WISD-AD: REAL WORLD SEMI SUPERVISED BENCHMARK FOR DEFECT DETECTION OF SOFT MAGNETIC WOUND INDUCTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing public datasets for industrial defect detection generally face two major limitations. Cross-domain benchmarks such as MVTec AD and VisA cover diverse but scattered categories and scenarios, which differ from practical product-level inspection needs, and performance on several commonly used benchmarks has gradually approached saturation. Although domain-specific datasets such as NEU-DET and GC10-DET are closer to real industrial scenarios, their data splits and evaluation protocols are usually inconsistent across studies, hindering fair comparison. More importantly, most existing datasets focus mainly on unsupervised or fully supervised settings, whereas a unified 2D benchmark for semi-supervised settings remains lacking. To address this gap, we propose WISD-AD, a real-world dataset for surface defect detection of soft-magnetic wound inductors. It contains five defect categories with practical challenges such as small defects, weak textures, and class imbalance, and supports fully and semi-supervised classification and detection with unified data splits and evaluation protocols. In addition, we further propose M^2D^3 as a benchmark framework for WISD-AD. Extensive experiments reveal that existing methods still struggle under sparse annotations, and further show that representation quality before self-training critically affects subsequent pseudo-label reliability. M^2D^3 substantially alleviates this limitation, providing a strong reference for annotation-efficient defect inspection.

1 INTRODUCTION

Existing industrial defect datasets have greatly promoted the development of defect detection algorithms. General-purpose industrial defect datasets, represented by MVTec AD Bergmann et al. (2019), VisA, Zou et al. (2022), and Real-IAD Wang et al. (2024), cover various industrial objects and have become widely used benchmarks in industrial inspection. However, these datasets are usually designed for cross-object or cross-domain evaluation, emphasizing the general defect perception ability of models rather than their stable detection capability for specific industrial products. When the research focus shifts to practical inspection of a specific product, such scenario datasets cannot fully capture the defect distributions, background textures, and process-related characteristics of real industrial applications. Meanwhile, the performance of many methods on common benchmarks such as MVTec AD has gradually approached saturation, making it increasingly difficult to distinguish their advantages and limitations in real industrial scenarios. In contrast, representative domain-specific defect detection datasets, such as NEU-DET Song & Yan (2013) and RSDD Niu et al. (2020), are more closely aligned with real-world scenarios. Nevertheless, they still have several limitations. First, some datasets, such as NEU-DET, are deliberately constructed with relatively balanced data scales and defect categories, making it difficult to capture the long-tailed defect distributions commonly observed in industrial production. Second, existing domain-specific datasets are typically tailored to a single task setting, such as fully supervised classification or object detection, and rarely provide unified support for multiple fine-grained tasks. In addition, the use of inconsistent data splits, annotation ratios, and evaluation protocols across studies makes fair comparison difficult.

More importantly, existing defect datasets mainly focus on either unsupervised defect detection or fully supervised learning. In real industrial scenarios, fully supervised methods require high-quality bounding-box annotations from domain experts, leading to high annotation costs and long annota-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

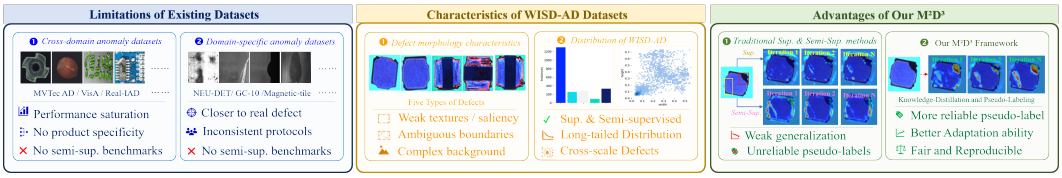


Figure 1: Overview of WISD-AD and M^2D^3 relative to representative industrial defect datasets and methods. WISD-AD supports fully and semi-supervised classification and detection, while M^2D^3 enables annotation-efficient defect recognition and localization.

tion cycles. Although unsupervised methods avoid labeled data, their accuracy and stability often fall short of practical requirements. In contrast, semi-supervised defect detection better matches industrial needs by leveraging limited labeled data and abundant unlabeled data. However, public 2D benchmarks specifically designed for semi-supervised industrial defect detection remain limited.

To mitigate these gaps, we propose WISD-AD, a real-world 2D dataset for surface defect detection of soft-magnetic wound inductors. WISD-AD targets practical product-level production-line inspection. It covers five typical surface defect categories of inductors, while exhibiting common challenges in defect detection, such as small defects, weak textures, strong background interference, and imbalanced class distributions. Unlike existing datasets designed for a single task or supervision paradigm, WISD-AD supports fully and semi-supervised defect classification and detection with unified data splits, annotation ratios, and evaluation protocols, enabling fair and reproducible evaluation across different annotation budgets. These controlled low-label settings also reveal that weak initial representations under extremely sparse annotations can lead to unreliable pseudo labels and error propagation during self-training. Motivated by this observation, we propose $\mathrm{M}^2\mathrm{D}^3$, which first transfers complementary knowledge from heterogeneous foundation models to strengthen representation initialization and then performs EMA-based pseudo-label self-training for annotation-efficient defect learning. Overall, our contributions are summarized as follows:

- We propose WISD-AD, a product-specific industrial dataset with box and class-level annotations. The dataset reflects practical challenges such as small defects, weak textures, background interference, and class imbalance.
- We propose $\mathrm{M}^2\mathrm{D}^3$, which strengthens representation initialization through heterogeneous foundation-model distillation before EMA-based pseudo-label self-training, establishing strong reference performance across annotation budgets.
- Comprehensive experiments show that existing methods still struggle with small defects under limited annotations, show the value of WISD-AD for annotation efficient inspection.

2 Related Work

2.1 Cross-Domain defect detection Datasets

Cross-domain defect detection aims to improve model generalization across diverse objects and industrial scenarios, and has been widely studied in reconstruction-based, few-shot, and multimodal settings Li et al. (2025). MVTec AD Bergmann et al. (2019) is a widely used unsupervised dataset with defect-free training data and both normal and anomalous test samples. BTAD Mishra et al. (2021) further introduces real-world industrial products with image-level and pixel-level annotations, while MVTec LOCO AD Bergmann et al. (2022) extends defect detection to logical defects, such as structural inconsistencies and object arrangement errors. VisA Zou et al. (2022) and Real-IAD Wang et al. (2024) further introduce complex objects and multi-instance scenarios. Meanwhile, MVTec3D-AD Bergmann et al. (2021) and Eyecandies Bonfiglioli et al. (2022) extend defect detection to 3D or multimodal settings for evaluating geometric and appearance defects. Recent datasets such as MVTec AD2 Heckler-Kram et al. (2026) and Real-IAD D3 Zhu et al. (2025) further emphasize fine-grained evaluation with subtle defects and multimodal annotations.

Despite their importance, cross-domain datasets mainly focus on general defect inspection, while product specific and semi-supervised industrial inspection scenarios remain insufficiently explored.

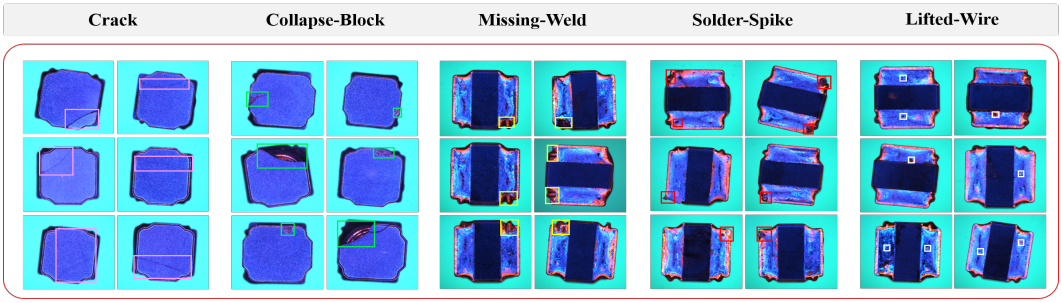


Figure 2: Representative defects of WISD-AD. WISD-AD contains five typical defect categories of wound inductors, including Crack, Collapse Block, Missing Weld, Solder Spike, and Lifted Wire.

2.2 Domain-Specific Defect Datasets

To better meet practical defect detection needs, recent studies have introduced domain-specific datasets for fabric, rail surface, and other defect inspection tasks Ni et al. (2026); Niu et al. (2020); Zhang et al. (2021). Compared with cross-domain datasets, these datasets emphasize material-specific defect patterns. For example, NEU-DET Song & Yan (2013) is widely used as a fully supervised benchmark, while GC10-DET Lv et al. (2020) contains more diverse and weakly defects, better reflecting subtle abnormalities. Similarly, DeepPCB Tang et al. (2019) also emphasizes small and weak defects in dense circuit patterns. Beyond object-level, some datasets focus on defect classification or segmentation. ELPV Deitsch et al. (2019) targets photovoltaic cell inspection, where defect are usually cross-scale variations. AITEX Silvestre-Blanes et al. (2019) focuses on fabric defect inspection, emphasizing texture irregularities and morphological differences in anomalous regions. For segmentation task, magnetic tile datasets Huang et al. (2020), KolektorSDD2 Božič et al. (2021), and DAGM Wieler & Hahn (2007) provide pixel-level annotations for localizing defects with irregular shapes and background interference. However, although semi-supervised learning is practical for annotation-limited scenarios, corresponding datasets remain largely lacking.

3 WISD-AD Dataset Description

The WISD-AD dataset is a real-world 2D industrial defect detection dataset constructed for winding inductor surface inspection. Unlike general cross-domain defect detection datasets, WISD-AD is designed for a product-specific inspection scenario with fixed imaging conditions, structured component appearances, and production-related defect distributions. WISD-AD contains five defects: Crack, Collapse Block, Missing Weld, Solder Spike, and Lifted Wire and presents practical challenges such as small defects, low-saliency patterns, cross-scale variations and long-tailed distribution. These characteristics make WISD-AD suitable for evaluating fine-grained detection and classification. Following representative domain-specific defect detection benchmarks such as NEU-DET and GC10-DET, WISD-AD adopts a defect-centric setting that focuses on fine-grained recognition and localization of predefined defect categories rather than normal versus anomalous screening.

3.1 Data Collection and Construction

Data Collection and Imaging Setup. The WISD-AD images were collected using an industrial vision acquisition platform composed of a high-resolution area-scan camera, an FA lens, and LED ring light sources. The tested winding inductors were placed vertically with respect to the camera and illumination system, and a frontal lighting strategy was adopted to enhance defect visibility and suppress unnecessary background variations. The dataset contains 2,521 images at 1440×1080 resolution. To prevent information leakage caused by similar samples, images from the same production batch are assigned exclusively to the training or test set.

Defect Categories and Annotation. WISD-AD covers five defect types: Crack, Collapse Block, Missing Weld, Solder Spike, and Lifted Wire, as illustrated in Fig. 2. These defects vary in morphology, scale, and visual saliency, covering practical challenges such as small targets, ambiguous

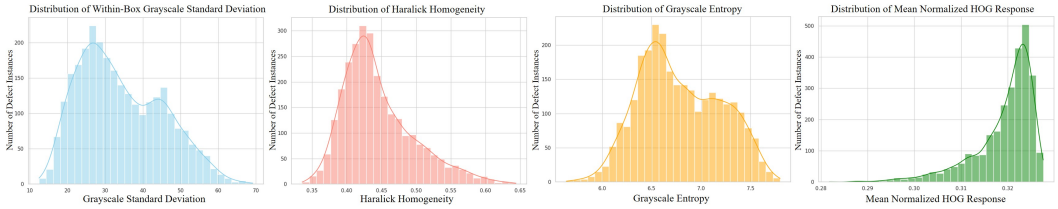


Figure 3: Morphological feature distributions of defects in WISD-AD, including grayscale contrast, Haralick homogeneity, information entropy, and HOG responses.

Table 1: Statistics of images and instances under different annotation ratios on the WISD dataset.

Ratio	Images					Instances					Total Images	Avg per Image
	Cr	CB	MW	SS	LW	Cr	CB	MW	SS	LW		
1%	14	3	2	3	3	14	3	3	4	3	25	1.080
5%	74	14	6	18	14	76	14	8	19	17	126	1.063
10%	147	30	10	36	28	156	31	13	40	35	251	1.096
15%	222	44	15	55	42	239	47	21	62	52	378	1.114
20%	296	60	19	73	56	320	63	26	80	72	504	1.113
25%	370	74	24	92	70	403	79	33	102	88	630	1.119
40%	593	118	36	148	113	648	124	50	167	145	1008	1.125
80%	1174	253	70	289	230	1297	254	90	330	277	2016	1.115
100%	1462	321	88	368	282	1604	325	115	407	353	2521	1.112

boundaries, and low-contrast regions. All images were annotated with image-level class labels and bounding boxes in VOC and YOLO formats. The dataset splits are defined separately for classification and detection. For classification, 40% of the images form the training-validation pool, while the remaining 60% constitute a fixed test set. Five-fold cross-validation is performed on the training-validation pool. The final classification results are reported as the average over the five folds. For detection, 80% of the images form the training pool and the remaining 20% form a fixed held-out evaluation set shared by all methods. Annotation ratios are defined relative to all 2,521 images. Accordingly, the full classification and detection training pools contain 1,008 (40%) and 2,016 (80%) images, respectively. WISD-AD also provides fixed semi-supervised settings at annotation ratios of 1%, 5%, 10%, 15%, 20%, and 25%. In each setting, only selected training samples are labeled and the rest are treated as unlabeled, enabling evaluation from limited to sufficient supervision.

3.2 Dataset Characteristics

Defect Morphological Characteristics To quantitatively characterize the morphological properties of defects in WISD-AD, four metrics were computed for each annotated defect instance: grayscale standard deviation (range 0-255), Haralick homogeneity (range 0-1), grayscale entropy for 8-bit images (range 0-8), and mean normalized HOG response. Their distributions are visualized in Fig. 3. First, the grayscale standard deviation is mainly distributed between 20 and 50, indicating moderate intensity variation within most defect regions and suggesting that many defects are manifested as subtle local changes. Second, Haralick homogeneity is concentrated around 0.45, reflecting an intermediate level of texture regularity with both locally smooth regions and irregular transitions. Third, grayscale entropy is primarily distributed between 6 and 7, indicating rich local intensity information and relatively complex grayscale distributions. Finally, the mean normalized HOG response mainly lies between 0.28 and 0.33, indicating relatively dispersed gradient-orientation patterns without a strongly dominant direction. This characteristic is consistent with the irregular and weakly structured boundaries observed in several defect categories. These analyses show that WISD-AD is characterized by low contrast, low saliency, high complexity, and diverse morphology, presenting significant challenges for defect detection models.

Defect Distribution Characteristics Beyond the morphological characteristics of different defects, WISD-AD also presents realistic distribution challenges that reflect practical industrial scenarios. First, WISD-AD exhibits a clear long-tailed category distribution. As shown in Table 1, under the full annotation setting, common categories such as Crack contain 1,462 images and 1,604 instances, while rare categories such as Missing Weld include only 88 images and 115 instances. Second, defect instances are sparsely distributed, with approximately 1.06 to 1.13 instances per image across different annotation ratios. This limits the number of positive training instances available under

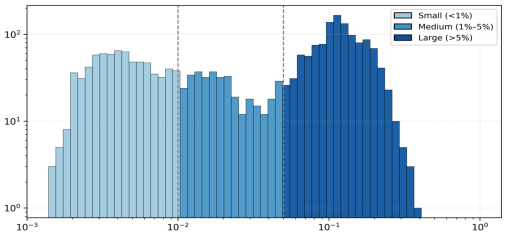


Figure 4: Distribution of defect scales in WISD-AD, showing cross-scale variations.

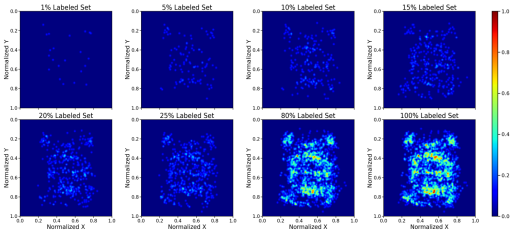


Figure 5: Spatial distribution heatmaps of defect locations under different annotation ratios.

Table 2: Comparison of WISD-AD and Representative 2D defect datasets

Characteristic	WISD-AD	MVTec AD	NEU-DET	VisA	BTAD	KolektorSDD2	Eyecandies	DAGM
Standardized semi-supervised protocol	✓	✗	✗	✗	✗	✗	✗	✗
Long-tailed defect distribution	✓	✗	✗	✗	✗	✗	✗	✗
High small-object proportion	✓	✗	✗	✓	✗	✓	✗	✗
Intra-class Cross-scale Variation	✓	✗	✗	✓	✗	✓	✓	✗
Standardized Dataset split	✓	✓	✗	✓	✓	✓	✓	✓
Include Synthetic Data	✗	✗	✗	✗	✗	✗	✓	✓
Fine-grained annotations	✓	✓	✓	✓	✓	✓	✓	✓
Cross-Domain Defect	✗	✓	✗	✓	✗	✗	✓	✗
Complex background interference	✓	✗	✗	✓	✗	✗	✗	✓

low-annotation settings, making it more difficult to learn reliable defect representations. Third, WISD-AD exhibits pronounced cross-scale variations. As shown in Fig. 4, defect regions range from very small areas, occupying less than 1% of the image, to relatively large local abnormalities, exceeding 5%. This requires models to remain sensitive to small defects while being robust to larger structural anomalies. Finally, the location frequency maps in Fig. 5 show that defects are mainly concentrated in the central inductor region, their locations scattered and uncertain. As the annotation ratio decreases, the distribution maps become sparser, but the overall spatial distribution remains largely consistent. This indicates that low-annotation subsets retain representative spatial patterns, requiring models to generalize from limit labeled data.

In summary, WISD-AD poses several practical distribution challenges, including long-tailed distributions, sparse defect instances, cross-scale sizes, and spatially scattered locations. providing a challenging benchmark for both fully supervised and semi-supervised defect detection task.

3.3 Comparison with Representative 2D Defect Dataset

To evaluate the uniqueness and challenge of the WISD-AD dataset, we compare it with several widely used 2D defect benchmarks, including MVTec AD Bergmann et al. (2019), NEU-DET Song & Yan (2013), VisA Zou et al. (2022), BTAD Mishra et al. (2021), KolektorSDD2 Božič et al. (2021), Eyecandies Bonfiglioli et al. (2022), and DAGM Wieler & Hahn (2007). The detailed comparison is summarized in Table 2. As shown in Table 2, WISD-AD differs from existing datasets in several aspects. First, it supports multi-ratio annotations for semi-supervised defect detection, enabling evaluation under different levels of labeled data, whereas most existing datasets provide only fully supervised or unsupervised setting. Second, WISD-AD includes multiple real-world challenges, such as long-tailed distribution, cross-scale variations, complex backgrounds, and texture noise, making it suitable for evaluating robust defect inspection models in practical industrial scenarios. Finally, it provides both box-level and class-level annotations with standardized data splits, supporting reproducible experiments and fair comparisons. Overall, WISD-AD fills a gap in existing defect inspection datasets by providing standardized semi-supervised classification and detection protocols while capturing complex real-world defect characteristics, making it a challenging and versatile resource for future research.

4 Benchmark

To address the challenges of WISD-AD under limited annotations, we propose Multi-Model Distilled Defect Diagnosis (M^2D^3). A central consideration is that pseudo-label self-training relies

heavily on the quality of the initial representation: under extremely sparse supervision, weak representations may generate unreliable pseudo labels and limit subsequent self-training. As shown in Fig. 6, M^2D^3 adopts a two-stage pipeline that first strengthens the downstream representation through multi-model heterogeneous distillation and then performs EMA-based pseudo-label self-training for target-domain adaptation. The core configurations are provided in the **Appendix A**.

Multi-Model Heterogeneous Distillation. To exploit complementary knowledge from different pretrained foundation models, we introduce a multi-model heterogeneous distillation strategy to initialize the downstream model. Given an input image x, two frozen visual foundation models, DINOv3Siméoni et al. (2025) and EUPEZhu et al. (2026), are employed as heterogeneous teachers. DINOv3 provides dense self-supervised structural representations that are effective in capturing local textures, edges, and fine-grained defect patterns, while EUPE complements it with more balanced semantic-spatial representations obtained from its universal perception capability. For each image, the two teachers extract patch-level feature maps:

$$F_D = T_D(x), \quad F_E = T_E(x), \tag{1}$$

where T_D and T_E denote the DINOv3 and EUPE encoders, respectively. To transfer these teacher representations to the downstream model, The first N backbone layers serve as the student feature extractor, as defined in Eq. 2.

$$F_S = B_\theta^{1:N}(x) \tag{2}$$

Considering the different output dimensions of pretrained models, projection heads are introduced to align the student features with each teacher feature space, as defined in Eq. 3.

$$\hat{F}_D = P_D(F_S), \quad \hat{F}_E = P_E(F_S). \tag{3}$$

The F_S is optimized by matching the projected features with the corresponding teacher features:

$$\mathcal{L}_{\text{MHD}} = \lambda_D \mathcal{L}_D + \lambda_E \mathcal{L}_E, \tag{4}$$

where $\mathcal{L}_D$ and $\mathcal{L}_E$ denote the cosine-distance losses between the projected student features and the spatially aligned features from DINOv3 and EUPE, respectively.

Pseudo-Label Based Semi-Supervised Self-Training. Although multi-model heterogeneous distillation provides strong generic visual priors, the downstream model still needs to adapt to the target defect distribution with limited annotations. Since abundant unlabeled images from the same production line are available in semi-supervised inspection scenarios, we introduce EMA teacher–student self-training to convert them into additional supervision after distilled-backbone initialization. Specifically, the downstream model is first trained on the labeled defect images to obtain an initial supervised student model. An exponential moving average (EMA) teacher is then constructed from the student model to generate pseudo labels for unlabeled images. For classification, the pseudo labels correspond to image-level defect categories, whereas for detection, they consist of defect categories and bounding boxes. Only predictions confidence exceed a predefined threshold τ are retained. The student model is jointly optimized using origin labeled and pseudo-labeled data:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \alpha \mathcal{L}_{\text{pseudo}}, \tag{5}$$

where $\mathcal{L}_{\text{sup}}$ is the supervised loss computed on labeled samples, $\mathcal{L}_{\text{pseudo}}$ is the loss computed on pseudo-labeled samples, and α controls the weights of pseudo supervision. After each self-training iteration, the EMA teacher is updated as Eq.6:

$$\theta_T \leftarrow \mu\theta_T + (1-\mu)\theta_S, \tag{6}$$

where θ_T and θ_S are the teacher and student parameters, and μ is the EMA decay coefficient. By updating the teacher from the student parameters, the framework maintains a more stable ability for pseudo-label generation, improving the reliability of semi-supervised learning on unlabeled images.

5 Experiments

5.1 Fully-Supervised Analysis

We first conduct a fully supervised analysis on WISD-AD to assess their performance upper bounds under sufficient supervision. Under full supervision, the semi-supervised components of all methods

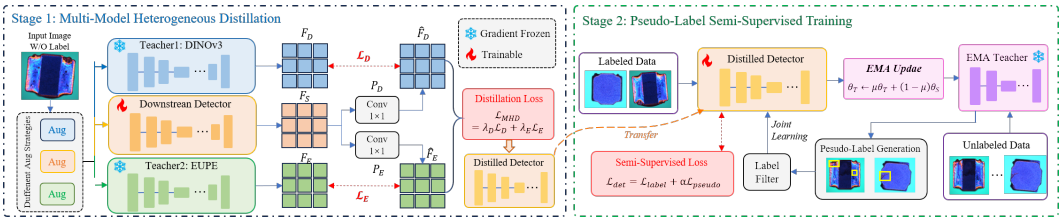


Figure 6: Overview of M^2D^3 framework with multi-teacher distillation and pseudo self-training.

Table 3: Comparison of class-wise accuracy, mean class accuracy, and class-wise accuracy standard deviation of different models on WISD-AD under the fully supervised setting. Results are reported as mean ± standard deviation over 3 random seeds.

Method	Class-wise Accuracy					Mean Acc	Std.
	Cr	CB	LW	MW	SS		
ResNet-50He et al. (2016)	97.01 ± 0.17	63.02 ± 0.21	**99.39 ± 0.16**	70.66 ± 0.28	80.01 ± 0.33	82.02 ± 0.24	14.29 ± 0.32
EdgeNextMaaz et al. (2022)	99.77 ± 0.19	11.05 ± 0.52	77.58 ± 0.34	72.03 ± 0.27	93.95 ± 0.25	70.88 ± 0.30	31.60 ± 0.37
MobileNetV4-SQin et al. (2024)	99.54 ± 0.22	35.47 ± 0.46	75.15 ± 0.33	71.70 ± 0.29	71.63 ± 0.30	70.70 ± 0.26	20.47 ± 0.31
GhostNet-V3Liu et al. (2024)	99.30 ± 0.17	51.74 ± 0.38	83.03 ± 0.25	71.70 ± 0.32	84.65 ± 0.24	78.08 ± 0.22	15.83 ± 0.29
FasterNet-SChen et al. (2023b)	50.91 ± 0.41	**95.68 ± 0.20**	94.36 ± 0.27	77.25 ± 0.31	91.95 ± 0.23	82.03 ± 0.25	16.90 ± 0.34
ConvNextV2Woo et al. (2023)	**99.97 ± 0.15**	27.59 ± 0.43	90.91 ± 0.30	88.89 ± 0.21	**98.61 ± 0.18**	81.19 ± 0.28	27.14 ± 0.35
Swin-TinyLiu et al. (2021b)	99.88 ± 0.19	15.30 ± 0.55	91.52 ± 0.26	**96.23 ± 0.17**	39.49 ± 0.27	68.48 ± 0.23	34.51 ± 0.26
ViT-BaseDosovitskiy et al. (2021)	99.88 ± 0.21	6.21 ± 0.32	51.92 ± 0.44	72.90 ± 0.30	87.91 ± 0.22	63.76 ± 0.16	32.94 ± 0.21
SwiftFormerShaker et al. (2023)	98.37 ± 0.25	77.95 ± 0.32	92.79 ± 0.24	69.58 ± 0.33	70.56 ± 0.31	81.85 ± 0.23	11.71 ± 0.27
MobileFormerChen et al. (2022)	98.84 ± 0.20	16.28 ± 0.51	83.64 ± 0.32	77.36 ± 0.26	90.23 ± 0.28	73.27 ± 0.31	29.37 ± 0.38
M^2D^3(Ours)	98.73 ± 0.20	70.93 ± 0.28	87.88 ± 0.22	73.58 ± 0.23	95.35 ± 0.20	**85.29 ± 0.18**	**11.24 ± 0.19**

are disabled, trained using only labeled data. For classification and detection, M^2D^3 uses ResNet-50 He et al. (2016) and YOLO11-n Ultralytics (2024) as downstream models, respectively.

Defect Classification We first conduct fully supervised classification to evaluate whether WISD-AD contains distinguishable defect patterns at the image level. Representative CNN-based and Transformer-based models are trained and evaluated using accuracy and standard deviation metrics. The results are shown in Table.3. The results reveal substantial class-wise performance disparity on WISD-AD. For example, Swin-Tiny achieves accuracies of 99.88% and 96.23% on Crack and Missing Weld, respectively, whereas its accuracies on Collapse Block and Solder Spike decrease to 15.30% and 39.49%, resulting in the largest class-wise accuracy standard deviation of 34.51. Similar gaps appear for EdgeNext and ViT. A high Class Std. indicates class imbalanced recognition and may reflect a preference for particular classes under visual ambiguity and class imbalance. In contrast, M^2D^3 achieves the highest mean class accuracy (85.29%) and lowest Class Std. (11.24), indicating stronger and more balanced recognition across defect classes. These results highlight the need for fine-grained discrimination and class-balanced performance on WISD-AD.

Defect Detection. Table 4 presents the results of several representative fully supervised models on the WISD-AD dataset. The evaluated models include two-stage CNN-based detectors such as the Faster RCNN series, one-stage real-time detectors in the YOLO series, and transformer-based DETR variants. It can be observed that defects such as Lifted Wire and other small-scale anomalies pose significant challenges to most CNN and transformer detectors due to their limited size and sparse data, resulting in lower AP values. In addition, Solder Spike defects are usually accompanied by complex background interference, and the presence of visually similar structures to Lifted Wire further increases inter-class confusion, lowering detection accuracy for both categories. The proposed M^2D^3 achieves the highest overall performance, demonstrating the effectiveness of the multi-teacher distillation scheme in capturing complex defect characteristics. These results further highlight the challenges of WISD-AD, particularly for small and low-saliency defects.

5.2 Semi-Supervised Analysis

Semi-supervised learning is important for defect inspection, where dense annotations are costly and sparse-label benchmarks remain limited. Therefore, we further evaluate semi-supervised classification and detection on WISD-AD with 1%, 5%, 10%, 15%, 20%, and 25% labeled data.

Defect Classification. We first evaluate defect classification on WISD-AD by comparing representative fully supervised and semi-supervised methods with different CNN and Transformer-based

Table 4: Class-wise AP and overall metrics of different supervised models on WISD-AD under fully-supervised setting. Results are reported as mean ± standard deviation over 3 random seeds.

Methods	Class AP_{50}					mAP		F1
	Cr	CB	LW	MW	SS	@50	@50:95	
Faster R-CNNRen et al. (2016)	83.26 ± 0.30	89.15 ± 0.27	74.33 ± 0.35	85.72 ± 0.31	78.61 ± 0.33	82.21 ± 0.30	43.65 ± 0.34	78.13 ± 0.32
Sparse R-CNNSun et al. (2021)	85.74 ± 0.28	91.43 ± 0.25	77.29 ± 0.33	87.94 ± 0.29	81.05 ± 0.31	84.69 ± 0.28	46.82 ± 0.32	80.46 ± 0.30
YOLO11Ultralytics (2024)	91.03 ± 0.24	$\mathbf{97.96 \pm 0.20}$	84.01 ± 0.30	93.62 ± 0.25	87.38 ± 0.27	90.79 ± 0.23	53.92 ± 0.26	88.07 ± 0.25
YOLO26Chakrabarty (2026)	90.66 ± 0.25	91.32 ± 0.29	89.77 ± 0.23	$\mathbf{97.09 \pm 0.19}$	87.68 ± 0.26	91.30 ± 0.22	53.46 ± 0.27	87.55 ± 0.26
NanoDet-PlusLyu (2021)	87.19 ± 0.27	93.08 ± 0.24	79.65 ± 0.32	89.47 ± 0.28	83.22 ± 0.30	86.52 ± 0.27	48.74 ± 0.31	83.17 ± 0.29
RT-DETRZhao et al. (2024)	90.60 ± 0.24	94.75 ± 0.22	87.54 ± 0.26	95.51 ± 0.21	84.02 ± 0.29	90.48 ± 0.23	52.03 ± 0.28	88.95 ± 0.24
Co-DETRZong et al. (2023)	90.92 ± 0.23	97.36 ± 0.21	89.17 ± 0.24	94.28 ± 0.23	88.46 ± 0.25	92.04 ± 0.21	52.19 ± 0.27	89.72 ± 0.22
DINOZhang et al. (2022)	90.42 ± 0.24	96.65 ± 0.22	85.10 ± 0.28	93.06 ± 0.24	90.12 ± 0.23	91.07 ± 0.22	52.25 ± 0.26	87.26 ± 0.25
DEIMv2Huang et al. (2025a)	91.15 ± 0.22	97.23 ± 0.21	89.48 ± 0.23	95.19 ± 0.20	90.60 ± 0.21	92.73 ± 0.20	53.12 ± 0.24	88.19 ± 0.23
M^2D^3(Ours)	$\mathbf{91.71 \pm 0.18}$	$\mathbf{97.79 \pm 0.17}$	$\mathbf{93.57 \pm 0.16}$	$\mathbf{95.68 \pm 0.15}$	$\mathbf{90.95 \pm 0.17}$	$\mathbf{93.94 \pm 0.14}$	$\mathbf{53.97 \pm 0.16}$	$\mathbf{90.01 \pm 0.15}$

Table 5: Comparison of supervised and semi-supervised classification models on WISD-AD under different annotation ratio. Results are reported as mean ± standard deviation over 3 random seeds.

Methods	Accuracy under different Annotation Ratio						
	1%	5%	10%	15%	20%	25%	40%
ResNet-50He et al. (2016)	49.57 ± 0.33	62.80 ± 0.28	70.64 ± 0.24	77.39 ± 0.21	78.48 ± 0.20	80.13 ± 0.22	82.02 ± 0.21
ConvNextV2-TWoo et al. (2023)	46.71 ± 0.37	50.12 ± 0.34	56.28 ± 0.30	61.94 ± 0.27	67.22 ± 0.25	76.27 ± 0.23	81.19 ± 0.28
Swin-TinyLiu et al. (2021b)	45.33 ± 0.39	57.22 ± 0.31	62.45 ± 0.29	64.90 ± 0.26	66.73 ± 0.24	67.51 ± 0.25	68.48 ± 0.33
ViT-BaseDosovitskiy et al. (2021)	52.57 ± 0.35	56.79 ± 0.32	60.72 ± 0.30	61.42 ± 0.28	62.63 ± 0.27	63.36 ± 0.26	63.76 ± 0.36
SwiftFormerShaker et al. (2023)	46.45 ± 0.36	62.14 ± 0.29	70.08 ± 0.25	72.39 ± 0.23	76.05 ± 0.22	78.56 ± 0.21	81.85 ± 0.23
MixMatchBerthelot et al. (2019)	50.89 ± 0.32	63.25 ± 0.27	69.93 ± 0.24	76.62 ± 0.22	78.05 ± 0.21	79.44 ± 0.20	80.61 ± 0.24
SoftMatchChen et al. (2023a)	51.72 ± 0.30	64.19 ± 0.26	70.86 ± 0.23	77.47 ± 0.21	79.27 ± 0.20	80.35 ± 0.20	81.79 ± 0.22
InfoMatchHan et al. (2024)	52.30 ± 0.29	65.41 ± 0.25	71.52 ± 0.22	78.15 ± 0.20	79.84 ± 0.19	81.12 ± 0.19	82.43 ± 0.21
PEPL Tian et al. (2025)	52.95 ± 0.27	66.43 ± 0.24	75.06 ± 0.21	78.38 ± 0.20	82.39 ± 0.19	83.87 ± 0.18	84.03 ± 0.20
SemiVisBoosterZhang et al. (2025)	53.17 ± 0.26	66.99 ± 0.23	75.68 ± 0.20	79.04 ± 0.19	82.96 ± 0.18	84.21 ± 0.18	84.46 ± 0.19
M^2D^3(Ours)	$\mathbf{54.02 \pm 0.24}$	$\mathbf{67.82 \pm 0.21}$	$\mathbf{76.11 \pm 0.19}$	$\mathbf{79.66 \pm 0.18}$	$\mathbf{83.47 \pm 0.17}$	$\mathbf{84.94 \pm 0.17}$	$\mathbf{85.29 \pm 0.19}$

Table 6: Comparison of supervised and semi-supervised detection models on WISD-AD under different annotation ratio. Results are reported as mean ± standard deviation over 3 random seeds.

Methods	mAP_{50} under different Annotation Ratio						
	1%	5%	10%	15%	20%	25%	80%
YOLO11Ultralytics (2024)	12.23 ± 0.37	30.08 ± 0.33	59.36 ± 0.29	70.89 ± 0.26	74.02 ± 0.24	79.47 ± 0.22	90.79 ± 0.21
YOLO26Chakrabarty (2026)	8.23 ± 0.40	21.18 ± 0.36	52.17 ± 0.31	62.08 ± 0.28	69.24 ± 0.25	74.29 ± 0.23	91.30 ± 0.22
RT-DETRZhao et al. (2024)	0.16 ± 0.44	12.27 ± 0.38	54.03 ± 0.30	62.20 ± 0.27	68.12 ± 0.25	74.98 ± 0.23	90.48 ± 0.21
DINOZhang et al. (2022)	1.09 ± 0.42	7.04 ± 0.39	42.56 ± 0.33	56.68 ± 0.29	65.34 ± 0.26	74.29 ± 0.24	91.07 ± 0.22
DEIMv2Huang et al. (2025a)	4.48 ± 0.41	9.68 ± 0.38	48.45 ± 0.32	59.72 ± 0.28	68.41 ± 0.25	77.37 ± 0.23	92.73 ± 0.20
Soft TeacherXu et al. (2021)	24.80 ± 0.35	45.45 ± 0.31	55.23 ± 0.28	66.17 ± 0.25	73.89 ± 0.23	79.56 ± 0.21	85.64 ± 0.20
Unbiased TeacherLiu et al. (2021a)	26.37 ± 0.34	46.78 ± 0.30	53.15 ± 0.29	64.02 ± 0.26	71.75 ± 0.24	77.38 ± 0.22	84.18 ± 0.21
Efficient TeacherXu et al. (2023)	29.17 ± 0.33	50.23 ± 0.29	58.67 ± 0.27	69.54 ± 0.24	76.38 ± 0.22	81.92 ± 0.20	88.87 ± 0.19
ARSLLiu et al. (2023)	28.78 ± 0.34	49.15 ± 0.30	57.32 ± 0.28	68.21 ± 0.25	75.04 ± 0.23	80.87 ± 0.21	87.45 ± 0.20
Semi-DETRZhang et al. (2023)	31.91 ± 0.32	51.89 ± 0.28	61.25 ± 0.26	71.98 ± 0.23	78.72 ± 0.21	83.56 ± 0.20	90.27 ± 0.19
Sparse DETRShehzadi et al. (2024)	34.20 ± 0.31	53.67 ± 0.27	60.89 ± 0.26	71.54 ± 0.24	78.25 ± 0.22	83.12 ± 0.20	89.87 ± 0.19
SSDDetLiu et al. (2025)	28.42 ± 0.34	50.56 ± 0.29	57.89 ± 0.27	68.67 ± 0.25	75.41 ± 0.23	81.15 ± 0.21	87.72 ± 0.20
STEP-DETRShehzadi et al. (2025)	34.38 ± 0.30	53.92 ± 0.27	61.37 ± 0.25	72.68 ± 0.23	79.45 ± 0.21	84.21 ± 0.19	90.55 ± 0.18
M^2D^3(Ours)	$\mathbf{44.60 \pm 0.25}$	$\mathbf{68.17 \pm 0.22}$	$\mathbf{75.04 \pm 0.20}$	$\mathbf{81.33 \pm 0.18}$	$\mathbf{84.07 \pm 0.17}$	$\mathbf{88.25 \pm 0.16}$	$\mathbf{93.94 \pm 0.15}$

models under varying annotation ratios, aiming to analyze their performance trends with different amounts of labeled data. The results are shown in Table 5.

It can be seen that under extremely limited annotations, existing supervised methods suffer from a sharp performance drop. Although semi-supervised methods can exploit unlabeled data, they are still vulnerable to pseudo-label noise, where visually similar backgrounds and subtle defect patterns usually cause class confusion. M^2D^3 achieves the best performance across different annotation ratios, but its accuracy still decreases noticeably compared with the full-annotation setting(40%) when labeled data are severely limited. These results indicate that WISD-AD poses a challenging low-label learning scenario, requiring models to learn robust defect representations and reliable pseudo labels.

Defect Detection We further compare our method with both mainstream fully supervised and semi-supervised detection approaches. The fully supervised models include real-time CNN-based YOLO series, and transformer-based DETR variants. Semi-supervised methods include teacher-student consistency-based methods and transformer-based semi-supervised detectors, covering the current mainstream approaches for leveraging unlabeled data in defect detection. All models are evaluated using mAP_{50} metric, and the results are summarized in Table 6.

From Table 6, M^2D^3 achieves the best performance across all annotation ratios. In particular, with extremely low 1% labeled data, M^2D^3 outperforms the best supervised and semi-supervised model mAP_{50} by 32.37 and 10.22 percentage points, respectively. As the annotation ratio increases, M^2D^3



Figure 7: Feature clustering of representative models on WISD-AD under the 1% and full-annotation settings.

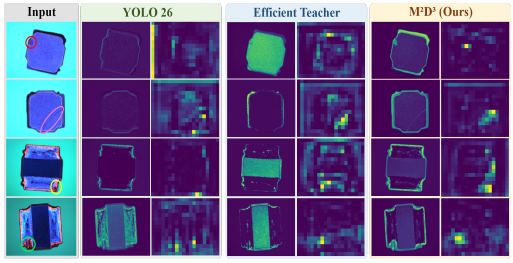


Figure 8: Visualization of feature maps of representative supervised and semi-supervised models on WISD-AD under 1% annotation setting.

consistently improves and achieves comparable performance to other SOTA supervised models with only 25% labeled data. Under full supervision (80%), it further surpasses the best supervised and semi-supervised methods by 1.21 and 3.39 mAP_{50} points, respectively, demonstrating strong generalization ability.

5.3 Feature Visualization Analysis of WISD-AD

Feature Clustering Analysis. To analyze the separability of defect representations, we visualize representative fully supervised and semi-supervised methods under the 1% labeled and fully labeled detection settings. For all methods, features are extracted before the prediction head from the same test samples and aggregated into fixed-length vectors using an identical feature-processing strategy. The resulting feature vectors are ℓ_2-normalized and projected into a two-dimensional space using PCA, as shown in Fig. 7. As shown in Fig. 7, under the 1% labeled setting, the baseline methods exhibit substantial overlap among different defect categories, indicating limited feature separability under extremely scarce supervision. In contrast, M^2D^3 produces more compact intra-class distributions and clearer inter-class boundaries, demonstrating stronger discriminative representation learning from limited annotations. Nevertheless, overlaps remain among several visually similar defect categories even under full annotation. This observation highlights the inherent difficulty of fine-grained defect discrimination on WISD-AD and indicates the need for stronger representation learning methods, even when sufficient supervision is available.

Feature Extraction Analysis. To further analyze the feature representation ability of different models, we visualize their shallow and deep layer feature maps under the 1% annotation settings, as shown in Fig. 8. It can be observed that under the 1% annotation setting, shallow feature maps from fully supervised methods commonly fail to capture fine defect details and exhibit weak defect responses, while semi-supervised methods are more susceptible to background noise. In contrast, M^2D^3 benefits from stronger initial representations and responds more clearly to defect regions. In deeper feature maps, noise accumulated from shallow features in YOLO26 and Efficient Teacher causes substantial shifts in activations associated with defect morphology and location. In contrast, benefiting from foundation-model distillation and EMA teacher, M^2D^3 more reliably preserves defect structures and spatial information. Overall, these results highlight the need for robust representations of low-saliency defects under sparse annotations and complex backgrounds on WISD-AD.

6 Conclusion

In this work, we present WISD-AD, a real-world 2D industrial defect dataset covering five typical defect categories and practical challenges, including small defects, low saliency, background interference, and long-tailed class distributions. With standardized data splits and multiple annotation ratios, WISD-AD enables fair and reproducible evaluation of fully supervised and semi-supervised methods for defect classification and detection. We further propose M^2D^3, a semi-supervised framework for annotation-efficient defect inspection. Extensive experiments demonstrate its effectiveness and reveal that, under extremely sparse supervision, representation initialization critically affects pseudo-label reliability and subsequent self-training. These findings highlight the importance of strengthening representations before self-training for annotation-efficient industrial detection.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and improving the clarity of technical descriptions. They were not used to generate experimental data, annotations, or results, nor did they replace model training, evaluation, or empirical analysis. All AI-assisted content was carefully reviewed and verified by the authors. The authors take full responsibility for the final content of the paper, including all text, experimental results, and conclusions.

ETHICS STATEMENT

This work focuses on industrial visual defect inspection and does not involve human subjects or personally identifiable information. The WISD-AD dataset consists of images of soft-magnetic wound inductors collected using an industrial vision acquisition system, together with defect category and bounding-box annotations. The dataset is intended solely for research on industrial visual inspection, defect recognition, and annotation-efficient learning. No human-related or privacy-sensitive information is included in the collected data, and the proposed benchmark and method are not intended for privacy-invasive or harmful applications.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the WISD-AD dataset construction, defect categories, data splits, annotation-ratio settings, and evaluation protocols in the main paper. We also describe the complete M^2D^3 framework, including multi-model heterogeneous distillation, EMA-based pseudo-label self-training, training objectives, and implementation settings. Detailed hyperparameters, ablation studies, pseudo-label reliability analysis, and additional experimental results are provided in the appendix and supplementary materials. We additionally provide the core implementation code, model configurations, and benchmark settings in the supplementary materials to facilitate reproduction of the reported results.

REFERENCES

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9592–9600, 2019.

Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. *arXiv preprint arXiv:2112.09045*, 2021.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130(4):947–969, 2022.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

Luca Bonfiglioli, Marco Toschi, Davide Silvestri, Nicola Fioraio, and Daniele De Gregorio. The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In *Proceedings of the Asian Conference on Computer Vision*, pp. 3586–3602, 2022.

Jakob Božič, Domen Tabernik, and Danijel Skočaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129:103459, 2021.

Sudip Chakrabarty. Yolo26: An analysis of nms-free end to end framework for real-time object detection. *arXiv preprint arXiv:2601.12882*, 2026.

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*, 2023a.

Jierun Chen, Shiu-hong Kao, Hao He, Weipeng Zhuo, Song Wen, Chul-Ho Lee, and S-H Gary Chan. Run, don't walk: chasing higher flops for faster neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12021–12031, 2023b.

Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobile-former: Bridging mobilenet and transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5270–5279, 2022.

Sergiu Deitsch, Vincent Christlein, Stephan Berger, Claudia Buerhop-Lutz, Andreas Maier, Florian Gallwitz, and Christian Riess. Automatic classification of defective photovoltaic module cells in electroluminescence images. *Solar Energy*, 185:455–468, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Qi Han, Zhibo Tian, Chengwei Xia, and Kun Zhan. Infomatch: entropy neural estimation for semi-supervised image classification. *arXiv preprint arXiv:2404.11003*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Lars Heckler-Kram, Jan-Hendrik Neudeck, Ulla Scheler, Rebecca König, and Carsten Steger. The mvtec ad 2 dataset: Advanced scenarios for unsupervised anomaly detection. *International Journal of Computer Vision*, 134(4):175, 2026.

Shihua Huang, Yongjie Hou, Longfei Liu, Xuanlong Yu, and Xi Shen. Real-time object detection meets dinov3. *arXiv preprint arXiv:2509.20787*, 2025a.

Shihua Huang, Zhichao Lu, Xiaodong Cun, Yongjun Yu, Xiao Zhou, and Xi Shen. Deim: Detr with improved matching for fast convergence. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15162–15171, 2025b.

Yibin Huang, Congying Qiu, and Kui Yuan. Surface defect saliency of magnetic tile. *The Visual Computer*, 36(1):85–96, 2020.

Zhuo Li, Yuhao Yan, Xiangheng Wang, Yifei Ge, and Lin Meng. A survey of deep learning for industrial visual anomaly detection. *Artificial Intelligence Review*, 58(9):279, 2025.

Chang Liu, Weiming Zhang, Xiangru Lin, Wei Zhang, Xiao Tan, Junyu Han, Xiaomao Li, Errui Ding, and Jingdong Wang. Ambiguity-resistant semi-supervised learning for dense object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15579–15588, 2023.

Guanghu Liu, Maoxiang Chu, and Rongfen Gong. Semi-ssddet: An adaptive semi-supervised steel surface defect detection. *Signal, Image and Video Processing*, 19(10):1–11, 2025.

Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021b. doi: 10.1109/ICCV48922.2021.00986.

Zhenhua Liu, Zhiwei Hao, Kai Han, Yehui Tang, and Yunhe Wang. Ghostnetv3: Exploring the training strategies for compact models. *arXiv preprint arXiv:2404.11202*, 2024.

Xiaoming Lv, Fajie Duan, Jia-jia Jiang, Xiao Fu, and Lin Gan. Deep metallic surface defect detection: The new benchmark and detection network. *Sensors*, 20(6):1562, 2020.

Rangi Lyu. Nanodet-plus: Super fast and high accuracy lightweight anchor-free object detection model. *URL: https://github. com/RangiLyu/nanodet*, 2021.

Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *European conference on computer vision*, pp. 3–20. Springer, 2022.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. *arXiv preprint arXiv:2104.10036*, 2021.

Yan-Qin Ni, Pei-Kai Huang, Wei-Jen Wang, Deron Liang, and Chia-Yu Lin. Tsfabrics: A time-series fabric dataset for real-time defect detection on circular knitting machines. *Scientific Data*, 2026.

Menghui Niu, Kechen Song, Liming Huang, Qi Wang, Yunhui Yan, and Qinggang Meng. Unsupervised saliency detection of rail surface defects using stereoscopic images. *IEEE Transactions on Industrial Informatics*, 17(3):2271–2281, 2020.

Danfeng Qin, Chas Leichner, Manolis Delakis, Marco Fornoni, Shixin Luo, Fan Yang, Weijun Wang, Colby Banbury, Chengxi Ye, Berkin Akin, et al. Mobilenetv4: universal models for the mobile ecosystem. In *European Conference on Computer Vision*, pp. 78–96. Springer, 2024.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 17425–17436, 2023.

Tahira Shehzadi, Khurram Azeem Hashmi, Didier Stricker, and Muhammad Zeshan Afzal. Sparse semi-detr: Sparse learnable queries for semi-supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5840–5850, 2024.

Tahira Shehzadi, Khurram Azeem Hashmi, Shalini Sarode, Didier Stricker, and Muhammad Zeshan Afzal. Step-detr: Advancing detr-based semi-supervised object detection with super teacher and pseudo-label guided text queries. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3069–3079, 2025.

Javier Silvestre-Blanes, Teresa Albero-Albero, Ignacio Miralles, Rubén Pérez-Llorens, and Jorge Moreno. A public fabric database for defect detection methods and results. *Autex Research Journal*, 19(4):363–374, 2019.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020.

Kechen Song and Yunhui Yan. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Applied Surface Science*, 285:858–864, 2013.

Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14454–14463, 2021.

Sanli Tang, Fan He, Xiaolin Huang, and Jie Yang. Online pcb defect detector on a new pcb defect dataset. *arXiv preprint arXiv:1902.06197*, 2019.

Bowen Tian, Songning Lai, Lujundong Li, Zhihao Shuai, Runwei Guan, Tian Wu, and Yutao Yue. Pepl: Precision-enhanced pseudo-labeling for fine-grained image classification in semi-supervised learning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Ultralytics. Yolo11. `https://github.com/ultralytics/ultralytics`, 2024.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22883–22892, 2024.

Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM symposium in*, volume 6, pp. 11, 2007.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16133–16142, 2023. doi: 10.1109/CVPR52729.2023.01548.

Bowen Xu, Mingtao Chen, Wenlong Guan, and Lulu Hu. Efficient teacher: Semi-supervised object detection for yolov5. *arXiv preprint arXiv:2302.07577*, 2023.

Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3060–3069, 2021.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

Jiacheng Zhang, Xiangru Lin, Wei Zhang, Kuo Wang, Xiao Tan, Junyu Han, Errui Ding, Jingdong Wang, and Guanbin Li. Semi-detr: Semi-supervised object detection with detection transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23809–23818, 2023.

Wenjin Zhang, Xinyu Li, Chenyang Gao, and Ivan Marsic. Semivisbooster: Boosting semi-supervised learning for fine-grained classification through pseudo-label semantic guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1195–1204, 2025.

Zihao Zhang, Shaozuo Yu, Siwei Yang, Yu Zhou, and Bingchen Zhao. Rail-5k: A real-world dataset for rail surface defects detection. *arXiv preprint arXiv:2106.14366*, 2021.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16965–16974, 2024.

Chenchen Zhu, Saksham Suri, Cijo Jose, Maxime Oquab, Marc Szafraniec, Wei Wen, Yunyang Xiong, Patrick Labatut, Piotr Bojanowski, Raghuraman Krishnamoorthi, et al. Efficient universal perception encoder. *arXiv preprint arXiv:2603.22387*, 2026.

Wenbing Zhu, Lidong Wang, Ziqing Zhou, Chengjie Wang, Yurui Pan, Ruoyi Zhang, Zhuhao Chen, Linjie Cheng, Bin-Bin Gao, Jiangning Zhang, et al. Real-iad d3: A real-world 2d/pseudo-3d/3d dataset for industrial anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15214–15223, 2025.

Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6748–6758, 2023.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision*, pp. 392–408. Springer, 2022.

A Implementation Details

Basic Environment Settings

All experiments were conducted on a computer with Linux operating system, also having an Intel Core I9-10900X CPU, 64G of RAM and an NVIDIA RTX A6000 GPU. The equipped software runtime environment was also set up with Ubuntu 22.04.4 LTS, python 3.9, PyTorch 2.1.1, CUDA 11.7 and cuDNN 8.6.0.

Model Hyperparameter Settings

To ensure the reproducibility of M^2D^3, Table 7 summarizes the main hyperparameters used in its two-stage training pipeline. In Stage I, the frozen DINOv3 and EUPE models are employed as heterogeneous teachers with separate projection heads to transfer complementary representations into the YOLO backbone. Stage II loads the distilled backbone and performs EMA-based teacher-student self-training to exploit unlabeled defect images.

B Additional Details of WISD-AD

Task Scope. WISD-AD is designed as a defect-centric benchmark for fine-grained defect recognition and localization rather than a normal-versus-anomalous inspection benchmark. Similar to representative domain-specific defect detection benchmarks such as NEU-DET and GC10-DET, it focuses on distinguishing and localizing predefined defect categories in defective samples. Accordingly, the current dataset contains five defect categories and does not introduce an additional defect-free category.

This task definition is complementary to industrial anomaly detection benchmarks that primarily evaluate the discrimination between normal and anomalous samples. The objective of WISD-AD is instead to investigate fine-grained defect discrimination and localization, particularly under limited annotation budgets, where small-scale, low-saliency, and visually similar defects remain difficult to recognize. Extending the benchmark toward joint normal-defect screening and fine-grained defect localization is a potential direction for future work.

Annotation Procedure and Quality Control All images were manually annotated with defect categories and bounding boxes. After the initial annotation, the annotations were independently reviewed to identify incorrect categories, inaccurate boundaries, duplicated boxes, and missing instances. Ambiguous samples were re-examined before being included in the final benchmark.

Dataset Split and Leakage Prevention To reduce potential information leakage caused by highly similar samples collected from the same production batch, samples belonging to one production batch are never distributed across the training and evaluation sets. The held-out evaluation set remains fixed across all annotation ratios and methods. Only the amount of labeled data within the training pool is varied, while the remaining training samples are treated as unlabeled data.

Semi-supervised Subset Construction For each annotation ratio, the labeled subset is independently constructed from the corresponding training pool using class-wise stratified sampling, while approximately preserving the original long-tailed category distribution of WISD-AD. The resulting subset is fixed and shared by all compared methods under the same annotation ratio to ensure a fair comparison. The remaining samples in the training pool are treated as unlabeled data, and their ground-truth category and bounding-box annotations are never accessed during training, pseudo-label generation, or pseudo-label filtering. The held-out evaluation set remains unchanged across all annotation ratios and methods. We provide the fixed subset indices for each annotation ratio together with the benchmark to facilitate reproducible evaluation.

Evaluation Protocol and Hyperparameter Selection All methods are evaluated using the same fixed training and held-out evaluation partitions under each annotation ratio. The held-out evaluation set is used exclusively for final performance reporting and is never involved in model training, pseudo-label generation, pseudo-label filtering, or hyperparameter selection. For M^2D^3, the training hyperparameters are fixed beforehand based on the default configurations of the corresponding training framework and are kept unchanged across all annotation ratios. In particular, the pseudo-label

Table 7: Core hyperparameter configurations of M^2D^3 in Stage I and Stage II.

Stage	Module	Configuration	Value
Stage I	DINOv3	Teacher model	`Dinov3_vit_b16`
		Teacher input size	256×256
		Distillation weight λ_D	1.0
	EUPE	Teacher model	`eupe_vit_b16`
		Teacher input size	256×256
		Distillation weight λ_E	1.0
	Shared	Student model	ResNet-50 / YOLO11-n
		Input size	256×256
		batch size	128
		Training epochs	200
		warm-up epochs	5
		Optimizer	SGD
		learning rate	0.02
		Weight decay	1×10^{-4}
		Distillation loss	Cosine distance
Stage II	Classification	Student model	ResNet-50
		Supervised training epochs	150
		Self-training rounds	3
		epochs per round	50
		Input size	224×224
		batch size	64
		Pseudo-label threshold	0.3
		EMA momentum μ	0.999
		Pseudo-label TTA	False
		Optimizer	AdamW
		Learning rate	5×10^{-4}
		Weight decay	5×10^{-2}
		warm-up epochs	5
	Detection	Student model	YOLO11-n
		Warm-up epochs	300
		Self-training rounds	10
		epochs per round	50
		Input size	640×640
		batch size	64
		Pseudo-label threshold	0.2
		NMS IoU	0.3
		Pseudo-label TTA	True
		EMA momentum μ	0.999
		Optimizer	SGD
		Initial learning rate	0.01
		final learning rate	0.0001
		Momentum	0.937
		weight decay	0.0005
		Mosaic / MixUp	1.0 / 0.0

confidence threshold, EMA momentum, optimization settings, data augmentation, and self-training schedule are shared across different annotation budgets rather than being individually tuned for each setting. For competing methods, we follow their official implementations and recommended configurations, while using the same WISD-AD data partitions and evaluation protocol for fair comparison.

C Ablation Analysis of M^2D^3

To investigate the contribution of each component in M^2D^3, we conduct ablation experiments on multi-model heterogeneous distillation (MHD) and pseudo-label self-training (PST). All variants use the same training settings and are evaluated under the 1% annotation ratio using classification accuracy and detection mAP_{50}. As shown in Table 8 and Table 9, both single-teacher variants outperform the epoch-matched baseline, confirming that the representations transferred from DINOv3 and EUPE are beneficial for defect recognition. Combining the two teachers further improves perfor-

Table 8: Ablation results for defect classification on WISD-AD under the 1% annotation setting. All variants use the same training epochs.

MHD		PST	Accuracy (%)
DINOv3	EUPE		
-	-	-	39.54
✓	-	-	50.94
-	✓	-	50.75
✓	✓	-	52.05
-	-	✓	46.88
✓	-	✓	51.37
-	✓	✓	52.05
✓	✓	✓	**54.02**

Table 9: Ablation results for defect detection on WISD-AD under the 1% annotation setting. All variants use the same training epochs.

MHD		PST	mAP_{50} (%)
DINOv3	EUPE		
-	-	-	12.23
✓	-	-	34.02
-	✓	-	33.74
✓	✓	-	36.85
-	-	✓	13.31
✓	-	✓	37.08
-	✓	✓	37.66
✓	✓	✓	**44.60**

Table 10: Robustness analysis across independently sampled labeled subsets on WISD-AD under 1% and 5% annotation settings. Results are reported as mAP_{50} (%).

Method	**1% Annotation**				**5% Annotation**			
	Split-1	Split-2	Split-3	Mean $\pm$ Std.	Split-1	Split-2	Split-3	Mean $\pm$ Std.
YOLO11	12.23	12.21	11.68	12.04 ± 0.31	30.08	29.83	29.45	29.79 ± 0.32
STEP-DETR	34.38	33.92	33.75	34.02 ± 0.33	53.92	53.77	53.34	53.68 ± 0.30
M^2D^3 (Ours)	**44.60**	**44.12**	**43.94**	$\mathbf{44.22 \pm 0.34}$	**68.17**	**68.03**	**67.55**	$\mathbf{67.92 \pm 0.33}$

mance, indicating that their structural and semantic representations are complementary. In contrast, PST alone provides only limited gains, as the weak initial representation learned from extremely limited labeled data may produce noisy pseudo labels. By combining heterogeneous distillation with pseudo-label self-training, the complete M^2D^3 achieves the best performance, demonstrating that strong representation initialization is essential for reliable pseudo-label generation and effective semi-supervised learning.

Robustness to Labeled Subset Sampling

Under extremely limited annotation budgets, model performance may be sensitive to the particular samples selected as labeled data. To evaluate whether the improvements of M^2D^3 depend on a specific favorable labeled subset, we further conduct a robustness analysis under the 1% and 5% annotation settings. For each annotation ratio, we independently construct three labeled subsets using the same class-wise stratified sampling strategy adopted in the main experiments. Within each subset, all compared methods use exactly the same labeled and unlabeled samples. The held-out evaluation set, training hyperparameters, data augmentation, and training random seed are kept fixed across different subsets, such that the primary source of variation is the selection of labeled samples. No hyperparameter is re-tuned for individual subsets.

We compare M^2D^3 with representative supervised and semi-supervised baselines and report the mean and standard deviation across three independently sampled labeled subsets. The results are summarized in Table 10. As expected, the absolute performance varies slightly across different labeled subsets, reflecting the sensitivity of low-label learning to the composition of labeled data. Nevertheless, the variations remain small across all methods and annotation settings. More importantly, M^2D^3 consistently outperforms STEP-DETR across all subset realizations, with average improvements of approximately 10.2 and 14.2 mAP_{50} under the 1% and 5% annotation settings, respectively. These performance margins are substantially larger than the variations caused by labeled-subset sampling, indicating that the gains of M^2D^3 are consistent across different labeled subsets rather than being attributable to a particular favorable split.

D PSEUDO-LABEL RELIABILITY ANALYSIS

To investigate the effect of heterogeneous multi-teacher distillation on pseudo-label reliability, we evaluate the predictions generated by the EMA teacher on the unlabeled training set after each self-training round. Ground-truth annotations are used only for offline evaluation and are never involved in training, pseudo-label filtering, or hyperparameter selection. We compare random initialization, conventional ImageNet and COCO pretraining, single-teacher distillation using DINOv3 or EUPE, and the proposed heterogeneous mutil-teacher distillation. All variants share the same labeled subset, confidence threshold, EMA update strategy, data augmentation, and self-training schedule, and pseudo-label quality is measured using $_{50}$. As shown in Fig. 9, random initialization produces the

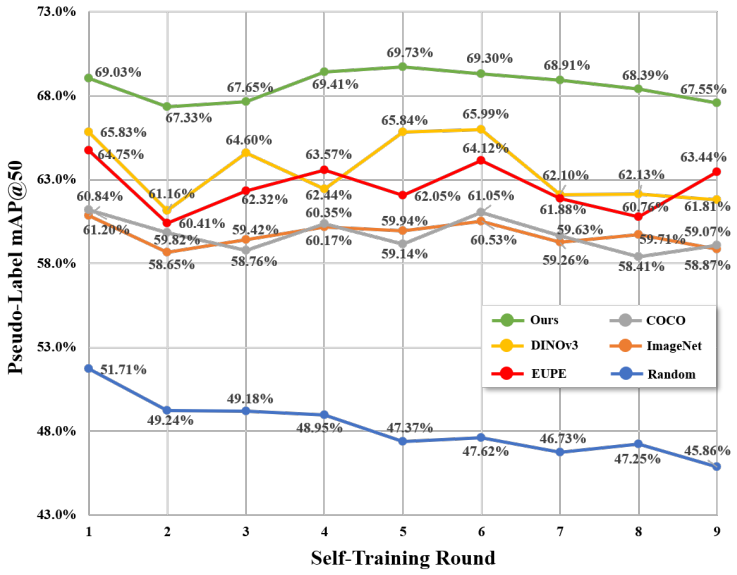


Figure 9: Pseudo-label $_{50}$ across self-training rounds under different initialization strategies.

lowest pseudo-label accuracy and gradually decreases from 51.71% to 45.86%, indicating poor convergence and increasing sensitivity to pseudo-label noise. ImageNet and COCO pretraining improve the initial representation but remain around 58%–61%, suggesting that general-purpose pretraining provides limited adaptation to fine-grained defects and complex industrial backgrounds. Single-teacher distillation from DINOv3 or EUPE further improves pseudo-label quality, but both exhibit noticeable fluctuations because the knowledge transferred from an individual teacher remains incomplete and may generalize poorly to specific defect patterns. In contrast, the proposed heterogeneous multi-teacher distillation consistently achieves the highest mAP_{50} throughout self-training. By transferring complementary structural and semantic knowledge from different teachers, it provides a more task-adaptive initialization for WISD-AD and alleviates the representation limitations of single-teacher distillation, thereby generating more accurate and stable pseudo labels.

E FURTHER COMPARATIVE ON PUBLIC INDUSTRIAL DATASET

To further assess the generalizability of M^2D^3 beyond WISD-AD, we conduct additional experiments on the public industrial defect dataset NEU-DET. Following the semi-supervised protocol adopted on WISD-AD, we construct labeled subsets with annotation ratios ranging from 1% to 100%, while the remaining training samples are treated as unlabeled data in the semi-supervised settings. All methods are evaluated using the same data partitions and evaluation protocol to ensure a fair comparison.

As shown in Table 11, M^2D^3 shows clear advantages under limited supervision. With only 1% and 5% labeled data, it achieves 51.1 and 62.7 mAP_{50}, outperforming the strongest competing methods by 6.6 and 7.3 points, respectively. As the annotation ratio increases, the performance gap gradually

Table 11: Comparison of supervised and semi-supervised detection methods on the public NEU-DET dataset under different annotation ratios. Results are reported in terms of mAP_{50} (%).

Methods	Annotation Ratio							
	1%	5%	10%	15%	20%	25%	50%	100%
YOLO11Ultralytics (2024)	21.8	54.4	61.6	65.4	68.4	69.1	73.9	75.7
YOLO26Chakrabarty (2026)	21.3	43.3	47.7	54.4	58.4	62.7	68.1	76.3
RT-DETRZhao et al. (2024)	2.3	26.2	52.9	53.1	60.8	67.4	68.2	70.4
DINOZhang et al. (2022)	3.8	29.5	55.4	57.8	63.2	69.8	71.5	72.8
DEIMHuang et al. (2025b)	24.2	55.4	63.2	64.3	67.1	69.0	74.3	76.7
STACSohn et al. (2020)	32.4	44.1	57.0	60.6	61.1	63.5	68.5	70.6
Soft TeacherXu et al. (2021)	36.9	37.0	42.8	43.1	67.5	69.7	71.9	73.4
Unbiased TeacherLiu et al. (2021a)	30.4	40.8	50.3	56.9	64.2	68.4	70.8	73.1
Efficient TeacherXu et al. (2023)	29.7	38.0	47.7	54.0	66.5	68.3	70.9	72.4
ARSLLiu et al. (2023)	28.7	37.4	46.0	56.7	60.5	67.5	70.1	72.8
Semi-DETRZhang et al. (2023)	39.5	45.2	52.8	59.4	68.1	70.5	72.6	74.5
Sparse Semi-DETRShehzadi et al. (2024)	42.1	49.5	56.4	62.1	69.5	70.9	73.8	75.2
Semi-SSDDetLiu et al. (2025)	38.5	48.7	56.0	60.3	64.2	68.8	71.5	73.1
STEP-DETRShehzadi et al. (2025)	44.5	52.8	59.1	64.3	68.7	69.0	74.4	76.1
M^2D^3 (Ours)	**51.1**	**62.7**	**65.3**	**66.4**	**69.5**	**70.2**	**75.0**	**77.2**

Table 12: Deployment efficiency comparison under the 1% annotation setting.

Method	Params. (M)	FLOPs (G)	Latency (ms)	mAP_{50} (%)
YOLO11-nUltralytics (2024)	2.6	6.5	3.80	12.23
RT-DETRZhao et al. (2024)	42.8	130.5	4.97	0.16
YOLO26 Chakrabarty (2026)	13.7	3.7	3.42	8.23
Efficient TeacherXu et al. (2023)	7.2	16.5	4.26	29.17
Unbiased TeacherLiu et al. (2021a)	39.5	204.1	5.73	26.37
M^2D^3(Ours)	2.6	6.5	3.87	44.60

narrows, while M^2D^3 remains competitive across all settings and achieves the best or tied-best performance in seven of the eight annotation ratios. These results indicate that the benefit of M^2D^3 is most pronounced in low-label scenarios, while its effectiveness is maintained as more labeled data become available.

The performance trend is consistent with the design motivation of M^2D^3. Under extremely limited annotations, heterogeneous foundation-model distillation provides the downstream detector with stronger initial representations, which improves the quality of subsequent pseudo-label generation and makes self-training more reliable. As the amount of labeled data increases, stronger task-specific supervision reduces the relative dependence on transferred priors, and the performance gap between different methods naturally narrows. Nevertheless, M^2D^3 remains competitive across all annotation ratios and achieves the best or tied-best performance in seven of the eight settings. Furthermore, the strong low-label performance on NEU-DET indicates that the effectiveness of M^2D^3 is not confined to the specific data distribution or acquisition conditions of WISD-AD. The consistent improvements observed across two distinct industrial defect benchmarks suggest that the proposed framework can generalize across different defect characteristics and visual domains. These results provide further support for the applicability of M^2D^3 to annotation-efficient industrial defect detection.

F Deployment Efficiency Analysis of M^2D^3

To evaluate the practical applicability of M^2D^3, we analyze whether the additional training components introduce extra computational overhead during deployment. We measure the number of parameters, FLOPs, inference latency, and detection performance on the same hardware platform, using batch size 1 and the same input resolution for all models.

As shown in Table 12, M^2D^3 substantially improves detection accuracy while maintaining nearly the same parameter count, FLOPs, and inference latency as the original downstream model. DINOv3,

EUPE, the projection heads, and the EMA teacher are used only during training, while only the compact student model is retained for inference. Therefore, M^2D^3 improves annotation efficiency without introducing noticeable additional inference overhead and preserves the inference efficiency of the downstream model on the evaluated hardware.

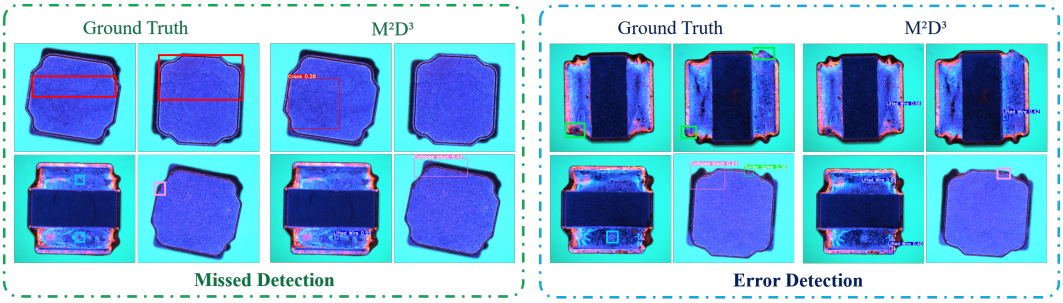


Figure 10: Representative challenging and failure cases of M^2D^3 on WISD-AD. Missed detections mainly occur for small and low-saliency defects, while erroneous detections are often caused by visually similar categories and complex background interference.

G Challenging and Failure Case Analysis

Although M^2D^3 achieves strong overall performance across different annotation ratios, WISD-AD still contains challenging scenarios. To further investigate these difficulties and the limitations of M^2D^3 under extremely low-label conditions, we analyze representative challenging and failure cases in Fig. 10. As shown in Fig 10, in the first row of the Missed Detection examples, defects with extremely low saliency exhibit severe missed detections and inaccurate bounding-box localization. Additionally, small-scale defects in the second row are also prone to omission and incomplete localization, highlighting the need to strengthen the model's ability to detect weakly salient and cross-scale anomalies. Furthermore, in the Error Detection examples, some defects such as Lifted Wire and Solder Spike exhibit high inter-class similarity, making it difficult for the model to distinguish visually similar categories. In addition, Lifted Wire, Solder Spike, and Missing Weld are often accompanied by complex background interference, which further disrupts the model's perception and leads to erroneous detections. These challenges are particularly pronounced for rare and long-tailed defect categories, further emphasizing the importance of enhanced feature representation and context modeling in WISD-AD for capturing subtle and challenging defect characteristics.

H Future Extensions of WISD-AD

Although WISD-AD provides a unified benchmark for fully supervised and semi-supervised defect inspection, several directions remain to be explored. First, future versions will extend the benchmark to unsupervised and anomaly-detection settings, enabling models to learn defect representations without relying on category-level or bounding-box annotations. Second, multimodal data, such as depth, thermal, and production-process signals, can be incorporated to improve the perception of weakly visible and structurally ambiguous defects. Third, the diversity of defect categories will be further expanded to cover additional appearance, structural, and process-related anomalies. Finally, we plan to continuously increase the dataset scale by collecting images from more production batches, devices, and operating conditions, thereby improving its representativeness and supporting more reliable evaluation of model generalization.

I Recommendations for the Application of the Proposed M^2D^3 and WISD-AD Dataset

Based on the above extensive analyses and experiments, we provide practical recommendations for applying M^2D^3 and using WISD-AD under different research and deployment scenarios.

- **Application of M^2D^3.** For benchmark comparison, we recommend strictly following the default two-stage configuration reported in this work, including the teacher models, training schedules, pseudo-label thresholds, EMA momentum, and downstream architectures, to ensure fair and reproducible comparison.

  When applying M^2D^3 to a novel model or dataset, the two-stage pipeline should be retained, while the projection heads, distillation schedule, and self-training parameters can be adapted to the target setting. In extremely low-annotation or large-domain-gap scenarios, Stage I heterogeneous distillation should be prioritized to provide reliable initialization. When pseudo labels are unstable because of class imbalance, visually similar defects, or distribution-shifted unlabeled data, more conservative confidence thresholds and self-training schedules are recommended.

- **Application of WISD-AD.** For benchmark comparison, researchers should directly use the official training and evaluation splits, annotation ratios, and evaluation protocols. In particular, the predefined annotation ratios of 1%, 5%, 10%, 15%, 20%, and 25% should be retained, and samples from the same production batch should remain exclusively in either the training or evaluation set.

  For customized data partitioning, production-batch isolation should still be maintained to avoid information leakage. The labeled subsets should preserve representative class, defect-scale, and spatial distributions, especially under extremely low annotation ratios. Any newly constructed split or extended task protocol should be clearly distinguished from the official benchmark, and its results should not be directly compared with the reported benchmark results unless the same evaluation setting is used.

  These recommendations serve as practical references, while final configurations should be determined according to the characteristics of the target model, dataset, and application scenario.