome these limitations of the conventional methods, we propose data-efficient learning methods that leverage domain-specific knowledge of semiconductor images for TEM segmentation and SEM defect classification in three different applications as shown in Figure 2. A brief overview of each method is described below.

Transfer learning of a foundation model for semiconductor image segmentation. As Figure 1 (a) shows, annotating TEM images to segment them based on material requires significant effort and time. Therefore, if a universal segmentation model that can be applied to diverse semiconductor devices and structures is developed, and its inference values are used as initial segmentation masks, this could drastically reduce annotation time. As shown in Figure 2 (a), we develop transfer learning of the foundation model, Segment Anything Model (SAM) [9], originally developed for natural images, to semiconductor TEM data to adapt a foundation model for TEM segmentation.

Semi-supervised learning with TEM-specialized augmentation. Although annotation costs have been reduced through our foundation model, emerging new devices make it challenging to secure sufficient labeled data for training segmentation models. To address this, as shown in Figure 2 (b), we propose a semi-supervised learning method that leverages the characteristics of TEM images via TEM-specialized augmentation to improve model performance using unlabeled data.

Self-supervised pre-training with SEM-specialized contrastive learning. As shown in Figure 1 (b), SEM images are used for defect inspection and classification. To rapidly adapt classification models to new defect data arising from new devices or manufacturing processes, we propose a self-supervised learning method utilizing unlabeled datasets as illustrated in Figure 2 (c). The proposed method is trained to extract scale-invariant features by leveraging the unique characteristics of SEM images.

2. DATA-EFFICIENT LEARNING FOR SEMICONDUCTOR IMAGE ANALYSIS

A task-specific methodology, which is tailored to the specific data constraints of each problem, should be applied to analyze semiconductor images. From the perspective of semiconductor MI using electron microscopy images, we focus on three primary issues and propose data-efficient learning methods tailored to each application. First, for fast and efficient annotation of segments across diverse devices, a domain-specific segmentation model is constructed through transfer learning of a foundational model. Second, to address model performance degradation due to insufficient training data, we propose semi-supervised learning based on TEM-specialized data augmentation. Lastly, to improve the performance of defect classification models using limited labeled data, we introduce self-supervised learning techniques based on SEM-specialized contrastive learning. Each methodology is described in detail in the following subsections.

2.1 Transfer learning of a foundation model for semiconductor image segmentation

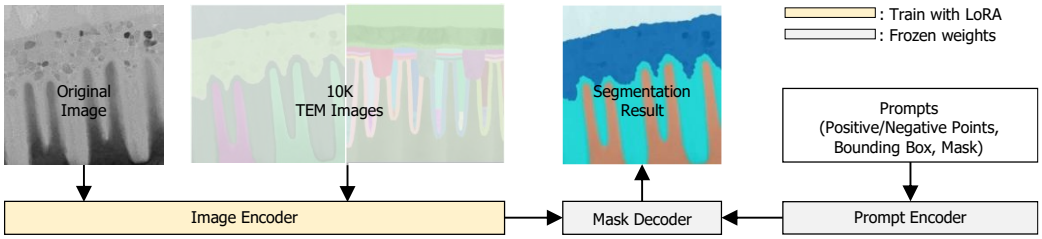
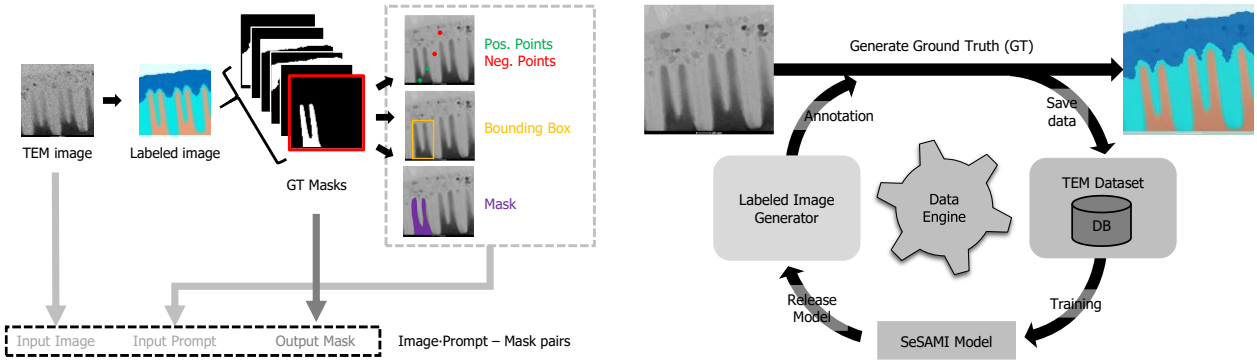


Figure 3. Overview of the proposed SeSAMI model which consists of image encoder, prompt encoder, and mask decoder. Masked decoder generates segmentation outputs using encoded image and prompts. Image encoder of SeSAMI is fine-tuned by low-rank adaptation (LoRA) which allows parameter-efficient fine-tuning of large foundation models.



(a) Dataset: TEM image-mask pairs

(b) Data Engine: virtuous cycle system to create labeled images

Figure 4. Dataset generation procedure. (a) TEM image-mask pairs are generated from labeled images for transfer learning of SAM. (b) Labeled TEM images are obtained through virtuous cycle system.

Recently, a foundation model for image segmentation named Segment Anything Model (SAM) [9] has been released, which is trained on a vast amount of labeled datasets and can be used universally. While the model offer versatility, their performance on TEM images is hindered by domain gaps (e.g., data distribution, noise, ambiguity). We present Semiconductor Segment Anything for Metrology and Inspection (SeSAMI), a refined SAM variant fine-tuned using Low-Rank Adaptation (LoRA) [17] on semiconductor dataset. By freezing SAM's backbone and updating only 1–2% of parameters, SeSAMI outperforms SAM and baseline model trained with supervised learning, and generalizes to novel materials without full retraining. This approach leverages the foundation model trained on a large-scale natural image dataset, enabling efficient adaptation to the specific context of semiconductor images even with limited data.

Figure 3 illustrates structure of the SeSAMI model, where the image encoder generates image features, and the prompt encoder generates the sparse or dense embedding according to prompt types. The mask decoder receives encoded image and encoded prompt as input to create the desired segmentation masks. While the image encoder is fine-tuned, the prompt encoder and mask decoder are frozen to prevent them from being trained.

As a transfer learning methodology, SeSAMI adopt LoRA to fine-tune the image encoder with a pre-trained vision transformer (ViT) [18]. LoRA enables efficient adaptation of a model to semiconductor image domain where training data and resources are limited. It updates the adapter weights while keeping the baseline model weights fixed, thereby facilitating rapid and resource-efficient specialization of the model in semiconductor image domain. It can be expressed as an equation as follows:

$$W_0 + \Delta W = W_0 + BA,$$
$$\text{where } B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times d}, r \ll d. \quad (1)$$

W_0 corresponds to pre-trained weights with a dimension of $\mathbb{R}^{d\times d}$, and ΔW denotes the amount of transferred weights and has the same dimension as pre-trained weights. Therefore, the output $\mathcal{h}$ and the input x satisfy the following equation:

$$h = W_0 x + \Delta x = W_0 x + \Delta W x = W_0 x + BAx. \quad (2)$$

As a re-parameterization, A and B matrices are initialized to random Gaussian and zero, respectively. As in [17, 26], SeSAMI applies LoRA only to query and value of transformer blocks.

MI in semiconductor manufacturing requires high accuracy and consistency. SeSAMI ensures coherent segmentation results by training on a variety of input prompt-output mask pairs, as depicted in Figure 4 (a). To simulate user inputs, diverse prompt for training is generated from ground truth (GT) masks. Specifically, point prompts are created by randomly inserting coordinates into the images, bounding box prompts are derived by outlining bounding boxes around objects based on GT masks, and mask prompts are modeled using randomly generated partial masks that are smaller in size than the original GT masks. By incorporating this variety of prompts into training dataset, models can robustly adapt to different input types, ensuring consistent performance across diverse prompt configurations.

Training data consists of TEM images including the core structures of products such as DRAM, Flash, and Logic. To generate the training data, segmentation GT of TEM images are produced by elaborately defined criteria from the consultation of engineers. To generate the labeled data, super pixel-based annotation algorithm and drawing tools are used, which is provided by our labeled image generator. After training, SeSAMI can infer the segmentation masks from the newly produced TEM images. If the result shows low coherence with the GT, it is refined by manual correction to be used as new training data. As repeating the generation and re-training process, the segmentation accuracy of SeSAMI is enhanced. The data engine represents the cyclical data generation and re-training process as shown in Figure 4 (b).

2.2 Semi-supervised learning with TEM-specialized augmentation

Using the SeSAMI methodology introduced in the previous section, we are able to achieve rapid annotation of segmentation masks. However, since a vast amount of data is continuously generated, an enormous quantity of unlabeled data remains. Building on success of semi-supervised learning methods [10, 19] on natural image datasets, we propose a semi-supervised learning framework for TEM image segmentation in semiconductor devices to leverage unlabeled images. By combining limited labeled data (20–100 images) with additional unlabeled TEM data, our method introduces TEM-specialized noise perturbations in weak-to-strong consistency regularization [10, 19] to improve segmentation performances by expanding training dataset.

As a semi-supervised learning, consistency regularization assumes that features perturbed from the same data should have similar probability distribution. By applying consistency regularization using variously perturbed unlabeled data, the decision boundary becomes closer to the real data distribution. Figure 5 illustrates the overview of the proposed semantic segmentation by utilizing weak-to-strong consistency. This framework uses pseudo labeling and consistency regularization based on weak and strong perturbations from unlabeled data. While weak-to-strong consistency was originally designed for natural images, we adapt this method to TEM image segmentation by leveraging TEM-specialized image augmentation.

The process of weak-to-strong consistency regularization is expressed as follows. For the unlabeled data x^u, the probability of weak perturbation p^w is calculated by using model F after applying weak image perturbation $\mathcal{A}^w$. On the other hand, the probability of a strong image perturbation p^s is calculated by first applying weak perturbations followed sequentially by strong perturbations and then utilizing model F for computation.

$$p^w = F\left(\mathcal{A}^w(x^u)\right), \quad p^s = F\left(\mathcal{A}^s\left(\mathcal{A}^w(x^u)\right)\right). \quad (3)$$

If the maximum value of p^w is bigger than threshold τ, it is determined as the pseudo label of semi-supervised learning, and the semi-supervised loss $\mathcal{L}_u$ is calculated by comparing p^s with p^w as follows:

$$\mathcal{L}_u = \frac{1}{\mathcal{B}_u} \sum \mathbb{I}(\max(p^w) \geq \tau) \mathrm{H}(p^w, p^s), \quad (4)$$

where $\mathcal{B}_u$, $\mathbb{I}$, and H is the batch size for unlabeled data, unit step function, and cross entropy, respectively.

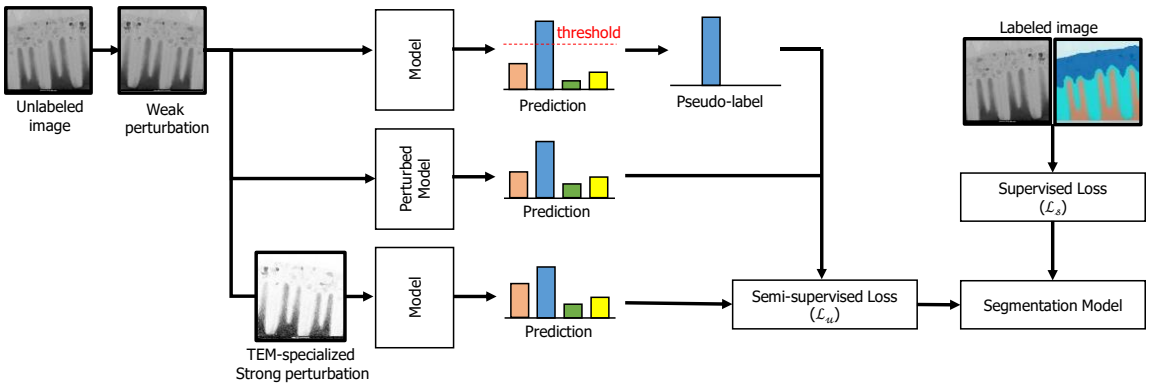


Figure 5: Overview of the proposed semi-supervised learning framework for semantic segmentation based on weak-to-strong consistency. The framework generates two distinct augmented images from a single unlabeled input to form positive sample pairs, enabling consistency regularization. To generate strong perturbation, we introduce TEM-specialized augmentations, designed to simulate diverse TEM-specific noise patterns.

Finally, for the model training, the final loss function $\mathcal{L}$ is obtained by combining $\mathcal{L}_s$, obtained b supervised learning, and $\mathcal{L}_u$ as follows:

$$\mathcal{L} = \frac{1}{2}(\mathcal{L}_s + \mathcal{L}_u). \tag{5}$$

Our framework consists of similar structure with UniMatch [19], with respect to simultaneously proceeding feature perturbation and data perturbation. However, strong perturbation set of UniMatch is composed of color jittering and CutMix [23] which consider characteristics of natural images, not TEM images. To specialize augmentation in the TEM images, the proposed method uses more than two strong perturbation set based on TEM-specialized perturbation. The model selectively engages the perturbation set which shows the highest loss in training. Combining feature perturbation into Equation 4, semi-supervised loss can be described as follows:

$$\mathcal{L}_u = \frac{1}{\mathcal{B}_u}\sum \mathbb{I}(\max(p^w) \geq \tau) \times (\lambda \cdot \sum_i \mathrm{H}(p^w, p^{s_i}) + (1\text{-}\lambda) \cdot \mathrm{H}(p^w, p^{fp})), \tag{6}$$

where, p^{fp} denotes the probability of feature perturbation, λ is weight of cross entropy, which is experimentally set to 0.5.

For TEM-specialized augmentation, we subdivide the TEM noise into three categories: sample noise, TEM machine noise, and post processing noise. Subdivided TEM noises are applied to each augmentation stage to improve the segmentation accuracy of the model. Sample noise arises from contamination and deformation of a sample during sample preparation process using Focused Ion Beam (FIB). TEM machine noise, which includes dark current noise, gain noise, and shot noise, arises from variations in sensor scaling of the TEM equipment. Post-processing noise is exemplified by artifacts such as the scale bar superimposed on a TEM image.

We can generate strong perturbations applying TEM-specialized image augmentation to unlabeled data. By using the model's inference results on the unlabeled data as pseudo-labels, we can incorporate a consistency regularization term into the loss function by leveraging these image pairs, thereby enhancing model performance.

2.3 Self-supervised pre-training with SEM-specialized contrastive learning

In semiconductor manufacturing, SEM images are essential for defect inspection and classification. They enable training of the auto defect classification (ADC) models to detect and classify defects in semiconductor wafers. However, in a dynamic R&D environment, frequent domain changes occur due to the introduction of new devices, inspection steps, or process modifications. These changes lead to shifts in data distributions, necessitating recurring model retraining.

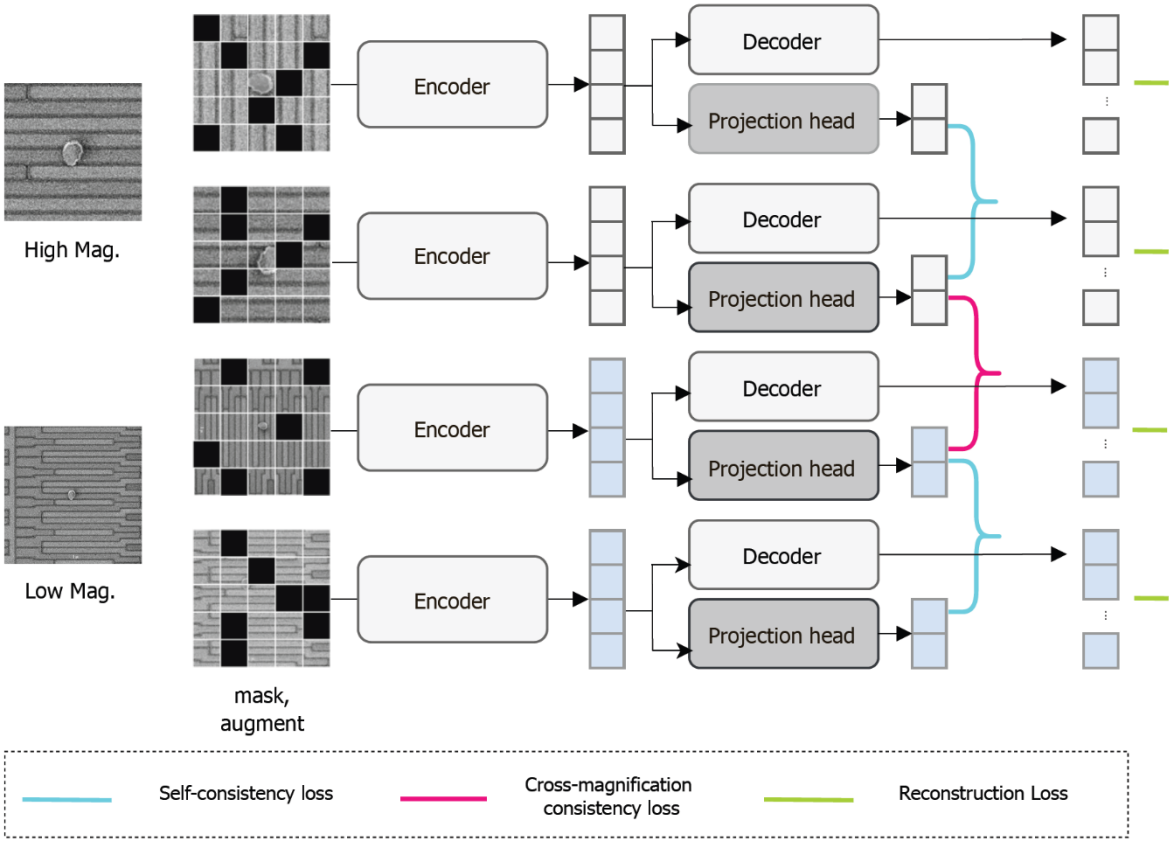


Figure 6. Overview of the proposed self-supervised learning framework. Masked autoencoder [11] is trained by reconstructing defect images to learn low-level features, and contrastive learning [12] is deployed via self-consistency loss and cross-magnification consistency loss to learn augment- and scale-invariant high-level features. High- and low-magnification images derived from a single defect are used to compute cross-magnification consistency loss, enabling the model to learn scale-robust features.

To rapidly adapt models to new domains, a large amount of labeled data is required which is labor-intensive and time-consuming to gather, often leading to delays in model deployment. One promising method to tackle data scarcity issue is using pre-trained model from a large amount of labeled dataset. However, most pre-trained models for image classification [18, 24] are trained on natural images which differ significantly from SEM images in characteristics and data distribution. As a result, directly applying these models for SEM analysis yields limited performance benefits.

To address this, we propose a self-supervised learning method that efficiently adapts models to new target domains using minimal labeled data combined with unlabeled data. Building on the work by Mishra et al. [20], which combines masked autoencoder [11] and contrastive learning [12] techniques for self-supervised learning, we introduce a novel pre-training approach tailored specifically for semiconductor SEM images. As shown in Figure 6, proposed self-supervised learning approach combines masked autoencoder and contrastive learning.

In the masked autoencoder framework, the input image is partitioned into fixed-size patches, and a predefined ratio of these patches is randomly masked, while the remaining unmasked patches are fed through an encoder to produce embedded features. These features are concatenated with dummy placeholders representing the masked patches, forming the input to the decoder. The decoder then reconstructs masked patches. As specified in Equation 8, the model is trained by minimizing the L2-loss between the reconstructed values of the masked patches and their corresponding original patches. SEM images exhibit small-scale defects with highly variable patterns depending on the device structures and inspection steps. By training the model to accurately reconstruct both defective and background regions, the model learns to extract low-level

features that capture the essential characteristics of SEM imagery. This approach not only restores pixel-level context but also enables the model to adapt to various defect patterns, improving its performance across different types of datasets.

In the contrastive learning framework, two augmentations are applied to a single image to create two independent views, which are then input into an encoder to extract embedded features. The embedded features are transformed into high-dimensional feature vectors through a projection head. The feature vectors from the same image are treated as positive pairs, while feature vectors extracted from different images are treated as negative pairs. The model is trained such that the extracted features of positive pairs are close to each other in feature space, while the negative pairs are far apart.

However, in semiconductor SEM equipment, if the same defect is captured in both high- and low- magnification modes, two views can be created from a single defect without additional augmentation. By applying further augmentation to each view, as shown in Figure 6, a total of four augmented images can be generated from a single defect. Therefore, we can use these four images as a positive pair to calculate the consistency loss. By considering the consistency between the feature vectors extracted from images with different magnifications, we can train the model to extract scale-invariant features, which is named cross-magnification consistency loss. Additionally, we can also compute the self-consistency loss by generating two augmented images from each magnification image.

Therefore, the final loss function for the proposed self-supervised learning is composed of three components: reconstruction loss, self-consistency loss, and cross-magnification consistency loss as follows:

$$\mathcal{L} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{con}\mathcal{L}_{con}, \tag{7}$$

$$\mathcal{L}_{rec} = \frac{1}{2N}\sum_{m\in\{H,L\}}\sum_{i=1}^{2N}(x_i^m - y_i^m)^2, \tag{8}$$

$$\mathcal{L}_{con} = -\frac{1}{4}\sum_{m,n\in\{H,L\}}\frac{1}{2N}\{\sum_{i=1}^{N}\log\frac{\exp(f_i^m\cdot f_{j(i)}^n/\tau)}{\sum_{a\in I}\exp(f_i^m\cdot f_a^n/\tau)} + \sum_{i=N+1}^{2N}\log\frac{\exp(f_i^n\cdot f_{j(i)}^m/\tau)}{\sum_{a\in I}\exp(f_i^n\cdot f_a^m/\tau)}\}, \tag{9}$$

where x and y denote reconstructed and GT image patches, respectively, and f represents the features extracted from the projection head. Each image i in batch size of $2N$ has its corresponding positive pair $j(i)$. For simplicity, the negative samples associated with image i in the batch are represented as I. High- and low-magnification images are denoted as H and L, respectively. Final loss function for self-supervised learning is denoted as $\mathcal{L}$, which is weighted sum of reconstruction loss $\mathcal{L}_{rec}$ and contrastive loss $\mathcal{L}_{con}$ based on balancing weights λ.

3. EXPERIMENTAL RESULT

We describe the experimental setup and dataset configuration for each proposed method, and present the experimental results using real TEM/SEM images of semiconductor devices, comparing them with conventional methods.

3.1 Transfer learning of a foundation model for semiconductor image segmentation

Experimental setup. To evaluate the proposed foundation model for segmentation, SeSAMI, 96 types of semiconductor TEM images are collected as training datasets including DRAM, Logic, and Flash devices, which consists of 3,873 training images with 23,861 GT masks. Mean intersection over union (mIoU) [21] is used as an evaluation metric, which quantifies the overlap between the estimated segmentation result and the GT, to compare segmentation accuracy with existing benchmark method and base model of SeSAMI. The semantic segmentation method based on DeepLab V3+ [22] has been applied to automate measurements via image segmentation. Since this approach is a supervised learning-based method that demonstrates general performance for device module TEM images, it serves as a benchmark. Additionally, we evaluate the extent to which the proposed foundation model improves upon SAM [9].

Evaluation on various devices. As shown in Figure 7, mIoU is compared for 96 types of device modules which are same types of training dataset. For each image of a device module, mIoU is calculated and averaged for all images. Improvement is observed for 97.96% of all modules, and few degradation is also observed in 2.04% of modules. According to qualitative analysis, degradation of accuracy occurs when the two materials share a thin boundary, thus it can be improved by properly given prompts such as negative points. It indicates that human aid is still necessary for SeSAMI to be a comprehensive solution to segmentation without any external intervention.

Table 1. Comparison of segmentation accuracy for all modules and detection-related modules.

mIoU		Segmentation model		
		DeepLab V3+ [22]	SAM [9]	SeSAMI
All modules (96)	Avg.	0.9517	0.7614	**0.9534**
	Max.	0.9940	0.9744	**0.9961**
	Min.	**0.8374**	0.3165	0.7167
	Std.	**0.0297**	0.1548	0.0394
Detection-related (36)	Avg.	0.9429	0.7702	**0.9634**
	Max.	0.9915	0.9483	**0.9961**
	Min.	0.8374	0.5065	**0.9045**
	Std.	0.0345	0.1341	**0.0200**

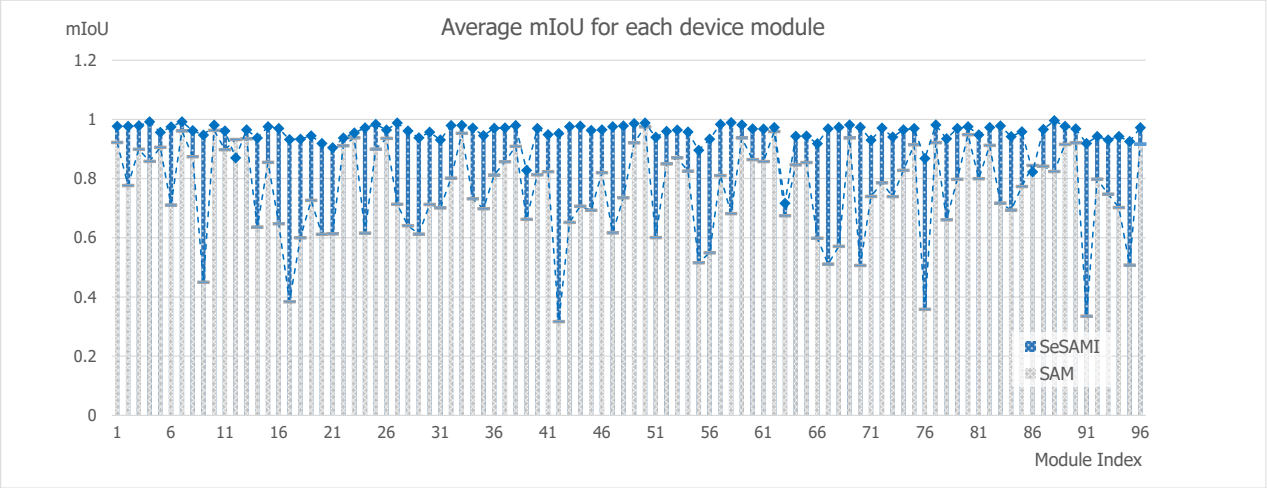


Figure 7. Comparison of average mIoU on each device module between the base model (SAM) and the proposed model (SeSAMI). The proposed SeSAMI outperforms the base model in 97.96% of all modules.

In a comprehensive analysis, mIoU for all modules is compared with supervised learning and SAM to evaluate how improved the foundation model is, compared to the supervised learning-based overfitting training method. As shown in Table 1, the average mIoU is improved by 0.0017 and 0.1920 compared to supervised learning and SAM, respectively. In particular, SeSAMI shows a similar segmentation accuracy compared to supervised learning, which means that the foundation model, SeSAMI, is comparable to supervised learning based overfitted model. Because the segmentation results of semiconductor TEM images are similar to GT, which means SeSAMI has same consistency with the criteria of existing engineers, measurement automation based on segmentation is feasible.

Evaluation on detection-related modules. TEM image analysis on device module includes detecting particular structures as well as measuring critical dimension. Among the 96 modules, 36 modules include structures for small object detection. To evaluate segmentation accuracy from the perspective of object detection, SeSAMI is compared with supervised learning and SAM using mIoU on these detection-related modules, as shown in Table 1. Even segmentation has a different purpose from object detection, it shows somewhat higher overall detection and segmentation accuracy than supervised learning in detecting small objects. The average mIoU is improved by 0.0205 and 0.1932 compared to supervised learning and SAM, respectively.

Adaptability of structurally similar devices The primary objective of constructing a segmentation foundation model for TEM images of diverse semiconductor device modules is to develop a robust framework that adapts to process variations in high-volume manufacturing (HVM) or R&D products. It can minimize the human and material resources required for training new models, reducing a process from obtaining TEM images to generating GT data. Structural variations in TEM images, including differences in scale, perspective, and fabrication steps, significantly affect segmentation reliability. Therefore, SeSAMI which is trained with TEM images with specific structure, is validated to assess segmentation performance under structural changes.

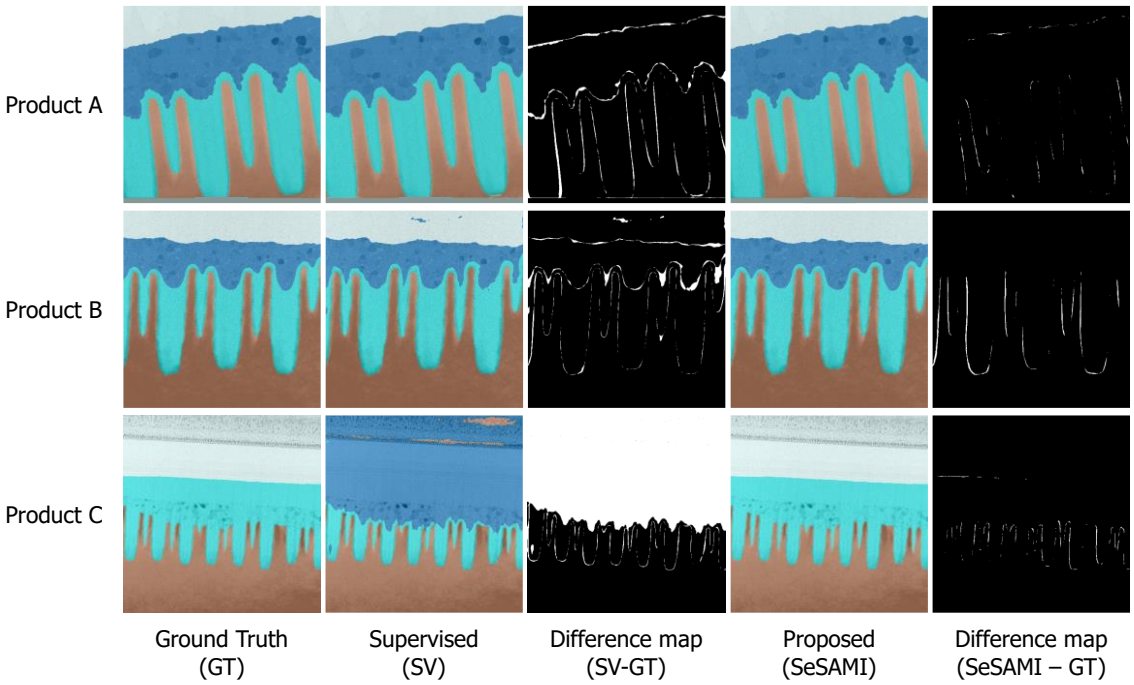


Figure 8. Comparison of segmentation results for three products. SeSAMI demonstrates a minimal difference from the GT, indicating that it is sufficiently trained as a robust foundation model capable of handling TEM images derived from various products.

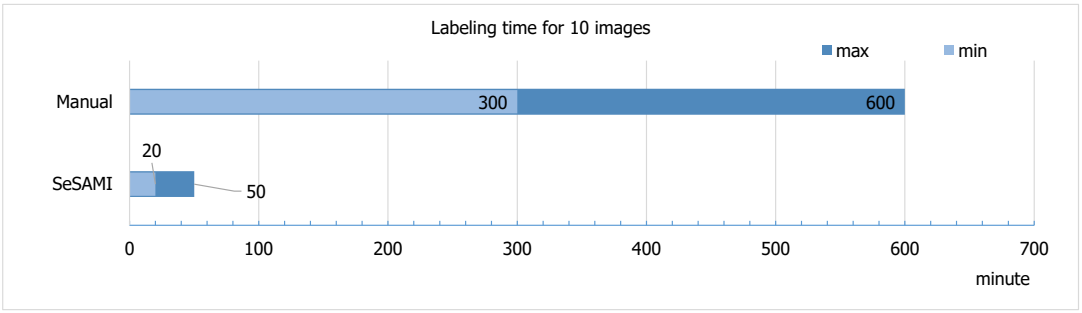


Figure 9. Time efficiency of annotation using SeSAMI-generated initial masks. The model produces high-quality preliminary annotations, enabling a significant reduction in manual labeling efforts compared to traditional method.

The proposed model was trained using TEM images from identical semiconductor modules across different product generations, where the most varied datasets were collected. Both the existing supervised learning model and the proposed SeSAMI were trained exclusively on images from a single reference product, product A. To evaluate their adaptability, each trained model was qualitatively compared using TEM images with distinct device IDs (product B) and significantly different process steps (product C), as illustrated in Figure 8. The difference maps demonstrate that while the supervised learning model shows minimal error for product A, it exhibited substantial discrepancies for products B and C. In contrast, SeSAMI consistently aligned its segmentation results with GT images, confirming its robustness as a foundation model capable of adapting to diverse device structures or generations.

These results indicate that transfer learning applied to SeSAMI achieves segmentation accuracy comparable to fully supervised learning across various device modules, ensuring consistent standards even when analyzing TEM images from different generations or modules. Additionally, the model's stability enables efficient cross-type TEM image labeling. By leveraging SeSAMI's prediction outputs as initial annotation masks, the time required to assemble training datasets is significantly reduced, which is empirically demonstrated in Figure 9.

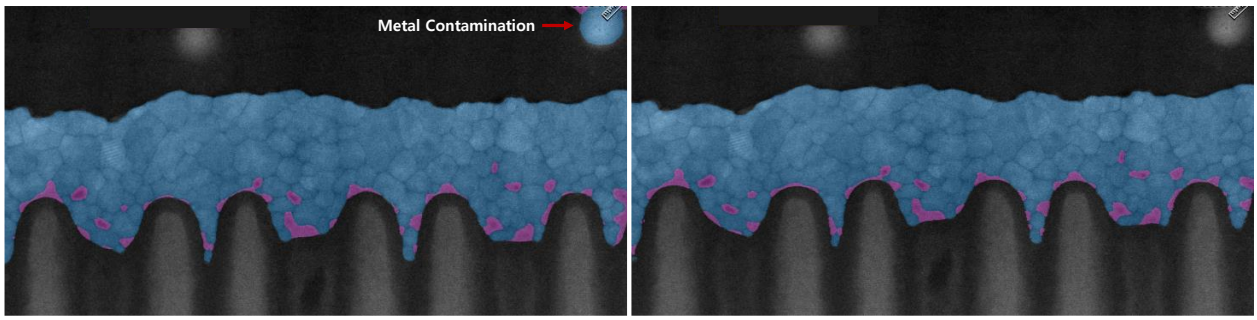
3.2 Semi-supervised learning with TEM-specialized augmentation

Table 2. mIoU comparison for few-shot case: all segments, material-only, and defect-only. The left side shows the result using 20 labeled data from device A, while the right side shows the result using 60 labeled data from device A. For semi-supervision, additional 70 unlabeled samples from device A were used.

Methods	Device A_20			Device A_60		
	All	Material	Defect	All	Material	Defect
Supervised	0.852	0.974	0.592	0.867	0.976	0.634
CutMix [23]	0.867	0.978	0.630	0.868	0.978	0.634
UniMatch [19]	0.795	0.958	0.440	0.827	0.968	0.521
Proposed	**0.874**	**0.979**	**0.650**	**0.877**	**0.979**	**0.661**

Table 3. mIoU comparison on device B: all segments and defect-only. A scenario without labeled data from device B was assumed. For semi-supervision, 100 unlabeled data from device B were used.

Methods	Device A _20		Device A_60	
	All (Device B)	Defect (Device B)	All (Device B)	Defect (Device B)
Supervised	0.646	0.091	0.813	0.515
CutMix [23]	0.784	0.440	0.843	0.599
Proposed	**0.825**	**0.556**	**0.845**	**0.604**



(a) (b)

Figure 10. Comparison of segmentation accuracy: (a) supervised learning and (b) the proposed model. The supervised learning approach frequently misclassifies regions affected by metal contamination. In contrast, the proposed method exhibits robustness to such noise patterns, reliably maintaining accurate predictions

Experimental setup. For validation of the proposed semi-supervised modeling approach, 150 labeled data of device A was collected. Among the labeled data, 60 samples were used as the training dataset. In addition, extreme scenario was assumed where a training set contains only 20 samples. For the unlabeled data used for semi-supervised learning, 70 samples of device A and 100 samples of device B were used. Segmentation accuracy is compared with benchmark models using mIoU as the evaluation metric. The experiments evaluate supervised learning, CutMix [23], UniMatch [19], and the proposed model, all utilizing DeepLab V3+ [22] as the backbone structure for the semantic segmentation model.

Evaluation on few-shot case. According to the experimental result shown in Table 2, the proposed model exhibits the highest performance with both 20 and 60 labeled data of device A. Particularly, it demonstrates significant accuracy improvement in environments with fewer data. Another noteworthy finding in the results is that when the UniMatch is used without modifications, even it is also based on weak-to-strong consistency similar to our proposed model, it delivers significantly lower accuracy compared to conventional supervised learning. This can be attributed to the lack of suitable perturbations for TEM images. Specifically, color jittering shows significant negative impact on prediction performance.

In Table 2, when comparing the mIoU values between the supervised learning method and the proposed method, the difference is trivial, with a 0.022 difference for device A_60 and a 0.003 difference for device A_20. However, it is important to note that the test data used in this experiment was collected around the same time as the training data, which differs from a real-world scenario where the model, trained on TEM data, would be applied to data collected after the

training process. Additionally, various noises which occur obtaining TEM images were excluded from this test set. Figure 10 compares the results of applying supervised learning and the proposed method to the TEM images, considering such real-world test scenarios. As highlighted by the red arrow in Figure 10 (a), the supervised learning method tends to incorrectly predict areas with metal contamination. In contrast, as shown in Figure 10 (b), the proposed method demonstrates robust performance against these types of noise.

Evaluation on unlabeled data. Furthermore, experiments are expended to determine whether it is possible to improve performance using only unlabeled data for new devices without labeled data of the target device. The experiment was conducted by preparing labeled data consisting of device A, with 20 and 60 samples respectively. Additionally, 100 samples of unlabeled data of device B and 70 samples of unlabeled data of device A were used for proposed semi-supervised learning.

As listed in Table 3, experimental results show that regardless of the number of labeled data, the proposed semi-supervised learning approach achieves the highest mIoU. In particular, the proposed network demonstrates high accuracy with respect to mIoU over 0.8 even in situations where only 20 labeled data samples were available.

Particularly, the model accuracy trained with device A_20 data was improved by 0.179 and 0.041 compared to benchmark methods. In addition, in the case of defect in device B, which is the main measurement of B product, the accuracy was improved by 0.465 and 0.116 compared to benchmark methods. In addition, if model is trained with device A_60 data, it shows slightly improved segmentation accuracy compared to benchmark methods. Based on these results, we conclude that our proposed semi-supervised learning method achieves performance comparable to supervised approaches in the target domain, even when no labeled data is available in that domain, by leveraging only a small amount of labeled data from a different source domain and a few unlabeled samples from the target domain.

3.3 Self-supervised pre-training with SEM-specialized contrastive learning

Experimental setup. In this study, we evaluated the proposed self-supervised learning method against conventional approaches in semiconductor defect classification. All experiments leveraged SEM images captured from DRAM devices, with high- and low-magnification images paired to form each sample. The backbone model chosen is the Vision Transformer (ViT) [18], a state-of-the-art architecture known for its superior performance on image classification applications. The ViT's attention mechanisms are well-suited for capturing both global structural patterns and localized defect details in SEM images. For reconstruction, we adopted the decoder structure from [11]. The loss function combined the reconstruction loss from masked autoencoder (MAE) and a specialized contrastive learning (CL) loss. Loss weights λ_{rec} and λ_{con} in Equation 7 are set to 0.97 and 0.03, respectively, to calculate the final loss function in proposed pre-training scheme.

The experimental datasets were constructed to reflect realistic industrial scenarios. For device A, the pre-training dataset contained 320K unlabeled high- and low-magnification pairs, and 80K labeled pairs were reserved for fine-tuning. A distinct test set of 60K labeled pairs were prepared. These datasets are temporally disjoint, ensuring no overlap between unlabeled data (used for pre-training), labeled training data (for fine-tuning), and test data (for evaluation). For device B, which is a novel device with extremely limited data, no unlabeled images were available. Therefore, only 20K labeled pairs were used for both pre-training and fine-tuning, and the test set comprised 2K held-out samples, which is an extremely constrained environment mimicking real-world semiconductor R&D conditions.

Evaluation on device *A*. Results for device A, as listed in Table 4, demonstrate the importance of domain-specific pre-training. Transfer learning from pre-trained models with ImageNet [25] using cross-entropy (CE) loss achieved a relatively low accuracy of 84.65%, due to the significant domain mismatch between natural images and SEM images. In contrast, MAE pre-training on SEM data boosted accuracy by 2%, which aligns with MAE's strength in recovering delicate structural features, such as small defects often obscured by noise or background interference. Consequently, we demonstrate that pre-training with a relatively small amount of semiconductor SEM images, only 320K, which is just 1% of the tens of millions of natural images typically used for pre-training, is still effective in improving performance.

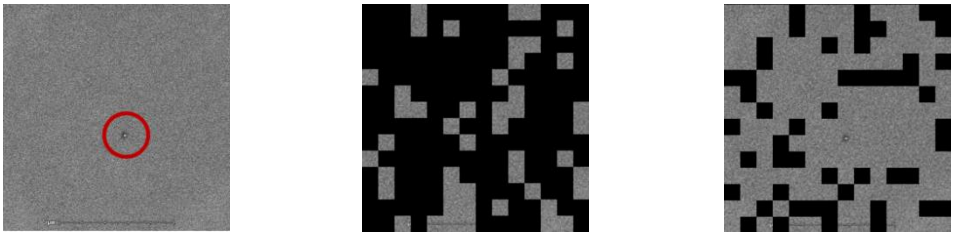
A striking finding emerged with masking ratios: reducing the ratio to 0.25 (vs. the standard 0.75 in natural image tasks [11]) increased accuracy further. This result differs from previous studies because SEM defects are pixel-wise tiny compared to relatively big objects in natural images. As depicted in Figure 11, a higher masking ratio risks obscuring entire defects, forcing the model to learn background textures instead. The lower masking ratio allows the model to observe sufficient defect features during reconstruction, thereby focusing on learning discriminative patterns.

Table 4. Comparison of classification accuracy for device *A*. Proposed self-supervised learning which deploys cross-magnification consistency loss shows the best performance.

Pre-train loss	Masking ratio	Self-consistency	Cross-mag.-consistency	Accuracy
CE (ImageNet)	.	.	.	84.65
MAE	0.75	.	.	86.78
MAE	0.25	.	.	87.64
MAE+CL	0.25	O	X	87.65
MAE+CL	0.25	X	O	87.89
MAE+CL (Proposed)	0.25	O	O	**88.09**

Table 5. Comparison of classification accuracy for device *B*. It assumes a scenario where both labeled and unlabeled data in the target domain for training are extremely limited. The proposed method achieves considerable model performance with only a small amount of data.

Pre-train loss	Masking ratio	Self-consistency	Cross-mag.-consistency	Accuracy
CE (other device)	.	.	.	84.3
MAE+CL (Proposed)	0.25	O	O	**85.9**



(a) Original image (b) masking ratio=0.75 (c) masking ratio=0.25

Figure 11. In semiconductor SEM images, defects are often small in size. Using a high masking ratio can lead to removing the entire defect.

The addition of contrastive learning (CL) further elevates performance, particularly when using cross-magnification consistency loss which is a novel loss designed to align feature representations between high- and low-magnification images. This loss function enforces that the model's latent embeddings for a given defect remain consistent even under different magnifications. By ensuring cross-scale feature alignment, the model generalizes better to unseen defects. Combined with the proposed framework, this yielded a 3.3% accuracy gain over ImageNet pre-training method, proving the method's effectiveness in adapting to the semiconductor domain with minimal labeled data.

Evaluation on device *B*. device *B* exemplified a critical challenge in semiconductor R&D: extremely limited labeled data (only 20K samples) with no pre-training data. To address this, we explored two strategies: transfer learning by initializing the model with weights pre-trained on other semiconductor devices and fine-tuning on device *B*'s 20K labeled pairs and the proposed method to pre-train the model by self-supervised manner using the same labeled data (treated as unlabeled) and then fine-tuning with the same dataset. The results listed in Table 5 showed that even in this data-scarce scenario, the proposed approach achieved a 1.6% accuracy improvement over naive transfer learning. This success hinges on the self-supervised framework's ability to fully exploit the labeled data. By masking and reconstructing corrupted patches during pre-training, the model implicitly learns the unique texture variations and defect characteristics of device *B*'s images without requiring explicit labels. Contrastive loss further refines representation learning by emphasizing spatial relationships between defect patterns. In conclusion, we demonstrate that the proposed self-supervised learning framework effectively enhances model performance even without requiring extensive pre-training datasets, as it achieves remarkable results using only a limited amount of data from the target device.

Efficiency of the proposed method is paramount in industries where new semiconductor devices are rapidly developed and data collection is expensive. Conventional supervised learning risks overfitting due to insufficient samples, while transfer learning from unrelated domains like other devices struggles to overcome the domain gap. In contrast, the proposed approach maximizes the utility of scarce labeled data by first extracting unlabeled-style features (via reconstruction and

contrastive learning) before focusing on task-specific learning during fine-tuning, which is effectively mimicking semi-supervised learning in low-data conditions.

4. CONCLUSION AND FUTURE WORK

In this work, we have proposed data-efficient machine learning approaches for semiconductor image analysis. Among numerous semiconductor image analysis applications, we addressed two major and highly challenging tasks: TEM image segmentation and SEM image defect classification. First, for TEM image segmentation (pixel-level material classification), we developed a versatile foundation model named SeSAMI to mitigate the high labeling costs associated with this task. It adopts low-rank adaptation as a transfer learning of SAM, a foundation model which is trained with a large-scale natural image dataset, to enable effective segmentation of TEM images, and it demonstrated robust performance across various semiconductor devices and structures, accelerating the creation of labeled images from the unlabeled TEM images. Second, to further enhance segmentation performance by leveraging unlabeled TEM data, we introduced a semi-supervised learning framework. This method incorporated tailored augmentation strategies based on an analysis of TEM-specific noise characteristics, improving segmentation accuracy and ensuring stability against TEM noise. Finally, for SEM image defect classification in scenarios with limited labeled data such as early development phases, we proposed a self-supervised learning approach. By training the model to extract scale-invariant features of SEM defect images, we achieved superior classification performance even under data constraints.

The methodologies presented in this study have already demonstrated practical value in semiconductor R&D, powering automated metrology tools and defect classification software that reduce manual analysis time by up to 90%. Looking ahead, we envision extending these advancements beyond TEM/SEM to other critical semiconductor imaging modalities, such as optical inspections. To achieve this, our future work will focus on integrating our domain-specific transfer learning, semi-supervised augmentation, and contrastive pre-training strategies into a unified foundation model for semiconductor image analysis. By optimizing the foundation model and achieving cross-equipment adaptability, we aim to accelerate R&D workflows, enabling rapid analysis of novel processes and materials with minimal labeling cost.

REFERENCES

[1] Orji, Ndubuisi G., et al. "Metrology for the next generation of semiconductor devices," Nature Electronics, vol. 1, no. 10, pp. 532-547, 2018.

[2] Dey, B., et al. "SEM image denoising with unsupervised machine learning for better defect inspection and metrology," Metrology, Inspection, and Process Control for Semiconductor Manufacturing XXXV, vol. 11611, SPIE, 2021.

[3] Moly, A., et al. “Self-supervised deep learning neural network for CD-SEM image denoising using reduced dataset,” Metrology, Inspection, and Process Control XXXVIII, vol. 12496, SPIE, 2023.

[4] Lee, W. and Chen, L. “AI-guided optical-model-based superresolution for semiconductor CM metrology,” Metrology, Inspection, and Process Control XXXVIII, vol. 12496, SPIE, 2023.

[5] Kim, D., et al. "Interactive image annotation and AI-assisted segmentation of TEM images for automatic CD measurement," Metrology, Inspection, and Process Control XXXVIII, vol. 12955, SPIE, 2024.

[6] Phua, C. and Theng, L. B. "Semiconductor wafer surface: automatic defect classification with deep CNN," 2020 IEEE Region 10 Conference (TENCON), IEEE, 2020.

[7] Kim, E., et al. “Deep learning-based automatic defect classification for semiconductor manufacturing,” Metrology, Inspection, and Process Control XXXVIII, vol. 12496, SPIE, 2023.

[8] Xie, Q. “Towards data-efficient machine learning.” Ph.D. thesis, 2020

[9] Kirillov, A., et al. “Segment anything,” In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015-4026, 2023.

[10] Sohn, K., et al. "FixMatch: Simplifying semi-supervised learning with consistency and confidence," Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 596-608, 2020.

[11] He, K., et al. “Masked autoencoders are scalable vision learners,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16000-16009, 2022.

[12] Chen, T., et al. "A simple framework for contrastive learning of visual representations," International Conference on Machine Learning (ICML), PmLR, 2020.

[13] Lee, Y., et al. “Weakly supervised image segmentation for detecting defects from scanning electron microscopy images in semiconductor,” IEEE Access, vol.12, 2024.

[14] Cai, L., et al. “Exploring active learning for semiconductor defect segmentation,” IEEE International Conference on Image Processing (ICIP), 2022.

[15] Geng, H., et al. “When wafer failure pattern classification meets few-shot learning and self-supervised learning,” IEEE/ACM International Conference on Computer Aided Design (ICCAD), 2021.

[16] Kwak, M., Lee, Y., and Kim, S., “SWaCo: Safe wafer bin map classification with self-supervised contrastive learning,” IEEE Transactions on Semiconductor Manufacturing, vol. 36, no. 3, pp. 416–424, 2023.

[17] Hu, E. J., et al. "Lora: Low-rank adaptation of large language models," In Proceedings of the International Conference on Learning Representations (ICLR), 2022.

[18] Dosovitskiy, A., et al. "An image is worth the 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[19] Yang, L., et al. "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[20] Mishra, S., et al. "A simple, efficient and scalable contrastive masked autoencoder for learning visual representations," arXiv preprint arXiv:2210.16870, 2022.

[21] Jaccard, P., "The Distribution of the Flora in the Alpine Zone," New Phytologist, 1912.

[22] Chen, L. C., et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation," In Proceedings of the European conference on computer vision (ECCV), 2018.

[23] Yun, S., et al. "Cutmix: Regularization strategy to train strong classifiers with localizable features," In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[24] Liu, Z., et al. "Swin transformer: Hierarchical vision transformer using shifted windows," In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)2021.

[25] Deng, J., et al. "Imagenet: A large-scale hierarchical image database," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[26] Hu, J., et al. "Computational limits of low-rank adaptation (LoRA) for transformer-based models," arXiv preprint arXiv:2406.03136, 2024.