IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 May 2025, accepted 26 July 2025, date of publication 31 July 2025, date of current version 7 August 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3594581

RESEARCH ARTICLE

Robust Semantic Segmentation of Wafer Transmission Electron Microscopy Image Using Multi-Task Learning With Edge Detection

YONGWON JO[1], JINSOO BAE[1], HANSAM CHO[1], SUNGSU KIM[1], HEEJOONG ROH[2], KYUNGHYE KIM[2], MUNKI JO[2], MUNUK KIM[2], JAEUNG TAE[2], AND SEOUNG BUM KIM[1]

[1]Department of Industrial and Management Engineering, Korea University, Seongbuk-gu, Seoul 02841, Republic of Korea
[2]SK Hynix, Icheon-si 17336, Republic of Korea

Corresponding author: Seoung Bum Kim (sbkim1@korea.ac.kr)

This work was supported by SK Hynix Inc.

ABSTRACT Semantic segmentation for wafer transmission electron microscopy (TEM) images plays a crucial role in the automated measurement of semiconductors. However, the automated measurement of wafer TEM images presents three main challenges: difficulty in image acquisition, significant noise, and ambiguous object boundaries. While existing methods for automated measurement have used semantic segmentation algorithms, they often lead to inaccurate object boundary detection, resulting in over- or under-estimation. In this study, we propose a multi-task pre-training with masked autoencoder and virtual edge detection (MTP-MAVED) to improve object recognition and boundary detection in wafer TEM images with consideration of three challenges. Our approach includes three primary components: a pre-training phase using self-supervised representation learning to extract meaningful representations from unlabeled wafer TEM images, a multi-task learning-based fine-tuning phase incorporating both semantic segmentation and edge detection tasks, and a boundary-aware loss function to enhance boundary recognition accuracy. We demonstrate that MTP-MAVED outperforms existing methods in both object recognition and boundary detection, even with limited labeled data. This framework offers a more robust solution for addressing the complexities of wafer TEM image analysis and advances the field of automated semiconductor analysis.

INDEX TERMS Wafer transmission electron microscope image, self-supervised representation learning, multi-task learning, semantic segmentation, edge detection.

I. INTRODUCTION

Deep learning models have become essential for optimizing semiconductor manufacturing, a wide range of applications from equipment diagnostics [1], [2], [3] to defect prediction and classification on wafer bin maps (WBM) [4]. As semiconductor technology advances to smaller nanoscales, the demand for precise quality control at these scales has intensified [5]. To meet this demand, scanning electron microscopy (SEM) and transmission electron microscopy (TEM) images are used for high-resolution analysis of microscopic structures [5]. However, manual inspection of these images is not only time-consuming and costly but also prone to variation based on the skill of the inspector and the quality of the images [6]. Given the fine-scale nature of these structures, even small variances in inspection can lead to significant deviations in product quality. To overcome these limitations, there is an increasing need for automated structural analysis systems that minimize human intervention and ensure consistent, accurate interpretation of nanostructures during semiconductor production.

TEM images provide highly detailed visualizations by revealing internal cross-sections of semiconductor structures collected from processes that inherently require destructive testing [6]. The analysis of wafer TEM images poses three major challenges [7]. First, because TEM images are obtained

The associate editor coordinating the review of this manuscript and approving it for publication was Jeon Gwanggil.

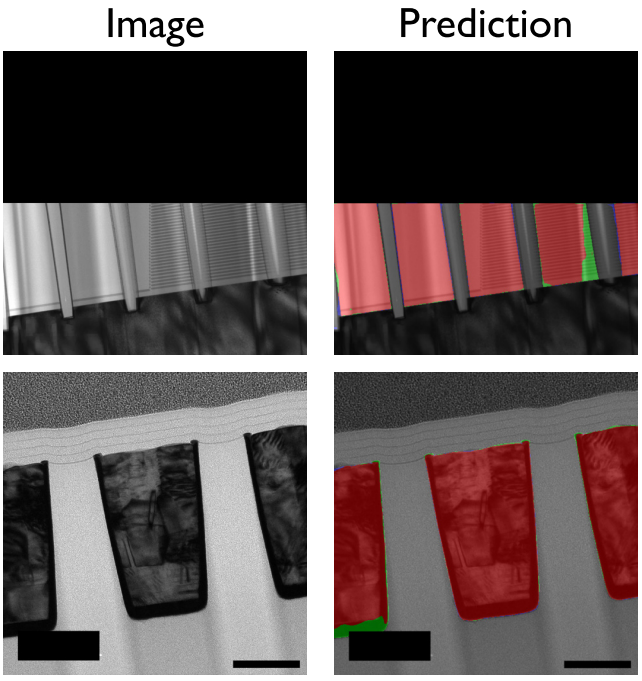

FIGURE 1. Prediction results of WTEM-SST. Red pixels represent correctly predicted pixels. Green pixels indicate under-estimated pixels where the object boundaries are predicted to be smaller than their actual size, while blue pixels show over-estimated areas where the boundaries are predicted to extend beyond their actual size.

through destructive testing that results in production loss, acquiring a large volume of wafer TEM images is inherently difficult, thereby complicating the annotation of ground truth (GT) labels. Second, wafer TEM images frequently exhibit substantial noise because of interactions between the electron beam and the material. Finally, the elevated noise levels further obscure object boundaries, making their precise delineation challenging. These issues called limited data availability, high noise levels, and ambiguous object boundaries, render the analysis of wafer TEM images particularly complex.

Despite these challenges, several image segmentation algorithms have been developed to identify objects of interest in wafer TEM images for automated measurement. Segmentation algorithms play an important role in this context because they precisely visualize objects by distinguishing their boundaries from surrounding noise and background artifacts. However, previous studies exhibit fundamental limitations in addressing the specific characteristics of wafer TEM images. For example, Baderot et al. used an instance segmentation algorithm to facilitate automated measurements by detecting objects of interest, but their method failed to adequately address the specific challenges inherent to wafer TEM images, including limited data availability, high noise levels and ambigous object boundaries [8]. To overcome these limitations, Jo et al. proposed the wafer TEM-specific semantic segmentation and transfer learning (WTEM-SST) framework, which was specifically designed to address the aforementioned three challenges of wafer TEM image analysis [7]. Although WTEM-SST improves the recognition of objects of interest with considering three challenges, it still struggles to accurately recognize object boundaries. As illustrated in Fig. 1, WTEM-SST frequently exhibits underestimation (green) or overestimation (blue) of boundaries. Underestimation refers to the incomplete detection of object edges, whereas overestimation occurs when predicted boundaries extend beyond the actual object. Furthermore, WTEM-SST heavily relies on GT annotations during pre-training, thereby limiting its ability to use the abundant volume of unlabeled wafer TEM images accumulated during production. This inherent dependency on labeled data ultimately constrains the model's ability to generalize across diverse manufacturing processes.

In this study, we propose a framework called multi-task pre-training with masked autoencoder and virtual edge detection (MTP-MAVED) to address the specific challenges inherent in wafer TEM image analysis, including limited data availability, high noise levels, and the difficulty in accurately identifying object boundaries. Unlike previous approaches that primarily rely on supervised learning with limited labeled wafer TEM images, MTP-MAVED adopts a fundamentally different strategy that uses self-supervised pre-training and boundary-aware optimization to overcome the limitations of existing methods. MTP-MAVED is designed to effectively recognize objects of interest and their boundaries through three key components: (1) a pre-training phase using unlabeled wafer TEM images from various manufacturing processes, (2) a supervised fine-tuning phase using a multi-task learning strategy that combines semantic segmentation and edge detection, and (3) the incorporation of a boundary-aware loss function to

The pre-training phase of MTP-MAVED is crucial for handling the limited availability of labeled data. During this phase, the model is trained using unlabeled wafer TEM images to reconstruct the original wafer TEM images and detect virtual edges that likely correspond to real object boundaries. This allows the model to learn meaningful representations of the wafer structures without reliance on GT annotations. Following this phase, the model is trained on a small number of labeled samples within a multi-task learning framework, where semantic segmentation and edge detection tasks are jointly optimized. We hypothesize that these tasks are closely aligned with the wafer TEM image reconstruction and the virtual edge detection (VED) established during pre-training, so ensuring continuity between pre-training and fine-tuning stages.

To further improve boundary recognition, we incorporate the active boundary loss (ABL) [9] function into the fine-tuning phase. The integration of multi-task learning with ABL enables the model to be jointly optimized for both semantic segmentation and edge detection tasks, thereby effectively addressing the principal challenges associated with wafer TEM image analysis. The contribution of our study can be summarized as follows:

- We propose the MTP-MAVED framework, specifically designed to address the key challenges in wafer TEM image analysis, including the limited data availability, high noise levels, and the difficulty of accurately identifying ambiguous object boundaries.
- We propose a self-supervised pre-training method that uses image reconstruction and VED tasks, enabling the model to learn meaningful representations from unlabeled wafer TEM images without relying on GT annotations.
- We propose a multi-task fine-tuning strategy that improves object and boundary recognition in wafer TEM images, even under high noise levels.
- We propose the incorporation of ABL to further improve edge detection peformances, enabling the model to more precisely capture nano-scale structural details critical for automated measurements.

II. RELATED WORK

A. SELF-SUPERVISED REPRESENTATION LEARNING

Self-supervised representation learning (SSRL) is a type of unsupervised learning that uses large amounts of unlabeled examples to extract intrinsic features without explicit label information [10]. The key idea behind SSRL is to pre-train neural networks to extract meaningful representations from the unlabeled data, which can then be fine-tuned for downstream tasks to improve overall performance. In SSRL, there are two primary approaches: pre-training only the encoder, which focuses on feature extraction from the input data, or pre-training both the encoder and decoder, where the decoder learns to reconstruct the input based on the encoded representations [10]. In methods that only pre-train the encoder, two data augmentation techniques are used to the single input simultaneously, encouraging the networks to learn intrinsic representations that are invariant to such transformations, resulting in contrastive learning approaches [11]. Non-contrastive methods, by contrast, use augmentations without considering relationships between different unlabeled data, aiming for similar feature extraction [12]. Other methods, called the information maximization method, seek to ensure that the elements in the feature making up the learned representations reflect diverse and meaningful features [13]. In addition, cluster-based learning approaches hypothesize that features are grouped into different clusters in the representation space [14]. Encoder-only approaches rely heavily on contrastive or non-contrastive learning with strong data augmentations to enforce invariance in the learned representations. However, these methods are not inherently optimized for dense prediction tasks such as semantic segmentation and edge detection, where spatial precision, boundary localization, and structural consistency are crucial [15], [16]. The lack of pixel-level supervision during pre-training limits their applicability to tasks that require fine-grained spatial understanding like dense prediction tasks. Consequently, these methods are suboptimal for wafer TEM image analysis, which requires detailed boundary recognition under high noise conditions.

In contrast, encoder-decoder based SSRL approaches are better aligned with the requirements of dense prediction tasks. Among them, the autoencoder (AE) is the most representative, where the model compresses the input into a lower-dimensional latent space and then reconstructs the input from that representation [17]. Variants such as the denoising autoencoder (DAE) [18] and context autoencoder (CAE) [19] enhance robustness by training models to reconstruct the original input from corrupted or masked versions, thereby encouraging the extraction of more generalizable features. A more recent and effective extension is the masked autoencoder (MAE) [15], where input images are divided into patches, and a subset of patches is masked during training. The model learns to reconstruct the masked patches, encouraging the encoder to extract semantically meaningful and spatially aware representations while training the decoder to reconstruct fine-grained structure. These encoder-decoder SSRL formulations are particularly suited to tasks involving lots of noise and missing information, both of which are common in wafer TEM images. In this study, we propose MTP-MAVED that combines MAE-based self-supervised learning with the VED task. Unlike encoder-only methods, our approach jointly trains the encoder and decoder to learn spatial and boundary-aware representations. The integration of VED further enhances boundary recognition, making it well-suited for our wafer TEM image segmentation task under high noise and limited annotation conditions.

B. BOUNDARY-AWARE SEMANTIC SEGMENTATION

Semantic segmentation, a core task in computer vision, assigns each pixel in an image to a specific class information [20]. Cross-entropy (CE) is commonly used as the loss function for this task. However, relying solely on CE often results in poor boundary recognition for objects of interest [21]. This problem occurs because CE focuses exclusively on pixel-level classification and does not account for the similarity between predicted and actual boundaries. To address this challenge, research has evolved in two directions: multi-task learning and the incorporation of boundary-aware loss functions [9]. Multi-task learning involves training models to perform additional and complementary tasks such as edge detection or depth estimation [22] along with semantic segmentation. The underlying hypothesis is that by learning to detect object edges or depth variations around boundaries, the model can generate more accurate segmentation predictions that align closely with real object boundaries [23]. By using shared representations between different tasks, multi-task learning enables models to generalize better and achieve higher accuracy, making it an effective approach for addressing complex problems in areas such as robotic grasping, autonomous driving, and human activity recognition [24], [25], [26].

Another approach is to introduce loss functions that explicitly account for the distance between the predicted

and actual boundaries [9]. These loss functions guide the model to minimize the difference between predicted and GT boundaries during training, hence integrating real object boundary information directly into the learning process [27]. Such boundary-aware loss functions can be easily combined with standard semantic segmentation loss functions, offering a flexible yet powerful way to improve boundary accuracy. In this study, we propose the integration of both multi-task learning and a boundary-aware loss function to enhance boundary detection in wafer TEM images, which are often characterized by noisy and ambiguous boundaries. Specifically, we define edge detection as an auxiliary task for multi-task learning. Additionally, we adopt ABL [9] function to refine the model's ability to recognize object boundaries, demonstrating superior performance over traditional semantic segmentation losses for wafer TEM images. By addressing these challenges of ambiguous object boundaries, MTP-MAVED significantly improves the performance of boundary recognition and semantic segmentation. This superiority is shown in the experiments in Section IV, where MTP-MAVED consistently outperforms existing methods in accurately identifying objects and their boundaries.

C. WAFER IMAGE ANALYSIS IN SEMICONDUCTOR MANUFACTURING

The analysis of images collected from semiconductor manufacturing processes has become a crucial research area, driven by the growing complexity of these processes and the high precision required for effective defect detection [28]. In this domain, several imaging modalities are frequently used, including WBM, SEM, and TEM. WBMs visualize the test results of chips on a wafer, primarily focusing on the automatic defect patterns classification. Notably, Kahng and Kim proposed an SSRL method designed to extract representations from WBMs, reducing the costs associated with annotating defect types [4]. In addition, Kwak et al. proposed an SSRL approach that improves classification performance by facilitating the recognition of both previously seen and unseen defect patterns [29]. Despite many studies leveraging machine learning techniques for WBMs, these analyses typically provide a broad overview of defects rather than in-depth insights into their characteristics. To obtain more granular information about wafer defects, SEM images are frequently used [30]. These high-resolution images, scanned and captured during the semiconductor manufacturing process, enable the identification of small-scale defects on the wafer surface. Research efforts in this area have primarily focused on the automatic classification of defect patterns from SEM images [30], with some studies also addressing the precise localization of defects [31]. However, while both WBM and SEM provide valuable information regarding wafers, they fall short in offering a comprehensive understanding of the internal structures of wafers.

TEM images, operating at a nanoscale resolution compared to SEM, facilitate in-depth investigations into the internal structures of semiconductor materials [32]. Although TEM provides more comprehensive structural information, its acquisition is more challenging because of the nature of the destructive testing process. In this context, Jo et al. proposed a wafer TEM images-specific semantic segmentation and transfer learning method, called WTEM-SST, aimed at recognizing objects of interest within wafer TEM images [7]. Their approach involved initially pre-training a semantic segmentation model using images collected from various manufacturing processes, along with pixel-level annotations to classify the source process of each object. To enhance the model's performance, they proposed wafer-specific data augmentation methods specifically tailored for wafer TEM images, including fill object (FO), fill background (FB), and object structure distortion (OSD). In addition, they introduced a loss function that assigns weights greater than one to object boundaries, aiming to improve the boundary recognition of interesting objects within wafer TEM images.

However, the use of pixel-level annotations during pre-training restricts the full usage of unlabeled images previously collected, and simply using larger weights to object boundaries does not guarantee accurate recognition of these objects and their boundaries. To address these limitations, the proposed method, MTP-MAVED, involves a two-stage approach based on SSRL: pre-training to extract representations from wafer TEM images using large amounts of unlabeled data, followed by fine-tuning through multi-task learning with ABL. In Section IV, we provide empirical evidence demonstrating the efficacy of MTP-MAVED in comparison to WTEM-SST.

III. PROPOSED METHOD

A. PRE-TRAINING WITH MASKED AUTOENCODER AND VIRTUAL EDGE DETECTION

Our proposed method, MTP-MAVED, involves a two-phase approach for training semantic segmentation models for each individual manufacturing process: (1) pre-training phase and (2) supervised multi-task fine-tuning phase. The overall process of MTP-MAVED, designed for process-specific semantic segmentation models, is illustrated in Fig 2. In the initial phase, the multi-task pre-training phase, a model with a shared encoder and dual decoders is trained to both reconstruct original images and detect virtual edge information. In the following supervised multi-task fine-tuning phase, the pre-trained model is fine-tuned to identify objects of interest within wafer TEM images for each individual manufacturing process. In addition, we use the ABL function to detect ambiguous edge regions in wafer TEM images more accurately. In this section, we provide comprehensive explanations of the MTP-MAVED training strategies, highlighting how our proposed method considers challenges in the analysis of wafer TEM images.

To address the challenge of data acquisition for wafer TEM images, we propose a SSRL approach, which uses images collected from various manufacturing processes. SSRL

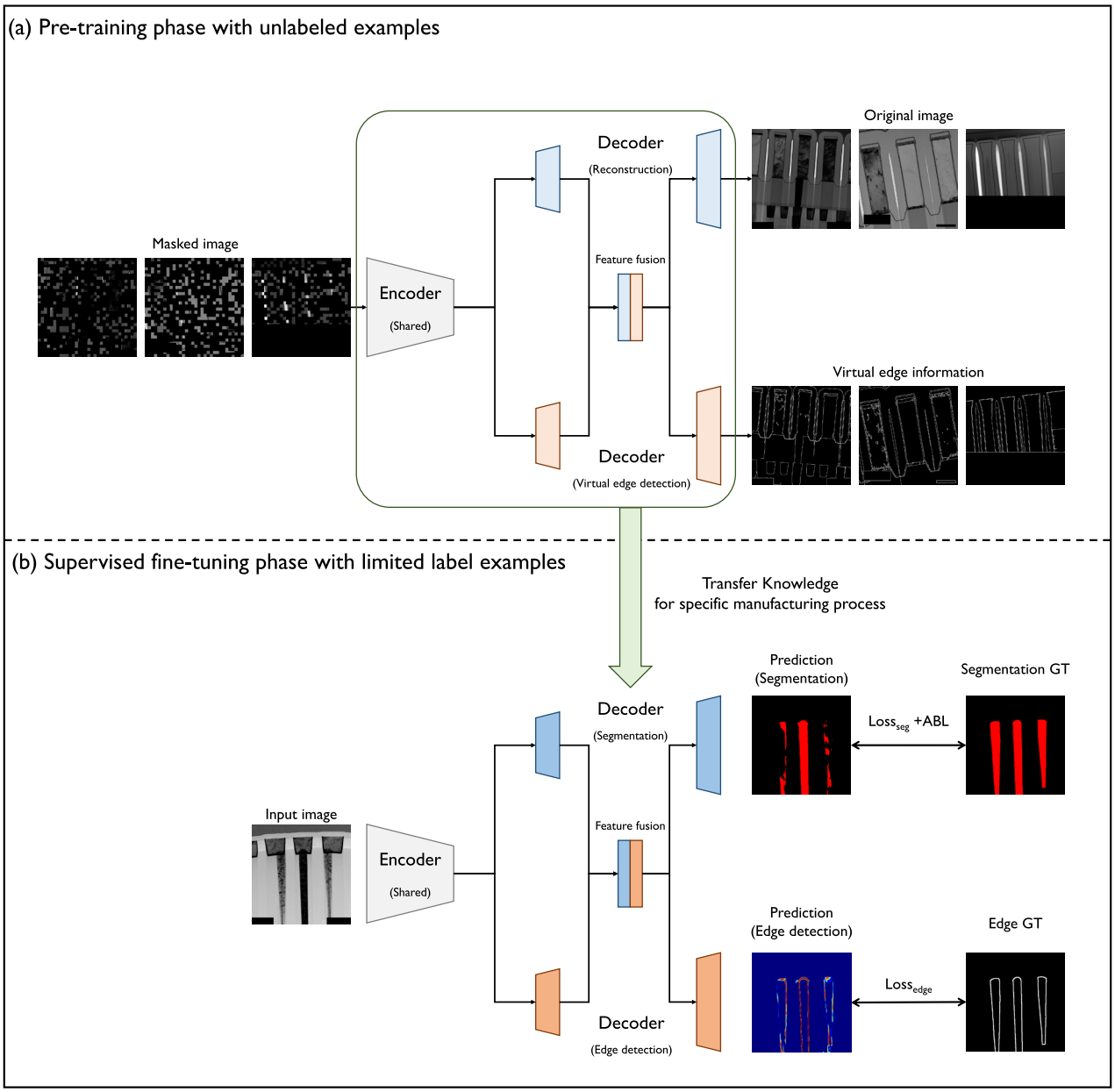

FIGURE 2. The overall process of MTP-MAVED. (a) The multi-task pre-training phase, the initial phase of MTP-MAVED. (b) The supervised multi-task fine-tuning phase using pre-trained weights and the ABL function.

comprises two stages: a pre-training phase that learns intrinsic representations from abundant unlabeled wafer TEM images, and a fine-tuning phase that uses the pre-trained weights as the initial parameters to train task-specific models [4]. In this study, we focus on pre-training with both the encoder and decoder, because our goal is to achieve well-trained process-specific semantic segmentation models. By using SSRL, we aim to improve performance while reducing the dependence on large amounts of labeled examples.

Our goal for the pre-training stage in SSRL is to train an encoder-decoder model that can learn intrinsic representations and reconstruct input wafer TEM images from various manufacturing processes. To achieve this goal, we minimize two loss functions: mean-squared error (MSE) loss for the image reconstruction task and cross-entropy (CE) loss for the VED inspired with [33]. Fig 2 (a) shows our multi-task pre-training with MAE and VED, called MTP-MAVED.

In the context of our semiconductor manufacturing processes, we formulate our MTP-MAVED approach as follows. Let X_i^p be an $i-th$ input wafer TEM image of size (H,W), collected from manufacturing process p $(p = 1,2,\ldots,P)$. We divide X_i^p into non-overlapping patches with size (k,k). Next, we randomly sample r % of the total patches without duplication and mask them by replacing their values with

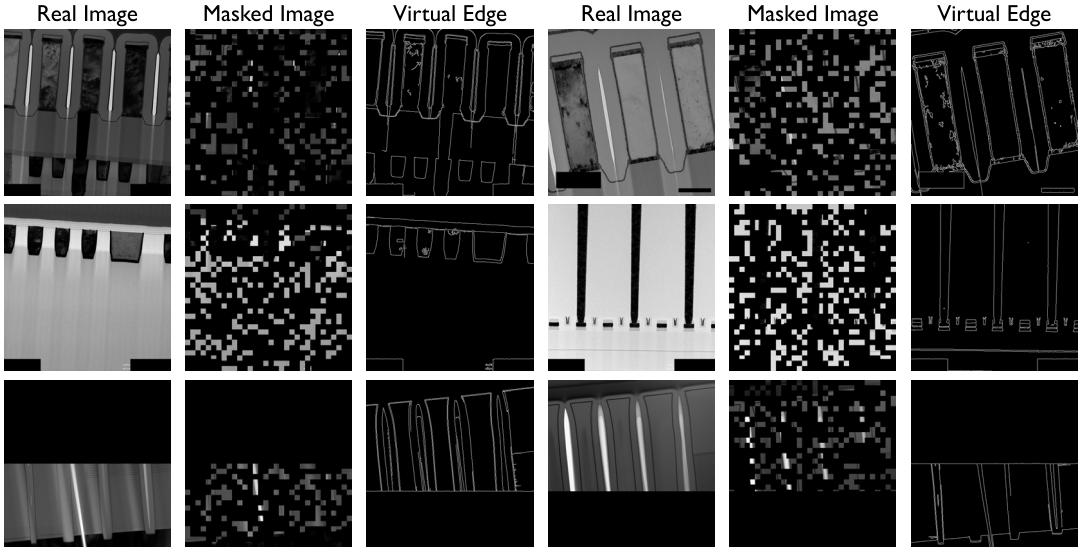

FIGURE 3. Examples of the real images, the masked images, and virtual edges generated by the canny algorithm.

zero [15]. Finally, masked patches are recombined into an image $\tilde{X}_i^p$ of size (H, W). Using wafer TEM images across overall manufacturing processes, we pre-train a single encoder and single decoder architecture, like [34], to reconstruct masked patches by minimizing the following MSE loss:

$$\mathcal{L}_{rec} = \frac{1}{B}\sum_{b=1}^{B}\sum_{m=1}^{M} \mathbb{1}(m) \cdot (X_{b,m} - \hat{X}_{b,m})^2, \tag{1}$$

where B represents the batch size, M is the number of pixels in the X_i^p, and m is the pixel index. $\hat{X}_{b,m}$ is the reconstructed pixel of X_i^p. The indicator function $\mathbb{1}(m)$ equals one if pixel m is masked, and zero otherwise.

We posit that training only the image reconstruction model is insufficient to learn representations for wafer TEM images. To refine the reconstruction model, we hypothesize that incorporating VED will enhance representation quality, making it more suitable for process-specific semantic segmentation. VED addresses the challenge of unclear object boundaries, often obscured by noise from the electron beam. To implement this additional task, we generate virtual binary edge information v_i^p using the canny edge detection method [35]. Fig 3 shows the wafer TEM images X_i^p, masked images $\tilde{X}_i^p$, and virtual edges v_i^p. For training the model on VED, we define the following loss function $\mathcal{L}_{ved}$:

$$\mathcal{L}_{ved} = \frac{1}{B}\sum_{b=1}^{B}\sum_{m=1}^{M} v_{b,m} \cdot log(\hat{v}_{b,m}), \tag{2}$$

where v_b is the b-th virtual edge information in the batch, $v_{b,m}$ is the true class label of pixel m in the v_b, and $\hat{v}_{b,m}$ is a predicted probability which belongs to the virtual edge at pixel m. To address both tasks simultaneously, we modify our single encoder and single decoder architecture to a single encoder paired with dual decoders. One decoder performs the image reconstruction task, and the other recognizes virtual edges. In addition, we fuse intermediate features in the dual decoders using convolutional neural networks (CNN) to enhance the performance of each individual task by allowing them to use complementary information [36]. These fused features are then fed into the remaining layers of each decoder. The combined loss function $\mathcal{L}_{pre-train}$ for training the single encoder-dual decoder model during pre-training is given by:

$$\mathcal{L}_{pre-train} = \mathcal{L}_{rec} + \lambda_{ved} \cdot \mathcal{L}_{ved}, \tag{3}$$

where λ_{ved} is a pre-defined hyperparameter controlling the influence of the VED loss.

B. SUPERVISED FINE-TUNING PHASE WITH MULTI-TASK LEARNING

In the supervised fine-tuning phase, we aim to train semantic segmentation models for various manufacturing processes to recognize interesting objects in wafer TEM images. We hypothesize that models trained using a multi-task learning approach, which includes both semantic segmentation and edge detection, will achieve higher performance than single-task models focused only on segmentation. This hypothesis is based on the strong relationship between the two tasks: segmentation requires accurate recognition of object boundaries, and edge detection explicitly identifies these boundaries.

For the multi-task training, we fully use pre-trained weights from MTP-MAVED. First, we modify the pre-trained reconstruction decoder by replacing its final layer with a CNN layer followed by a softmax activation function to solve the semantic segmentation task. To incorporate edge detection into the training process, we generate GT

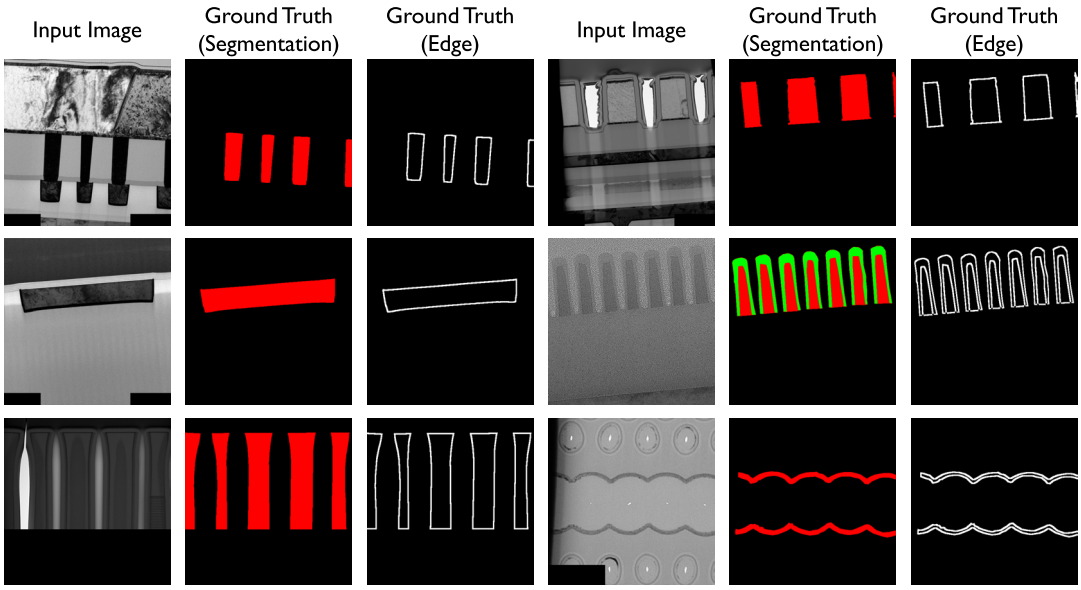

FIGURE 4. The examples of our wafer TEM images, and their semantic segmentation GT, and edge detection GT generated by the morphological edge detection.

information for the edge detection task based on the existing GT for semantic segmentation. We use the morphological edge detection algorithm [37] to produce object boundaries from the segmentation GT. However, this algorithm generates boundary regions both inside and outside the objects. To refine the boundaries and ensure its accuracy, we extract regions that intersect to both the segmentation GT and the generated object boundaries.

This generation process is formulated as follows: Let y_i^p be a semantic segmentation GT of for the input image X_i^p. Then, we use the morphological algorithm to generate the boundary information e_i^p. Finally, we update $\tilde{e}_i^p$ using the intersection operation $\tilde{e}_i^p = e_i^p \odot y_i^p$, where $\tilde{e}_i^p$ represents the refined object boundaries, and $\odot$ is the element-wise product. We define $\tilde{e}_i^p$ as the GT for the edge detection task. Fig 4 shows our wafer TEM images X_i^p, their semantic segmentation GTs y_i^p, and their edge detection GTs $\tilde{e}_i^p$, generated by the above operations.

To define an appropriate loss function for multi-task learning, we incorporate the commonly used CE loss and intersection over union (IoU) loss for semantic segmentation tasks. However, we hypothesize that these functions alone may not be sufficient to accurately identify interesting objects and their boundaries, particularly in the presence of heavy noise. To address this issue, we propose the integration of an additional loss function known as ABL [9], which is specifically designed to improve boundary recognition. In our approach, we use CE and IoU loss in combination with ABL. The semantic segmentation loss $\mathcal{L}_{seg}$ is expressed as follows:

$$\mathcal{L}_{seg} = \frac{1}{B}\sum_{b=1}^{B}\sum_{m=1}^{M} CE(\hat{y}_{b,m}, y_{b,m}) + IoU(\hat{y}_{b,m}, y_{b,m}) + \lambda_{ABL} \cdot ABL(\hat{y}_{b,m}, y_{b,m}), \quad (4)$$

where y_b represents the $b-th$ GT of the segmentation, $y_{b,m}$ represents the true class label of pixel m in the y_b, $\hat{y}_{b,m}$ is its predicted probability, and λ_{ABL} is a hyperparameter controlling the influence of the ABL. For the edge detection decoder, we train the decoder using the following CE loss function:

$$\mathcal{L}_{edge} = \frac{1}{B}\sum_{b=1}^{B}\sum_{m=1}^{M} CE(\hat{e}_{b,m}, e_{b,m}), \quad (5)$$

where e_b is the $b-th$ GT of the edge detection, $e_{b,m}$ is the real class label of pixel m in the e_b, $\hat{e}_{b,m}$ is its predicted probability. The combined loss function $\mathcal{L}_{fine-tune}$ for training the multi-task model during fine-tuning is given by:

$$\mathcal{L}_{fine-tune} = \mathcal{L}_{seg} + \lambda_{edge} \cdot \mathcal{L}_{edge} \quad (6)$$

where λ_{edge} is a pre-defined hyperparameter that controls the contribution of the edge detection loss during fine-tuning.

IV. EXPERIMENTS

A. DATASET

To assess the effectiveness of our proposed method, MTP-MAVED, we experimented with nine manufacturing process datasets. Wafer TEM images of these datasets were collected from manufacturing processes at SK hynix Inc., a prominent semiconductor manufacturing company in South Korea. Domain experts annotated semantic segmentation GTs of these images. Table 1 provides a summary of the number of wafer TEM images and the number of corresponding interesting objects across each manufacturing dataset. All wafer TEM images in these datasets are uniformly sized at (H = 2,048, W = 2,048). For pre-training and fine-tuning

models, we resized these images to ($H = 512$, $W = 512$), which were used as our input images. During the evaluation of MTP-MAVED, we upsampled the outputs back to the original resolution with ($H = 2{,}048$, $W = 2{,}048$).

TABLE 1. Summary of collected wafer TEM images.

Process	The number of object types in interest	The number of wafer TEM images
A	2(Object and Background)	121
B	2(Object and Background)	108
C	2(Object and Background)	155
D	2(Object and Background)	134
E	2(Object and Background)	149
F	3(Object1, Object 2, and Background)	63
G	2(Object and Background)	28
H	2(Object and Background)	29
I	2(Object and Background)	36

Each dataset was split into training (60%), validation (20%), and test (20%) sets. In the pre-training phase, we trained MTP-MAVED using only wafer TEM images from the training and validation sets across all manufacturing processes. In the supervised fine-tuning phase, we trained a process-specific semantic segmentation model using only the corresponding manufacturing dataset and reported the performance of the model on its respective test dataset.

B. IMPLEMENTATION DETAILS

We used SegFormerB3 [34] as the base model for assessing our MTP-MAVED, initializing it with pre-trained weights from the Cityscape dataset [38], [39]. This selection was motivated by SegFormer's patch-based processing, which inherently aligns with the masking strategy used in MAE. By using SegFormer's capability to handle global and local features simultaneously, we aimed to create a robust pre-trained model for integrating advanced MAE-based learning. During pre-training, we trained MTP-MAVED for 200 epochs with a batch size of two. In addition, we optimized our model with the AdamW optimizer [40], with a learning rate to 0.00006 and the weight decay factor of 0.01. We applied data augmentation methods, including vertical and horizontal flips, safe rotation, and color jitter, as described in [41]. Following the experimental settings in [15], which originally proposed MAE, the patch size k and the percent r of masking patches were set to 16 and 75, respectively. The weighting hyperparameter λ_{ved} between $\mathcal{L}_{rec}$ and $\mathcal{L}_{ved}$ was fixed at one. The best pre-trained model was selected based on the loss $\mathcal{L}_{pre-train}$ of validation datasets.

In the supervised fine-tuning phase, we fully used pre-trained MTP-MAVED to simultaneously address both segmentation and edge detection tasks. We replaced the final layer of the reconstruction decoder of the MTP-MAVED model with a CNN layer, where the number of output channels corresponded to the number of interesting objects for each specific manufacturing process. The VED decoder remained unchanged because its output channels for virtual and edge detection were fixed at two. Process-specific multi-task models were trained for 200 epochs with a batch size of two, using the same optimization hyperparameters as in the pre-training phase. In addition to the pre-training data augmentations, we applied wafer TEM-specific augmentations, including FO, FB, and OSD as described in [7]. We selected these augmentations because they were robust in noisy environments, and were specifically designed to address the unique challenges of wafer TEM image analysis, including the difficulty of collecting labeled examples, prevalent noise, and ambiguous object boundaries. The weighting hyperparameter λ_{edge} between $\mathcal{L}_{seg}$ and $\mathcal{L}_{edge}$ was fixed at one. The model with the lowest validation loss $\mathcal{L}_{fine-tune}$ was selected, and evaluation metrics were reported for the test dataset of each manufacturing process. All implementations were conducted using Python 3.10.8, Pytorch 1.13.1, and Transformers 4.41.1. The experiments were performed on a workstation equipped with an Intel i9-10900KF CPU@3.70GHz, 64GB DDR4 RAM 3,200MHz, and NVIDIA RTX 3090 GPU.

C. RESULTS

We evaluated the proposed method MTP-MAVED using three metrics: mean intersection over union (MIoU), optimal image scale (OIS), and optimal dataset scale (ODS). MIoU is widely used for the evaluation metric for semantic segmentation tasks because it measures the alignment between segmentation GTs and predicted segmentation results. MIoU is defined as follows:

$$MIoU = \frac{1}{C}\sum_{c=1}^{C}\frac{n_{cc}}{n_{c.} + n_{.c} + n_{cc}}, \tag{7}$$

where C represents the number of classes in each dataset, n_{cc} is the number of correctly predicted pixels, $n_{c.}$ is the number of pixels belonging to class c in GT, and $n_{.c}$ is the number of predicted pixels to be in class c.

OIS and ODS are widely used evaluation metrics for edge detection tasks [42]. These metrics measure the accuracy of predicted edges in comparison to GT edge detection using different aggregation methods. Because our goal was to evaluate how accurately MTP-MAVED recognizes edges compared to traditional semantic segmentation models, we derived edge predictions from the semantic segmentation outputs. Specifically, we converted the predicted probability $\hat{y}_i^p$ into a segmentation map $\bar{y}_i^p$ and extracted predicted edge regions $\bar{e}_i^p$ using the method in [37]. $\bar{e}_i^p$ was updated by performing an intersection operation between the predicted segmentation map $\bar{y}_i^p$ and the predicted edge $\bar{e}_i^p$. Finally, we obtained the predicted edge probability $\hat{e}_i^p$ by calculating the intersection between $\hat{y}_i^p$ and the updated $\bar{e}_i^p$. OIS is then calculated as follows:

$$OIS = \frac{1}{N}\sum_{i=1}^{N}\max_{\theta_i} F_1(\hat{e}_i^p, \tilde{e}_i^p|\theta_i), \tag{8}$$

where N represents the number of images in each dataset, and θ_i is the probability threshold for the $i-th$ image in the dataset. $F_1(\hat{e}_i^p, \tilde{e}_i^p|\theta_i)$ is the F1 score between $\hat{e}_i^p$ and e_i^p

TABLE 2. Average MIoU, OIS, and ODS along with standard deviations (in parentheses) from five repeated repetitions to compare the MTP-GT with our MTP-MAVED. The best score for each scenario and manufacturing process is highlighted in bold.

Metric: MIoU											
The ratio of training dataset	Method	A	B	C	D	E	F	G	H	I	Average
10%	MTP-GT	98.58 (0.49)	95.04 (0.80)	96.82 (0.43)	95.66 (1.54)	**97.89 (0.37)**	85.55 (0.17)	**92.46 (1.15)**	94.83 (1.44)	93.46 (0.50)	94.48 (3.79)
	MTP-MAVED (Proposed)	**98.59 (0.57)**	**95.73 (0.81)**	**97.97 (0.48)**	**97.61 (0.14)**	**97.89 (0.28)**	**85.83 (0.29)**	91.02 (1.43)	**95.31 (0.77)**	**93.53 (0.42)**	**94.83 (4.03)**
25%	MTP-GT	98.98 (0.13)	**97.85 (0.23)**	97.99 (0.48)	96.46 (1.54)	**98.81 (0.31)**	**87.91 (0.22)**	**95.08 (0.27)**	**97.67 (0.45)**	95.08 (0.17)	96.20 (3.31)
	MTP-MAVED (Proposed)	**99.12 (0.06)**	97.64 (0.21)	**98.77 (0.10)**	**97.96 (0.15)**	98.53 (0.36)	87.81 (0.16)	93.77 (0.49)	**97.67 (0.42)**	**95.17 (0.11)**	**96.27 (3.47)**
100%	MTP-GT	**99.38 (0.04)**	98.04 (0.14)	**99.26 (0.02)**	97.74 (1.44)	**99.34 (0.05)**	89.53 (0.04)	**98.11 (0.8)**	**99.16 (0.18)**	95.62 (0.07)	**97.35 (3.06)**
	MTP-MAVED (Proposed)	99.36 (0.02)	**98.32 (0.08)**	**99.26 (0.04)**	**98.45 (0.08)**	99.32 (0.05)	**89.54 (0.10)**	97.21 (0.91)	99.02 (0.03)	**95.65 (0.11)**	**97.35 (3.04)**
Metric: OIS											
The ratio of training dataset	Method	A	B	C	D	E	F	G	H	I	Average
10%	MTP-GT	**90.25 (1.10)**	74.85 (0.75)	88.93 (1.80)	79.87 (1.00)	84.99 (2.62)	78.80 (0.49)	66.14 (3.88)	80.13 (2.67)	86.89 (1.04)	81.21 (7.49)
	MTP-MAVED (Proposed)	89.49 (2.43)	**75.08 (2.29)**	**89.83 (0.45)**	**80.53 (0.69)**	**87.01 (1.81)**	**79.50 (0.53)**	**67.20 (1.84)**	**83.38 (1.28)**	**87.42 (0.73)**	**82.16 (7.27)**
25%	MTP-GT	92.04 (0.83)	**83.94 (0.46)**	91.83 (0.68)	**82.89 (0.68)**	89.96 (1.09)	**82.93 (0.24)**	**82.60 (1.69)**	**88.20 (0.84)**	90.13 (0.30)	**87.17 (3.93)**
	MTP-MAVED (Proposed)	**92.48 (0.48)**	83.34 (0.69)	**91.94 (0.20)**	81.84 (0.98)	**91.20 (0.49)**	82.60 (0.29)	79.33 (1.30)	87.26 (1.19)	**90.36 (0.20)**	86.71 (4.83)
100%	MTP-GT	**95.04 (0.11)**	**86.58 (0.42)**	**94.36 (0.17)**	84.08 (0.47)	**93.29 (0.40)**	**85.64 (0.16)**	**90.19 (1.72)**	**94.62 (0.56)**	91.05 (0.17)	**90.54 (4.04)**
	MTP-MAVED (Proposed)	94.85 (0.21)	**86.58 (0.24)**	94.24 (0.16)	**84.74 (0.95)**	92.97 (0.44)	85.47 (0.17)	88.41 (0.95)	93.29 (0.26)	**91.23 (0.20)**	90.20 (3.79)
Metric: ODS											
The ratio of training dataset	Method	A	B	C	D	E	F	G	H	I	Average
10%	MTP-GT	**89.90 (1.28)**	74.05 (0.56)	88.32 (1.85)	79.44 (1.00)	84.70 (2.70)	78.56 (0.51)	65.40 (3.57)	79.69 (2.76)	86.73 (1.05)	80.76 (7.60)
	MTP-MAVED (Proposed)	89.20 (2.33)	**74.50 (2.51)**	**89.14 (0.56)**	**80.08 (0.75)**	**86.55 (2.06)**	**79.17 (0.60)**	**66.48 (1.83)**	**83.17 (1.33)**	**87.02 (0.81)**	**81.70 (7.36)**
25%	MTP-GT	91.78 (0.89)	**83.53 (0.72)**	90.91 (0.83)	**82.46 (0.62)**	89.63 (1.02)	**82.70 (0.27)**	**81.97 (1.47)**	**88.12 (0.88)**	90.09 (0.29)	**86.80 (3.95)**
	MTP-MAVED (Proposed)	**92.26 (0.54)**	82.99 (0.82)	**91.59 (0.17)**	81.32 (1.12)	**90.89 (0.52)**	82.39 (0.30)	78.63 (1.26)	87.10 (1.28)	**90.31 (0.21)**	86.39 (4.95)
100%	MTP-GT	**94.92 (0.10)**	86.37 (0.42)	**94.09 (0.22)**	83.81 (0.53)	**93.05 (0.48)**	**85.55 (0.16)**	**89.77 (1.76)**	**94.60 (0.56)**	90.99 (0.18)	**90.35 (4.07)**
	MTP-MAVED (Proposed)	94.76 (0.19)	**86.43 (0.23)**	94.02 (0.18)	**84.50 (1.06)**	92.77 (0.47)	85.37 (0.18)	88.13 (0.93)	93.26 (0.25)	**91.17 (0.22)**	90.05 (3.82)

at threshold θ_i. OIS is the average of the highest F1 score for each image. ODS is defined as:

$$ODS = \max_{\theta} \sum_{i=1}^{N} F_1(\hat{e}_i^p, \tilde{e}_i^p|\theta), \tag{9}$$

where θ is a global threshold used uniformly to all images in the dataset. ODS captures the highest F1 score across the entire dataset. Higher OIS and ODS values indicate better alignment between predicted and true edges.

First, we evaluated the performance of MTP-MAVED in comparison with a pre-training method based on labeled examples under conditions where the availability of training data was extremely limited. To this end, we defined three experimental protocols, in which the supervised fine-tuning phase was conducted using only 10%, 25%, and 100% of the available training examples for each manufacturing process. This design allowed us to systematically assess the effectiveness of different pre-training strategies in data-scarce scenarios. For each process, we conducted five experimental trials and reported the mean and standard deviation of the MIoU, OIS, and ODS metrics on the nine test datasets. In addition, we investigated the mean and standard deviation of the 45 performances across experiments and reported these values in the "Average" column. We summarized the results in Table 2, which presents a performance comparison between the methods with and without GT information.

A comparative method, termed MTP-GT, represents a multi-task pre-training approach that excludes the masked patch generation step, directly performing semantic segmentation and edge detection during pre-training. MTP-GT uses

TABLE 3. Average MIoU, OIS, and ODS along with standard deviations (in parentheses) from five repeated trials for the ablation study. The best score for each scenario and manufacturing process is highlighted in bold.

Metric	Method	A	B	C	D	E	F	G	H	I	Average
MIoU	MTP-MAVED w/o MTP	98.38 (0.40)	94.30 (1.48)	96.74 (0.44)	96.05 (0.28)	97.97 (0.14)	85.44 (0.50)	**93.36 (1.08)**	93.52 (0.53)	92.96 (0.91)	94.30 (3.77)
	MTP-MAVED w/o MTL	98.06 (0.46)	94.97 (1.10)	97.03 (1.05)	97.03 (0.21)	97.48 (0.26)	85.20 (0.50)	87.72 (0.24)	93.64 (1.32)	92.73 (0.50)	93.76 (4.39)
	MTP-MAVED w/o ABL	97.43 (1.50)	95.06 (0.63)	96.98 (1.11)	97.36 (0.44)	97.62 (0.32)	85.30 (0.59)	90.12 (0.48)	94.47 (0.27)	92.71 (1.76)	94.02 (4.08)
	MTP-MAVED (Proposed)	**98.59 (0.57)**	**95.73 (0.81)**	**97.97 (0.48)**	**97.61 (0.14)**	**97.89 (0.28)**	**85.83 (0.29)**	91.02 (1.43)	**95.31 (0.77)**	**93.53 (0.42)**	**94.83 (4.03)**
OIS	MTP-MAVED w/o MTP	88.34 (1.56)	73.05 (2.85)	89.47 (0.87)	79.86 (1.02)	86.02 (1.13)	78.56 (0.47)	**70.57 (4.76)**	79.02 (1.36)	86.13 (1.33)	81.22 (6.67)
	MTP-MAVED w/o MTL	87.74 (1.31)	70.16 (3.24)	88.27 (1.70)	77.85 (1.02)	86.55 (1.32)	78.23 (0.72)	55.43 (1.95)	74.97 (1.28)	85.87 (1.35)	78.34 (10.31)
	MTP-MAVED w/o ABL	86.85 (3.38)	73.41 (2.21)	87.84 (1.94)	79.96 (1.09)	85.76 (0.75)	78.65 (0.92)	64.47 (2.00)	82.76 (2.20)	85.50 (3.18)	80.43 (7.57)
	MTP-MAVED (Proposed)	**89.49 (2.43)**	**75.08 (2.29)**	**89.83 (0.45)**	**80.53 (0.69)**	**87.01 (1.81)**	**79.50 (0.53)**	67.20 (1.84)	**83.38 (1.28)**	**87.42 (0.73)**	**82.16 (7.27)**
ODS	MTP-MAVED w/o MTP	**87.81 (1.77)**	72.63 (3.16)	88.85 (1.11)	79.51 (1.11)	85.84 (1.14)	78.35 (0.45)	**69.76 (5.03)**	78.69 (1.30)	85.65 (2.09)	80.79 (6.76)
	MTP-MAVED w/o MTL	87.43 (1.12)	69.83 (3.52)	87.75 (1.65)	77.33 (0.95)	86.15 (1.35)	77.90 (0.74)	54.67 (1.94)	74.74 (1.24)	85.42 (1.37)	77.91 (10.39)
	MTP-MAVED w/o ABL	86.15 (3.61)	72.81 (1.89)	87.09 (1.81)	79.52 (1.16)	85.72 (0.74)	78.32 (0.95)	63.67 (2.11)	82.50 (2.35)	85.36 (3.22)	79.97 (7.68)
	MTP-MAVED (Proposed)	**89.20 (2.33)**	**74.50 (2.51)**	**89.14 (0.56)**	**80.08 (0.75)**	**86.55 (2.06)**	**79.17 (0.60)**	66.48 (1.83)	**83.17 (1.33)**	**87.02 (0.81)**	**81.70 (7.36)**

TABLE 4. Average MIoU, OIS, and ODS along with standard deviations (in parentheses) from five repeated trials for experiments to assess the suitability of MAE for our reconstruction task. The best score for each scenario and manufacturing process is highlighted in bold.

Metric	Reconstruction	A	B	C	D	E	F	G	H	I	Average
MIoU	AE	98.22 (0.60)	90.14 (1.94)	94.69 (0.12)	94.96 (1.18)	97.67 (0.07)	84.67 (0.29)	88.96 (0.99)	93.69 (1.09)	**93.55 (0.52)**	92.95 (4.22)
	DAE	97.80 (0.76)	94.37 (0.77)	95.08 (0.43)	96.50 (0.30)	97.61 (0.30)	85.19 (0.55)	89.18 (0.88)	92.81 (0.62)	93.05 (0.81)	93.51 (3.99)
	CAE	98.12 (0.76)	93.37 (0.99)	95.38 (1.22)	97.82 (0.56)	97.82 (0.56)	84.32 (0.36)	88.13 (1.29)	93.23 (0.37)	92.65 (0.49)	93.31 (4.42)
	MTP-MAVED (Proposed)	**98.59 (0.57)**	**95.73 (0.81)**	**97.97 (0.48)**	**97.61 (0.14)**	**97.89 (0.28)**	**85.83 (0.29)**	**91.02 (1.43)**	**95.31 (0.77)**	93.53 (0.42)	**94.83 (4.03)**
OIS	AE	88.50 (2.09)	66.66 (1.96)	87.50 (0.67)	79.74 (0.23)	86.43 (1.18)	77.91 (0.63)	55.66 (5.02)	77.53 (2.20)	**87.60 (0.89)**	78.51 (10.71)
	DAE	88.50 (2.56)	70.68 (2.09)	87.50 (0.67)	78.91 (1.32)	**87.60 (0.72)**	78.52 (0.77)	55.59 (2.86)	76.22 (2.68)	86.65 (1.09)	78.91 (10.36)
	CAE	88.66 (1.02)	69.29 (1.58)	87.09 (1.01)	78.14 (1.24)	86.68 (1.64)	76.71 (0.78)	60.42 (2.81)	77.62 (1.40)	85.94 (0.98)	78.95 (9.11)
	MTP-MAVED (Proposed)	**89.49 (2.43)**	**75.08 (2.29)**	**89.83 (0.45)**	**80.53 (0.69)**	87.01 (1.81)	**79.50 (0.53)**	**67.20 (1.84)**	**83.38 (1.28)**	87.42 (0.73)	**82.16 (7.27)**
ODS	AE	87.53 (1.79)	65.66 (2.43)	86.17 (0.90)	79.30 (0.18)	86.08 (1.22)	77.64 (0.66)	54.58 (4.70)	77.20 (2.18)	**87.18 (0.96)**	77.93 (10.88)
	DAE	87.81 (2.51)	69.67 (1.97)	86.96 (0.65)	78.11 (1.58)	**87.33 (0.77)**	78.28 (0.76)	54.88 (2.84)	75.86 (2.81)	86.26 (1.32)	78.35 (10.47)
	CAE	87.83 (1.48)	68.52 (1.71)	86.43 (1.06)	77.64 (1.40)	86.24 (1.76)	76.38 (0.73)	59.29 (3.07)	77.48 (1.39)	85.40 (0.86)	78.36 (9.24)
	MTP-MAVED (Proposed)	**89.20 (2.33)**	**74.50 (2.51)**	**89.14 (0.56)**	**80.08 (0.75)**	86.55 (2.06)	**79.17 (0.60)**	**66.48 (1.83)**	**83.17 (1.33)**	87.02 (0.81)	**81.70 (7.36)**

the same deep neural network architecture as MTP-MAVED but is trained only with 10% and 25% pairs of wafer TEM images and GTs for the 10% and 25% scenarios, respectively. The experimental results revealed that MTP-GT and MTP-MAVED achieve comparable MIoU scores across all scenarios. In terms of OIS and ODS, MTP-GT outperformed MTP-MAVED when 25% or more of the training examples are available. However, in the most data-constrained setting (10% of labeled data), MTP-MAVED showed superior performance, achieving approximately a 1% improvement in OIS and ODS compared to MTP-GT. These findings suggest that while label-based pre-training methods such as MTP-GT are advantageous when a moderate amount of labeled data is available, MTP-MAVED offers greater robustness and

TABLE 5. Average MIoU, OIS, and ODS along with standard deviations (in parentheses) from five repeated trials for experiments to measure the effectiveness of VED. The best score for each scenario and manufacturing process is highlighted in bold.

Metric	Method	A	B	C	D	E	F	G	H	I	Average
MIoU	MTP-MAVED w/o VED	98.15 (1.01)	95.79 (0.73)	95.81 (1.20)	97.01 (0.31)	97.47 (0.37)	85.12 (0.26)	88.24 (0.77)	93.25 (0.77)	92.31 (1.25)	93.68 (4.32)
	MTP-MAVED (Proposed)	**98.59 (0.57)**	**95.73 (0.81)**	**97.97 (0.48)**	**97.61 (0.14)**	**97.89 (0.28)**	**85.83 (0.29)**	**91.02 (1.43)**	**95.31 (0.77)**	**93.53 (0.42)**	**94.83 (4.03)**
OIS	MTP-MAVED w/o VED	89.17 (1.21)	74.29 (1.70)	88.57 (1.27)	77.51 (1.17)	86.54 (1.83)	78.67 (0.46)	60.78 (1.65)	78.95 (1.40)	85.40 (2.27)	79.99 (8.64)
	MTP-MAVED (Proposed)	**89.49 (2.43)**	**75.08 (2.29)**	**89.83 (0.45)**	**80.53 (0.69)**	**87.01 (1.81)**	**79.50 (0.53)**	**67.20 (1.84)**	**83.38 (1.28)**	**87.42 (0.73)**	**82.16 (7.27)**
ODS	MTP-MAVED w/o VED	88.67 (1.36)	73.92 (1.61)	87.60 (1.49)	76.81 (1.24)	86.14 (1.85)	78.41 (0.50)	59.85 (1.89)	78.52 (1.50)	85.12 (2.24)	79.45 (8.72)
	MTP-MAVED (Proposed)	**89.20 (2.33)**	**74.50 (2.51)**	**89.14 (0.56)**	**80.08 (0.75)**	**86.55 (2.06)**	**79.17 (0.60)**	**66.48 (1.83)**	**83.17 (1.33)**	**87.02 (0.81)**	**81.70 (7.36)**

effectiveness in highly data-scarce environments. To highlight the effectiveness of our proposed framework in highly data-scarce scenarios, all experiments reported below were conducted using only 10% of the available labeled training data.

Our proposed method consists of three key components: the pre-training with the MAE and VED (MTP), supervised fine-tuning with multi-task learning (MTL), and the incorporation of the ABL function. To evaluate the contribution of each component to the overall performance, we conducted an ablation study, summarized in Table 3. In the 1^{st} row (MTP-MAVED w/o MTP), we initialized model weights using pre-trained weights from the Cityscape dataset, with separate decoders for segmentation and edge detection initialized by replicating these pre-trained decoder weights. In the 2^{nd} row (MTP-MAVED w/o MTL), we fine-tuned the segmentation model with a single encoder and single decoder by initializing it with pre-trained weights of the encoder and reconstruction decoder. In the 3^{rd} row (MTP-MAVED w/o ABL), we defined the segmentation loss function using CE instead of ABL.

Our results demonstrated that the MTP-MAVED, which includes all three components, achieves the highest performance across all metrics. Analyzing the performance gains of each component, we observed that MTL (-3.79%) contributes the most, followed by ABL (-1.73%) and MTP (-0.91%) in terms of ODS. In especially, comparing the 2^{nd} and 4^{th} rows, we found that MTL enhances OIS and ODS metrics of object boundary recognition compared to only segmentation models. This supports our hypothesis regarding the effectiveness of multi-task learning. Furthermore, a comparison between the 3^{rd} and 4^{th} rows indicates that incorporating ABL into the traditional segmentation loss further improves boundary recognition accuracy. Finally, the results confirmed that pre-training with MTP, combining MAE and VED, improves performance by learning weights tailored to wafer TEM images, as opposed to using pre-trained weights from Cityscape only.

To assess the suitability of the MAE-based reconstruction in our approach, we conducted experiments by replacing MAE with different autoencoder methods for reconstruction during the pre-training phase. Specifically, we considered three methods, AE [17], DAE [18], and CAE [19], which are commonly used for extracting representations from unlabeled examples and reconstruction tasks. For a fair comparison, we integrated VED with each of these methods instead of using MAE approach. The results, presented in Table 4, indicated that the MTP-MAVED method pre-trained with MAE reconstruction task achieves the best performance.

Comparing 1^{st} row (AE) with the other rows, we observed that using distortion to the input images during the pre-training leads to significant performance improvements. In addition, when comparing 2^{nd} row (DAE) with 3^{rd} and 4^{th} rows, the results indicated that masking regions of the input image is a more effective form of distortion for learning representations of wafer TEM images than adding noise. Moreover, we observed that reconstructing small and masked regions scattered across the image (MAE) proves more beneficial for representation learning than restoring a single, larger masked region (CAE).

These findings highlight the advantages of MAE-based pre-training in improving the model's ability to extract meaningful representations from complex semiconductor images. This improvement likely arises because training MTP-MAVED to reconstruct masked patches enables it to accurately reconstruct images despite prevalent noise, while also effectively recognizing virtual edges.

We conducted an experiment to evaluate the effectiveness of VED on our model's performance. For a fair comparison, we pre-trained a version of the model initialized with weights from the Cityscape dataset [38], using only the image reconstruction task based on MAE (MTP-MAVED without VED). Table 5 presents the performance comparison between MTP-MAVED and MTP-MAVED without VED. The results show that MTP-MAVED consistently outperforms across all manufacturing processes, regardless of the performance metric.

This suggests that solely relying on image reconstruction is insufficient for extracting meaningful representations from

TABLE 6. Average MIoU, OIS, and ODS along with standard deviations (in parentheses) from five repeated trials for performance comparison based on the definition of the loss function. The best score for each scenario and manufacturing process is highlighted in bold.

Metric	Segmentation Loss	A	B	C	D	E	F	G	H	I	Average
MIoU	CE	97.43 (1.50)	95.06 (0.63)	96.98 (1.11)	97.36 (0.44)	97.62 (0.32)	85.30 (0.59)	90.12 (0.48)	94.47 (0.27)	92.71 (1.76)	94.02 (4.08)
	BF	98.55 (0.30)	94.95 (0.87)	96.87 (0.75)	97.08 (0.77)	97.45 (0.23)	85.00 (0.56)	89.78 (0.24)	94.08 (0.36)	91.62 (0.90)	93.73 (4.29)
	MTP-MAVED (Proposed)	**98.59 (0.57)**	**95.73 (0.81)**	**97.97 (0.48)**	**97.61 (0.14)**	**97.89 (0.28)**	**85.83 (0.29)**	**91.02 (1.43)**	**95.31 (0.77)**	**93.53 (0.42)**	**94.83 (4.03)**
OIS	CE	86.85 (3.38)	73.41 (2.21)	87.84 (1.94)	79.96 (1.09)	85.76 (0.75)	78.65 (0.92)	64.47 (2.00)	82.76 (2.20)	85.50 (3.18)	80.43 (7.57)
	BF	**89.68 (1.07)**	74.05 (2.41)	89.56 (0.51)	80.13 (0.80)	**87.97 (1.78)**	78.64 (0.94)	63.40 (1.99)	81.64 (1.70)	84.20 (1.34)	80.85 (8.34)
	MTP-MAVED (Proposed)	89.49 (2.43)	**75.08 (2.29)**	**89.83 (0.45)**	**80.53 (0.69)**	87.01 (1.81)	**79.50 (0.53)**	**67.20 (1.84)**	**83.38 (1.28)**	**87.42 (0.73)**	**82.16 (7.27)**
ODS	CE	86.15 (3.61)	72.81 (1.89)	87.09 (1.81)	79.52 (1.16)	85.72 (0.74)	78.32 (0.95)	63.67 (2.11)	82.50 (2.35)	85.36 (3.22)	79.97 (7.68)
	BF	88.98 (1.32)	73.38 (2.33)	88.42 (0.60)	79.29 (0.92)	**87.61 (1.86)**	78.18 (0.98)	62.84 (2.00)	81.40 (1.81)	83.91 (1.53)	80.27 (8.31)
	MTP-MAVED (Proposed)	**89.20 (2.33)**	**74.50 (2.51)**	**89.14 (0.56)**	**80.08 (0.75)**	86.55 (2.06)	**79.17 (0.60)**	**66.48 (1.83)**	**83.17 (1.33)**	**87.02 (0.81)**	**81.70 (7.36)**

TABLE 7. Average MIoU, OIS, and ODS along with standard deviations (in parentheses) from five repeated trials for comparison MTP-MAVED with WTEM-SST. The best score for each scenario and manufacturing process is highlighted in bold.

Metric	Method	A	B	C	D	E	F	G	H	I	Average
MIoU	WTEM-SST	96.42 (0.63)	84.25 (2.79)	95.74 (1.26)	93.55 (1.04)	97.19 (1.04)	81.53 (0.65)	78.80 (2.46)	90.51 (1.23)	79.84 (0.94)	88.65 (7.33)
	MTP-MAVED (Proposed)	**98.59 (0.57)**	**95.73 (0.81)**	**97.97 (0.48)**	**97.61 (0.14)**	**97.89 (0.28)**	**85.83 (0.29)**	**91.02 (1.43)**	**95.31 (0.77)**	**93.53 (0.42)**	**94.83 (4.03)**
OIS	WTEM-SST	83.73 (1.70)	56.33 (1.94)	85.72 (1.39)	67.94 (0.32)	81.81 (3.29)	72.18 (0.75)	42.71 (7.11)	71.18 (4.34)	68.72 (2.15)	70.04 (13.46)
	MTP-MAVED (Proposed)	**89.49 (2.43)**	**75.08 (2.29)**	**89.83 (0.45)**	**80.53 (0.69)**	**87.01 (1.81)**	**79.50 (0.53)**	**67.20 (1.84)**	**83.38 (1.28)**	**87.42 (0.73)**	**82.16 (7.27)**
ODS	WTEM-SST	82.00 (1.76)	54.32 (2.01)	84.70 (1.35)	66.49 (0.35)	81.18 (3.16)	71.84 (0.71)	41.21 (6.95)	70.44 (4.28)	66.10 (2.02)	68.70 (13.68)
	MTP-MAVED (Proposed)	**89.20 (2.33)**	**74.50 (2.51)**	**89.14 (0.56)**	**80.08 (0.75)**	**86.55 (2.06)**	**79.17 (0.60)**	**66.48 (1.83)**	**83.17 (1.33)**	**87.02 (0.81)**	**81.70 (7.36)**

wafer TEM images. In contrast, pre-training with VED enables the model, initially pre-trained on a different dataset, to learn useful representations for wafer TEM images. This can be observed from the significant improvements in OIS (+2.17%) and ODS (+2.35%).

In previous experiments, we explored appropriate methods for pre-training on wafer TEM images. We now focused on optimizing the semantic segmentation task during supervised fine-tuning with multi-task learning using ABL as the loss function for semantic segmentation. To evaluate ABL's effectiveness in detecting object boundaries in wafer TEM images, we compared it against the boundary-focused (BF) loss from [7] and the traditional cross-entropy (CE) loss. The results, presented in Table 6, show that MTP-MAVED using ABL consistently outperformed other approaches across all performance metrics.

In addition, we observed that edge detection metrics, such as OIS and ODS, improved when using BF loss compared to CE, highlighting BF's capability to accurately recognize object boundaries. However, ABL proved even more effective. These experiments confirmed the advantages of using ABL for enhanced boundary detection in our method. While CE and BF focus on maximizing pixel-level similarity, ABL allows the predicted boundaries to learn to approximate the GT boundaries more closely. Therefore, we conclude that MTP-MAVED achieves superior performance when ABL is used.

We conducted experiments to compare the effectiveness of the proposed MTP-MAVED framework against WTEM-SST developed for wafer TEM image analysis. In WTEM-SST, various backbone architectures — including UNet [43], DeepLabV3+ [44], and SegFormer [34] — were used to perform semantic segmentation during pre-training using wafer TEM images and their corresponding GT annotations. In contrast, MTP-MAVED uses wafer TEM images only during pre-training, relying instead on self-supervised tasks. The comparison results, summarized in Table 7, clearly showed that MTP-MAVED consistently outperforms WTEM-SST, particularly under the scenario with extremely limited training data. Especifically,

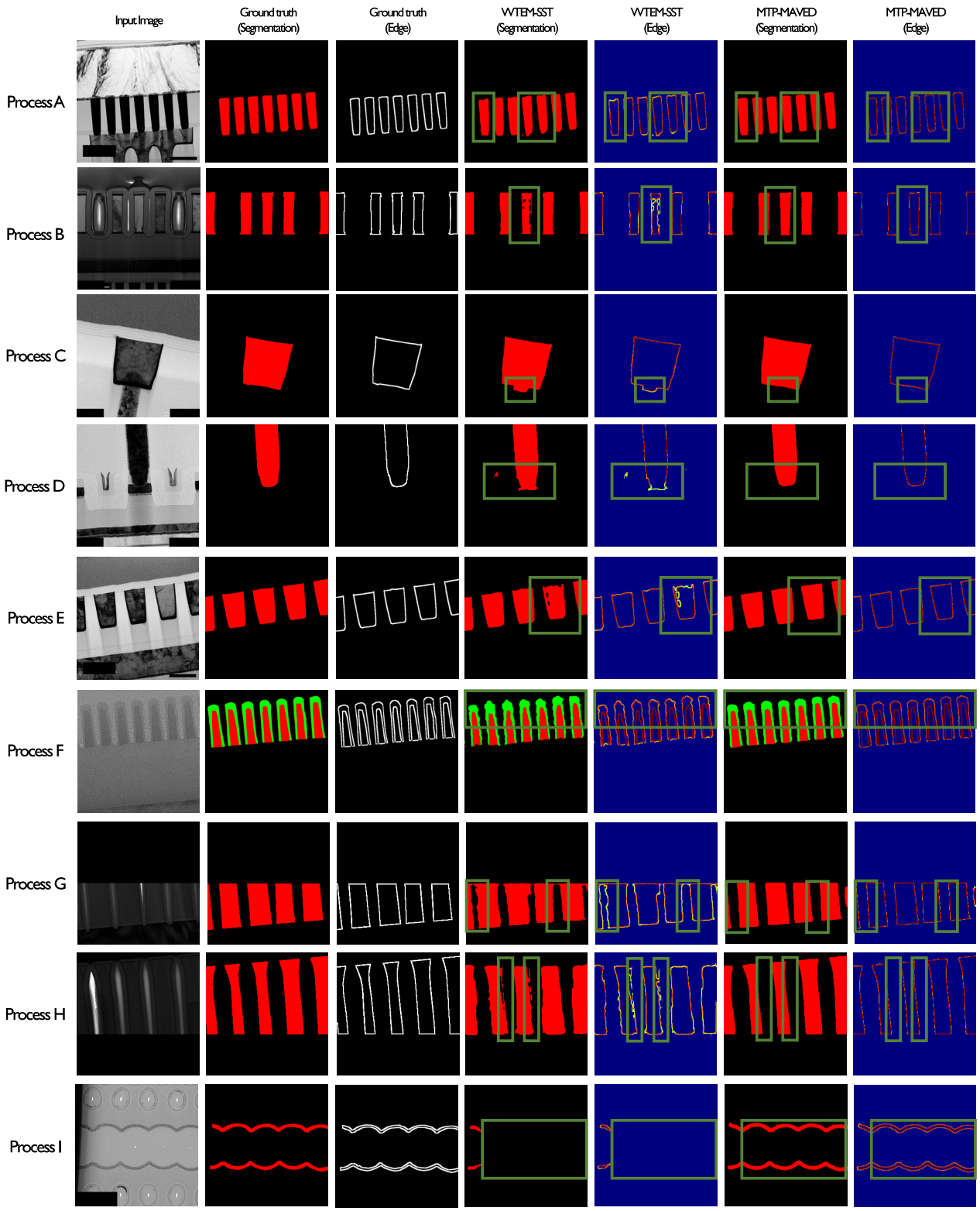

FIGURE 5. Comparison of semantic segmentation and edge detection prediction results between WTEM-SST and MTP-MAVED. The green boxes highlight regions where the predictions of WTEM-SST and MTP-MAVED differ significantly, illustrating the improvement achieved by MTP-MAVED.

MTP-MAVED achieved an improvement of +6.18 % in MIoU, +12.12 % in OIS, and +13.00 % points in ODS on average compared to WTEM-SST. These results highlight the superiority and robustness of MTP-MAVED in highly data-scarce environments, even without access to GT labels during pre-training. Our method further benefits from reformulating

the segmentation problem as a multi-task learning task by integrating edge detection, which provides complementary structural information and enhances segmentation performance. This multi-task strategy improves the model's understanding of object boundaries, leading to more accurate delineation. In addition, the incorporation of the ABL enables the model to better align predicted boundaries with the actual object boundaries, showing superior effectiveness compared to the BF loss in WTEM-SST. These combined strategies, including self-supervised pre-training, multi-task fine-tuning with edge detection, and boundary-aware optimization, highlight the strength of MTP-MAVED in accurately recognizing both objects and their boundaries, as reflected in consistently higher MIoU, OIS, and ODS scores compared to WTEM-SST.

Fig 5 illustrates the prediction results for both semantic segmentation and edge detection using WTEM-SST and MTP-MAVED. WTEM-SST failed to clearly delineate interesting object boundaries because of the limited amount of training examples. In contrast, MTP-MAVED effectively identified interesting objects and their boundaries, indicating a notable performance improvement over existing methods.

V. CONCLUSION

In this study, we propose the MTP-MAVED framework to improve object recognition and boundary detection in wafer TEM images. Our framework addresses three primary challenges of wafer TEM image analysis: the difficulty in image acquisition and its GT annotation, significant noise, and ambiguous object boundaries. By leveraging SSRL combined with VED during pre-training, incorporating multi-task learning for both semantic segmentation and edge detection, and integrating a boundary-aware loss function, the proposed MTP-MAVED demonstrates superior performance over existing methods. Our results show that MTP-MAVED not only enhances the accuracy of semantic segmentation but also improves boundary detection, even in situations with limited labeled examples. The combination of pre-training on unlabeled images and fine-tuning with multi-task learning proves effective for addressing the complexities of wafer TEM images. Furthermore, the inclusion of the ABL function enables more precise boundary recognition, overcoming the limitations of traditional approaches that often result in over- or under-estimation. Compared to the previous approach using labeled images during pre-training, our method enables more effective representation learning by using diverse unlabeled wafer TEM images during pre-training, thereby enhancing the model's adaptability to various manufacturing processes. Furthermore, the multi-task learning with edge detection allows the model to precisely identify both objects of interest and their boundaries, leading to improved segmentation accuracy over conventional methods. The proposed framework contributes to the advancement of automated measurement in the semiconductor industry by providing a more robust and accurate solution for analyzing wafer TEM images. Nevertheless, our study did not directly address the prevalent noise issue in wafer TEM images. To address this limitation, we aim to jointly train deep neural networks for image denoising and semantic segmentation model within a unified multi-task learning framework, enabling the model to simultaneously remove noise and accurately recognize interesting objects in wafer TEM images. Moreover, we plan to expand our research by developing a wafer vision foundation model based on the segment anything approach [45] that can integrate images from various stages of the wafer manufacturing process.

REFERENCES

[1] K. B. Lee, S. Cheon, and C. O. Kim, "A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes," *IEEE Trans. Semicond. Manuf.*, vol. 30, no. 2, pp. 135–142, May 2017.

[2] F. Zhu, X. Jia, M. Miller, X. Li, F. Li, Y. Wang, and J. Lee, "Methodology for important sensor screening for fault detection and classification in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 34, no. 1, pp. 65–73, Feb. 2021.

[3] J. Choi and S. B. Kim, "Multi-stage process diagnosis networks in semiconductor manufacturing," *IEEE Access*, vol. 12, pp. 39495–39504, 2024.

[4] H. Kahng and S. B. Kim, "Self-supervised representation learning for wafer bin map defect pattern classification," *IEEE Trans. Semicond. Manuf.*, vol. 34, no. 1, pp. 74–86, Feb. 2021.

[5] H. Wang, J. Fang, J. Arjavac, and R. Kellner, "Scanning transmission electron microscopy for critical dimension monitoring in wafer manufacturing," *Microsc. Today*, vol. 16, no. 1, pp. 24–27, Jan. 2008.

[6] J. S. Lee, S. H. Choi, S. J. Yun, Y. I. Kim, S. Boandoh, J.-H. Park, B. G. Shin, H. Ko, S. H. Lee, Y.-M. Kim, Y. H. Lee, K. K. Kim, and S. M. Kim, "Wafer-scale single-crystal hexagonal boron nitride film via self-collimated grain formation," *Science*, vol. 362, no. 6416, pp. 817–821, Nov. 2018.

[7] Y. Jo, J. Bae, H. Cho, H. Roh, K. Kim, M. Jo, J. Tae, and S. B. Kim, "Semantic segmentation for noisy and limited wafer transmission electron microscope images," *IEEE Trans. Semicond. Manuf.*, vol. 37, no. 3, pp. 345–354, Aug. 2024.

[8] J. Baderot, M. Grould, D. Misra, N. Clément, A. Hallal, S. Martinez, and J. Foucher, "Application of deep-learning based techniques for automatic metrology on scanning and transmission electron microscopy images," *J. Vac. Sci. Technol. B*, vol. 40, no. 5, Sep. 2022, Art. no. 054003.

[9] C. Wang, Y. Zhang, M. Cui, P. Ren, Y. Yang, X. Xie, X.-S. Hua, H. Bao, and W. Xu, "Active boundary loss for semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 2397–2405.

[10] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023.

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[12] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.

[13] J. Žbontar, J. Li, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12310–12320.

[14] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 9912–9924.

[15] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15979–15988.

[16] Z. Huang, X. Jin, C. Lu, Q. Hou, M.-M. Cheng, D. Fu, X. Shen, and J. Feng, "Contrastive masked autoencoders are stronger vision learners," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2506–2517, Apr. 2024.

[17] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.
[18] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 110, pp. 3371–3408, 2010.
[19] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.
[20] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, "Multi-task learning for dense prediction tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3614–3633, Jul. 2022.
[21] S. Borse, Y. Wang, Y. Zhang, and F. Porikli, "InverseForm: A loss function for structured boundary-aware segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5897–5907.
[22] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7482–7491.
[23] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape CNNs for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5229–5238.
[24] S. Duan, G. Tian, Z. Wang, S. Liu, and C. Feng, "A semantic robotic grasping framework based on multi-task learning in stacking scenes," *Eng. Appl. Artif. Intell.*, vol. 121, May 2023, Art. no. 106059.
[25] A. T. M. Nakamura, V. Grassi, and D. F. Wolf, "An effective combination of loss gradients for multi-task learning applied on instance segmentation and depth estimation," *Eng. Appl. Artif. Intell.*, vol. 100, Apr. 2021, Art. no. 104205.
[26] Y. Zhou and P. Y. Mok, "Knowledge enhanced multi-task learning for simultaneous optimization of human parsing and pose estimation," *Eng. Appl. Artif. Intell.*, vol. 138, Dec. 2024, Art. no. 109413.
[27] H. Kervadec, J. Bouchtiba, C. Desrosiers, É. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *Proc. Int. Conf. Med. Imag. Deep Learn.*, vol. 67, 2020, pp. 285–296.
[28] H. Do, C. Lee, and S. B. Kim, "A hierarchical spatial-test attention network for explainable multiple wafer bin maps classification," *IEEE Trans. Semicond. Manuf.*, vol. 35, no. 1, pp. 78–86, Feb. 2022.
[29] M. G. Kwak, Y. J. Lee, and S. B. Kim, "SWaCo: Safe wafer bin map classification with self-supervised contrastive learning," *IEEE Trans. Semicond. Manuf.*, vol. 36, no. 3, pp. 416–424, Aug. 2023.
[30] E. Kim, M.-C. Shin, H.-J. Ahn, S. Park, D. R. Lee, H. Park, M. Shin, and D. Ihm, "Deep learning-based automatic defect classification for semiconductor manufacturing," *Proc. SPIE*, vol. 12496, pp. 404–411, Mar. 2023.
[31] M. Kim, H. Jo, M. Ra, and W.-Y. Kim, "Weakly-supervised defect segmentation on periodic textures using CycleGAN," *IEEE Access*, vol. 8, pp. 176202–176216, 2020.
[32] J. Xu, Z. Ren, B. Dong, X. Liu, C. Wang, Y. Tian, and C. Lee, "Nanometer-scale heterogeneous interfacial sapphire wafer bonding for enabling plasmonic-enhanced nanofluidic mid-infrared spectroscopy," *ACS Nano*, vol. 14, no. 9, pp. 12159–12172, Sep. 2020.
[33] Y. Li, T. Zhou, K. He, Y. Zhou, and D. Shen, "Multi-scale transformer network with edge-aware pre-training for cross-modality MR image synthesis," *IEEE Trans. Med. Imag.*, vol. 42, no. 11, pp. 3395–3407, Nov. 2023.
[34] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Álvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.
[35] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vols. PAMI–8, no. 6, pp. 679–698, Nov. 1986.
[36] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3994–4003.
[37] B. Kaur and A. Garg, "Mathematical morphological edge detection for remote sensing images," in *Proc. 3rd Int. Conf. Electron. Comput. Technol.*, vol. 5, Apr. 2011, pp. 324–327.
[38] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[39] [Online]. Available: https://huggingface.co/nvidia/segformer-b3-finetuned-cityscapes-1024-1024
[40] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[41] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020.
[42] X. Soria, E. Riba, and A. Sappa, "Dense extreme inception network: Towards a robust CNN model for edge detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1912–1921.
[43] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.
[44] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 833–851.
[45] A. M. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jul. 2023, pp. 4015–4026.

YONGWON JO received the Ph.D. degree from the Department of Industrial and Management Engineering, Korea University, in 2025. His research interests include semantic segmentation, its industrial applications, and domain generalization.

JINSOO BAE received the B.S. degree, in 2020. He is currently pursuing the Ph.D. degree with the Department of Industrial and Management Engineering, Korea University, Seoul, South Korea. His research interests include uncertainty-aware deep neural networks and incomplete data analysis.

HANSAM CHO received the B.S. degree in industrial and management engineering from Korea University, in 2020, where he is currently pursuing the Ph.D. degree with the Department of Industrial and Management Engineering. His research interests include diffusion models, image editing, and their industrial applications.

SUNGSU KIM received the B.S. degree in industrial and systems engineering from Kyung Hee University, in 2022. He is currently pursuing the Ph.D. degree with the Department of Industrial and Management Engineering, Korea University. His research interests include computer vision, incomplete data analysis, and their industrial applications.

HEEJOONG ROH received the master's degree in industrial and system engineering from Korea Advanced Institute of Science and Technology. He is currently a Data Scientist with SK Hynix. He specializes in researching vision algorithms for semiconductor industry. He leads the development of a system that offer solutions in the field of semiconductor image processing.

KYUNGHYE KIM received the M.S. degree from the Department of Industrial Engineering, Yonsei University, in 2019. She is currently a Data Scientist with SK Hynix. Her current research interests include computer vision, multi-modal analysis, artificial intelligence, and their industrial applications.

MUNKI JO received the master's degree in industrial and system engineering from Korea Advanced Institute of Science and Technology, in 2022. He is currently a Data Scientist with SK Hynix. He is interested in providing solutions to the semiconductor industry through computer vision algorithm.

MUNUK KIM received the M.S. degree in information statistics, Korea University, in 2014. He is currently a Data Scientist with SK Hynix. He is interested in charge of image MLOps solution construction. In particular, he working to simplify complex workflows and processes from model development to deployment to shorten the deployment period of AI solutions by designing MLOps functions.

JAEUNG TAE received the master's degree in industrial and system engineering from Hanyang University. He is currently a Data Scientist and an ETCH Process Engineer with SK Hynix. He is also the Head of the Research and Development and Manufacturing Analysis Team with SK Hynix and is developing AI systems in the semiconductor industry in research, development, and production.

SEOUNG BUM KIM received the M.S. and Ph.D. degrees in industrial and systems engineering from Georgia Institute of Technology, in 2001 and 2005, respectively. From 2005 to 2009, he was a Professor with the Department of Industrial and Manufacturing Systems Engineering, The University of Texas at Arlington. He is currently a Professor with the School of Industrial and Management Engineering, Korea University. He is also the Director of the Artificial Intelligence Engineering Center, Korea University. He has published more than 150 internationally recognized journals and refereed conference proceedings. His research interests utilize machine learning algorithms to create new methods for various problems appearing in engineering and science.