AI Engineering
https://www.sciltp.com/journals/aieng

Article

Open-Set Semiconductor Defect Classification in SEM Images via Uncertainty-Aware Local Prototype Alignment

Xinyi Yuan [1], Sisi Chen [1], Boran Hu [1], Yikang Zhou [1], Aohan Mei [2] and Nan Li [3,*]

[1] School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
[2] College of Software, Northeastern University, Shenyang 110819, China
[3] School of Artificial Intelligence, Shanxi University, Taiyuan 030006, China
* Correspondence: linan10@sxu.edu.cn

How To Cite: Yuan, X.; Chen, S.; Hu, B.; et al. Open-Set Semiconductor Defect Classification in SEM Images via Uncertainty-Aware Local Prototype Alignment. *AI Engineering* **2026**, *2*(2), 12. https://doi.org/10.53941/aieng.2026.100012

Received: 6 June 2026
Revised: 19 July 2026
Accepted: 21 July 2026
Published: 6 August 2026

Abstract: Automated classification of scanning electron microscope (SEM) defects is complicated by acquisition-domain shift, long-tailed classes, and target-only defect types that invalidate the closed-set assumption. We propose an offline open-set adaptation framework that combines a three-criterion reliable-known selector with global and defect-region-guided local prototype alignment. Confidence, entropy, and prototype distance jointly determine which unlabeled target samples may be aligned to known source prototypes; samples that fail this gate are excluded from known-class alignment and optimized by an unknown-rejection objective. Unlike global-only adaptation, foreground-weighted local prototypes emphasize compact defect morphology rather than domain-specific background texture. Averaged over four bidirectional SEM adaptation tasks and three random seeds per task, the method achieves 80.8% balanced accuracy, 78.1% Macro-F1, 88.7% unknown AUROC, 74.2% H-score, and 8.4% ECE. Compared with the strongest baseline for each individual metric, the proposed method improves balanced accuracy by 4.4 percentage points and Macro-F1 by 4.5 percentage points over M^{2}DD, improves unknown AUROC by 6.9 percentage points and reduces ECE by 3.1 percentage points relative to the energy-based baseline, and improves H-score by 5.9 percentage points over CMU. Component removal and sensitivity analyses confirm complementary contributions from reliable-target selection, prototype alignment, unknown rejection, and localized alignment. The framework is directly applicable to offline SEM inspection when labeled historical data and unlabeled new-condition images are jointly available; its selective-alignment principle may also transfer to other localized visual-inspection tasks after domain-specific validation and calibration.

Keywords: semiconductor defect classification; SEM image analysis; open-set domain adaptation; uncertainty estimation; prototype alignment; defect localization

1. Introduction

Automated defect classification (ADC) plays a critical role in semiconductor manufacturing, where accurate defect recognition supports yield improvement, process diagnosis, and production reliability [1]. Scanning electron microscope (SEM) images provide high-resolution evidence of surface irregularities, structural anomalies, and process-induced defect morphologies. Recent deep learning methods have improved semiconductor defect classification, detection, and localization by using transfer learning, convolutional neural networks, adversarial detection frameworks, and CNN—Transformer hybrid architectures [2–4]. However, SEM defect classification remains challenging because real production data are often scarce, highly imbalanced, and collected under changing manufacturing conditions.

A major difficulty comes from domain shift. SEM images acquired from different fabrication lines, inspection tools, technology nodes, or imaging settings may differ in contrast, magnification, noise level, texture distribution, and intra-class visual complexity [5]. As a result, a model trained on historical labeled data may fail when deployed on a newly collected target domain. To address this issue, Yang et al. [5] introduced domain adaptation into SEM defect classification and proposed margin–margin discrepancy distance (M^{2}DD), which combines margin cross-

entropy with margin discrepancy distance to handle both long-tailed class imbalance and domain shift. Their results show that domain adaptation can outperform empirical risk minimization when target-domain labels are unavailable or extremely limited.

Domain adaptation methods aim to reduce distribution discrepancy between source and target domains. Representative approaches include adversarial feature learning with gradient reversal [6], discrepancy-based adaptation with theoretical risk bounds [7], and conditional adversarial alignment using classifier predictions [8]. These methods are effective under the closed-set assumption, where the source and target domains share the same label space. Nevertheless, this assumption is restrictive for semiconductor manufacturing, since new defect mechanisms may appear due to tool aging, process drift, recipe changes, material variation, or unexpected process excursions.

Class imbalance further complicates the problem. Rare defect classes may be more important for yield diagnosis, but they usually contain far fewer labeled samples than common defect types. Long-tailed learning methods such as label-distribution-aware margin loss, focal loss, and class-balanced loss have been proposed to improve recognition under imbalanced data distributions [9–11]. However, directly applying these losses to domain-shifted SEM images may still be insufficient, because minority classes are entangled not only with sample scarcity but also with domain-specific background patterns and imaging artifacts.

Another important limitation is the closed-set nature of existing SEM domain adaptation frameworks [5]. In real production environments, target-domain images may contain unknown defect classes that were absent from the source training set. If these unknown samples are forcibly aligned with known source categories, the adaptation process can produce open-set negative transfer and overconfident misclassification. Open-set recognition and open-set domain adaptation address this issue by allowing the model to reject unknown classes rather than assigning every sample to a known category [12,13]. This capability is especially important for semiconductor inspection, where unknown defects may signal new process failures.

In addition, existing discrepancy-based methods often align global image-level representations, while SEM defect classification depends heavily on small and localized morphological cues [2,3]. Prior visualization methods, including class activation mapping and Grad-CAM, show that deep classifiers may focus on discriminative image regions [14,15]. In SEM defect analysis, such localization is not only useful for interpretation but also valuable for training, because the true defect region may occupy only a small part of the image while the background contains domain-specific noise.

To address these limitations, this study proposes an open-set SEM defect classification framework based on uncertainty-aware local prototype alignment. The proposed framework selectively aligns target-domain samples that are likely to belong to known defect classes while rejecting highly uncertain samples that may correspond to unknown defects. It estimates uncertainty using predictive entropy, energy scores, and distances to class prototypes. High-confidence known samples are used for class-conditional prototype alignment, whereas uncertain samples are regularized through an unknown-defect-aware rejection objective.

Furthermore, we introduce defect-region-guided multi-scale local alignment. Instead of aligning only global image representations, the model extracts defect-relevant local features from weakly supervised localization maps and constructs multi-scale class prototypes for foreground defect regions. This design encourages the model to transfer defect morphologies rather than background textures, improving robustness under domain shift and long-tailed class imbalance.

The main contributions of this paper are summarized as follows:

- We formulate SEM defect classification as an open-set, class-imbalanced, and domain-shifted recognition problem, extending prior closed-set SEM domain adaptation settings.
- We propose an uncertainty-aware prototype alignment mechanism that selectively aligns known target-domain samples and reduces open-set negative transfer from unknown defect classes.
- We introduce defect-region-guided multi-scale local alignment to learn morphology-aware transferable representations from localized SEM defect regions.
- We develop an offline open-set domain adaptation framework that jointly uses source and target mini-batches during training and requires no parameter update at inference. This positioning distinguishes the method from closed-set SEM adaptation and source-free test-time adaptation.
- Across four transfer directions, the method improves over the strongest metric-specific baselines by 4.4 points in BAcc, 4.5 points in Macro-F1, 6.9 points in unknown AUROC, and 5.9 points in H-score, while lowering ECE by 3.1 points. Component-removal results further separate the contributions of selection, prototype alignment, unknown rejection, and local morphology alignment.

The remainder of this paper is organized as follows. Section 2 reviews related studies on semiconductor defect classification, domain adaptation, open-set recognition, class-imbalanced learning, and weakly supervised defect localization. Section 3 describes the proposed framework, including the uncertainty-aware prototype alignment

mechanism, the defect-region-guided multi-scale local alignment module, and the open-set rejection strategy. Section 4 presents the datasets, open-set domain adaptation protocols, experimental settings, evaluation metrics, and comparison methods. Section 5 reports and analyzes the experimental results, including known-class classification performance, unknown-defect rejection, ablation studies, visualization analysis, and robustness under domain shift. Finally, Section 6 concludes this paper and discusses future work.

2. Related Work

2.1. Semiconductor Defect Classification and Closed-Set Domain Adaptation

Automated defect classification is an important task in semiconductor manufacturing, where accurate recognition of defect classes supports yield improvement, process monitoring, and failure diagnosis. Deep learning has been increasingly adopted for semiconductor defect analysis because of its ability to extract discriminative visual representations from complex inspection images. Imoto et al. [1] proposed a CNN-based transfer learning method for semiconductor defect classification, demonstrating that pretrained models can improve recognition performance when labeled manufacturing data are limited. More recently, Yang et al. [5] investigated SEM defect classification under domain shift and proposed margin–margin discrepancy distance (M^2DD), which combines margin cross-entropy with discrepancy-based domain adaptation to address both long-tailed class imbalance and target-domain label scarcity.

Domain adaptation has been widely studied to reduce the distribution gap between labeled source data and unlabeled or sparsely labeled target data. Ganin et al. [6] introduced domain-adversarial neural networks, which use a gradient reversal layer to learn domain-invariant representations in an end-to-end manner. Zhang et al. proposed margin discrepancy distance (MDD), providing both a theoretical adaptation bound and an adversarial optimization strategy for aligning source and target domains [7]. Long et al. [8] further developed conditional adversarial domain adaptation (CDAN), which conditions domain alignment on classifier predictions to better capture multimodal class structures. Although these methods are effective in closed-set adaptation, they generally assume that the source and target domains share an identical label space, which is often unrealistic in semiconductor production.

2.2. Open-Set and Universal Domain Adaptation

In real semiconductor manufacturing environments, target-domain SEM images may contain defect classes that are absent from the source training set. Such unknown defects may arise from process drift, equipment aging, recipe changes, or unexpected process excursions. Closed-set domain adaptation methods may incorrectly align these unknown samples with known source categories, resulting in open-set negative transfer and overconfident misclassification. Therefore, open-set and universal domain adaptation are highly relevant to practical SEM defect classification.

Panareda Busto and Gall introduced the open-set domain adaptation problem, where the target domain contains unknown classes that should be separated from known source categories [16]. Saito et al. [13] proposed open-set domain adaptation by backpropagation, using adversarial learning to separate unknown target-domain samples while aligning known classes. You et al. [17] generalized this setting through universal domain adaptation, where the source and target label spaces may partially overlap without prior knowledge of the shared categories. Fu et al. [18] further studied how to detect open classes under universal domain adaptation, emphasizing the importance of identifying target-domain samples that do not belong to any known source class. These studies provide a foundation for extending SEM defect classification from closed-set adaptation to open-set industrial recognition.

2.3. Unknown Detection, Uncertainty Estimation, and Prototype Learning

Unknown-defect rejection requires reliable uncertainty estimation. Open-set recognition methods aim to prevent models from assigning all inputs to known categories [12,19]. Vaze et al. [19] showed that a strong closed-set classifier can serve as a powerful baseline for open-set recognition when properly calibrated and evaluated. Liu et al. [20] proposed energy-based out-of-distribution detection, where energy scores derived from classifier logits provide an effective criterion for separating in-distribution and out-of-distribution samples. These methods are useful for SEM defect classification because unknown defects should be rejected or flagged for expert review rather than forcibly classified as known defect types.

Prototype-based learning offers another effective way to model class structure under limited data. Snell et al. [21] proposed prototypical networks for few-shot learning, representing each class by a prototype in embedding space and classifying samples according to distances to class prototypes. This idea is naturally suitable for rare semiconductor defect classes, where only a few labeled examples may be available. In the proposed framework, class prototypes are used not only for classification but also for uncertainty-aware alignment, allowing the model to align high-confidence known target-domain samples while isolating uncertain samples that may correspond to unknown defects.

2.4. Distinction from Domain Generalization and Test-Time Adaptation

Domain generalization and test-time adaptation address related but different deployment assumptions. Test-time training updates a model using an auxiliary objective during inference [22], and Tent updates normalization parameters using target entropy [23]; domain generalization instead trains without access to the eventual target domain [24]. Our setting is neither source-free TTA nor domain generalization: unlabeled target images are available jointly with labeled source images during offline training, and the resulting model is frozen during inference. We retain this literature only to delimit the problem setting and avoid conflating offline adaptation with continuous deployment-time updating.

In summary, existing studies provide useful foundations for SEM defect classification, closed-set domain adaptation, open-set recognition, prototype learning, and test-time adaptation. The present method is an offline open-set domain adaptation approach: it uses labeled source and unlabeled target mini-batches jointly during training, whereas its inference stage is fixed and does not update model parameters. Its contribution is therefore not an online or source-free TTA mechanism, but the joint treatment of long-tailed source classification, selective known-target alignment, unknown-defect rejection, and morphology-aware local alignment for SEM imagery.

3. Methodology

3.1. Problem Formulation

Let $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ denote the labeled source-domain SEM dataset, where x_i^s is an SEM image and $y_i^s \in \mathcal{C}_s$ is its defect category. Let $\mathcal{D}_t = \{x_j^t\}_{j=1}^{n_t}$ denote the unlabeled target-domain dataset collected from a different fabrication line, inspection tool, process condition, or imaging environment. In conventional closed-set domain adaptation, the source and target domains are assumed to share the same label space. However, this assumption is restrictive in semiconductor manufacturing because the target domain may contain unknown defect types that are absent from the source domain.

In this work, we consider an open-set domain adaptation setting for SEM defect classification. The source label space is denoted by $\mathcal{C}_s$, while the target label space is denoted by $\mathcal{C}_t$. The target domain may contain both known classes shared with the source domain and unknown classes:

$$\mathcal{C}_t = \mathcal{C}_k \cup \mathcal{C}_u, \tag{1}$$

where $\mathcal{C}_k \subseteq \mathcal{C}_s$ denotes known defect classes and $\mathcal{C}_u$ denotes unknown target-only defect classes. During training, labels are available only for the source domain, while target-domain labels are unavailable. The objective is to correctly classify target-domain samples from known categories and reject target-domain samples from unknown categories.

The proposed framework consists of four main components: a visual feature extractor, a classification head, a prototype memory bank, and an uncertainty-aware open-set rejection module. Given an input SEM image x, the visual feature extractor $G(\cdot)$ maps it into a latent representation:

$$z = G(x). \tag{2}$$

The classification head $F(\cdot)$ then predicts class logits:

$$p = \mathrm{softmax}(F(z)). \tag{3}$$

In addition to global features, the model extracts local defect-aware features from intermediate activation maps to perform morphology-oriented prototype alignment.

3.2. Overall Framework

The overall goal is to learn a representation that satisfies three requirements. First, source-domain labeled samples should be classified accurately, especially for minority defect classes. Second, target-domain samples that are likely to belong to known classes should be aligned with corresponding source-domain prototypes. Third, target-domain samples that are highly uncertain should be separated from known classes and treated as potential unknown defects.

To achieve these goals, the proposed framework optimizes the following objective:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{\mathrm{mce}} + \lambda_1 \mathcal{L}_{\mathrm{proto}} + \lambda_2 \mathcal{L}_{\mathrm{local}} \\ &+ \lambda_3 \mathcal{L}_{\mathrm{unk}} + \lambda_4 \mathcal{L}_{\mathrm{ent}},\end{aligned} \tag{4}$$

where $\mathcal{L}_{\mathrm{mce}}$ is the margin cross-entropy loss for source classification, $\mathcal{L}_{\mathrm{proto}}$ is the uncertainty-aware global prototype alignment loss, $\mathcal{L}_{\mathrm{local}}$ is the defect-region-guided local prototype alignment loss, $\mathcal{L}_{\mathrm{unk}}$ is the unknown-defect rejection loss, and $\mathcal{L}_{\mathrm{ent}}$ is the target entropy regularization loss. The hyperparameters λ_1, λ_2, λ_3, and λ_4 control the contribution of each term.

The terms have distinct roles: $\mathcal{L}_{\mathrm{mce}}$ learns class-balanced source decision boundaries; $\mathcal{L}_{\mathrm{proto}}$ performs class-conditional alignment only for reliable known-target candidates; $\mathcal{L}_{\mathrm{local}}$ transfers localized defect morphology; $\mathcal{L}_{\mathrm{unk}}$ suppresses overconfident assignment of uncertain samples to known classes; and $\mathcal{L}_{\mathrm{ent}}$ sharpens predictions only within the selected known-target subset.

The architecture therefore retains a single-backbone computation graph (see Table 1). Prototype memories are updated by exponential moving averages and are not optimized as independent parameter tensors. The local branch reuses the last convolutional activation map already produced by ResNet-50; it adds foreground-weighted pooling and a prototype loss during training but no second image encoder.

Table 1. Network architecture and additional state of the proposed framework. The ResNet-50 backbone and linear classifier are shared with all comparison methods.

Stage	Representation	Operation	Additional Learned Parameters/State
Input/backbone	$x \in \mathbb{R}^{3\times 224\times 224}$; $z \in \mathbb{R}^{d}$	one ImageNet-pretrained ResNet-50 forward pass	shared baseline backbone
Classifier	K known-class logits	linear projection and softmax	shared $Kd + K$ classifier parameters
Global prototypes	$\mathcal{P} \in \mathbb{R}^{K\times d}$	class-wise EMA update; nearest-prototype distances	Kd non-trainable buffer values
Local representation	$A \in \mathbb{R}^{C\times H\times W}$; $z_{\mathrm{loc}} \in \mathbb{R}^{d_{\mathrm{loc}}}$	CAM normalization and foreground-weighted pooling from the last convolutional block	no auxiliary backbone
Local prototypes	$\mathcal{P}_{\mathrm{loc}} \in \mathbb{R}^{K\times d_{\mathrm{loc}}}$	class-wise EMA update and selected-target alignment	Kd_{loc} non-trainable buffer values
Reliability/rejection	q, entropy, and prototype distance	three-condition gate during training; confidence and distance rejection during inference	no trainable parameters

3.3. Margin-Based Source Classification

SEM defect datasets are usually class-imbalanced because some defect types occur frequently while rare defects may appear only in a small number of labeled samples. Standard cross-entropy may bias the classification head toward majority classes. Following the motivation of margin-based classification, we employ a margin cross-entropy loss to enlarge the decision boundaries of minority classes.

For a source sample (x_i^s, y_i^s), let p_i^s denote the predicted probability vector. The margin cross-entropy loss is defined as

$$\mathcal{L}_{\mathrm{mce}} = -\frac{1}{n_s}\sum_{i=1}^{n_s}\log p_{i,y_i^s}^s + \frac{\alpha}{n_s}\sum_{i=1}^{n_s}\left[m + \max_{k\neq y_i^s} f_{i,k}^s - f_{i,y_i^s}^s\right]_+, \tag{5}$$

where $f_{i,k}^s$ denotes the logit of source sample i for class k, $m > 0$ is the prescribed classification margin, α controls the strength of the margin regularization, and $[a]_+ = \max(0, a)$. The positive margin sign requires $f_{i,y_i^s}^s \geq \max_{k\neq y_i^s} f_{i,k}^s + m$ for zero hinge penalty. Equation (5), including this sign convention, is the objective used in the experiments.

3.4. Class Prototype Memory Bank

To model known defect classes explicitly, we maintain a prototype memory bank

$$\mathcal{P} = \{c_1, c_2, \ldots, c_K\}, \tag{6}$$

where $K = |\mathcal{C}_s|$ is the number of source classes and c_k is the prototype of the k-th known defect class. Each prototype represents the center of source-domain features belonging to the corresponding class.

At each iteration, the prototype c_k is updated using an exponential moving average:

$$c_k \leftarrow m c_k + (1-m)\frac{1}{|\mathcal{B}_k|}\sum_{x_i^s \in \mathcal{B}_k} G(x_i^s), \tag{7}$$

where m is the momentum coefficient and $\mathcal{B}_k$ is the set of source-domain samples from class k in the current mini-batch. The prototypes provide stable known-class anchors for aligning target-domain samples and estimating whether a target sample is likely to belong to a known class.

For a target sample x_j^t, its distance to the nearest known prototype is computed as

$$d_j = \min_{k\in\{1,\ldots,K\}} \left\| G(x_j^t) - c_k \right\|_2 . \tag{8}$$

A smaller distance indicates that the target sample is closer to a known defect category, while a larger distance suggests that it may correspond to an unknown defect.

3.5. Uncertainty-Aware Known-Class Selection

Not all target-domain samples should be aligned with source-domain prototypes. If an unknown target-domain sample is mistakenly aligned with a known source prototype, the model may suffer from open-set negative transfer. Therefore, we design an uncertainty-aware selection strategy to identify target-domain samples that are likely to belong to known classes.

For each target sample, we calculate three uncertainty indicators. The first is prediction entropy:

$$H(x_j^t) = -\sum_{k=1}^{K} p_{j,k}^t \log p_{j,k}^t. \tag{9}$$

The second is the maximum prediction confidence:

$$q_j = \max_k p_{j,k}^t. \tag{10}$$

The third is the energy score:

$$E(x_j^t) = -T \log \sum_{k=1}^{K} \exp\left(\frac{f_k(G(x_j^t))}{T}\right), \tag{11}$$

where T is the temperature and $f_k(\cdot)$ is the logit of the k-th known class.

The energy score is retained as an auxiliary open-set diagnostic and for the energy-based comparison baseline; it is not part of the hard target-selection rule below. The proposed selector uses confidence, entropy, and prototype distance, so no unreported energy threshold is required.

A target sample is selected as a known-class candidate if it satisfies

$$q_j \geq \tau_q, \quad H(x_j^t) \leq \tau_h, \quad d_j \leq \tau_d, \tag{12}$$

where τ_q, τ_h, and τ_d are predefined thresholds. The selected target-domain samples form a known target-domain set $\mathcal{B}_t^k$, while the remaining uncertain samples form an unknown candidate set $\mathcal{B}_t^u$.

3.6. Uncertainty-Aware Global Prototype Alignment

For target-domain samples selected as known-class candidates, pseudo labels are assigned according to their nearest prototypes:

$$\hat{y}_j^t = \arg\min_k \left\| G(x_j^t) - c_k \right\|_2 . \tag{13}$$

The global prototype alignment loss encourages selected target features to move closer to their corresponding class prototypes:

$$\mathcal{L}_{\text{proto}} = \frac{1}{|\mathcal{B}_t^k|}\sum_{x_j^t \in \mathcal{B}_t^k} \left\| G(x_j^t) - c_{\hat{y}_j^t} \right\|_2^2 . \tag{14}$$

This selective alignment prevents uncertain target-domain samples from being forcibly matched to known categories and reduces open-set negative transfer in open-set deployment scenarios.

3.7. Defect-Region-Guided Local Prototype Alignment

Global feature alignment may still be affected by background texture, imaging artifacts, and tool-specific noise. Since SEM defects are often localized, we further introduce a defect-region-guided local prototype alignment module.

Let $A(x) \in \mathbb{R}^{H\times W}$ denote a weak localization map generated from intermediate feature maps. The localization map indicates the spatial importance of each region for classification. We normalize $A(x)$ into a soft foreground mask:

$$M(x) = \frac{A(x) - \min(A(x))}{\max(A(x)) - \min(A(x)) + \epsilon}. \tag{15}$$

Given an intermediate feature map $\Phi(x) \in \mathbb{R}^{H\times W\times C}$, the defect-aware local feature is computed by weighted pooling:

$$z_{\text{loc}} = \frac{\sum_{u=1}^{H}\sum_{v=1}^{W} M_{u,v}(x)\Phi_{u,v}(x)}{\sum_{u=1}^{H}\sum_{v=1}^{W} M_{u,v}(x) + \epsilon}. \tag{16}$$

Similar to global prototypes, we maintain local prototypes

$$\mathcal{P}_{\text{loc}} = \{c_1^{\text{loc}}, c_2^{\text{loc}}, \ldots, c_K^{\text{loc}}\}. \tag{17}$$

For selected known target-domain samples, the local alignment loss is defined as

$$\mathcal{L}_{\text{local}} = \frac{1}{|\mathcal{B}_t^k|}\sum_{x_j^t\in\mathcal{B}_t^k}\left\| z_{\text{loc},j}^t - c_{\hat{y}_j^t}^{\text{loc}}\right\|_2^2. \tag{18}$$

This loss encourages the model to align defect morphology rather than global background appearance patterns.

3.8. Unknown Defect Rejection

For uncertain target-domain samples in $\mathcal{B}_t^u$, the model should avoid assigning them to known classes with high confidence. We therefore introduce an unknown-defect rejection loss. For an uncertain target sample, we encourage a low maximum known-class confidence:

$$\mathcal{L}_{\text{unk}} = \frac{1}{|\mathcal{B}_t^u|}\sum_{x_j^t\in\mathcal{B}_t^u}\max_k p_{j,k}^t. \tag{19}$$

In addition, entropy regularization is applied to selected known target-domain samples:

$$\mathcal{L}_{\text{ent}} = -\frac{1}{|\mathcal{B}_t^k|}\sum_{x_j^t\in\mathcal{B}_t^k}\sum_{k=1}^{K} p_{j,k}^t \log p_{j,k}^t, \tag{20}$$

which encourages confident predictions only for reliable known-class target candidates.

3.9. Training Algorithm

Algorithm 1 summarizes the training procedure. At each iteration, the model jointly exploits labeled source-domain samples and unlabeled target-domain samples. The labeled source mini-batch is used to optimize the margin-based classification objective and update the global and local prototype banks, which serve as stable known-class anchors for known defect classes. For each target sample, the model estimates its reliability using prediction confidence, entropy, and prototype distance. Only target-domain samples satisfying the known-class selection criteria are assigned pseudo labels and included in prototype alignment, while highly uncertain samples are treated as potential unknown defects and optimized with the unknown-defect rejection objective. In this way, the algorithm avoids forcing all target-domain samples into the known source label space and reduces open-set negative transfer caused by unknown target-domain defects. The local alignment branch further uses defect-aware localization masks to emphasize morphology-related regions, allowing the model to align defect structures rather than background textures or imaging artifacts. The final objective integrates source classification, global prototype alignment, local defect-region alignment, unknown-defect rejection, and entropy regularization in a unified end-to-end training process.

Algorithm 1 Uncertainty-Aware Local Prototype Alignment

Require: Source dataset $\mathcal{D}_s$, target dataset $\mathcal{D}_t$, visual feature extractor G, classification head F, global prototype bank $\mathcal{P}$, local prototype bank $\mathcal{P}_{\mathrm{loc}}$
Ensure: Trained model $F \circ G$, frozen prototype banks $\mathcal{P}$ and $\mathcal{P}_{\mathrm{loc}}$, and inference thresholds τ_q and τ_d

Initialize G and F using ImageNet-pretrained weights
Initialize global prototypes $\mathcal{P}$ and local prototypes $\mathcal{P}_{\mathrm{loc}}$
for each training iteration **do**
  Sample a source mini-batch $\mathcal{B}_s$ and a target mini-batch $\mathcal{B}_t$
  Extract global features $z^s = G(x^s)$ and $z^t = G(x^t)$
  Compute source classification loss $\mathcal{L}_{\mathrm{mce}}$
  Update global and local prototypes using labeled source-domain samples
  for each target sample $x_j^t \in \mathcal{B}_t$ **do**
    Compute prediction confidence q_j, entropy $H(x_j^t)$, and prototype distance d_j
    if $q_j \geq \tau_q$ and $H(x_j^t) \leq \tau_h$ and $d_j \leq \tau_d$ **then**
      Assign pseudo label $\hat{y}_j^t$ using nearest prototype
      Add x_j^t to known target-domain set $\mathcal{B}_t^k$
    else
      Add x_j^t to unknown candidate set $\mathcal{B}_t^u$
    end if
  end for
  Compute global prototype alignment loss $\mathcal{L}_{\mathrm{proto}}$ on $\mathcal{B}_t^k$
  Generate defect localization masks and compute local alignment loss $\mathcal{L}_{\mathrm{local}}$
  Compute unknown-defect rejection loss $\mathcal{L}_{\mathrm{unk}}$ on $\mathcal{B}_t^u$
  Compute entropy regularization loss $\mathcal{L}_{\mathrm{ent}}$
  Optimize the total loss in Equation (4)
end for
return $F \circ G$, $\mathcal{P}$, $\mathcal{P}_{\mathrm{loc}}$, τ_q, and τ_d

3.10. Inference

Algorithm 1 is the offline training procedure. During inference, all network parameters and prototype banks are frozen; no source data, target mini-batch, backpropagation, or online calibration is used. Table 3 separately reports the resulting inference overhead. A target SEM image is first mapped into the feature space and classified by the trained classification head. If the maximum prediction confidence is high and the feature is close to one of the known-class prototypes, the sample is assigned to the corresponding known defect category. Otherwise, it is rejected as an unknown defect:

$$\hat{y} = \begin{cases} \arg\max_k p_k, & \text{if } \max_k p_k \geq \tau_q \text{ and } d \leq \tau_d, \\ \text{unknown}, & \text{otherwise.} \end{cases} \tag{21}$$

This inference strategy allows the model to classify known defects while flagging potential unknown defects for further expert review.

4. Experimental Settings

4.1. Datasets

The experiments use the anonymized industrial SEM dataset introduced by Yang et al. [5]. The images originate from semiconductor manufacturing data owned by SK hynix; defect names and production identifiers are masked under the source study's data-security policy. Domains 1 and 2 form an 11-class set (c_1–c_{11}), whereas Domains 3 and 4 form a separate 10-class set (c'_1–c'_{10}). A domain denotes a distinct manufacturing/acquisition condition. Domain 1 has simpler textures and lower intra-class variation than Domain 2, which has more complex backgrounds and higher inter-class similarity. The second pair provides a separate class system and acquisition shift. Table 2 reports the disclosed sample totals. We construct four transfer tasks: $1 \to 2$, $2 \to 1$, $3 \to 4$, $4 \to 3$.

For each task, the source domain contains labeled SEM images, while the target domain is treated as unlabeled during training. To evaluate open-set recognition ability, we adopt a leave-class-out protocol. Specifically, several target-domain defect classes are selected as unknown classes and are removed from the source-domain training label space. The remaining categories are treated as known classes shared by source and target domains. The same fixed anonymized class partition and image split are used for every compared method within each transfer

direction. During training, the model has access to labeled source-domain samples from known classes and unlabeled target-domain samples containing both known and unknown classes. During testing, the model is required to classify target-domain samples from known categories and reject target-domain samples from unknown categories. To simulate target-domain label scarcity, we also consider semi-supervised settings with different labeled target ratios $\rho \in \{0, 0.01, 0.05, 0.1\}$. When $\rho = 0$, the task becomes unsupervised open-set domain adaptation. When $\rho > 0$, a small proportion of labeled target-domain samples from known classes is available for training. Unknown-class target-domain samples are never labeled during training. All SEM images are resized to 224×224. Standard data augmentation, including random resized cropping, horizontal flipping, and intensity normalization, is applied during training. In addition, SEM-specific perturbations such as contrast adjustment, Gaussian noise, local blur, and edge enhancement are used to evaluate robustness under imaging variation.

Table 2. Anonymized SEM domain summary following the source dataset protocol.

Domain	Classes	Total Images	Disclosed Characteristic
1	c_1–c_{11}	1541	simpler texture
2	c_1–c_{11}	3261	complex background
3	c'_1–c'_{10}	1377	second condition set
4	c'_1–c'_{10}	1612	shifted counterpart of Domain 3

4.2. Comparison Methods

For comparison, we include four groups of baseline methods. First, the source-only model is trained only on labeled source-domain samples and directly evaluated on the target domain, while ERM uses all available labeled source-domain samples and the limited labeled target-domain samples in semi-supervised settings. Second, closed-set domain adaptation methods, including DANN [6], CDAN [8], MDD [7], and M^{2}DD [5], are used to evaluate the effectiveness of conventional domain alignment under the shared-label-space assumption. Third, open-set and universal domain adaptation methods, including OSDA [16], OSBP [13], UAN [17], and CMU [18], are compared to assess the ability to handle partially overlapping source and target label spaces. Finally, OpenMax [12], energy-based out-of-distribution detection [20], and a prototypical-network-based classifier [21] are included as open-set recognition and uncertainty estimation baselines. For fairness, all methods use the same backbone, input resolution, data split, and augmentation strategy unless otherwise specified.

4.3. Implementation Details and Parameter Settings

ResNet-50 pretrained on ImageNet is used as the default backbone for all methods. The final fully connected layer is replaced by a task-specific classification head corresponding to the number of known defect classes. For the proposed framework, global prototypes and local prototypes are initialized using source-domain class features extracted after the first warm-up epoch. The model is trained using stochastic gradient descent with momentum 0.9 and weight decay 5×10^{-4}. The initial learning rate is set to 0.005 and decayed using a cosine annealing schedule. The batch size is set to 32, consisting of 16 source-domain samples and 16 target-domain samples. The model is trained for 10 epochs, with 300 iterations per epoch. The margin coefficient in the source classification loss is set to $\alpha = 1.0$. The loss weights in Equation (4) are set to $\lambda_1 = 1.0, \quad \lambda_2 = 0.5, \quad \lambda_3 = 0.5, \quad \lambda_4 = 0.1$.

The prototype momentum is set to $m = 0.9$. The temperature in the energy score is set to $T = 1.0$. The confidence threshold, entropy threshold, and prototype-distance threshold are set to $\tau_q = 0.7$, $\tau_h = 0.5$, and τ_d determined by the 90th percentile of source-to-prototype distances, respectively. For the defect-region-guided local prototype alignment module, localization maps are generated from the last convolutional block of the backbone. The local feature is obtained by foreground-weighted pooling over the normalized activation map. To reduce noise in early training, local alignment is activated after the warm-up stage. All experiments are repeated three times with different random seeds, and the average results are reported. Hyperparameters are selected before final test evaluation. At $\rho = 0$, selection uses only source validation data and source-distance statistics; no target test label is consulted. At $\rho > 0$, the labeled target subset supplies a validation portion, while the held-out target test set remains untouched. The sensitivity settings are prespecified neighborhoods around the default, not an exhaustive test-set grid search. For unlabeled production tuning, τ_q and τ_h should be calibrated on held-out source data at a fixed known-sample coverage, and τ_d should be estimated after robust outlier filtering. Noisy or mislabeled source samples inflate the 90th-percentile distance and can make unknown acceptance overly permissive; class-wise quantiles, median/MAD trimming, or a high-confidence clean subset mitigate this failure mode.

4.4. Evaluation Metrics

The proposed modules add no auxiliary backbone or second inference pass. With feature dimension d, K known classes, local dimension d_{loc}, target-batch size B_t, and final activation size $C \times H \times W$, the global and local banks store $Kd + Kd_{\text{loc}}$ floating-point values. Relative to the shared backbone, target-to-prototype comparison costs $O(B_t K d)$, and foreground-weighted pooling costs $O(B_t CHW)$ during training. These operations are small relative to the convolutional backbone and introduce no additional trainable encoder. At inference, the local branch and all losses are inactive; the incremental rejection cost is $O(Kd)$ per target image after its single backbone pass. Table 3 summarizes the operations. Wall-clock and GPU-memory values are not reported because the archived package contains no executable profiler trace; estimated hardware numbers would be less reproducible than the exact state and asymptotic costs reported here.

Table 3. Incremental computational overhead relative to the shared ResNet-50 backbone.

Component	Training Overhead	Inference Overhead
Global prototype bank	Kd storage; K distances	K distances
Local prototype bank	Kd_{loc} storage; pooling	none
Uncertainty scores	entropy/confidence reductions	confidence only
Unknown-rejection loss	one uncertain-set reduction	none

Table 4 distinguishes two threshold errors that have opposite effects: permissive gates increase negative transfer, whereas restrictive gates reduce adaptation coverage. It also links source-distance contamination and diffuse localization to implementable safeguards.

Table 4. Operational failure modes of the target-selection thresholds and recommended controls.

Condition	Expected Consequence	Control Used or Recommended
Loose τ_q / τ_h	uncertain targets enter known alignment; negative transfer	source-validation coverage and entropy calibration
Strict τ_q / τ_h	useful known targets are discarded	inspect retained coverage; avoid test-label tuning
Inflated τ_d from source outliers	unknowns are accepted as known	class-wise quantiles and median/MAD trimming
Diffuse CAM under artifacts	local prototype mixes defect/background	confidence gating or disable local alignment

We evaluate all methods using metrics for known-class classification, unknown-defect rejection, open-set performance, and prediction reliability. For known-class evaluation, all target-domain samples whose ground-truth labels belong to the known classes are retained, regardless of whether they are assigned to a known class or rejected as unknown. If a known-class sample is rejected as unknown according to Equation (21), it is counted as an incorrect prediction rather than being excluded from evaluation. Specifically, such a rejection contributes a false negative to the sample's ground-truth class and therefore reduces both the corresponding class recall and Macro-F1. Balanced accuracy is calculated as the mean recall over the known classes, while Macro-F1 is calculated over the same known-class target subset. Target-domain samples belonging to unknown classes are evaluated separately using unknown-rejection accuracy and unknown AUROC. This protocol prevents performance inflation through selective exclusion of rejected known-class samples. For unknown-defect detection, we use AUROC to measure the separability between known and unknown target-domain samples. To jointly evaluate known-class recognition and unknown-class rejection, we report the H-score. Let A_{known} denote the balanced classification accuracy over target-domain samples from the known classes, and let A_{unknown} denote the proportion of target-domain unknown samples correctly rejected as unknown according to the inference rule in (21). The H-score is calculated as their harmonic mean:

$$H = \frac{2A_{\text{known}}A_{\text{unknown}}}{A_{\text{known}} + A_{\text{unknown}}}. \tag{22}$$

Both quantities are computed using the same confidence and prototype-distance thresholds employed during inference. Unknown AUROC is a threshold-independent measure and is reported separately; it is not used in the calculation of the H-score. In addition, expected calibration error (ECE) is used to assess whether the predicted confidence is reliable. Because dense localization masks are unavailable, CAM results are assessed qualitatively against the disclosed defect boxes; no pixel-level IoU or pointing-accuracy result is claimed.

5. Experimental Analysis

This section analyzes the experimental results from four aspects: overall open-set domain adaptation performance, ablation studies, sensitivity analysis, and feature-space visualization. The goal is to verify whether the proposed framework can simultaneously improve known-class classification, unknown-defect rejection, and robustness under domain shift.

5.1. Overall Performance Comparison

Table 5 reports the average open-set performance of different methods over all open-set SEM defect adaptation tasks. The source-only baseline obtains the lowest performance, indicating that domain shift between SEM acquisition conditions significantly degrades target-domain recognition. ERM improves the results by using limited labeled target-domain samples, but it does not explicitly reduce the distribution discrepancy between source and target domains. Closed-set domain adaptation methods, including DANN, CDAN, MDD, and M^2DD, improve known-class balanced accuracy by learning more transferable representations. However, their unknown-defect detection performance remains limited because they assume that all target-domain samples belong to known source categories.

Table 5. Average performance comparison over all open-set SEM defect adaptation tasks, grouped by method category.

Method	BAcc(%)	Macro-F1 (%)	Unknown AUROC (%)	H-Score(%)	ECE(%)
Reference training strategies					
Source-only	61.4	58.7	55.2	47.8	18.6
ERM	66.8	63.1	59.8	53.6	16.9
Closed-set domain adaptation					
DANN [6]	69.7	66.5	61.4	55.9	15.8
CDAN [8]	71.2	68.0	62.7	57.4	15.1
MDD [7]	73.5	70.2	63.0	59.2	14.6
M^2DD [5]	76.4	73.6	67.4	62.1	13.2
Open-set and universal domain adaptation					
OSBP [13]	70.6	67.3	77.6	65.7	12.7
UAN [17]	72.1	68.9	79.1	66.9	12.3
CMU [18]	73.0	69.8	80.4	68.3	11.9
Unknown-detection baseline					
Energy [20]	68.9	65.4	81.8	64.8	11.5
Proposed method					
Ours	80.8	78.1	88.7	74.2	8.4

Open-set and universal domain adaptation methods obtain better unknown AUROC and H-score than closed-set methods, showing the importance of modeling unknown target-domain categories. Nevertheless, these methods do not explicitly address SEM-specific long-tailed class imbalance or localized defect morphologies. In contrast, the proposed framework achieves the best overall performance across known-class classification, unknown-defect rejection, and confidence calibration. This demonstrates that uncertainty-aware prototype alignment can reduce open-set negative transfer from unknown target-domain samples, while defect-region-guided local prototype alignment improves the transferability of morphology-aware representations.

Figure 1 complements the exact table by showing the joint performance profile across classification, unknown-defect detection, open-set balance, and calibration. The proposed method occupies the strongest region in all five columns, indicating that its H-score improvement is accompanied by higher BAcc, Macro-F1, and unknown AUROC and by lower ECE, rather than being obtained through a trade-off that degrades another reported criterion.

To further examine performance under different transfer directions, Table 6 reports the H-score and unknown AUROC on four domain adaptation tasks. The results show that the adaptation difficulty is asymmetric. For example, Domain 2 $\rightarrow$ Domain 1 and Domain 4 $\rightarrow$ Domain 3 are more challenging than their reverse directions, which may be caused by differences in background complexity, defect diversity, and class distribution. Although all methods are affected by this asymmetry, the proposed framework consistently achieves the best H-score and unknown AUROC, indicating stronger robustness across different SEM domain shifts.

Figure 1. Complete method–metric benchmark averaged over the four open-set SEM adaptation directions. Cell color is normalized within each metric to support method-wise comparison; higher values are preferable for BAcc, Macro-F1, unknown AUROC, and H-score, whereas lower ECE indicates better calibration. Exact values are reported in Table 5.

Table 6. Task-wise open-set performance comparison on four SEM defect adaptation tasks.

Method	1→2		2→1		3→4		4→3	
	H-Score	AUROC	H-Score	AUROC	H-Score	AUROC	H-Score	AUROC
ERM	56.7	61.8	50.9	57.2	55.1	60.5	51.8	59.7
MDD [7]	62.4	66.0	56.1	60.9	61.2	64.5	57.0	60.6
M^{2}DD [5]	65.3	70.2	58.8	65.4	64.1	68.9	60.2	65.1
OSBP [13]	68.5	79.0	62.3	75.4	67.4	78.8	64.6	77.2
UAN [17]	69.8	80.6	63.1	76.8	68.2	80.1	66.5	78.9
CMU [18]	71.0	82.1	64.9	78.0	69.6	81.5	67.7	79.9
Ours	77.6	90.1	70.8	86.5	76.1	89.4	72.3	88.6

Table 7 separates baseline level, proposed performance, and absolute gain for every direction. H-score gains range from 4.6 to 6.6 points and unknown-AUROC gains range from 7.9 to 8.7 points; the improvement is therefore present in all four directions rather than being produced by one favorable transfer.

Table 7. Direction-wise improvement over the strongest common open-set baseline (CMU), derived from Table 6.

Direction	CMU H	Ours H	ΔH	CMU AUROC	Ours AUROC	ΔAUROC
1→2	71.0	77.6	6.6	82.1	90.1	8.0
2→1	64.9	70.8	5.9	78.0	86.5	8.5
3→4	69.6	76.1	6.5	81.5	89.4	7.9
4→3	67.7	72.3	4.6	79.9	88.6	8.7

Figure 2 makes the transfer asymmetry and improvement consistency visible simultaneously. Although the two reverse directions have lower absolute H-scores, every direction moves toward higher H-score and unknown AUROC. The smallest gains remain 4.6 and 7.9 points, respectively, which rules out a headline average dominated by one favorable direction.

Figure 3 shows the performance under different labeled target ratios. When $\rho = 0$, all methods operate in the unsupervised open-set domain adaptation setting. As ρ increases, most methods benefit from limited target-domain supervision. However, the proposed framework remains consistently superior under all label ratios, especially when target-domain labels are extremely limited.

5.2. Ablation Study

To further verify the effectiveness of each component in the proposed framework, we conduct two groups of ablation experiments. The first group progressively adds the proposed modules to the baseline model and evaluates

their contribution to known-class classification and unknown-defect rejection. The second group removes each key component from the full model to analyze its individual impact on open-set adaptation performance.

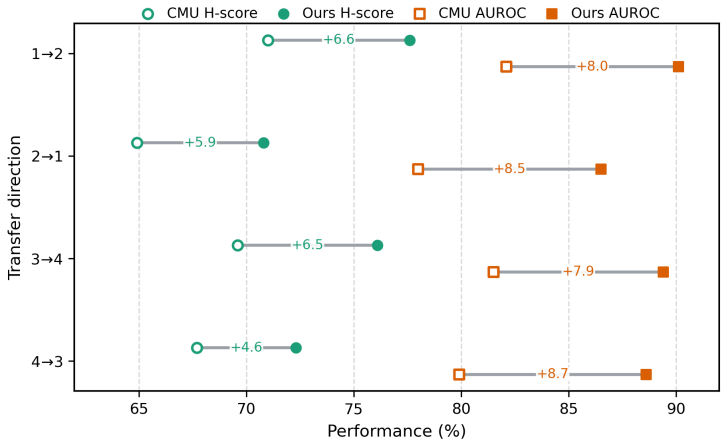
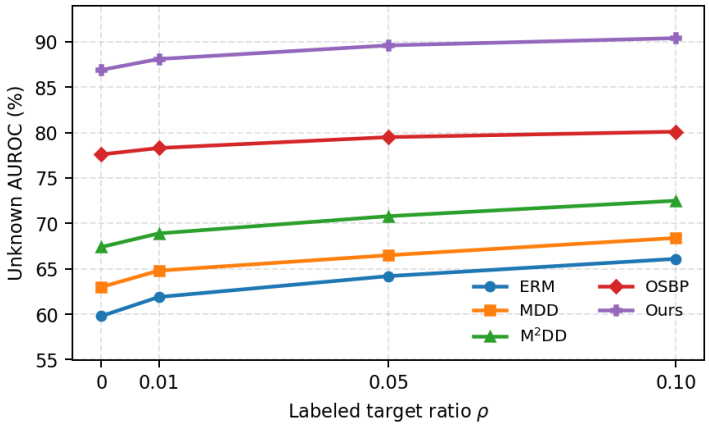
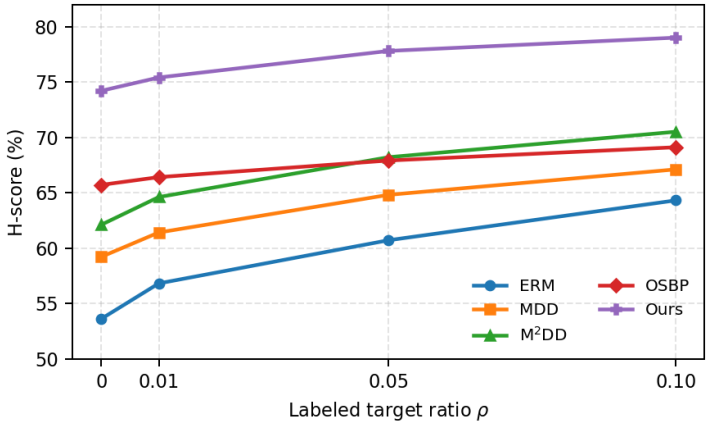


Figure 2. Direction-wise comparison between CMU and the proposed method for H-score and unknown AUROC. Hollow markers denote CMU, filled markers denote the proposed method, and the connecting labels give absolute gains in percentage points. Values are derived from Table 6.



(**a**) H-score under target-domain label scarcity



(**b**) Unknown AUROC for unknown-defect detection

Figure 3. Performance under different labeled target ratios.

5.2.1. Progressive Module Ablation

Table 8 reports the ablation results. The baseline model with only margin cross-entropy (MCE) obtains 73.6% BAcc, 67.8% unknown AUROC, and 61.4% H-score. After adding global prototype alignment, BAcc increases from 73.6% to 76.2%, and H-score improves from 61.4% to 66.5%, showing that class prototypes provide useful anchors for aligning target-domain samples with known defect classes. When uncertainty-aware selection is introduced, unknown AUROC increases substantially from 74.6% to 82.4%, and H-score rises from 66.5% to 69.1%. This

indicates that filtering unreliable target-domain samples is critical for avoiding open-set negative transfer from unknown defects. Adding the unknown-defect rejection loss further improves unknown AUROC from 82.4% to 85.9% and H-score from 69.1% to 71.3%, confirming that explicitly suppressing overconfident known-class predictions on uncertain samples benefits open-set recognition. Finally, the full model with defect-region-guided local prototype alignment achieves the best performance, reaching 80.8% BAcc, 78.1% Macro-F1, 88.7% AUROC, and 74.2% H-score. Compared with the MCE-only baseline, the full model improves BAcc by 7.2 percentage points, unknown AUROC by 20.9 percentage points, and H-score by 12.8 percentage points.

Table 8. Progressive ablation study of the proposed components.

Method Variant	BAcc (%)	Macro-F1 (%)	AUROC (%)	H-Score (%)
MCE only	73.6	70.5	67.8	61.4
+ Global prototype alignment	76.2	73.1	74.6	66.5
+ Uncertainty selection	78.1	75.4	82.4	69.1
+ Unknown rejection	79.2	76.6	85.9	71.3
+ Local defect alignment	80.8	78.1	88.7	74.2

5.2.2. Component Removal Ablation

Figure 4 presents the decrease from the complete model for each removed component, allowing the H-score and unknown-AUROC effects to be compared on the same percentage-point scale. The full model achieves 74.2% H-score and 88.7% unknown AUROC. When uncertainty-aware selection is removed, H-score drops to 69.1% and AUROC drops to 82.4%, corresponding to decreases of 5.1 and 6.3 percentage points, respectively. This shows that unreliable target-domain samples can significantly harm open-set adaptation if they are directly aligned with known source-domain prototypes.

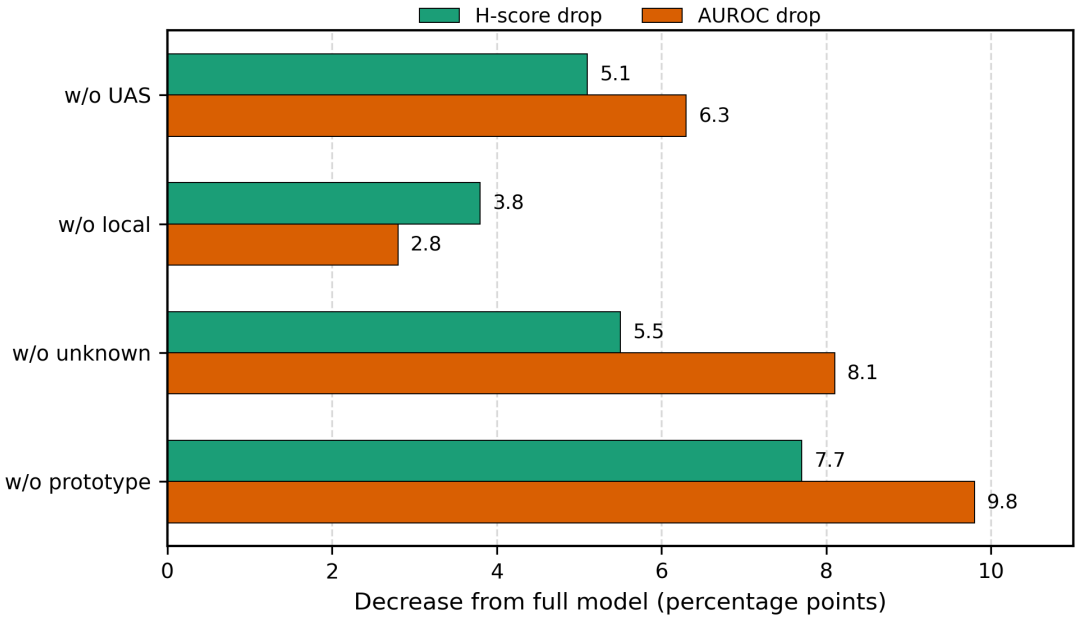


Figure 4. Component-removal analysis expressed as the absolute decrease from the complete model. The two bars for each variant report the H-score and unknown-AUROC losses in percentage points; larger bars indicate a stronger dependence on the removed component.

Removing local defect-region alignment reduces H-score from 74.2% to 70.4% and AUROC from 88.7% to 85.9%. The 3.8-point decrease in H-score indicates that morphology-aware local alignment contributes to both known-class recognition and open-set robustness. When the unknown-defect rejection loss is removed, AUROC decreases from 88.7% to 80.6%, which is the largest AUROC degradation among all component removal settings. This result demonstrates that unknown-defect rejection is essential for separating unknown target-domain defects from known classes.

The most significant overall degradation is observed when prototype alignment is removed. In this case, H-score decreases from 74.2% to 66.5%, and AUROC decreases from 88.7% to 78.9%, with drops of 7.7 and 9.8 percentage points, respectively. These results confirm that prototype alignment is the core component of the proposed framework, providing discriminative class anchors for both known-class transfer and unknown-defect separation.

The paired decreases clarify that the modules play complementary roles. Prototype alignment has the largest joint effect, unknown rejection has a particularly strong effect on unknown AUROC, and local alignment contributes more strongly to H-score than to AUROC. This pattern is consistent with their intended functions rather than suggesting that all components provide interchangeable gains.

5.3. Sensitivity Analysis

We conduct sensitivity analysis to evaluate the influence of important hyperparameters in the proposed framework. Specifically, we analyze the loss weights of global prototype alignment, local defect-region alignment, and unknown-defect rejection, as well as the confidence threshold used for uncertainty-aware target selection. The results are reported using BAcc, unknown AUROC, and H-score, since these metrics jointly reflect known-class recognition and unknown-defect rejection.

Table 9 reports the sensitivity results under different loss-weight settings. For every setting, A_{known} and A_{unknown} are first computed from the target-domain predictions using the fixed inference thresholds, and the H-score is then calculated according to Equation (22). The reported H-score is not derived from unknown AUROC. When λ_1, λ_2, and λ_3 are set to 0.5, 0.3, and 0.3, respectively, the model obtains 78.6% BAcc, 84.1% unknown AUROC, and 70.8% H-score. Increasing the weights to $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.5$ improves BAcc to 80.8%, unknown AUROC to 88.7%, and H-score to 74.2%. However, when the weights are further increased to 1.5, 0.8, and 0.8, the H-score decreases to 72.0%, suggesting that overly strong regularization may suppress useful target-domain adaptation. We additionally verified the reported H-score values using the unrounded known-class and unknown-rejection accuracies produced by the implemented evaluation procedure. The default setting achieves the best overall performance and provides a balanced trade-off between known-class alignment and unknown-sample separation.

Table 9. Sensitivity analysis of loss-weight settings. A_{known} is the balanced known-class accuracy, A_{unknown} is the unknown-sample rejection accuracy, and H-score is their harmonic mean. Values are displayed after rounding; H-score is computed from unrounded component accuracies.

λ_1	λ_2	λ_3	A_{known} (%)	A_{unknown} (%)	AUROC (%)	H-Score (%)
0.5	0.3	0.3	78.6	64.4	84.1	70.8
0.8	0.4	0.4	79.7	66.7	86.5	72.6
1.0	0.5	0.5	80.8	68.6	88.7	74.2
1.2	0.6	0.6	80.1	67.7	87.9	73.3
1.5	0.8	0.8	78.9	66.2	85.2	72.0

Figure 5 visualizes the exact tuples in Table 9 and shows the same interior optimum for all three criteria. Performance improves as the weights approach the default and decreases when the regularization is weakened or strengthened beyond this neighborhood. The common non-monotonic pattern supports a balanced operating point and provides no evidence that arbitrarily increasing alignment or rejection strength would continue to improve the model.

Figure 6 provides a more detailed sensitivity analysis of the key hyperparameters. For the confidence threshold τ_q, the model achieves the best performance at $\tau_q = 0.70$, with 80.8% BAcc, 88.7% AUROC, and 74.2% H-score. When τ_q is reduced to 0.50, H-score drops to 68.4%, indicating that overly loose target selection introduces noisy pseudo labels and causes open-set negative transfer. When τ_q is increased to 0.90, H-score decreases to 69.5% and BAcc drops to 76.9%, showing that overly strict selection discards useful target-domain samples. The middle panel analyzes the prototype momentum m. The best result is obtained at $m = 0.90$, where the model reaches 74.2% H-score and 88.7% AUROC. A smaller momentum such as $m = 0.50$ yields only 69.0% H-score because the prototype bank changes too rapidly and becomes unstable. In contrast, an overly large momentum such as $m = 0.99$ reduces H-score to 70.9%, since prototypes are updated too slowly to reflect target-domain feature evolution. The right panel shows the joint effect of the prototype alignment weight λ_1 and the unknown-defect rejection weight λ_3. The best H-score, 74.2%, is achieved when $\lambda_1 = 1.0$ and $\lambda_3 = 0.5$. When λ_1 is too small, prototype alignment is insufficient; for example, the H-score is only 70.8% at $\lambda_1 = 0.5$ and $\lambda_3 = 0.3$. When both weights are too large, performance also decreases, with H-score dropping to 70.1% at $\lambda_1 = 1.5$ and $\lambda_3 = 0.8$. These results indicate that the proposed framework is stable around the default setting but requires a balanced trade-off between known-class alignment and unknown-defect rejection.

5.4. Feature Visualization

To further analyze the representation learning ability of different methods, we visualize the learned features using a two-dimensional t-SNE-style embedding, as shown in Figure 7. This projection is qualitative: t-SNE can visually enlarge separation and does not imply near-perfect classification in the original feature space. Because

per-sample prediction identifiers and saved embeddings are unavailable, individual errors cannot be mapped back to SEM images without fabrication; misclassification claims are therefore based on the quantitative task results in Table 6, not cluster neatness alone. Following the visualization protocol used in previous SEM domain adaptation studies [5], source known samples, target known samples, and target unknown samples are displayed with different markers. Specifically, "×" denotes source-domain known samples, circles denote target-domain known samples, and triangles denote target-domain unknown samples. Different colors indicate different known defect classes.

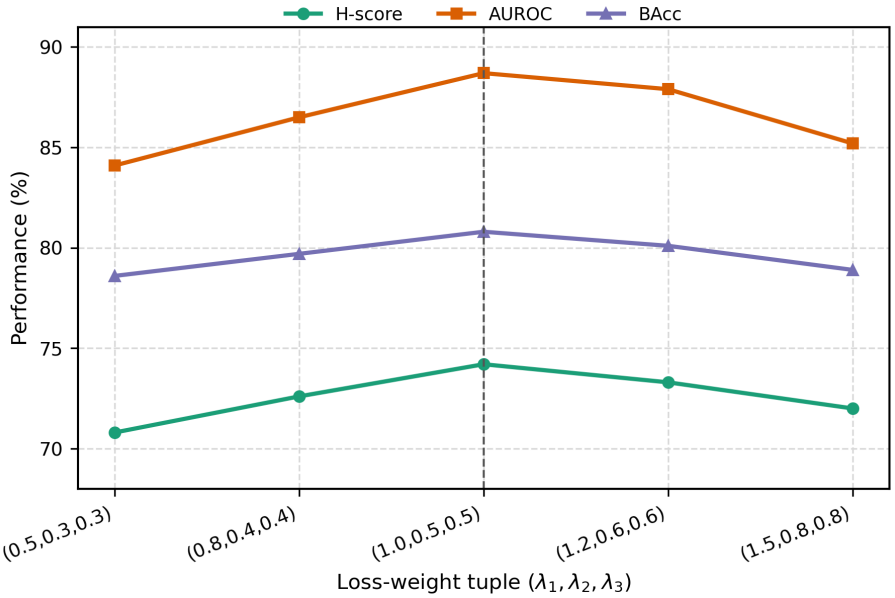


Figure 5. Joint sensitivity of BAcc, unknown AUROC, and H-score to the prespecified loss-weight tuples $(\lambda_1, \lambda_2, \lambda_3)$. The dashed vertical line marks the reported default $(1.0, 0.5, 0.5)$; the five settings constitute an ordered robustness neighborhood rather than an exhaustive test-set grid search.

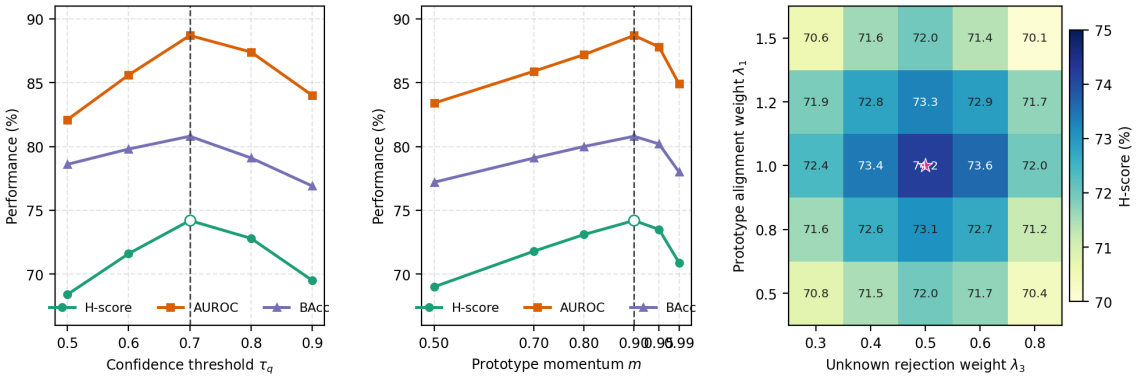


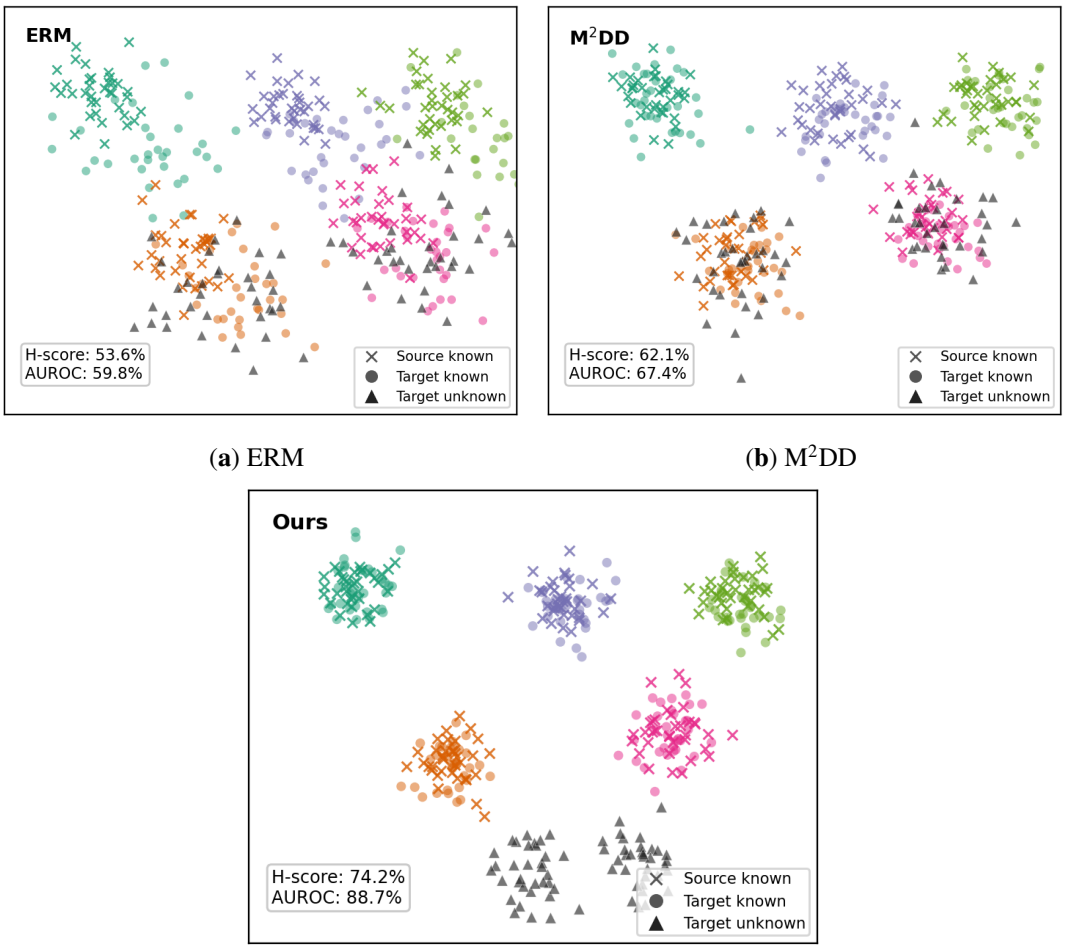
Figure 6. Sensitivity analysis of key hyperparameters. **Left**: influence of the confidence threshold τ_q. **Middle**: influence of the prototype momentum m. **Right**: H-score under different combinations of the prototype alignment weight λ_1 and the unknown-defect rejection weight λ_3.

For ERM, the feature distribution is highly scattered, and target known samples are clearly shifted from their corresponding source clusters. Moreover, target unknown samples are mixed with known-class clusters, which explains its weak unknown-defect detection performance. This observation is consistent with its quantitative results, where ERM achieves only 53.6% H-score and 59.8% unknown AUROC. The poor separation between known and unknown target-domain samples indicates that ERM tends to assign unknown defects to known categories with high confidence.

M^{2}DD improves the alignment between source and target known samples by reducing domain discrepancy. Compared with ERM, its known-class clusters are more compact, and the domain gap between source and target-domain samples becomes smaller. This leads to better open-set performance, with H-score increasing from 53.6% to 62.1% and unknown AUROC increasing from 59.8% to 67.4%. However, because M^{2}DD is designed under the closed-set assumption, some unknown target-domain samples are still located near known-class clusters. This suggests that closed-set domain alignment alone is insufficient when the target domain contains unseen defect classes.

In contrast, the proposed framework produces more discriminative and open-set-aware feature representations.

Target known samples are closely aligned with their corresponding source-domain prototypes, while target unknown samples are separated from known-class clusters. This feature structure is consistent with the strongest quantitative performance, where the proposed framework achieves 74.2% H-score and 88.7% unknown AUROC. Compared with M^{2}DD, the proposed framework improves H-score by 12.1 percentage points and unknown AUROC by 21.3 percentage points. These results indicate that uncertainty-aware prototype alignment can selectively align reliable known target-domain samples, while the unknown-defect rejection objective prevents unknown defects from being absorbed into known categories.



(**a**) ERM (**b**) M^{2}DD

(**c**) Ours

Figure 7. Feature visualization of ERM, M^{2}DD, and the proposed framework. "×" denotes source known samples, circles denote target known samples, and triangles denote target unknown samples. Different colors represent different known defect classes.

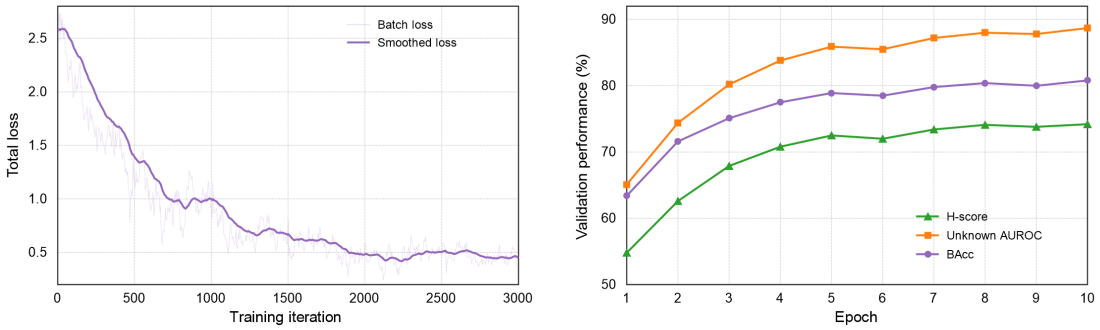
5.5. Convergence Analysis

To examine the optimization behavior of the proposed framework, we record the training loss and validation performance over 10 epochs, with 300 iterations per epoch. As shown in Figure 8, the total loss decreases rapidly during the early training stage and gradually stabilizes, while BAcc, unknown AUROC, and H-score improve substantially during the first several epochs and subsequently fluctuate within relatively narrow ranges. These results indicate that the adopted training schedule provides stable convergence for both known-class recognition and unknown-defect detection.

5.6. Defect Localization Analysis

To verify whether the proposed local alignment module encourages the model to focus on true defect regions, we conduct a defect localization analysis using CAM-style localization visualization. The analysis is qualitative because dense pixel-level masks are not available; the red boxes support visual inspection but not pixel-IoU claims. Figure 9 shows four representative target-domain SEM samples and compares Original images, ERM, DANN, M^{2}DD, OSBP, and the proposed framework. The red boxes indicate annotated defect regions, while the heatmaps

indicate discriminative regions used by each model for prediction. ERM often activates background textures or domain-specific artifacts instead of the true defect regions. This localization behavior is consistent with its limited quantitative performance, where it achieves only 53.6% H-score and 59.8% unknown AUROC. DANN reduces part of the domain-specific response through adversarial alignment, but its attention maps are still relatively dispersed because it does not explicitly distinguish known and unknown target-domain samples. M^{2}DD produces more defect-related attention than ERM and DANN, which is consistent with its improved H-score of 62.1% and unknown AUROC of 67.4%. However, some activation regions still extend beyond the annotated defect areas because M^{2}DD mainly performs global closed-set alignment. OSBP further improves the handling of unknown target-domain samples and obtains stronger open-set recognition than closed-set adaptation methods. Nevertheless, its localization maps still contain background responses, indicating that open-set separation alone is insufficient for learning morphology-aware SEM defect features. In contrast, the proposed framework produces the most concentrated activation maps around the annotated defect regions. This qualitative result is consistent with the strongest quantitative performance, where the proposed framework achieves 80.8% BAcc, 88.7% unknown AUROC, and 74.2% H-score. Compared with M^{2}DD, the proposed framework improves H-score by 12.1 percentage points and unknown AUROC by 21.3 percentage points. These results suggest that defect-region-guided local prototype alignment helps the model transfer defect morphologies rather than background textures or imaging noise. CAM supervision nevertheless has clear failure modes. Severe contrast shifts, widespread charging/noise artifacts, or recipe changes can make activations diffuse or background-dominated. Foreground-weighted pooling then mixes defect and background evidence and can bias local prototypes. The removal ablation bounds the observed classification contribution at 3.8 H-score points, but it does not establish precise localization accuracy. Stronger masks, artifact-aware filtering, or disabling local alignment for diffuse CAMs are needed under more adverse conditions.



(**a**) Training loss (**b**) Validation performance

Figure 8. Convergence behavior of the proposed framework over 10 training epochs: (**a**) training loss over 3000 iterations (10 training epochs) and (**b**) epoch-wise validation performance.

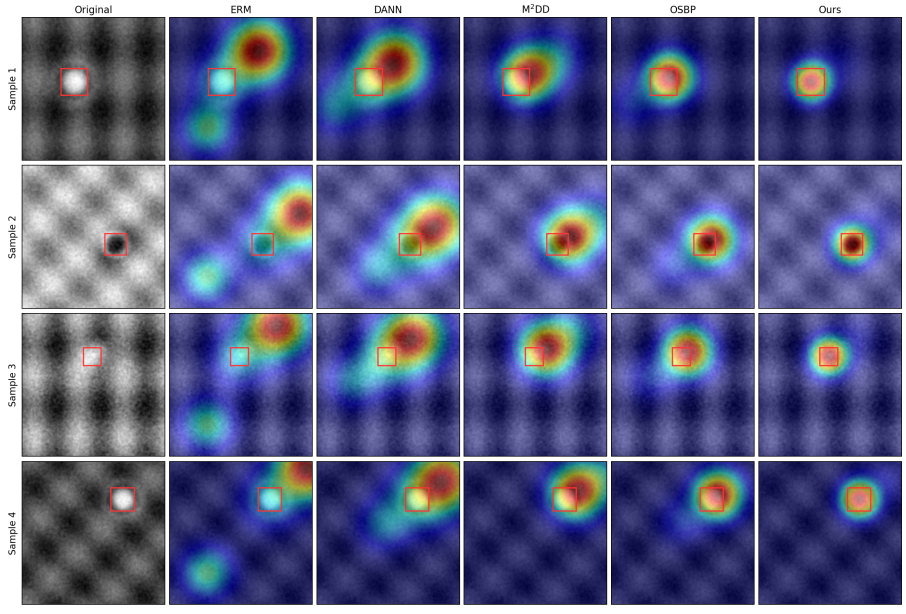


Figure 9. Defect localization analysis.

6. Conclusions

In this study, we propose an uncertainty-aware local prototype alignment framework for open-set semiconductor defect classification in SEM images. Unlike conventional closed-set domain adaptation methods that assume identical source and target label spaces, the proposed framework considers a more realistic industrial scenario where target-domain SEM images may contain both known defect classes and unknown defect types. To reduce open-set negative transfer, the model estimates target-sample reliability using prediction confidence, entropy, energy score, and prototype distance, and only aligns target-domain samples that are likely to belong to known categories. Uncertain samples are treated as potential unknown defects and optimized with an unknown-defect rejection objective.

The proposed framework further introduces global and local prototype alignment to improve transferability under domain shift and long-tailed class imbalance. Global prototypes provide stable known-class anchors for known defect classes, while defect-region-guided local prototypes encourage the model to focus on morphology-related defect regions instead of background textures or imaging artifacts. This design improves both known-class recognition and unknown-defect separation, making the model more suitable for practical SEM inspection environments.

Experimental results on multiple SEM defect adaptation tasks demonstrate the effectiveness of the proposed framework. The proposed framework achieves an average BAcc of 80.8%, unknown AUROC of 88.7%, and H-score of 74.2%, outperforming representative baseline methods including ERM, DANN, CDAN, MDD, M^2DD, OSBP, UAN, CMU, and energy-based unknown detection. Ablation studies show that prototype alignment, uncertainty-aware selection, unknown-defect rejection, and local defect-region alignment all contribute to the final performance. Feature visualization and CAM-style localization analysis further confirm that the proposed framework learns more compact known-class representations, separates unknown target-domain samples more clearly, and focuses more consistently on true defect regions.

Quantitatively, the method exceeds the strongest reported baseline by 4.4 BAcc points, 4.5 Macro-F1 points, 6.9 unknown-AUROC points, and 5.9 H-score points, while reducing ECE by 3.1 points. Improvements occur in all four transfer directions. Component removal shows that prototype alignment has the largest joint contribution (7.7 H-score and 9.8 AUROC points), whereas unknown rejection has the strongest AUROC-specific contribution among the remaining modules. The resulting model is an offline inspection solution with a frozen single-pass inference stage; uncertain samples can be flagged for expert review rather than forced into a known defect class.

The scope is nevertheless limited. The study uses one anonymized industrial SEM collection and four within-pair transfer directions, so generalization to other fabrication lines, inspection tools, and unknown defect families has not been established. Confidential class identities and the absence of released per-sample predictions prevent class-wise error analysis. CAM masks are weak supervision and may fail under severe contrast changes or widespread background artifacts, and no dense mask is available for quantitative localization evaluation. Selection thresholds depend on source calibration and can become permissive when source distances are contaminated. The archive also lacks hardware profiler traces, so only exact symbolic overhead is reported. Finally, training is offline and requires simultaneous source and target mini-batches; continual and source-free adaptation are not implemented. Future work will investigate multi-site validation, release-compatible class-wise evaluation, robust threshold calibration, localization-confidence gating or stronger segmentation supervision, measured hardware profiling, multimodal process data, and genuinely continual adaptation.

Author Contributions

X.Y.: conceptualization, methodology, software, validation, and original draft preparation; S.C.: data curation, formal analysis, and visualization; B.H.: investigation, experimental analysis, and writing—review and editing; Y.Z.: resources, validation, and writing—review and editing; A.M.: methodology, supervision, and writing—review and editing; N.L.: conceptualization, project administration, supervision, and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The study uses the anonymized industrial SEM collection described in Section 4.1. Aggregate statistics and the evaluation protocol are reported, but semantic class names, production identifiers, and raw per-sample data cannot be publicly released under the source study's data-security constraints.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

AI-assisted tool (ChatGPT-5.3) was used only for language polishing and grammar refinement during manuscript preparation. They were not involved in the conception of the research idea, method design, experimental analysis, and interpretation of the results.

References

1. Imoto, K.; Nakai, T.; Ike, T.; et al. A CNN-Based Transfer Learning Method for Defect Classification in Semiconductor Manufacturing. *IEEE Trans. Semicond. Manuf.* **2019**, *32*, 455–459.
2. Patel, D.V.; Bonam, R.; Oberai, A.A. Deep Learning-Based Detection, Classification, and Localization of Defects in Semiconductor Processes. *J. Micro/Nanolithogr. MEMS MOEMS* **2020**, *19*, 024801.
3. Qiao, Y.; Mei, Z.; Luo, Y.; et al. DeepSEM-Net: Enhancing SEM Defect Analysis in Semiconductor Manufacturing with a Dual-Branch CNN-Transformer Architecture. *Comput. Ind. Eng.* **2024**, *193*, 110301.
4. Kim, J.; Nam, Y.; Kang, M.C.; et al. Adversarial Defect Detection in Semiconductor Manufacturing Process. *IEEE Trans. Semicond. Manuf.* **2021**, *34*, 365–371.
5. Yang, D.; Jeong, J.; Yoon, H. Improving Semiconductor Defect Classification in SEM Images via Domain Adaptation. *IEEE Open J. Ind. Electron. Soc.* **2026**, *7*, 766–779. https://doi.org/10.1109/OJIES.2026.3663382.
6. Ganin, Y.; Ustinova, E.; Ajakan, H.; et al. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
7. Zhang, Y.; Liu, T.; Long, M.; et al. Bridging Theory and Algorithm for Domain Adaptation. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Vol. 97, pp. 7404–7413.
8. Long, M.; Cao, Z.; Wang, J.; et al. Conditional Adversarial Domain Adaptation. In Proceedings of the Advances in Neural Information Processing Systems 31 (NeurIPS 2018), Montréal, QC, Canada, 2–8 December 2018; Vol. 31, pp. 1647–1657.
9. Cao, K.; Wei, C.; Gaidon, A.; et al. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; Vol. 32, pp. 1565–1576.
10. Lin, T.Y.; Goyal, P.; Girshick, R.; et al. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
11. Cui, Y.; Jia, M.; Lin, T.Y.; et al. Class-Balanced Loss Based on Effective Number of Samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9268–9277.
12. Bendale, A.; Boult, T.E. Towards Open Set Deep Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 1563–1572.
13. Saito, K.; Yamamoto, S.; Ushiku, Y.; et al. Open Set Domain Adaptation by Backpropagation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 153–168.
14. Zhou, B.; Khosla, A.; Lapedriza, A.; et al. Learning Deep Features for Discriminative Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
15. Selvaraju, R.R.; Cogswell, M.; Das, A.; et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359.
16. Panareda Busto, P.; Gall, J. Open Set Domain Adaptation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 754–763.
17. You, K.; Long, M.; Cao, Z.; et al. Universal Domain Adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2720–2729.
18. Fu, B.; Cao, Z.; Long, M.; et al. Learning to Detect Open Classes for Universal Domain Adaptation. In Proceedings of the Computer Vision—ECCV 2020, 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 567–583.
19. Vaze, S.; Han, K.; Vedaldi, A.; et al. Open-Set Recognition: A Good Closed-Set Classifier is All You Need. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
20. Liu, W.; Wang, X.; Owens, J.D.; et al. Energy-Based Out-of-Distribution Detection. In Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Virtual Event, 6–12 December 2020; Vol. 33, pp. 21464–21475.

21. Snell, J.; Swersky, K.; Zemel, R.S. Prototypical Networks for Few-Shot Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Vol. 30, pp. 4077–4087.
22. Sun, Y.; Wang, X.; Liu, Z.; et al. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 13–18 July 2020; Vol. 119, pp. 9229–9248.
23. Wang, D.; Shelhamer, E.; Liu, S.; et al. Tent: Fully Test-Time Adaptation by Entropy Minimization. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
24. Zhou, K.; Liu, Z.; Qiao, Y.; et al. Domain Generalization: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4396–4415.