



How Many Normal Images Are Needed? Training-Free Few-Shot Anomaly Detection on Scanning Electron Microscope Images of Integrated Circuits

K. Susmitha Reddy¹, G. Lavanya Devi²

Department of Computer Science and Systems Engineering, Andhra University College of Engineering (A),
Visakhapatnam, India¹

Department of Computer Science and Systems Engineering, Andhra University College of Engineering (A),
Visakhapatnam, India²

Abstract: Anomaly detection on scanning electron microscope (SEM) images of integrated circuits is an essential element in the semiconductor manufacturing quality control process to ensure that no defective chip is delivered. Because SEM data is specific to each product and imaging tool, inspection models are usually built or tuned for each production line. Developing these models involves a large quantity of images and takes considerable effort to collect the data and refine the model until it works reliably. Also, most of the time, collecting data results in a disproportionate amount of normal images compared to anomalous images. Vision foundation models, pretrained on very large image collections, can describe images in great detail and have become strong tools for finding anomalies in photographs. Even so, the best anomaly detection models on the public MIIC dataset have been trained on 23,888 normal images, which is still a large number. In this research, we studied training-free few-shot anomaly detection on MIIC using frozen vision foundation backbones to assess whether their features transfer to SEM images. We extracted features from k normal reference images to form a memory bank, and assigned an anomaly score to each test feature based on its distance to the nearest stored reference feature in the memory bank. We tested six backbones with four values of k and three random seeds. DINOv2-B with 64 references yielded 94.53% image AUROC, which was better than full-data PatchCore (93.33%), while DINOv3 and CLIP were less transferable. A layer analysis revealed that averaging the last four transformer layers played a critical role in localization, raising the AUPRO of DINOv3-S to 95.2% compared with 76.2% for the last layer alone. When the 16 most similar normal references were retrieved for every test image, AUPRO reached 99.23%, nearly equivalent to full-data CFA (99.04%), and the nearest-neighbour search was 129 times quicker than searching through a 2,000-image pool. However, image-level AUROC (94.78%) lagged behind full-data CFA (99.01%). A decision layer calibrated solely on normal images rejected only 0.39% of normal images and detected every anomaly covering at least 0.25% of the image.

Keywords: Few-shot anomaly detection, foundation models, defect inspection, semiconductor manufacturing, electron microscopy.

I. INTRODUCTION

Automated inspection of scanning electron microscope (SEM) images has a direct effect on how quickly a manufacturing plant can locate and resolve a yield drop. Anomalies such as foreign material, dust, broken wires or contamination are infrequent and unpredictable. Therefore unsupervised anomaly detection, which only learns what normal images look like, is a logical formulation. The main limitation of unsupervised anomaly detection is data availability: every new product or device has to be onboarded, which means collecting new data and retraining the model.

MIIC [1] is a public benchmark of SEM images showing the metal layers of integrated circuits. On this dataset, Barusco et al. [2] evaluated ten previously published methods, all of which use ImageNet-pretrained convolutional networks and the entire training set of 23,888 normal images, and CFA [3] was the highest performing method with 99.01% image AUROC. At the same time, training-free methods employing self-supervised vision foundation models have become the strongest approach for few-shot anomaly detection on images of industrial objects [4], [5]. It is unclear if these features will translate to grayscale SEM images, which differ greatly from the photographs on which they were trained. Prior studies have shown varying degrees of success when transferring features from one imaging modality to another [6].



This paper questions how many normal images a foundation model needs to inspect SEM images of integrated circuits, and how to choose them. Our contributions are:

- To our knowledge, the first systematic evaluation of training-free few-shot anomaly detection on MIIC, covering six backbones, $k \in \{1, 4, 16, 64\}$ reference images and three seeds. With 64 images, DINOv2-B exceeds full-data PatchCore.
- A layer and resolution analysis showing that final-layer features localize poorly on SEM images, especially for DINOv3, and that averaging the last four layers recovers localization.
- A training-free, retrieval-based reference selection: the 16 most similar normal images per test image match full-pool and full-data localization at a fraction of the search cost.
- A calibrated, size-tolerant decision layer evaluated with overkill and underkill rates.

Each inspected image is compared only with its 16 most similar normal references

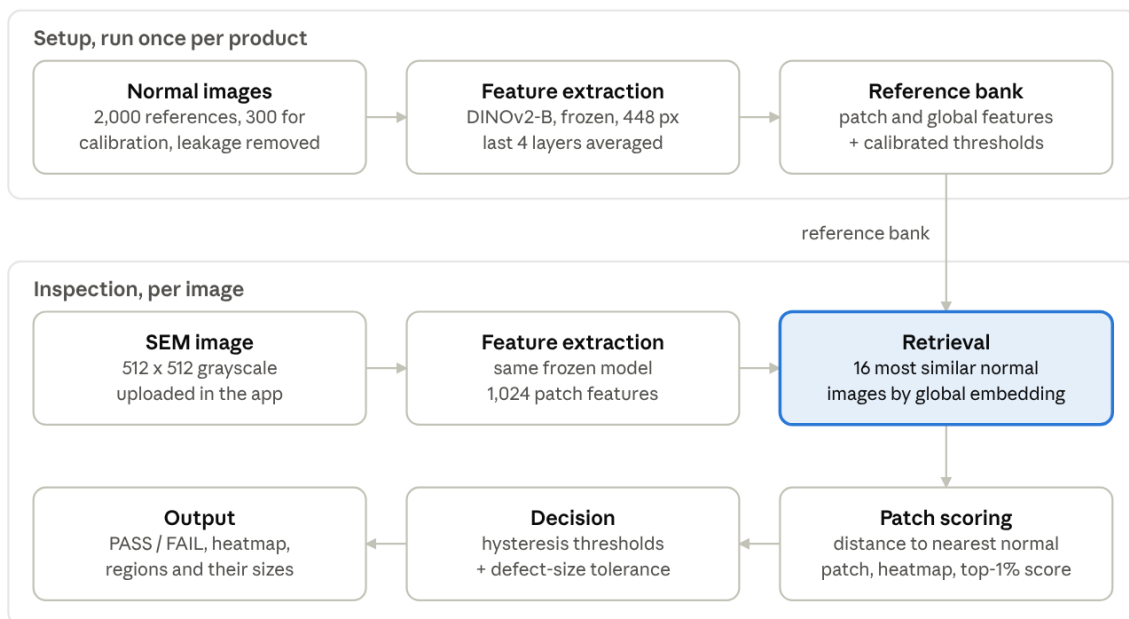


Fig. 1. Setup overview. The setup embeds a pool of normal images once and then establishes decision thresholds on additional normal images. Each image being inspected is embedded by the same frozen model and compared only to its 16 most similar normal references. A size tolerance is then applied to decide whether the image passes inspection. None of the components require training or anomalous images.

II. RELATED WORK

Industrial anomaly detection. MVTec AD [7] and VisA [8] established benchmarks where models only see normal images. Reconstruction [9], [10], knowledge-distillation [11], [12] and normalizing-flow [13] methods all require training. Memory-bank methods such as PatchCore [14] store patch features of normal images and score test patches by nearest-neighbour distance; CFA [3] adapts such features towards compact normal clusters.

Foundation models and few-shot detection. Self-supervised vision transformers such as DINOv2 [15] and DINOv3 [16], and image-text models like CLIP [17], provide general patch features of an image. WinCLIP [18] and AnomalyCLIP [19] use CLIP with text prompts. AnomalyDINO [4] showed that nearest-neighbour matching of frozen DINOv2 patches is a strong few-shot detector, and SubspaceAD [5] and DuoAD [20] refine this training-free line; DuoAD reports that DINOv3 needs multi-layer aggregation to surpass DINOv2. All are evaluated on photographs of objects.

Semiconductor and SEM inspection. Huang et al. [1] introduced MIIC with a self-supervised detection and inpainting model. Barusco et al. [2] benchmarked ten methods on it with full training data. SEM-CLIP [21] and IC-DefectNCD [22] apply CLIP and DINOv3 to SEM defects but require labelled defects or training and use private data. A recent SEM



foundation model [23] targets image restoration. Gharbage et al. [6] found that ImageNet pretraining can outperform DINOv3 under an X-ray modality shift.

Retrieval of references. RAID [24] retrieves normal references for each test image but trains a denoising module on top. Conventional wafer inspection compares a die with its neighbour or with the design layout. Our retrieval step is a training-free counterpart of that comparison.

III. METHODOLOGY

Fig. 1 depicts the overall methodology. With a restricted number of normal images, it generates an anomaly score, a heatmap and a PASS or FAIL verdict for each test image without training any networks.

A. Patch Features

A frozen encoder ϕ receives the grayscale image, copied to three channels and resized to 448×448 pixels, which yields a grid of 32×32 patches for DINOv2 (patch size 14). Let $\mathbf{h}^{(l)}(p)$ be the token of patch p at transformer layer l . We pass each token through the encoder's final layer normalization (LN), average over a set of layers $\mathcal{L}$ and normalize to unit length:

$$f(p) = \frac{\bar{f}(p)}{\|\bar{f}(p)\|}, \quad \bar{f}(p) = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \text{LN}(h^{(l)}(p)) \quad (1)$$

By default $\mathcal{L}$ contains only the final layer; Section V-B motivates using the last four. The class token, processed the same way, gives a global embedding $\mathbf{g}(I)$ of image I .

B. Nearest-Neighbour Scoring

The patch features of k normal reference images form a memory bank $\mathcal{M}$. Each test patch is scored by its cosine distance to the nearest stored patch, and the image score is the mean over the top 1% of patch scores $\mathcal{T}$:

$$s(p) = 1 - \max_{m \in \mathcal{M}} f(p)^\top m, \quad S(I) = \frac{1}{|\mathcal{T}|} \sum_{p \in \mathcal{T}} s(p). \quad (2)$$

Averaging the top 1% keeps a small defect from being diluted by hundreds of normal patches without relying on a single noisy patch. The anomaly map is $s(p)$ upsampled bilinearly to 256×256 and smoothed with a Gaussian filter ($\sigma = 4$).

C. Retrieval-Based Reference Selection

Integrated circuits have repeated layouts, so a pool $\mathcal{P}$ of normal images usually contains images of the same structures as a test image. Instead of a fixed random set of k references, we select for each test image the k pool images with the highest cosine similarity $\mathbf{g}(I)^\top \mathbf{g}(J)$ and build $\mathcal{M}$ from them only. This keeps the memory bank at $k \times 1,024$ patches regardless of the pool size, so the search cost grows with k , not with $|\mathcal{P}|$.

D. Calibrated Size-Tolerant Decision

Thresholds are set from a calibration set $\mathcal{C}$ of normal images disjoint from $\mathcal{P}$; no defect image is used. The high threshold τ_h is the q -quantile of the per-image maximum of the anomaly map over $\mathcal{C}$, which fixes the false-alarm rate on normal images in advance; the low threshold τ_l is the 0.999-quantile of all anomaly-map pixels of $\mathcal{C}$. A defect region is a connected component of $\{x : A(x) > \tau_l\}$ that contains at least one pixel above τ_h (hysteresis thresholding). An image fails if any region covers at least a tolerance t , expressed as a percentage of the image area, mirroring industrial acceptance rules that allow defects below a stated size.

IV. EXPERIMENTAL SETUP

Dataset. MIIC [1] contains 512×512 grayscale SEM images of IC metal layers. The training folder as released holds 25,276 images, including copies of all 1,388 test images. We removed them by filename and by exact file hash (MD5), leaving 23,888 normal training images, the number used in [2]. The test set has 1,272 normal and 116 defective images with pixel masks. Defects are small: the mean defect covers 1.23% of the image, and 42% of defective images contain no defect region as large as 0.1%.

Backbones. We compare an ImageNet-supervised WideResNet-50-2 (WRN50), using PatchCore-style features from layers 2 and 3 at 224 px with Euclidean distance and no coreset subsampling, against DINOv2-S/B (ViT/14), DINOv3-S/B (ViT/16) and CLIP ViT-B/16, all at the same 448 px.



Protocol. Random reference sets are pulled from the cleaned training set with seeds 0, 1 and 2; we report mean $\pm$ standard deviation. For retrieval, $\mathcal{P}$ is a fixed random sample of 2,000 training images and $\mathcal{C}$ another 300. We report image AUROC, AUPRO integrated to a 30% false-positive rate, and pixel F1, computed on 256×256 maps. Pixel AUROC exceeded 96% for every configuration and is omitted. With 116 defective images, differences below about 1–1.5 AUROC points are within noise. Note that the methods compared in [2] use 224-pixel inputs and choose checkpoints based on a validation split of the test set.

All experiments ran on one NVIDIA T4 GPU.

V. RESULTS

A. Effect of the Backbone and of k

As seen in Table I and Fig. 2, all foundation models outperform the ImageNet CNN from $k = 4$ onwards. DINOv2-B is the best performing backbone from $k = 4$ onwards. With 64 images (0.27% of the training set), DINOv2-B got 94.53% image AUROC and 96.50% AUPRO, and both metrics were higher than those achieved by full-data PatchCore. With one reference, results differed greatly based on which image was chosen, with standard deviations of up to 9 points, and here DINOv3-S got the top score (80.94%). CLIP-B stopped getting better after $k = 4$. DINOv3 showed poor localization at each k , getting less than 84% AUPRO, although its image-level scores remained close to the other transformers.

TABLE I
RESULTS WITH K RANDOM NORMAL REFERENCES. EACH CELL GIVES THE MEAN $\pm$ STANDARD DEVIATION OVER 3 SEEDS, IN %. THE BEST VALUE OF EACH METRIC FOR EVERY K IS IN BOLD.

Backbone	$k = 1$		$k = 4$		$k = 16$		$k = 64$	
	AUROC	AUPRO	AUROC	AUPRO	AUROC	AUPRO	AUROC	AUPRO
DINOv2-B	70.77 $\pm$ 5.85	77.86 $\pm$ 5.79	92.65$\pm$1.37	96.81$\pm$1.99	92.19$\pm$1.13	94.61$\pm$1.26	94.53$\pm$0.49	96.50$\pm$0.26
DINOv2-S	70.09 $\pm$ 8.94	75.88 $\pm$ 11.49	88.54 $\pm$ 1.32	89.84 $\pm$ 6.57	90.53 $\pm$ 0.41	90.61 $\pm$ 0.17	92.35 $\pm$ 0.75	93.46 $\pm$ 1.34
DINOv3-B	77.07 $\pm$ 7.13	69.77 $\pm$ 6.96	87.31 $\pm$ 3.01	81.23 $\pm$ 9.82	88.37 $\pm$ 1.11	79.77 $\pm$ 1.95	89.97 $\pm$ 0.42	83.53 $\pm$ 1.43
DINOv3-S	80.94$\pm$2.65	68.43 $\pm$ 3.82	87.58 $\pm$ 2.51	77.93 $\pm$ 8.69	88.10 $\pm$ 0.55	76.23 $\pm$ 0.71	89.78 $\pm$ 0.82	80.66 $\pm$ 0.67
CLIP-B	79.88 $\pm$ 5.58	82.42$\pm$7.54	89.38 $\pm$ 2.48	89.81 $\pm$ 4.96	88.11 $\pm$ 1.02	89.07 $\pm$ 2.43	88.17 $\pm$ 0.12	88.46 $\pm$ 1.09
WRN50	70.25 $\pm$ 1.67	69.58 $\pm$ 0.78	81.28 $\pm$ 4.42	82.05 $\pm$ 4.73	85.13 $\pm$ 0.99	84.14 $\pm$ 2.85	86.98 $\pm$ 1.84	88.86 $\pm$ 0.89

Full-data results from [2]: CFA 99.01 / 99.04 and PatchCore 93.33 / 93.03 (image AUROC / AUPRO).

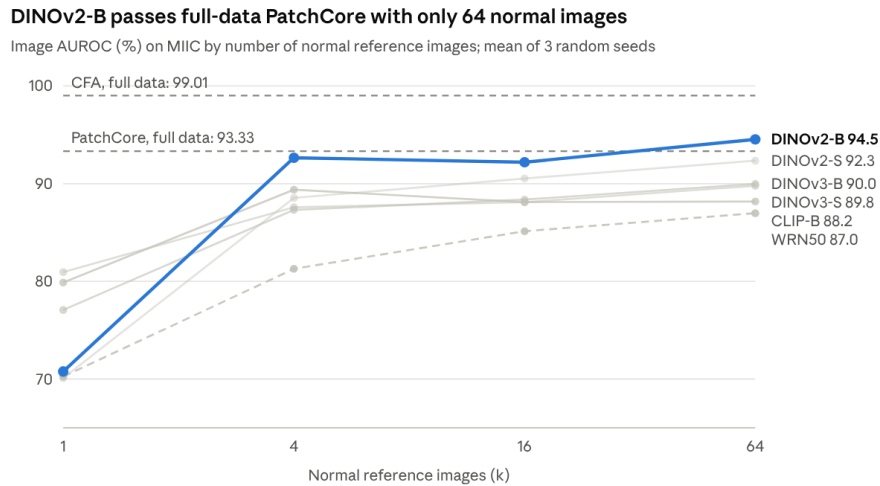


Fig. 2. Image AUROC as the number of random normal references grows (mean of 3 seeds). The dashed lines show full-data CFA and PatchCore [2].

B. Layer Selection and Input Resolution

Fig. 3 illustrates why DINOv3 performs poorly. Its final layer gives its highest image AUROC (88.10%) but its lowest AUPRO (76.23%), while layer 10 provides 95.97% AUPRO. When the final four layers are combined, DINOv3-S



reaches 95.17% AUPRO and DINOv2-S improves from 90.61% to 97.98%, while image AUROC hardly changed (it moved from 90.53% to 90.16% for DINOv2-S). These findings are similar to the observations reported by DuoAD on natural images [20]. Earlier layers (2 to 6) gave poorer image-level scores. Their AUPRO scores seemed reasonable (83 to 93%), but the pixel F1 of DINOv2-S at layers 4 and 6 was below 10%, which suggests that AUPRO alone can make blurry maps look better than they are.

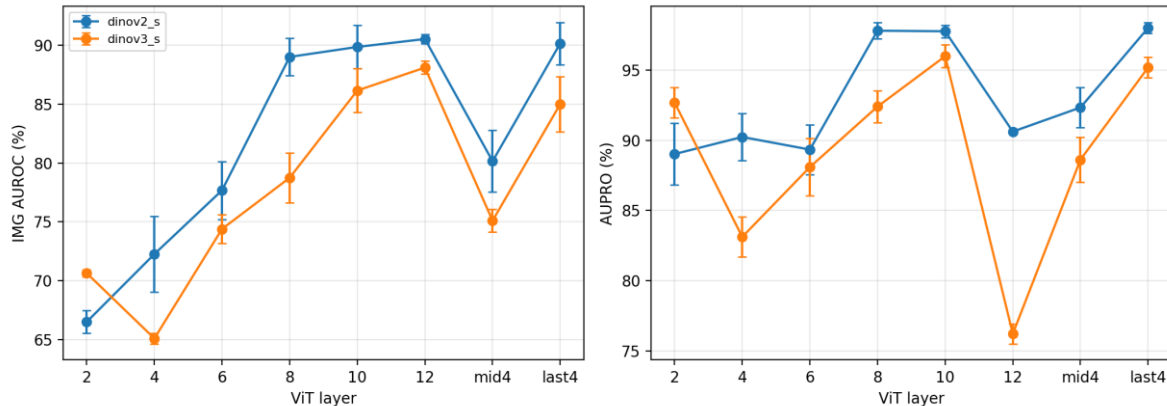


Fig. 3. How the chosen layer affects image AUROC (left) and AUPRO (right) for $k = 16$ over 3 seeds. “mid4” averages layers 5 to 8 and “last4” averages layers 9 to 12.

To check whether these gains came from the backbone or simply from the larger input, we also ran DINOv2-S at 224 pixels, the same resolution used for WRN50. At $k = 64$ it still gave a higher image AUROC than WRN50 (91.32% against 86.98%) and about twice its pixel F1 (53.25% against 29.90%). The AUPRO values, however, were almost the same (89.10% and 88.86%). So the foundation model is better at deciding whether an image is defective at either resolution, but part of its localization advantage comes from the 448-pixel input.

C. Random and Retrieved References

Table II compares random and retrieved references drawn from the same image pool. We observed that retrieval helps most when references are few: with a single reference it lifts image AUROC from 76.08% to 90.91% (DINOv2-S) and from 77.98% to 90.38% (DINOv2-B), and it removes the dependence on which images are drawn. For DINOv2-S, 16 retrieved references match the full pool (94.07% vs. 94.06%) with a nearest-neighbour search of 1.82 ms instead of 171.94 ms per image. Combined with the last-4 layer average, DINOv2-B with 16 retrieved references reaches 99.23% AUPRO, on par with full-data CFA (99.04%), and 94.78% image AUROC, 4.2 points below CFA; its search takes 1.90 ms against 245.93 ms for the full pool, a 129 $\times$ reduction. We use this configuration in the decision layer.

TABLE II

RANDOM AND RETRIEVED REFERENCES TAKEN FROM THE SAME POOL OF 2,000 NORMAL IMAGES (IMAGE AUROC AND AUPRO IN %). RANDOM SETS ARE THE MEAN $\pm$ STANDARD DEVIATION OVER 3 SEEDS; RETRIEVAL HAS NO RANDOMNESS. “POOL” MEANS THAT ALL 2,000 POOL IMAGES WERE USED AS REFERENCES.

Refs.	$k = 1$		$k = 4$		$k = 16$		$k = 64$	
	AUROC	AUPRO	AUROC	AUPRO	AUROC	AUPRO	AUROC	AUPRO
<i>DINOv2-S</i>								
Random	76.08±8.23	77.02±5.21	83.37±9.89	83.19±13.40	91.37±0.51	90.82±1.47	92.60±0.38	93.26±1.46
Retrieved	90.91	96.41	92.97	97.04	94.07	97.28	94.20	97.13
Pool	AUROC 94.06 / AUPRO 96.64 (all 2,000 pool images; independent of k)							
<i>DINOv2-B, last 4 layers</i>								
Random	77.98±6.18	92.35±1.02	85.31±7.34	94.33±4.63	90.51±2.73	96.81±1.63	94.43±0.64	98.01±0.15
Retrieved	90.38	97.98	93.12	98.94	94.78	99.23	95.85	99.15
Pool	AUROC 96.12 / AUPRO 99.05 (all 2,000 pool images; independent of k)							



D. Decision Layer

Table III reports overkill (good images rejected) and underkill (defective images passed) for the tolerance constraint. An image counts as defective under tolerance t if its ground truth contains a region of at least t . With $q = 0.99$ and zero tolerance, only 5 of 1,272 good images are rejected, but 46 of 116 defective images pass; at least 42 of these contain only defects under 0.1% of the image, a few dozen pixels. Every defect covering 0.25% or more is detected. Detected regions overestimate defect area by a median factor of 5.5 (Spearman $\rho = 0.56$ with true area), because of the coarse patch grid and the smoothing; this raises overkill slightly at larger tolerances, but the errors fall on the safe side. In our interactive application, a complete inspection, including feature extraction and retrieval, takes under 80 ms per image on a T4 GPU.

TABLE III
PASS/FAIL RESULTS FOR DIFFERENT CALIBRATION QUANTILES Q AND TOLERANCES T , USING DINOv2-B WITH THE LAST FOUR LAYERS AND 16 RETRIEVED REFERENCES.

q	t (% of image)	Accuracy (%)	Overkill (%)	Underkill (%)
0.95	0	96.47	1.65	24.14
0.99	0	96.33	0.39	39.66
0.99	0.1	98.85	0.91	5.97
0.99	0.25	97.84	2.23	0.00
0.99	0.5	96.97	3.10	0.00

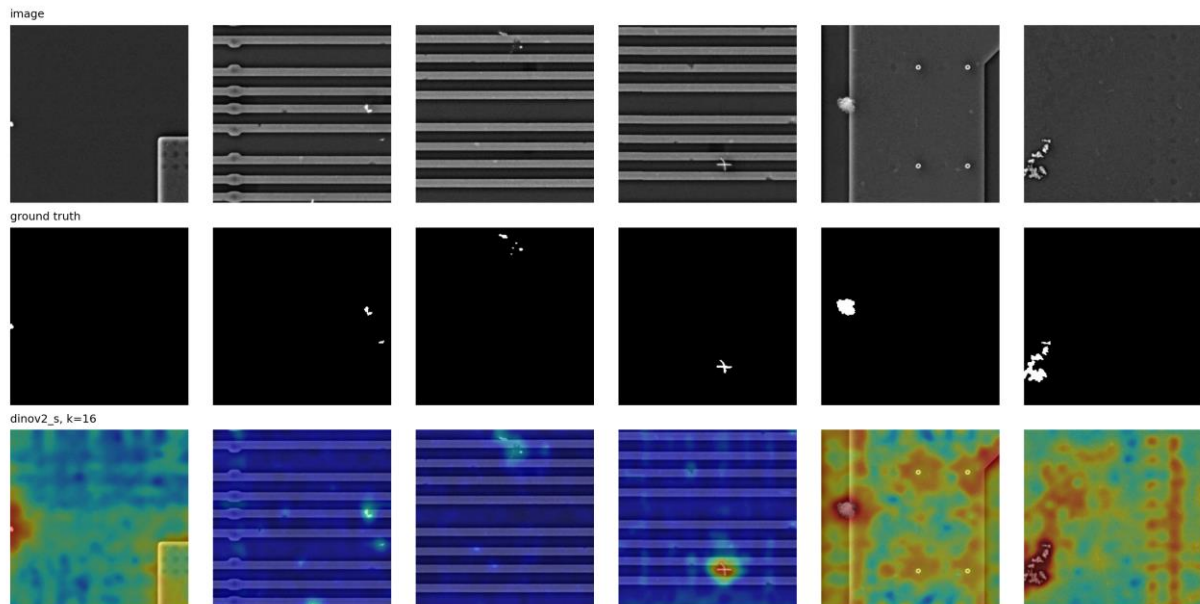


Fig. 4. Anomalous test images (top), their ground-truth masks (middle) and anomaly maps (bottom), produced by DINOv2-S with 16 random references and one shared color scale. Particles and a broken line are clearly marked in columns 2 and 4. Plain areas in columns 1, 5 and 6 also receive higher scores, which can lead to false alarms.

VI. DISCUSSION AND LIMITATIONS

Why do foundation features transfer? SEM images of metal layers consist of textures and line patterns rather than objects. A plausible explanation for the gap over WRN50, which holds at equal resolution, is that self-supervised features describe local structure, while ImageNet-supervised CNN features are tuned to separate object categories. Newer is not automatically better, however: under this change of modality DINOv3 needs multi-layer aggregation and still trails DINOv2, consistent with [6].

Why so few images are enough. Because image layouts are repetitive, four normal images already cover most local structures: DINOv2-B gains less than two points of image AUROC from $k = 4$ to $k = 64$. Retrieval pushes this further by choosing the references that match each test image, a learned analogue of die-to-die comparison. It assumes a pool of normal images and their stored features (about 3.1 GB for 2,000 images with DINOv2-B), so it is an efficiency result



rather than a strict few-shot result; the assumption is realistic because normal images are plentiful in a fab while labelled defects are not.

Limitations. Our results are based on a single dataset, and with only 116 defective images small differences are within noise. Retrieval was tested with a single pool, which may contain a few unlabelled anomalies, and one retrieved reference that we inspected showed a small particle. At a 1% false-alarm rate, anomalies smaller than 0.1% of the image were often missed, and the detected regions were larger than the actual anomalous regions.

VII. CONCLUSION

Frozen vision foundation models make few-shot inspection of SEM images of integrated circuits more practical. With 64 normal images, DINOv2-B surpasses full-data PatchCore on MIIC; averaging the last four layers is essential for localization, and retrieving the 16 most similar normal references matches the localization of the best full-data method at a small search cost, without training or defect labels. Future work will target the smallest defects with finer features, such as hairline fractures and chipping, and higher resolutions; curation of the reference pool; and evaluation on further SEM datasets and with larger or SEM-specific backbones [23].

ACKNOWLEDGMENT

The authors used AI assistants (Claude by Anthropic, and Consensus) for literature search and debugging, and Grammarly for editing and phrasing the manuscript text. All experiments were run by the authors, and the authors checked every result and claim and take full responsibility for the content.

REFERENCES

- [1]. L. Huang, D. Cheng, X. Yang, T. Lin, Y. Shi, K. Yang, B. H. Gwee, and B. Wen, "Joint anomaly detection and inpainting for microscopy images via deep self-supervised learning," in *Proc. IEEE ICIP*, 2021, pp. 3497–3501.
- [2]. M. Barusco, F. Borsatti, Y. Ben Khalifa, D. Dalle Pezze, and G. A. Susto, "Evaluating modern visual anomaly detection approaches in semiconductor manufacturing: A comparative study," *IFAC-PapersOnLine*, vol. 59, no. 26, pp. 103–108, 2025.
- [3]. S. Lee, S. Lee, and B. C. Song, "CFA: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization," *IEEE Access*, vol. 10, pp. 78446–78454, 2022.
- [4]. S. Damm, M. Laszkiewicz, J. Lederer, and A. Fischer, "AnomalyDINO: Boosting patch-based few-shot anomaly detection with DINOv2," in *Proc. IEEE/CVF WACV*, 2025, pp. 1319–1329.
- [5]. C. Lendering, E. Akdag, and E. Bondarev, "SubspaceAD: Training-free few-shot anomaly detection via subspace modeling," in *Proc. IEEE/CVF CVPR*, 2026, pp. 28557–28566.
- [6]. M. Gharbage, C. Teulière, P. Bouges, and T. Chateau, "Rethinking transfer learning for industrial inspection: DINOv3 vs. ImageNet pretraining across RGB and X-ray tasks," in *Proc. IEEE/CVF CVPR Workshops*, 2026, pp. 7260–7268.
- [7]. P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF CVPR*, 2019, pp. 9592–9600.
- [8]. Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "SPot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *Proc. ECCV*, 2022.
- [9]. S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. ACCV*, 2018.
- [10]. V. Zavrtanik, M. Kristan, and D. Skočaj, "DRAEM—A discriminatively trained reconstruction embedding for surface anomaly detection," in *Proc. IEEE/CVF ICCV*, 2021.
- [11]. G. Wang, S. Han, E. Ding, and D. Huang, "Student-teacher feature pyramid matching for anomaly detection," in *Proc. BMVC*, 2021.
- [12]. H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in *Proc. IEEE/CVF CVPR*, 2022.
- [13]. J. Yu, Y. Zheng, X. Wang, W. Li, Y. Wu, R. Zhao, and L. Wu, "FastFlow: Unsupervised anomaly detection and localization via 2D normalizing flows," arXiv:2111.07677, 2021.
- [14]. K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF CVPR*, 2022, pp. 14318–14328.
- [15]. M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, 2024.
- [16]. O. Siméoni *et al.*, "DINOv3," arXiv:2508.10104, 2025.



- [17]. A. Radford *et al.*, “Learning transferable visual models from natural language supervision,” in *Proc. ICML*, 2021, pp. 8748–8763.
- [18]. J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, “WinCLIP: Zero-/few-shot anomaly classification and segmentation,” in *Proc. IEEE/CVF CVPR*, 2023, pp. 19606–19616.
- [19]. Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, “AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection,” in *Proc. ICLR*, 2024.
- [20]. J.-Z. Tang, P.-H. Huang, M.-C. Chang, C.-F. Hsu, and J.-L. Li, “DuoAD: Leveraging [CLS] dual characteristics for training-free few-shot anomaly detection,” arXiv:2607.23924, 2026.
- [21]. Q. Jin, Y. Jiang, X. Lu, Y. Liu, Y. Chen, D. Gao, Q. Sun, and C. Zhuo, “SEM-CLIP: Precise few-shot learning for nanoscale defect detection in scanning electron microscope image,” in *Proc. IEEE/ACM ICCAD*, 2024.
- [22]. B. Zhao, X. Wang, S. Lyu, and Y. Lu, “Image-intrinsic priors for integrated circuit defect detection and novel class discovery via self-supervised learning,” arXiv:2511.03120, 2025.
- [23]. S. M. Ahmed *et al.*, “A mixture of experts foundation model for scanning electron microscopy image analysis,” arXiv:2604.05960, 2026.
- [24]. M. Cai, Z. Zhang, G. Wu, T. Chai, and X. Zhu, “RAID: Retrieval-augmented anomaly detection,” in *Proc. IEEE/CVF CVPR*, 2026, pp. 21367–21378.