applied sciences

MDPI

Article

Does Procedural Synthetic Pretraining Improve Few-Shot Industrial Defect Classification? An Evaluation with Frozen-Feature and Fine-Tuning Readouts

Jin W. Kim * and Ho-Sang Kim

Institute for Advanced Engineering (IAE), Yongin-si 17180, Gyeonggi-do, Republic of Korea; hoskim@iae.re.kr
* Correspondence: jinwkim153@iae.re.kr

Abstract

Procedural defect images are often added as an intermediate pretraining stage when labeled defects are scarce. We tested whether this stage (ImageNet → synthetic → real) improves few-shot defect classification over direct ImageNet → real fine-tuning on three public collections (NEU, MVTec-FS and GC10) and, as an exploratory case study, an eight-class optical wafer collection. Routes were compared in paired cells with matched heads, shuffled-label, equal-budget, head-only and generic-source controls, frozen and fine-tuned readouts, and a decision rule whose interval resamples training cells and test units together. Under this rule, synthetic pretraining did not improve the fine-tuned classifier on any collection. The largest gain, on GC10 with one labeled group per class (macro-F1 0.5569 to 0.5949 in the three pre-specified support selections), had an interval of −0.0128 to +0.0951 and was one of eleven unadjusted comparisons. Frozen readouts favored synthetic features on NEU and MVTec-FS without a corresponding fine-tuning gain. On wafers, an apparent support-dependent gain arose largely because one whole-wafer input carried two edge roll-off and two slip-line labels; after auditing the test units at the model-input level, no wafer difference met the rule. Synthetic pretraining should be evaluated with the intended classifier on audited units.

Keywords: synthetic data; industrial defect classification; few-shot learning; transfer learning; pseudoreplication; frozen features; model evaluation

Academic Editors: Pedro Couto, Ulises Orozco-Rosas, Kenia Picos and Diego Gragnaniello

Received: 6 October 2026
Revised: 7 November 2026
Accepted: 23 November 2026
Published: 27 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Industrial defect classification often has to be learned from a few labeled defects. Small real collections are enlarged by cropping, rotating or changing the contrast of a few defects, but this adds no distinct observations: treating such views as independent introduces pseudoreplication [1–3], and placing views of one defect in both training and test partitions causes leakage [4,5]. Synthetic pretraining offers another source of defect images: starting from ImageNet [6], a network is trained on generated defects and then adapted to real ones, whereas the reference is adapted directly from the same ImageNet weights. Because generating and training on synthetic images consumes resources that could be used to collect real defects, the synthetic stage should be judged on the classifier intended for use and on the number of distinct defects supporting the comparison.

The task is to classify a candidate with a supplied location; detection and images with several defects are outside its scope. A classification network has a backbone that extracts image features and a head that assigns classes. Frozen-feature evaluation fits a classification rule to unchanged features, whereas full fine-tuning updates both parts, so features that score well under one rule need not give the best fine-tuned model [7–10]. A rise in the synthetic-minus-ImageNet score can also come from a declining reference, unequal heads or schedules, a generator fault or test units that are not independent (repeated training seeds describe fitting variation without adding test defects [11]); these explanations must be examined before a change is attributed to synthetic pretraining.

We examine these explanations through three comparisons, formalized as hypotheses in Section 3.1: whether synthetic pretraining helps after full fine-tuning with matched target heads and schedules, whether frozen-readout gains predict fine-tuning gains, and whether a support-dependent gain reflects improvement of the synthetic model or decline of its comparator. The comparisons are made separately on NEU [12], MVTec-FS [13] and GC10-DET [14], which carry the pre-specified hypotheses and can be reproduced, and on an exploratory optical wafer collection that is reported as a secondary case study. The study proposes no new classifier or generator; it contributes a paired evaluation protocol with explicit hypotheses and decision criteria (Algorithm 1) and evidence on when an apparent synthetic-pretraining gain can be trusted. Three findings emerged. First, synthetic pretraining did not improve the fine-tuned classifier on any collection under the decision rule; the largest gain, on GC10 at the lowest support, was positive on average over ten support selections and in all eight source runs, but its interval included zero. Second, frozen-readout rankings did not predict fine-tuning outcomes. Third, the wafer case shows how evaluation units can create a gain: one whole-wafer image carrying two edge roll-off and two slip-line labels produced most of an apparent support-dependent gain, and no wafer difference met the decision rule once the test units were audited at the level of the model input. Section 2 reviews related work, Section 3 describes the methods, Section 4 reports the results, and Section 5.2 turns them into practical checks.

Algorithm 1. Paired evaluation of an additional synthetic pretraining stage.

Input: labeled pool with unit identities p and views V(p); synthetic source D; support levels K; target seeds; manifests M; readouts ϱ.

1. Group views, near-duplicates and candidates that give identical or nearly identical model inputs into one unit p, and split units into a support pool and a test set T; never split a unit.
2. For each manifest m, draw nested support $S(1, m) \subset \ldots \subset S(k_{max}, m)$ class by class without consulting T.
3. Train the source stages: S on D, Sh on D with labels shuffled per geometry and X on augmented views of S(k, m) at the budget of S.
4. For every (s, m, k), give all routes one random target head, schedule and augmentation stream, record unit-level predictions on T and fit the frozen readouts on S(k, m).
5. Compute F (Equation (1)), Δ(k; s, m) and each route's change between support levels (Equations (2) and (3)).
6. Summarize each endpoint by its cell mean, crossed range, class-stratified test-unit bootstrap and combined interval; for classes with fewer than five test units, add leave-one-out checks.
7. Apply the decision rule (combined interval) to H1–H4, report per-class F1 with test support and retain all conditions.

Output: endpoint estimates, intervals, per-class results and hypothesis verdicts (Section 4.7).

2. Related Work

Synthetic defects are used mainly in industrial anomaly detection, through cut-and-paste patches [15], simulated anomalies [16] or diffusion-synthesized anomalies [17,18] on benchmarks such as MVTec AD [19], sometimes reducing the gap between generated and target images explicitly [20]. Synthetic data also serve as pretraining sources, through domain randomization [21], formula-driven supervised learning (FDSL) on fractal images [22], synthetic wafer maps [23] or synthetic optical scratches [24]. In few-shot industrial classification, transfer depends on the number of labeled examples [25], and MVREC introduced the MVTec-FS benchmark used here [13].

The value of a pretrained representation depends on its use: ImageNet transfer varies with task and settings [26,27], extra source training can reduce target performance [28] or induce shortcuts [29], few-shot rankings change with the readout [7–9] and fine-tuning can distort pretrained features [10]. Pseudoreplication [1–3], leakage [4,5] and seed-only uncertainty [11] can each overstate a gain.

Table 1 summarizes representative studies; most report a gain for a single readout. To our knowledge, none compares frozen and fine-tuned readouts of the same synthetic source under a paired, unit-grouped evaluation with shuffled-label and equal-budget controls, which is the gap addressed here.

Table 1. Representative studies using synthetic data or transfer for industrial defect analysis.

Study	Task	Synthetic or Transfer Element	Relation to This Study
CutPaste, DRÆM, RealNet, AnomalyDiffusion [15–18]	Anomaly detection or anomaly generation	Cut-and-paste, simulated or diffusion-synthesized anomalies	Normal-versus-anomalous task or generator-centered; no few-shot multi-class comparison of routes
Domain randomization [21]; FDSL [22]	Pretraining sources	Randomized rendering; formula-generated fractal images	Synthetic source before real data; FDSL is the generic-source control here
Alqudah et al. [23]; Peñaloza et al. [24]	Wafer maps; optical scratch detection	Synthetic wafer maps or scratch images	Different input or task; one defect type
Liu et al. [25]; MVREC [13]	Few-shot anomaly and defect classification	Transfer of few-shot methods; multi-view region context	Effects change with labeled support; source of the MVTec-FS task
This study	Few-shot defect classification (NEU, MVTec-FS, GC10; wafer case study)	ImageNet → synthetic → real versus ImageNet → real	Paired frozen and fine-tuned readouts with shuffled-label, equal-budget, head-only and generic-source controls

3. Materials and Methods

3.1. Research Object, Notation and Hypotheses

The task is to assign a supplied defect candidate, given as a region of interest (ROI), to one of C classes. Test units T are public source-image groups or wafer parents; each unit p has a label y(p) and derived views V(p) that inherit its identity. Support S(k, m) holds k units per class from manifest m (a saved support selection), with $S(1, m) \subset S(2, m)$. A route r is ImageNet → real (I), ImageNet → synthetic → real (S), ImageNet → shuffled-label synthetic → real (Sh) or ImageNet → additional real-image training → real (X); with a target seed s and a readout ϱ (full fine-tuning, frozen ridge or frozen prototype) it yields a classifier f(r, ϱ; s, m, k). Table 2 explains each term and how it was implemented; Supplementary Note S1 gives the detailed methods and the training objectives.

Table 2. Key terms, what they mean and how they were implemented.

Term	Meaning	How It Was Implemented
Observation unit	One original public image with its duplicates, or one real wafer defect; only units count as observations.	Assigned to support or test before augmentation; all views keep the unit ID and are averaged to one prediction; wafer candidates with identical or nearly identical model inputs form one unit.
Derived view	An augmented copy of a unit; it adds no new defect.	Made on the fly from support units only, with a fixed random key, so every route sees the same views.
Support k; manifest	k labeled units per class; a manifest is one saved random choice of them.	Drawn per class without consulting the test set; a larger k adds to the same choice (nested). Three manifests; ten for the GC10 extension.
Route (arm); prior	The training stages before the target task; the prior is the backbone handed over.	I: ImageNet → real; S: ImageNet → synthetic → real; Sh: S with labels shuffled per geometry; X: extra augmented real training at the budget of S.
Paired cell; matched head	A like-for-like comparison of routes.	Same target seed, manifest, support, schedule and views, and one random target head copied to all routes; differences are averaged over the 3 × 3 = 9 cells (30 for the ten-manifest GC10 analyses).
Full fine-tuning	Target training of backbone and head.	AdamW, 15 epochs × 3200 balanced augmented rows, batch 16, learning rates 5×10^{-5} (backbone) and 3×10^{-4} (head) with cosine decay, weight decay 10^{-4}, label smoothing 0.05, gradient-norm clipping (5 public, 1 wafer), mixed precision; final epoch used.
Frozen readout	A classifier fitted on fixed features.	L2-normalized pooled 768-dimensional features; ridge regression ($\alpha = 1$) or nearest cosine class mean (prototype), closed form.
Head-only control	The fine-tuning head trained on a frozen backbone.	Same head, initialization, optimizer, schedule, loss and views as full fine-tuning.
Support interaction	Extra gain of the synthetic route over ImageNet as k increases.	$(S(k_2) - S(k_1)) - (I(k_2) - I(k_1))$, with each route's change reported (Equation (3)).
Crossed range	Spread from training seed and support choice, with the test set fixed.	2.5th–97.5th percentiles over seed-and-manifest resamples (all 729 for a 3 × 3 grid).
Test-set interval	Spread from which test units were sampled, with cells fixed.	Class-stratified bootstrap of test units (B = 2000) rescoring every classifier.
Combined interval	Both sources of spread together.	Each of 2000 replicates resamples test units, seeds and manifests at once.
Decision rule	When a difference counts as supported.	Predicted sign, and the combined interval excludes zero.

Performance is the macro-F1 over test units,

$$F(f) = \frac{1}{C}\sum_{c=1}^{C} F1_c(f), \quad F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c} \tag{1}$$

where TP_c, FP_c and FN_c are the true positives, false positives and false negatives of class c. Public regions are weighted so that each source-image group counts once, MVTec-FS scores are averaged over products, and the evaluation views of a wafer unit are averaged to one prediction. Macro-F1 weights every class equally, so rare defect types count as much as frequent ones. The paired effect in cell (s, m) is

$$\Delta(k;\, s,m) = F\big(f_S(s,m,k)\big) - F\big(f_I(s,m,k)\big) \tag{2}$$

with the same target head, schedule, augmentation stream and support for both routes; the support interaction between k_1 and k_2 separates the absolute changes of the two routes:

$$\Delta(k_2;\, s,m) - \Delta(k_1;\, s,m) = \left[F\left(f_S(k_2)\right) - F\left(f_S(k_1)\right)\right] - \left[F\left(f_I(k_2)\right) - F\left(f_I(k_1)\right)\right] \quad (3)$$

Reporting the two bracketed terms separately distinguishes an improving synthetic route from a declining comparator. Figure 1 summarizes the evaluation and Figure 2 summarizes the network and readouts.

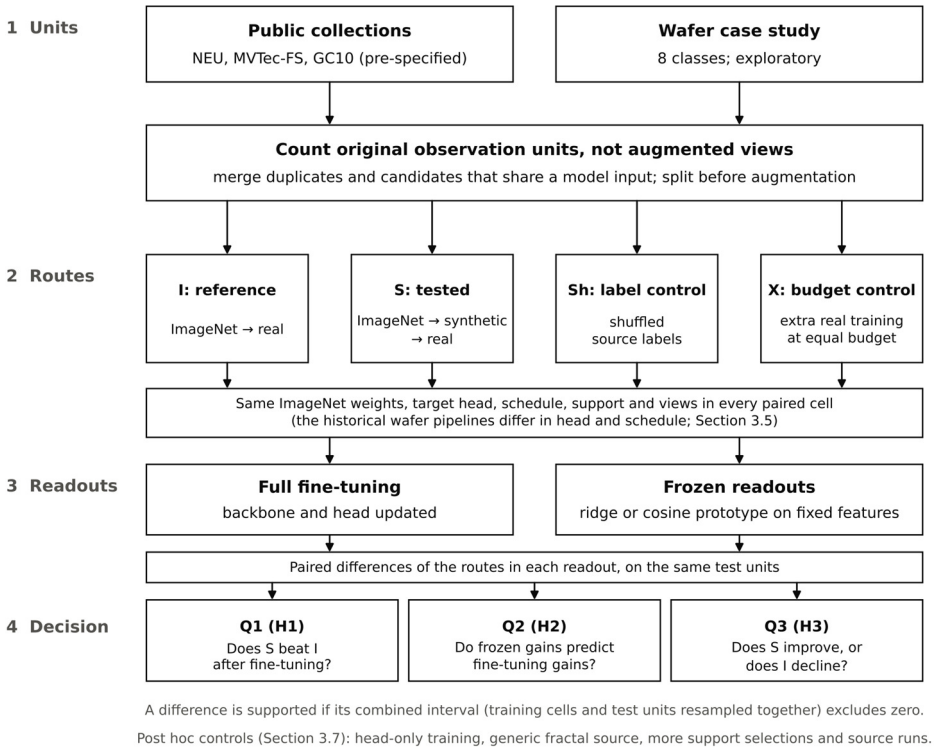

Figure 1. Overview of the evaluation. Original observation units, not augmented views, are counted and split first (1); all routes start from the same ImageNet weights, and routes S, Sh and X add a source stage before the same target head and schedule in every paired cell, except in the historical wafer pipelines (2); fine-tuned and frozen readouts are scored on the same test units (3); and each difference is judged by its combined interval (4). Questions Q1–Q3 correspond to hypotheses H1–H3; H4 asks whether correct source labels matter, and the post hoc controls are listed below the chart.

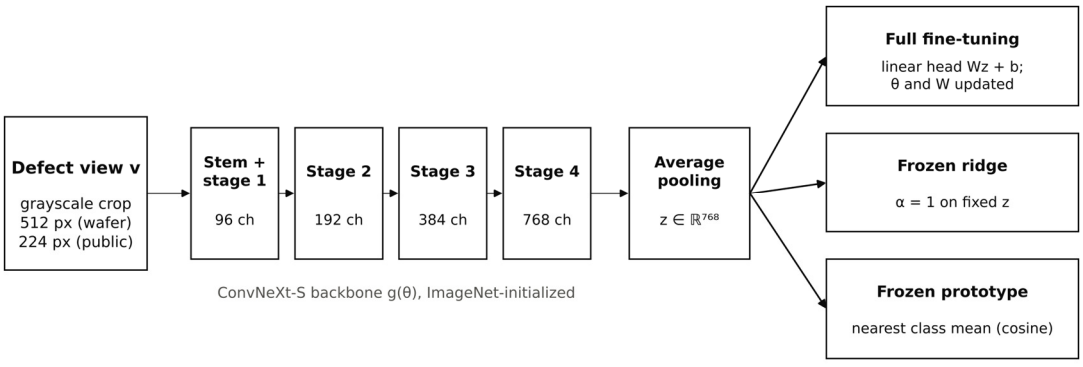

Figure 2. Network and readouts. The grayscale ConvNeXt-S backbone maps a view to a pooled 768-dimensional feature; full fine-tuning updates the backbone and a linear head, whereas frozen readouts keep the backbone fixed and fit a ridge or cosine-prototype classifier to the support features.

The backbone $g(\theta)$ is a grayscale ConvNeXt-S [30] that maps a view v to a pooled 768-dimensional feature $z = g(v; \theta)$; it was a design choice rather than a tested one (Supplementary Note S1). From the ImageNet weights θ_0, routes S, Sh and X add a source stage that minimizes label-smoothed cross-entropy on the synthetic source, on the same images with labels shuffled per geometry, or on augmented views of the support, respectively, whereas route I keeps θ_0. The source head is then discarded, and every route is fully fine-tuned from the same random head with AdamW [31] under one schedule and augmentation stream (except the historical wafer pipelines, Section 3.5). Frozen readouts keep the backbone fixed and fit ridge regression ($\alpha = 1$) or a nearest-class-mean classifier with cosine similarity to L2-normalized support features [7–9]. The objectives and readouts are written out as equations in Supplementary Note S1.

Four hypotheses were evaluated, where applicable, across the tasks and their relevant support levels; for the public collections, the support-averaged difference was specified as the endpoint before evaluation. A difference was supported only if its mean had the predicted sign and its combined interval excluded zero. The combined interval resamples training seeds and support selections together with the test units (Section 3.6), so a supported effect holds over both training randomness and test sampling. This rule replaces the one used in the first revision, which required the two sources of variation to exclude zero separately; Supplementary Table S66 lists every endpoint under both rules. H1, H2 and H4 formalize comparisons fixed in the public-data protocols before evaluation; H3 arose from the wafer results, was formalized during the first revision and was extended in the second revision to name the composition of the test units as an alternative explanation.

H1 (fine-tuning benefit; Q1) compares route S with route I under full fine-tuning with matched heads and schedules and is supported if $\Delta(k; s, m)$ of Equation (2) is positive under the decision rule.

H2 (frozen-feature validity; Q2) compares frozen-readout and fine-tuned differences in the same cells and is supported if a positive frozen-readout difference is accompanied by a positive fine-tuning difference.

H3 (attribution; Q3) examines the absolute changes in S and I between support levels (Equation (3)) and is supported if a positive support interaction arises from improvement of S rather than from decline of I, head policy or the composition of the test units.

H4 (source-label content) compares route S with route Sh at equal image exposure and is supported if correctly labeled pretraining scores higher under the decision rule.

3.2. Evaluation Protocol

The comparisons follow Algorithm 1. It fixes the observation unit before any split, at the level of the model input; pairs the routes on every source of training randomness, with identical target heads and schedules, so that a paired difference reflects the initialization alone; evaluates each readout that might be used; and reports uncertainty over training cells and test units together. The paired design was developed during the wafer study and applied unchanged to the public collections, so the wafer results motivated rather than confirmed it; the test-unit bootstrap and the decision rule (steps 6–7) were added during the first revision, and the input-level unit audit (step 1) and the combined-interval form of the rule were added during the second revision.

3.3. Evaluation Units and Study Scope

Observation units were, for the public collections, original images grouped with their exact or near duplicates before regions and views were assigned to a partition and, for the wafer case study, annotated candidates (parents) audited as described in Section 3.5; grouping prevents content overlap but cannot show that separate images come from separate specimens. Table 2 explains the terms and their implementation, and Table 3 the four tasks. The public collections cover whole-image steel textures (NEU), product-conditional defects (MVTec-FS) and box-localized, imbalanced metal-surface defects (GC10) and make the comparisons reproducible without the proprietary wafer images.

Table 3. Tasks and observation units; scores are interpreted within each row, not ranked across datasets.

Collection	Cohort and Unit	Support Levels	What Is Predicted
NEU	6 classes; 1081 support, 359 pilot and 359 reserved query image groups	1, 2, 5, 10	Steel-surface image class
MVTec-FS	14 products; 68 product–defect combinations; 1228 images and 1514 regions	1, 2, 5	Defect type within a known product
GC10 ROI	9 eligible classes; 3478 retained regions from the pinned annotated release	1, 2, 5, 10	Class of a supplied defect box
Wafer (case study)	8 optical classes; 116 development and 119 test candidates (111 test units after the input audit)	1, 2; secondary caps	Class of a supplied optical candidate

3.4. Public Collections and Their Controls

NEU contains six grayscale steel-defect classes [12]. Content grouping gave 1799 original-image groups (1081 support, 359 pilot and 359 reserved query groups); query images were decoded for integrity checks before the protocol was locked but not used for inference, tuning or selection. Grouping merged only one pair of images, and because NEU images are crops from a limited number of steel strips, crops of one strip are likely to occur in both support and query, which may inflate absolute scores.

A six-class procedural source (200 geometries per class, source seed 42) was drawn by hand-written rules from the official NEU class descriptions, without NEU images: crazing as random-walk polylines, inclusions as small elongated ellipses, patches as large irregular polygons, pitted surface as many small discs, rolled-in scale as compact polygonal flakes and scratches as long, slightly curved lines (Supplementary Table S57). Each blurred mask was added to a synthetic textured background with the same appearance distributions for all classes, and the same source was reused unchanged for MVTec-FS and GC10, with source heads discarded. The shuffled-label control used the same image pool with labels shuffled per geometry, keeping image exposure and low-level statistics while removing class meaning.

MVTec-FS was taken from a pinned version of the MVREC authors' repository [13] (14 products, 68 product-qualified defect types); product identity was supplied and probabilities were restricted to its defect types, so the task differs from anomaly detection and from evaluating MVREC. GC10 was taken from a pinned version of the dronefreak/GC10-DET mirror [14], whose train and valid partitions formed the support pool and whose test partition formed the query; classes needed at least five support and three query image groups, leaving nine classes (rolled pit excluded; Supplementary Table S41). After exact and near-duplicate checks on the original images, each region kept its image-group identity and was cropped with 1.25-times box context, converted to grayscale and resized to 224 pixels.

On all three collections, correct-label and shuffled-source checkpoints were compared with ImageNet at k = 1, 2, 5 and 10 (MVTec-FS: 1, 2 and 5) over target seeds 42–44 crossed with manifests 0–2, with matched heads and the fine-tuning schedule of Table 2 (bfloat16 mixed precision). Intensities were scaled with fixed constants, and neither query-based checkpoint selection nor test-time augmentation was used. The frozen readouts were fixed before query evaluation and need no tuning on query data. The extension completed 81 MVTec-FS and 108 GC10 full models and 54 and 72 frozen readouts; source-head retention, shuffled sources at higher support and a near-domain check on NEU are reported with them (Supplementary Tables S36–S49).

3.5. Wafer Case Study

Each of the ten 300 mm wafers was captured in a single top-down exposure covering the whole wafer, without raster scanning or stitching, on a custom bright-field optical inspection system (IOG Co., Ltd., Jeonju, Republic of Korea); the system specifications are proprietary, and the stored test-wafer images measure 3018–3176 pixels per side, so the results characterize this acquisition setting. Wafers 4–10 formed the development set and wafers 1–3 the retrospective test set. Eight optical classes were retained: contamination (CTM), dimple (DMP), edge damage (EDGD; crack and chip combined), edge roll-off (ERO), giant pit (GPIT), pit (PIT), slip line (SLP) and stain (STAIN), giving 116 development and 119 test candidates with no identifier or hash overlap between partitions (Supplementary Tables S1–S3). Because the same test candidates were reused across more than 300 models and in analyses designed after earlier results, all wafer results are exploratory.

Each candidate was cropped as a square around its box with context factors of 1.25, 1.50 and 2.00, clipped to the wafer image, resized to 512 × 512 pixels, converted to grayscale, contrast-equalized (contrast-limited adaptive histogram equalization) and scaled with fixed constants; evaluation averaged six views (three factors, with and without flipping) per candidate. Training views were generated from support crops only, with ranges set during earlier development, after the test wafers had been evaluated. Because crops of large candidates cover most of the wafer, the second-round audit compared the model inputs of every pair of candidates on each wafer (Supplementary Table S64). In the test set, two edge roll-off and two slip-line candidates on wafer 3 received the same whole-wafer input at the 2.00 factor and similar inputs at 1.50 (mean absolute difference 1.9–8.3 gray levels), and six pairs of same-label candidates had nearly coincident boxes (intersection over union 0.865–0.975), consistent with one defect annotated twice; in the development set, no identical inputs carried different labels, but four same-label groups were duplicates. The audited test set therefore merges each duplicate pair into one unit (each candidate weighted one half) and excludes the four whole-wafer candidates, leaving 111 units in seven classes (no edge roll-off unit remains); the published 119-candidate set is reported for comparison.

The synthetic source was rendered from hand-written, class-specific geometry rules (boundary-connected cracks or chips for edge damage, depressions for pits, directional traces for slip lines and diffuse contrast for stains), not from a physical simulator, and comprised 18,000 images (Supplementary Tables S2 and S3 and Figure S1). A defect of class c is rendered as $I(x) = b - s \cdot a \cdot H(x)/\max H$, where H is the class height map, b the mean gray level of the background image, $a \sim U(\max(2, \mu_c - 0.6\sigma_c), \mu_c + 0.8\sigma_c)$ the contrast and s = +1 (darkening) except for contamination, which is brightened with probability 0.7; overlapping defects are combined by a pixel-wise minimum (maximum for bright contamination), and Gaussian noise with the standard deviation of the background is added before clipping. The background (3024 × 3024 pixels) is itself a generated image, a uniform gray level with a radial vignetting model, independent Gaussian noise and a soft

edge, whose three parameters (center gray level, center noise and edge-to-center brightness ratio) were measured on a whole-wafer image of an eleventh, defect-free wafer that is not among the ten studied wafers; it contains no pixels of any real wafer (Supplementary Note S2).

For seven classes, μ_c and σ_c are the median and spread of $|\text{image} - \text{background}|$ over pixels that differ by more than eight gray levels, outside components larger than 3000 pixels, in 200 single-defect composites per class in which real defect patches had been pasted onto the background; no defect mask was used. The edge-chip value was raised by hand from 25 to 38, and the stain value was set by hand. Because the patches came from all ten wafers, the source had indirect access to test-wafer appearance statistics. These priors therefore differ from those re-measured directly inside the defect masks of real patches (Section 4.6; Supplementary Note S2).

Source pretraining used ConvNeXt-S with AdamW for 15 epochs (backbone/head learning rates $10^{-4}/3 \times 10^{-4}$, batch 16, label smoothing 0.05); the final checkpoint scored 0.9988 on held-out generated images, which indicates recognition within the generator distribution, not fidelity to real images. Three nested class-wise support selections (k = 1 and 2) were crossed with target seeds 42–44; small-class selections overlap strongly (Supplementary Table S54b). Development selection gave the synthetic initialization full fine-tuning with its source-trained head and ImageNet a standard schedule with a new random head, so the historically selected pipelines differ in adaptation procedure and head. Matched-head, common-schedule and exposure [32] studies and further sensitivity analyses (ConvNeXt-Tiny, ResNet-50 [33], a repaired contamination operation, background-only Bootstrap Your Own Latent pretraining [34], frozen OpenCLIP [35] and DINOv2 [36] features, development-fold selection [37,38] and confidence rules [39]) are reported in Supplementary Tables S9–S35.

3.6. Score Differences, Uncertainty and Multiplicity

Differences were calculated from unrounded scores, so differences obtained by subtracting displayed means can differ by 0.0001, and the decomposition of Equation (3) is arithmetic, not a learning mechanism. Crossed ranges use all 729 ordered resamples of each 3 × 3 seed-by-manifest grid [40] (50,000 resamples for 3 × 10 grids); the combined interval draws one such resample and one class-stratified bootstrap of test units in each of 2000 replicates. These intervals describe variation within the recorded grid and test set, not across new lots, wafers or products, and counts of positive cells are not independent replications; these caveats apply throughout and are not repeated. For wafer classes with fewer than five test units, each unit and each test wafer was also left out in turn. Tasks were not pooled because their units and scores differ (Table 3), and class and product comparisons are sensitivity analyses. On the public collections, eleven synthetic-versus-ImageNet fine-tuning comparisons were made (NEU and GC10 at four support levels, MVTec-FS at three) without adjustment for multiplicity.

3.7. Post Hoc Extensions and Controls

Further analyses were added after the primary public-data results, each with a plan fixed before its models were trained (Supplementary Protocols S1 and S2 and Note S1). All reuse the observed query or test sets and are therefore sensitivity analyses. Seven further GC10 manifests (3–9) gave ten manifests and 30 paired cells per support level (252 new fine-tuned models and 240 frozen readouts over the ten manifests). A matched real-training control replaced the synthetic stage with augmented training on the same support at the same budget (3000 optimizer steps and 48,000 augmented draws, not floating-point operations) on manifests 0–2 (72 models). A head-only control trained the fine-tuning head on the frozen backbone (549 models) and was classified, by a rule fixed beforehand,

as following the frozen-readout or the fine-tuning pattern according to the direction of its difference.

The correct- and shuffled-label sources were retrained with source seeds 43 and 44 on the same images, and five further GC10 sources were trained on newly generated images (150 models at k = 1 over manifests 0–9); the GC10 gain would count as present on average only if the interval of the mean over the eight source runs, resampling runs, seeds, manifests and test groups jointly, excluded zero. A generic source of six fractal categories was rendered with the appearance model, image count and training recipe of the public-data source, so that only geometry differs (99 fine-tuned models and both frozen readouts); formula-driven supervised learning (FDSL) [22] was chosen because it is the established synthetic pretraining baseline that needs no real images, whereas a diffusion generator trained on the target defects would also change appearance and require those real images; such target-conditioned generators were not compared. Finally, wafer gray-level priors were re-measured inside the defect masks of real patches from all wafers and from development wafers 4–10 only, and three paired 18,000-image sources (published, all-wafer and development-only priors) were transferred with matched heads (63 models).

3.8. Reproducibility and Computation

Planned methods, code and input hashes, stopping rules and completion status were recorded for each study; a fixed protocol does not make a study planned after earlier results prospective. All outcomes, paired matrices and verification code are in Supplementary Data 1. Computation used Python 3.12.3, PyTorch 2.7.0, CUDA 12.8, torchvision 0.22.0, timm 1.0.26 and NumPy 2.4.3 under Windows Subsystem for Linux 2 (Microsoft Corporation, Redmond, WA, USA) on an Intel Core i9-14900K (Intel Corporation, Santa Clara, CA, USA) with an NVIDIA GeForce RTX 5090 (32 GB; NVIDIA Corporation, Santa Clara, CA, USA). Runs interrupted by crashes resumed from saved epoch checkpoints with unchanged settings, without established bitwise equivalence; for every post hoc analysis, all metrics were recomputed from the saved predictions and equaled the run ledgers, and archived cells retrained before reuse reproduced their predictions exactly (Supplementary Tables S48–S49 and Protocols S1 and S2, which also list deviations). The second-round input audit recomputed the model inputs from the stored candidates, and all re-scored endpoints were computed from the saved predictions without retraining. Generative AI tools (OpenAI ChatGPT and Codex (GPT-5 and GPT-6 series models; OpenAI, San Francisco, CA, USA); Anthropic Claude (Claude Fable 5.1 and Claude Opus 5.5; Anthropic, San Francisco, CA, USA); products listed in the Acknowledgments) were used to discuss study and implementation choices, to draft and debug the training, analysis, verification and figure-rendering scripts, and to edit the English text. No image, label or reported number was produced by an AI tool directly: images came from the documented renderers, labels from the annotations, and every number from executed scripts whose outputs the authors checked against the recorded evidence.

4. Results

The results support three findings, detailed below and summarized in Section 4.7. (i) Synthetic pretraining did not improve the fine-tuned classifier under the decision rule on any collection; the largest gain, on GC10 with one labeled group per class, was positive on average over ten support selections and in all eight source runs, but its combined interval included zero. (ii) Frozen-readout rankings did not predict fine-tuning outcomes: frozen readouts favored synthetic features on NEU and MVTec-FS without a fine-tuning gain and disfavored them on GC10. (iii) On wafers (exploratory), an apparent support-

dependent gain arose largely from the evaluation units, and no wafer difference met the rule after the input audit.

4.1. NEU: Frozen-Readout Gains Did Not Give a Better Fine-Tuned Classifier

On NEU, synthetic pretraining helped frozen readouts but not the fine-tuned classifier. Under matched-head full fine-tuning, both routes improved from 1 to 10 images per class and neared 1 from five images (Figure 3a), so only k = 1 and 2 separate the routes; the support-averaged difference was −0.0040 (combined interval [−0.0181, +0.0126]; Supplementary Table S36), showing neither a consistent advantage nor equivalence.

With frozen features (Figure 3b), synthetic ridge reached 0.8522 against 0.7396 for ImageNet with one image per class (difference +0.1125 [+0.0562, +0.1720]), but the advantage shrank with support, reversed at ten images and stayed below fully fine-tuned ImageNet (0.9043; Supplementary Table S37). Higher frozen-readout scores [9] therefore did not identify the best training route, although this cannot show whether fine-tuning erased synthetic information [10].

Retaining the source-trained head did not improve the mean at any support level, and correct source labels exceeded shuffled labels after fine-tuning at two images (+0.0331 [+0.0075, +0.0616]) and on average over support levels (+0.0244 [+0.0009, +0.0538]); because networks can fit random labels [41] and such training affects later learning [42], this control does not separate all sources of useful information. A near-domain check with new geometries from the same generator showed that the pipeline functioned, not that the evaluation can detect a benefit on real defects (Supplementary Tables S38–S40 and Figures S9–S11).

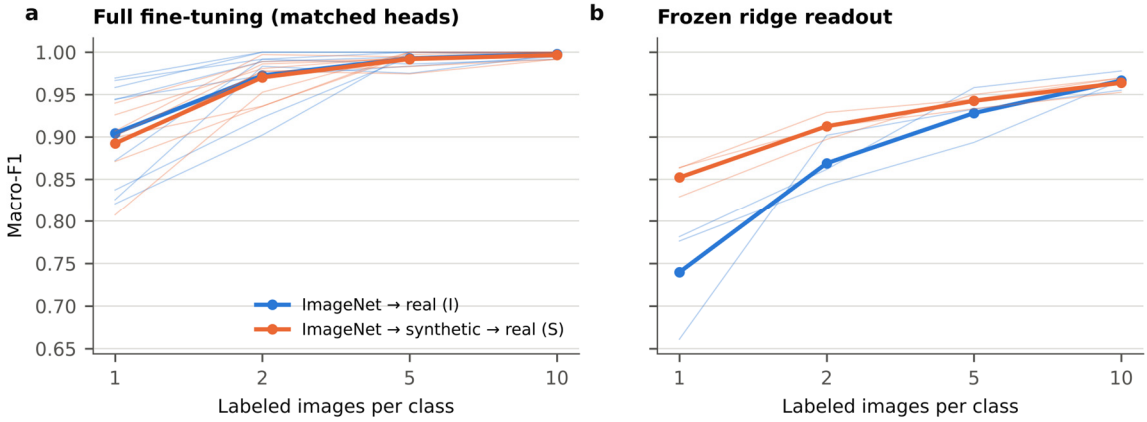

Figure 3. NEU results. (**a**) Full fine-tuning with matched heads (thin lines: nine cells; thick lines: means). (**b**) Frozen ridge readout (thin lines: three support selections). Both panels use the same reserved query images and vertical scale; lines are not confidence bands.

4.2. Fine-Tuning on MVTec-FS and GC10

After full fine-tuning, synthetic pretraining gave no consistent advantage on MVTec-FS, and on GC10 its largest gain was at one labeled group per class (Table 4, Figure 4). On MVTec-FS, both routes improved with support, and the support-averaged difference was −0.0029 [−0.0164, +0.0118] (Supplementary Table S42), so there was no consistent fine-tuning advantage, although similar means do not establish equivalence. At one group, the shuffled-source route scored highest (shuffled minus ImageNet +0.0136 [−0.0236, +0.0530]) but fell below both routes at two and five groups.

On GC10, macro-F1 with one labeled group per class rose from 0.5569 with ImageNet to 0.5949 with synthetic pretraining in the pre-specified three support selections, a paired

difference of +0.0380 with a combined interval of [−0.0128, +0.0951] (seven of nine cells positive), so it did not meet the decision rule; it is also one of the eleven unadjusted comparisons of Section 3.6. The pre-specified support-averaged difference was +0.0123 [−0.0120, +0.0361], and the difference became negative at ten groups (Supplementary Tables S43 and S47). Any gain applies to the recorded nine-class task, not to every class or a new production population.

Table 4. Fine-tuning results on MVTec-FS (known-product macro-F1) and GC10 (nine-class region macro-F1) are averaged over nine paired cells. Img., Syn. and Shuf.: ImageNet, synthetic and shuffled-source routes; Syn–Img: synthetic minus ImageNet with its combined interval; crossed ranges and test-set intervals are in Supplementary Table S53.

Task	k	Img.	Syn.	Shuf.	Syn–Img	Combined Interval	Positive
MVTec-FS	1	0.5896	0.5871	0.6031	−0.0025	[−0.0259, +0.0239]	3/9
MVTec-FS	2	0.7225	0.7161	0.7135	−0.0065	[−0.0253, +0.0117]	4/9
MVTec-FS	5	0.8477	0.8479	0.8349	+0.0003	[−0.0144, +0.0157]	4/9
GC10 ROI	1	0.5569	0.5949	0.5414	+0.0380	[−0.0128, +0.0951]	7/9
GC10 ROI	2	0.6997	0.7046	0.6324	+0.0049	[−0.0456, +0.0524]	5/9
GC10 ROI	5	0.7658	0.7804	0.7554	+0.0146	[−0.0211, +0.0591]	7/9
GC10 ROI	10	0.8396	0.8313	0.8185	−0.0083	[−0.0393, +0.0242]	3/9

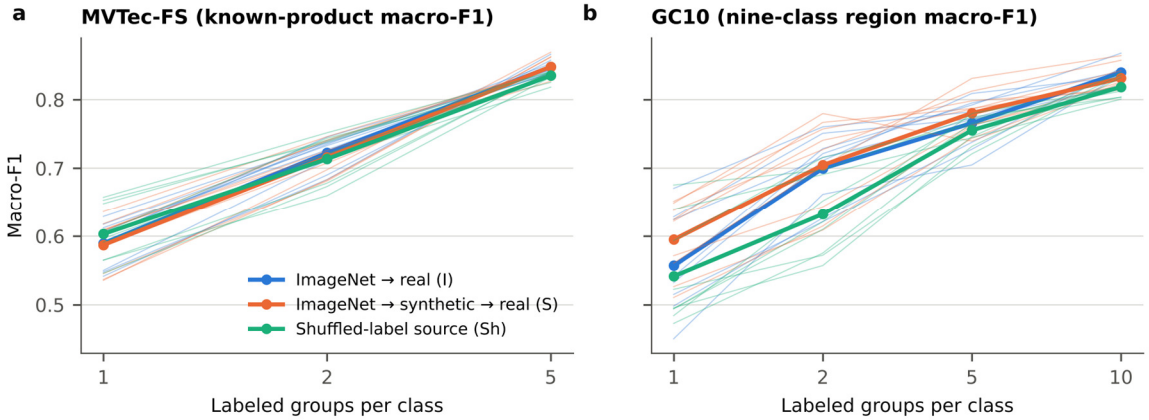

Figure 4. Fine-tuning results on MVTec-FS and GC10 (thin lines: nine paired cells; thick lines: means). All routes share target heads, support and schedules; the common vertical scale does not imply equal task difficulty.

4.3. Frozen Readouts Could Reverse the Interpretation

Frozen readouts ranked the routes differently from fine-tuning, in both directions (Figure 5). On MVTec-FS, synthetic features exceeded ImageNet features at one group for both frozen classifiers (ridge 0.4902 to 0.5470, +0.0568 [+0.0261, +0.0910]; prototypes +0.0688 [+0.0393, +0.0988]; Supplementary Table S44) despite no fine-tuning gain, and shuffled-source features scored still higher, so correct labels cannot account for this advantage. On GC10 the pattern reversed: over ten support selections, prototypes favored ImageNet at every level (with combined intervals excluding zero at two and five groups), and the one-group ridge difference was +0.0005, although the fine-tuned synthetic model scored higher. Selection by frozen scores would therefore have suggested a benefit on NEU and MVTec-FS that fine-tuning did not show and the opposite ranking on GC10.

The head-only control, which trains the fine-tuning head on the frozen backbone, gave one-group differences of +0.0187 (NEU), +0.0209 (MVTec-FS) and +0.0111 (GC10; Supplementary Table S58), none meeting the decision rule. By the pre-specified rule, all

three matched the frozen-readout pattern, so training the fine-tuning head on frozen features did not reproduce the fine-tuned results; the disagreement between readouts is consistent with an effect of updating the backbone rather than of how the classifier is fitted, although this neither identifies the mechanism nor establishes an advantage for the head.

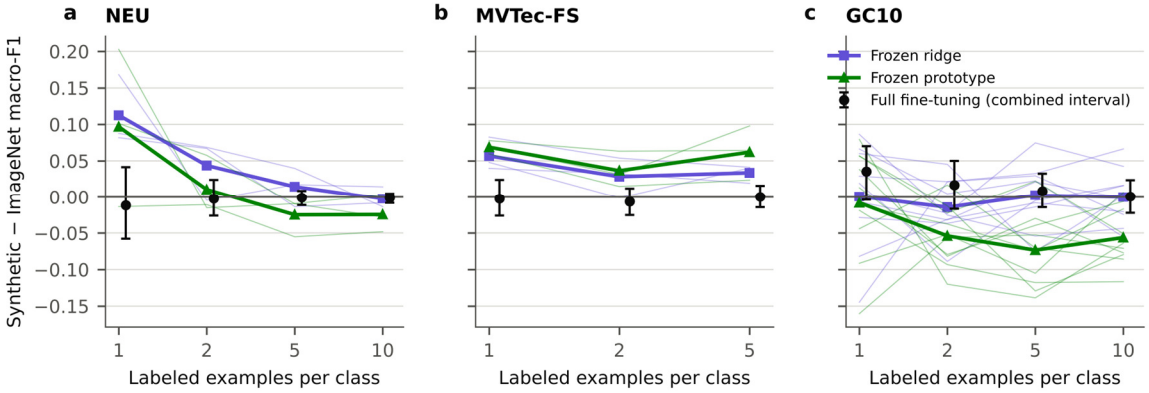

Figure 5. Readout dependence of the ranking: synthetic-minus-ImageNet macro-F1 for frozen ridge and frozen prototype readouts (thin lines: support selections; thick lines: means) and for full fine-tuning (points with combined intervals) on the same query sets and support selections (GC10: ten manifests for all readouts). The comparison identifies a readout-dependent ranking, not a loss of synthetic information during fine-tuning.

4.4. Labels and Class Failures Qualified the Mean Differences

Correct source labels mattered only at some support levels, and mean differences hid class-level failures. After fine-tuning, correct labels exceeded shuffled labels on GC10 at two groups (+0.0723 [+0.0280, +0.1227]; all nine cells positive) and on average (+0.0409 [+0.0094, +0.0788]), but not at any level on MVTec-FS (Supplementary Table S45); this control preserves image exposure but not optimization or source-training cost. From one to two groups both routes improved on both collections, and product and class summaries retain negative synthetic differences and zero-F1 cells (Supplementary Tables S46–S47 and Figure S8), so a mean difference is not a minimum F1 for every defect type.

4.5. Post Hoc Extensions: Support Selections, Real-Training Control, Source Runs and a Generic Source

These analyses were designed after the primary results (Section 3.7) and are reported separately from the primary hypothesis tests. The GC10 one-group difference stayed positive on average over ten support selections, in all eight source runs and against extra real-image training, but none of these comparisons met the decision rule.

Support selections. Over ten manifests, both routes improved with support, and the one-group difference was +0.0354 (combined interval [−0.0037, +0.0702]; 23 of 30 cells positive), falling to +0.0008 at ten groups; the support-averaged difference was +0.0152 [−0.0031, +0.0326] (Table 5). Because both absolute scores increased with support, the difference cannot be attributed to comparator decline; scores were lower than on manifests 0–2 for every route and level, so the original three selections were comparatively favorable, and support selection contributed more variation than the training seed (one-group manifest-level differences −0.0595 to +0.0719). Frozen readouts over the ten manifests showed no corresponding advantage (Supplementary Table S50).

Extra real-image training. In the matched real-training control (Supplementary Table S63), additional augmented training on the same support scored 0.5584 at one group, close to ImageNet (0.5569) and below the synthetic route (0.5949). The synthetic-minus-extra-

real difference fell with support (+0.0365 [−0.0301, +0.1069] at one group to −0.0132 at ten), so an equal optimization budget did not reproduce the synthetic score, but a synthetic-specific advantage was not established either.

Table 5. GC10 fine-tuning over ten support manifests (30 paired cells). Img., Syn. and Shuf.: ImageNet, synthetic and shuffled-source routes; Syn–Img: synthetic minus ImageNet with its combined interval; Syn–Shuf: synthetic minus shuffled source; the mean row gives the support-averaged difference.

k	Img.	Syn.	Shuf.	Syn–Img	Combined Interval	Positive	Syn–Shuf
1	0.5161	0.5515	0.5056	+0.0354	[−0.0037, +0.0702]	23/30	+0.0459
2	0.6519	0.6684	0.6071	+0.0165	[−0.0167, +0.0500]	19/30	+0.0614
5	0.7499	0.7582	0.7175	+0.0083	[−0.0144, +0.0324]	19/30	+0.0406
10	0.8183	0.8191	0.7979	+0.0008	[−0.0222, +0.0235]	13/30	+0.0212
Mean	–	–	–	+0.0152	[−0.0031, +0.0326]	25/30	–

Classes. With one labeled group per class, six of nine classes favored synthetic pretraining, and 17 of the 396 final models had a class with F1 = 0, with none excluded (Supplementary Tables S51–S52).

Source seeds and source runs. With source seeds 43 and 44 on manifests 0–2, the one-group difference fell to +0.0129 and +0.0158 (seed 42: +0.0380; Supplementary Table S59), whereas correct labels beat shuffled labels on GC10 at two groups and on average for all three seeds. Over ten manifests, the one-group difference was positive in all eight source runs (+0.0207 to +0.0427; Figure 6c; Supplementary Table S62), and its combined interval excluded zero in two of them; the mean over runs, +0.0294, had an interval of [−0.0030, +0.0626] when runs, seeds, manifests and test groups were resampled together. The GC10 difference was therefore consistent in sign, but it met the rule in only two of the eight individual sources and not on average.

Generic fractal source. At one example or group per class, the defect-shaped source exceeded the generic fractal source under the decision rule in three of nine dataset–readout combinations, the frozen ridge and prototype readouts on NEU and the frozen ridge readout on MVTec-FS (Figure 6b; Supplementary Table S61); for GC10 fine-tuning the difference was +0.0454 [−0.0029, +0.1010], and no combination favored the fractal source. The fractal source itself exceeded ImageNet only for the MVTec-FS frozen prototypes (+0.0436 [+0.0142, +0.0761]), so a generic synthetic stage accounted for part of the MVTec-FS frozen-readout gain.

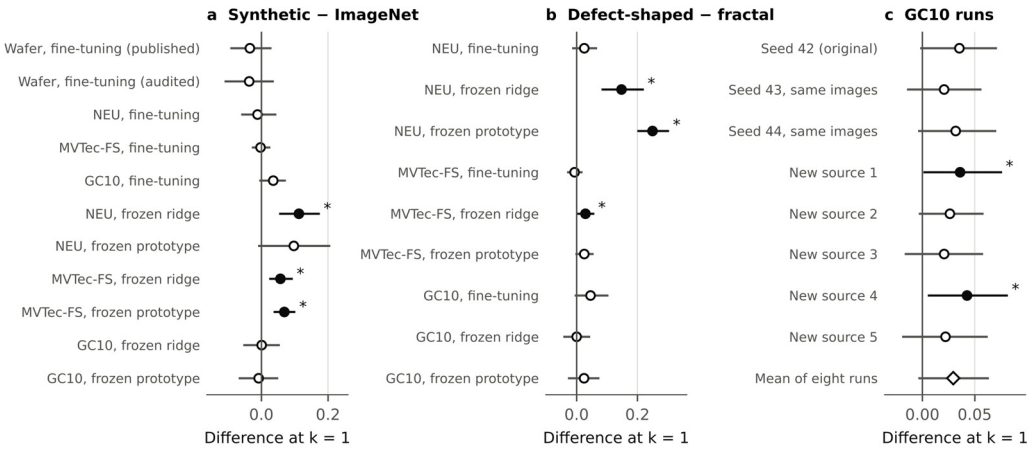

Figure 6. Main paired differences at one labeled example or group per class, with combined intervals: (**a**) synthetic minus ImageNet pretraining (wafer: matched heads on the published and the audited

test units; GC10: ten manifests); (**b**) defect-shaped minus generic fractal source; (**c**) GC10 fine-tuning in eight source runs and their mean (diamond). Filled markers and *: decision rule met.

4.6. Exploratory Wafer Case Study: An Apparent Gain Arose from the Evaluation Units

In the published 119-candidate comparison of the selected pipelines, synthetic macro-F1 was 0.0689 below ImageNet with one parent per class and 0.0389 above it with two, and the support interaction of +0.1078 arose mainly from a falling ImageNet score (−0.0695), 81.2% of it from edge roll-off with two test candidates (Table 6; Supplementary Table S55). The input audit explains this pattern (Section 3.5; Supplementary Table S64). The two edge roll-off and two slip-line candidates of wafer 3 received the same whole-wafer input at the largest context factor and similar inputs at 1.50, and an edge roll-off and a slip-line candidate received the same prediction in 71 of 72 models. With one parent per class, the ImageNet route called this image edge roll-off in all nine cells; with two, it called it giant pit in seven, whereas the synthetic route called it slip line in 16 of the 18 cells. A classifier that gives one image one label cannot be correct for both labels, so the ImageNet decline and much of the interaction reflected a changed prediction for a single image with conflicting labels.

On the audited test set (111 units in seven classes), no wafer difference met the decision rule (Table 6, Figure 7): synthetic minus ImageNet was −0.0208 with one parent and +0.0234 with two; the ImageNet change was −0.0022 and the support interaction was +0.0442 [−0.0751, +0.1842]. Excluding the whole-wafer candidates alone reduced the interaction to +0.0597, whereas merging the duplicates alone left it at +0.0924 (Supplementary Table S65). With matched heads, the audited differences were −0.0366 and −0.0200 and the interaction was +0.0165 [−0.0553, +0.0933]. Per-class F1 (Table 7) stayed low for the small classes, and F1 for a class with one or two units is not interpretable on its own.

The wafer controls did not reveal a synthetic advantage either (Supplementary Tables S9–S35 and S60, Figures S12–S13). Matched heads, a common schedule, a doubled training budget, a repaired contamination operation, background-only pretraining, higher support caps and frozen OpenCLIP and DINOv2 features left the synthetic route below or close to ImageNet. Re-measuring the priors inside the defect masks of real patches did not detect an effect of the test wafers: development-only minus all-wafer priors gave −0.0019 with one parent and +0.0347 with two on the published test set and +0.0314 and +0.0476 on the audited units, with combined intervals including zero. The direct measurements differed from the published priors (for example, pit 79 against 23) because the published values were taken over all thresholded pixels of composites without a defect mask (Section 3.5; Supplementary Note S2), and the development-only dimple and crack priors each rest on one patch, so the test is uninformative for those two classes.

Table 6. Selected wafer pipelines on the published 119 test candidates and on the 111 audited test units (seven classes; duplicate annotations merged, whole-wafer candidates with conflicting labels excluded). Published and Audited: the synthetic-minus-ImageNet difference (k rows) or the change in a route and the support interaction (other rows) on each test set; Audited Syn. and Audited Img.: mean macro-F1 of each route; the combined interval and the positive cells out of nine refer to the audited values.

Condition	Published	Audited Syn.	Audited Img.	Audited	Combined Interval	Positive
k = 1	−0.0689	0.2017	0.2225	−0.0208	[−0.1096, +0.0555]	4/9
k = 2	+0.0389	0.2437	0.2203	+0.0234	[−0.0635, +0.1155]	5/9
Synthetic $k_2 - k_1$	+0.0383			+0.0420	[−0.0401, +0.1186]	7/9
ImageNet $k_2 - k_1$	−0.0695			−0.0022	[−0.1015, +0.0841]	6/9

Support interaction	+0.1078	+0.0442	[−0.0751, +0.1842]	5/9

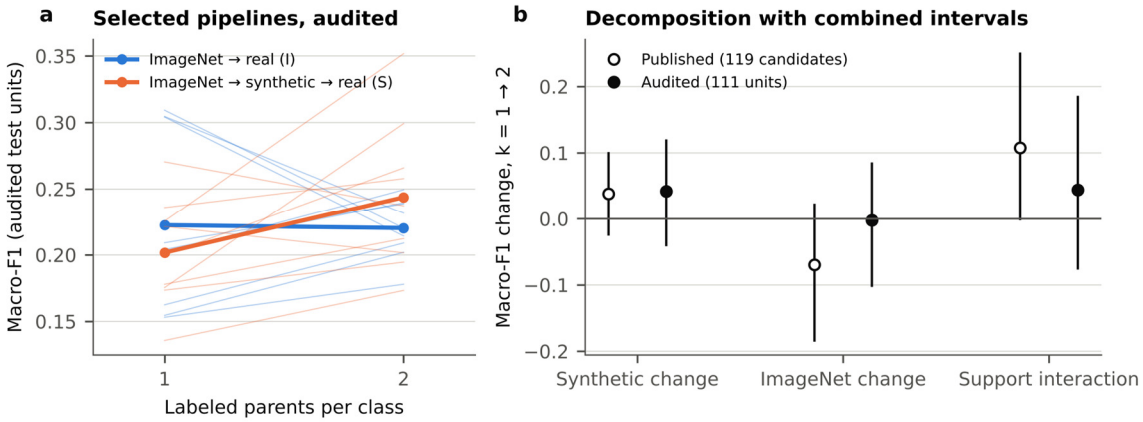

Figure 7. Wafer–support interaction before and after the input audit. (**a**) Audited macro-F1 of the selected pipelines at one and two parents per class (thin lines: nine cells; thick lines: means). (**b**) Change in each route and the support interaction on the published and the audited test set, with combined intervals.

Table 7. Per-class F1 of the selected wafer pipelines on the audited test units (mean of nine cells). F1 for SLP and STAIN (one or two units) is not interpretable on its own; no edge roll-off unit remains.

Class	Test units	ImageNet, k = 1	ImageNet, k = 2	Synthetic, k = 1	Synthetic, k = 2
CTM	40	0.324	0.442	0.269	0.494
DMP	5	0.342	0.307	0.202	0.216
EDGD	7	0.262	0.304	0.123	0.147
GPIT	3	0.266	0.215	0.489	0.475
PIT	53	0.280	0.274	0.283	0.255
SLP	2	0.000	0.000	0.000	0.063
STAIN	1	0.084	0.000	0.046	0.056

4.7. Summary of the Hypothesis Tests

Table 8 summarizes the evidence under the decision rule. H1 was not supported on any collection: the largest difference, GC10 at one group, had combined intervals including zero for the pre-specified three selections, the ten-selection extension and the mean over source runs. H2 was not supported, because frozen readouts favored synthetic features on NEU and MVTec-FS without a fine-tuning gain and disfavored them on GC10. H3 was not supported: the wafer interaction arose largely from the evaluation units, not from improvement in the synthetic route. H4 was supported after fine-tuning on NEU and GC10 and for frozen readouts on NEU, but not on MVTec-FS. Compared with the rule of the first revision, 12 of the 93 public-data endpoints no longer met the rule, among them the GC10 one-group fine-tuning differences, the MVTec-FS source-label difference at five groups and several frozen-readout differences at two groups; no endpoint changed in the other direction (Supplementary Table S66).

Table 8. Hypothesis verdicts under the decision rule (combined interval). Entries are mean paired differences, with the combined interval where it decides the verdict; * marks differences meeting the rule. All intervals are in Supplementary Tables S53, S58–S62, S65 and S66.

Hypothesis	Task	Key Differences	Verdict
H1 fine-tuning benefit	NEU; MVTec-FS	Support average −0.0040 [−0.0181, +0.0126]; −0.0029 [−0.0164, +0.0118]	Not supported
H1	GC10, manifests 0–2	k = 1: +0.0380 [−0.0128, +0.0951]; support average +0.0123 [−0.0120, +0.0361]	Not supported
H1	GC10, manifests 0–9 (post hoc)	k = 1: +0.0354 [−0.0037, +0.0702]; support average +0.0152 [−0.0031, +0.0326]	Not supported
H1	GC10, eight source runs (post hoc)	k = 1: +0.0207 to +0.0427 (2 runs *); mean +0.0294 [−0.0030, +0.0626]	Not supported
H1	Wafer, matched heads (audited)	k = 1: −0.0366; k = 2: −0.0200	Not supported
H2 frozen-feature validity	NEU; MVTec-FS; GC10 (manifests 0–9)	k = 1 frozen ridge +0.1125*, +0.0568*, +0.0005; fine-tuning −0.0119, −0.0025, +0.0354	Not supported
H2 head-only control	NEU; MVTec-FS; GC10	k = 1: +0.0187, +0.0209, +0.0111	Frozen-readout pattern; differences not supported
H3 attribution	Wafer	Interaction +0.1078 (published) and +0.0442 [−0.0751, +0.1842] (audited); ImageNet change −0.0695 and −0.0022	Not supported (evaluation-unit artifact)
Prior leakage (post hoc)	Wafer, matched heads (audited)	Development-only − all-wafer priors: +0.0314; +0.0476	No effect detected
H4 source-label content	NEU; GC10 (fine-tuning)	k = 2: +0.0331 *, +0.0723 *; support average +0.0244*, +0.0409*	Supported
H4	MVTec-FS (fine-tuning)	k = 5: +0.0130; support average −0.0002	Not supported
H4	Frozen readouts	NEU k = 1 ridge +0.1831 *, prototype +0.1518 *; none on MVTec-FS; negative on GC10 at some levels	Supported on NEU only
Geometry (post hoc)	NEU; MVTec-FS; GC10	Defect-shaped > fractal in 3 of 9 readouts at k = 1 (Figure 6b)	None favored fractal

5. Discussion

5.1. Readouts and Evaluation Units

Frozen and fine-tuned readouts disagreed because they answer different questions: a frozen readout measures how well the classes separate in the pretrained features, whereas full fine-tuning lets the backbone move away from them. Synthetic pretraining improved the first without improving the second on NEU and MVTec-FS, and the head-only control behaved like the frozen readouts, which is consistent with an effect of updating the backbone, although it does not identify the mechanism. A frozen-feature probe therefore answers a different question from the one asked by a user who will fine-tune [7–10].

The comparisons also show how a gain can be misread. On wafers, one whole-wafer image carrying conflicting labels was counted as four test units, which turned a single changed prediction into most of a support-dependent gain; the duplicate check had been made on the annotated regions, whereas the model received the context window, so units must be defined at the level of the model input [1–3]. On GC10, the difference depended on which few examples were labeled and on the source run, and on MVTec-FS, part of the frozen-readout gain came from generic image statistics. The public-data extension used an unchanged NEU-style source; class-conditioned diffusion, physically calibrated

rendering and other source distributions were not evaluated, and earlier synthetic anomaly and wafer studies address different tasks [13,15–18,20–25].

5.2. A Practical Comparison Before Adding Synthetic Training

The results suggest six checks before a synthetic stage is adopted; they extend Algorithm 1. First, define observation units at the level of the model input and merge views, duplicate annotations and candidates that share an input. Second, compare against direct ImageNet-to-real training with the classifier that will be used, because a frozen-feature probe suggested benefits that fine-tuning did not show. Third, inspect the comparator's absolute score before the difference. Fourth, confirm a low-support gain over several support selections and source runs; on GC10 it varied between runs and fell to +0.0008 by ten groups, so no universal threshold follows. Fifth and sixth, include an equal-budget real-training control and a generic synthetic source with the same appearance model.

These checks are a reporting guide, not a validated selector: each addresses a documented failure mode (pseudoreplication and leakage [1–5], seed-only uncertainty [11], readout-dependent rankings [7–10]) seen here in at least one task; their completeness beyond these tasks is not established, and because training, annotation and operating-point costs were not measured, a score gain alone does not establish a favorable deployment trade-off.

5.3. Limits of the Evidence

The wafer collection is small, single-source and retrospective: macro-F1 of about 0.2–0.3 reflects low-support behavior, not an inspection-ready classifier, and the test images were observed during development and reused across more than 300 models. The generator was not prospectively independent of the test wafers: its contrast priors were partly measured on test-wafer patches with a procedure that differs from a direct measurement, the edge-chip and stain priors were set by hand, the class geometry rules were written with test observations available, and the augmentation ranges were set after the test wafers had been evaluated. The re-measured priors addressed only the first of these channels and, for dimple and crack, whose development-only priors rest on single patches, not even that. Labels lacked metrological confirmation, same-label duplicates also occurred in the development set (the two slip-line support parents at $k = 2$ were one defect), candidates on one wafer are dependent, and additional seeds repeat model fitting, not wafer sampling [11].

The public collections provide no new wafer lots or physically identified specimens, and with product identity (MVTec-FS) and boxes (GC10) supplied, neither task evaluates discovery of unknown defects; the GC10 mirror differs from the original release, whose redistribution rights its metadata do not resolve, and MVTec-FS permits non-commercial research only. The main public comparisons used one defect-shaped source and seed, and no other defect-shaped generator was evaluated; NEU fine-tuning approached the ceiling; foundation-model exposure to public images is unknown, and the real-training control (manifests 0–2; steps and draws, not FLOPs, matched) does not separate source diversity from additional optimization.

The post hoc analyses reused the test or query sets after earlier results, so they are sensitivity analyses, not independent confirmations [37], although each protocol was fixed beforehand; a systematic human review of generated images has not been completed (Supplementary Table S35). The decision rule was changed to the combined interval during the second revision, after the results were known; this made the verdicts stricter, and both rules are reported (Supplementary Table S66). Prospective testing on new material is required before any synthetic stage is adopted.

6. Conclusions

Across three public defect collections and an exploratory wafer collection, synthetic pretraining did not improve the fine-tuned classifier under a decision rule that resamples training cells and test units together; the largest gain, on GC10 at one labeled group per class, was positive in all eight source runs but not established. Frozen-readout rankings did not predict fine-tuning outcomes, and on wafers an apparent support-dependent gain arose largely because one whole-wafer image carrying conflicting labels was counted as four test units. Synthetic pretraining should therefore be evaluated with the intended classifier, on observation units defined at the level of the model input and under matched training conditions (Algorithm 1).

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app162311816/s1, Table S1: Eligible support and overlap among manifests; Table S2: Procedural source parameters; Table S3: Locked historical synthetic checkpoint; Figure S1: Procedural synthetic examples and real development candidates; Table S4: Complete initial absolute and paired score matrices; Table S5: Arm decomposition and crossed sensitivity; Table S6: Classwise ImageNet change from one to two parents; Table S7: Rare-class and support-threshold sensitivity in the nested study; Table S8: Per-wafer and supported-class descriptions; Table S9: Initialisation by adaptation schedule factorial; Table S10: Matched head comparison under common full fine-tuning; Table S11: Architecture and head-policy sensitivity; Figure S2: Parent-level confusion matrices for the initially selected wafer pipelines; Table S12: All-parent evaluation and sensitivity to low-support classes; Figure S3: Sensitivity of the all-parent comparison to test-class support; Table S13: Historical calibration summaries; Table S14: Descriptive variance components of the observed grid; Figure S4: Paired transfer contrasts across training seeds and support selections; Table S15: Complete frozen encoder and classifier policies; Figure S5: Frozen ridge comparisons depend on the number of distinct labelled candidates; Table S16: Four initialisation histories with a common target head and schedule; Table S17: Higher parent caps for class-specific and shared-primitive sources; Table S18: Evidence scope and disposition; Table S19: Historical target learning protocol; Table S20: All eight support and exposure conditions; Table S21: Primary and all secondary exposure contrasts; Table S22: Class F1 in the support and exposure study; Table S23: Repair study retrospective means; Table S24: All repair-study retrospective paired comparisons; Table S25: Capped three-parent repair comparison; Table S26: Head-only and head-first repair comparisons; Table S27: Source-optimisation seed sensitivity; Table S28: Frozen readouts in the compositing diagnostic; Table S29: Nested retention-policy selection on 116 development parents; Table S30: Fixed epoch 15 confidence diagnostics; Figure S6: Confidence filtering does not remove the development-set error; Table S31: Class failures despite a favourable development mean; Table S32: Separate frozen-feature recovery study; Table S33: 84-model adapter recovery study on 114 development parents; Table S34: Optical-observation descriptor pilot outside the classification endpoint; Table S35: Retained historical integrity anchors; Table S36: Complete NEU primary support curves; Table S37: All 72 fixed-feature NEU readouts; Table S38: All 36 NEU head retention models; Table S39: Prespecified shuffled and near-domain controls; Table S40: NEU evidence scope and completion ledger; Figure S7: Six-class procedural source used in the NEU experiment; Table S41: Inventory summary; Table S42: All MVTec-FS primary fine-tuning conditions; Table S43: All GC10 ROI primary fine-tuning conditions; Table S44: Frozen means and all three synthetic-minus-ImageNet manifest contrasts; Table S45: Absolute changes, controls and support-averaged contrasts; Table S46a: All product means; Table S46b: Complete product-qualified class means and zero-F1 counts; Table S47: All included GC10 classes, including negative contrasts and zero-F1 cells; Figure S8: Complete product/class contrast plot; Table S48: Completed new blocks and interpretation boundaries; Table S49: Evidence archive index; Table S50: Group-weighted nine-class macro-F1 of the frozen readouts on GC10, averaged over manifests 0–9; Table S51: Synthetic-minus-ImageNet difference in class F1, each F1 averaged over the 30 paired cells, with the number

of test image groups per class; Table S52: All 17 of the 396 final models of the GC10 extension that contained at least one class with F1 = 0; Table S53: Mean paired differences, absolute changes between the two lowest support levels and support interactions; Table S54a: Wafer endpoints after leaving out each test parent of a class with fewer than five test parents and after leaving out each test wafer; Table S54b: Overlap of the nested support manifests; Table S55: Class F1 averaged over the paired cells of each route and support level, with the number of test units per class and the number of cells in which the class had F1 = 0; Table S56: Evaluation crop window of every wafer test parent compared with the candidate box of every other test parent on the same wafer; Table S57: Six-class morphology source; Table S58: Synthetic-minus-ImageNet macro-F1 under the head-only control, full fine-tuning and the frozen ridge readout in the same cells; Table S59: Synthetic-minus-ImageNet and synthetic-minus-shuffled macro-F1 after full fine-tuning on manifests 0–2 for each source-training seed; Table S60a: Wafer renderer gray-level priors used by the three paired sources; Table S60b: Wafer comparisons of the three prior arms and ImageNet under matched heads; Table S61: Generic fractal source versus the defect-shaped source and ImageNet; Table S62: GC10 synthetic-minus-ImageNet macro-F1 at one labeled group per class over manifests 0–9 for eight source runs; Figure S9: NEU confusion matrices; Figure S10: MVTec-FS confusion matrices per product (a, k = 1; b, k = 5); Figure S11: GC10 confusion matrices; Figure S12: Sensitivity of the support interaction to starting head and network; Figure S13: Training exposure and signed-contrast repair; Table S63: Matched real-training control on GC10; Table S64: Input-level audit of the wafer evaluation units; Table S65: Wafer endpoints under the input audit; Table S66: Hypothesis endpoints under the two decision rules; Protocol S1: Pre-specified design of the head-only and source-seed analyses; Protocol S2: Pre-specified design of the wafer-prior, generic-source and GC10 source-run analyses; Note S1: Detailed methods and training objectives; Note S2: Wafer background image and contrast priors; Supplementary Data 1: numerical results, paired matrices, predictions, protocols, integrity records and the organized code of all reported analyses. References [7–9,12–14,22,30–40,43–45] are cited in the Supplementary Materials.

Author Contributions: Software, data curation, investigation, formal analysis, visualization and writing—original draft preparation, J.W.K.; project administration (coordination with industrial collaborators) and writing—review and editing, H.-S.K. Both authors contributed to study discussions and interpretation review. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Research Equipment Technology Competitiveness Enhancement Project of the Commercialization Promotion Agency for R&D Outcomes, Republic of Korea, grant number RS-2025-02413262.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The wafer images, parent arrays and trained checkpoints are available from Jin W. Kim (jinwkim153@iae.re.kr) without author-imposed confidentiality restrictions (no repository accession yet; the equipment's proprietary design specifications remain undisclosed). Supplementary Data 1 contains the numerical results, protocols, integrity records and the organized code of all reported analyses, but no third-party images. A defect-placement routine in one historical script and the wafer renderer are covered by the pending patent applications named under Conflicts of Interest; the routine was replaced by a stub in the second revision, the renderer was not deposited, and both are available from Jin W. Kim on reasonable request under a confidentiality agreement. Every reported number can be recomputed from the deposited predictions. NEU is available at https://faculty.neu.edu.cn/songkc/en/zdylm/263265/list/ (accessed on 30 September 2026), MVTec-FS at https://github.com/ShuaiLYU/MVTec-FS (CC BY-NC-SA 4.0; accessed on 30 September 2026) and the pinned GC10 mirror at https://huggingface.co/datasets/dronefreak/GC10-DET (accessed on

30 September 2026) (original: https://github.com/lvxiaoming2019/GC10-DET-Metallic-Surface-Defect-Datasets, accessed on 30 September 2026); see Supplementary Tables S41 and S49.

Acknowledgments: The authors thank Young-tae Park (IOG Co., Ltd.), Geun-yeong Im (Daeheung Hi-Tech) and Gyeong-sik Seo (Foreworks) for technical discussions. The authors used OpenAI ChatGPT and Codex and Anthropic Claude for study and implementation discussions, software, analysis and verification scripts, English editing, figure code and layout, and Grammarly (web version; Grammarly Inc., San Francisco, CA, USA) for language checking; this went beyond spelling correction. Synthetic images came from the documented procedural renderers, and no annotation was generated by AI. The authors reviewed all assisted outputs and take full responsibility for the manuscript, software, numerical records and interpretations; AI assistance does not validate labels, optical realism or causal mechanisms.

Conflicts of Interest: J.W.K. and H.-S.K. are named inventors on pending patent applications filed by the Institute for Advanced Engineering on synthetic wafer-defect data generation and wafer inspection, including methods related to the wafer renderer used in this study; the Institute may license these applications. The authors declare no other conflicts of interest.

References

1. Hurlbert, S.H. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* **1984**, *54*, 187–211. https://doi.org/10.2307/1942661.
2. Lazic, S.E.; Mellor, J.R.; Ashby, M.C.; Munafò, M.R. A Bayesian predictive approach for dealing with pseudoreplication. *Sci. Rep.* **2020**, *10*, 2366. https://doi.org/10.1038/s41598-020-59384-7.
3. Lazic, S.E. Genuine replication and pseudoreplication. *Nat. Rev. Methods Primers* **2022**, *2*, 23. https://doi.org/10.1038/s43586-022-00114-w.
4. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. https://doi.org/10.1111/ecog.02881.
5. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. https://doi.org/10.1016/j.patter.2023.100804.
6. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.
7. Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C.F.; Huang, J.-B. A closer look at few-shot classification. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019. Available online: https://openreview.net/forum?id=HkxLXnAcFQ (accessed on 30 September 2026).
8. Luo, X.; Wu, H.; Zhang, J.; Gao, L.; Xu, J.; Song, J. A closer look at few-shot classification again. In *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, HI, USA, 23 – 29 July 2023; Volume 202, pp. 23103–23123. Available online: https://proceedings.mlr.press/v202/luo23e.html (accessed on 30 September 2026).
9. Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J.B.; Isola, P. Rethinking few-shot image classification: A good embedding is all you need? In *Computer Vision—ECCV 2020*; Springer: Cham, Switzerland, 2020; pp. 266–282. https://doi.org/10.1007/978-3-030-58568-6_16.
10. Kumar, A.; Raghunathan, A.; Jones, R.; Ma, T.; Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. In Proceedings of the Tenth International Conference on Learning Representations, Virtual, 25–29 April 2022. Available online: https://openreview.net/forum?id=UYneFzXSJWh (accessed on 30 September 2026).
11. Bouthillier, X.; Delaunay, P.; Bronzi, M.; Trofimov, A.; Nichyporuk, B.; Szeto, J.; Mohammadi Sepahvand, N.; Raff, E.; Madan, K.; Voleti, V.; et al. Accounting for variance in machine learning benchmarks. *Proc. Mach. Learn. Syst.* **2021**, *3*, 747–769. Available online: https://proceedings.mlsys.org/paper_files/paper/2021/hash/0184b0cd3cfb185989f858a1d9f5c1eb-Abstract.html (accessed on 30 September 2026).
12. Song, K.; Yan, Y. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Appl. Surf. Sci.* **2013**, *285*, 858–864. https://doi.org/10.1016/j.apsusc.2013.09.002.

13. Lyu, S.; Zhang, R.; Ma, Z.; Liao, F.; Mo, D.; Wong, W. MVREC: A general few-shot defect classification model using multi-view region-context. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 5937–5945. https://doi.org/10.1609/aaai.v39i6.32634.
14. Lv, X.; Duan, F.; Jiang, J.J.; Fu, X.; Gan, L. Deep metallic surface defect detection: The new benchmark and detection network. *Sensors* **2020**, *20*, 1562. https://doi.org/10.3390/s20061562.
15. Li, C.L.; Sohn, K.; Yoon, J.; Pfister, T. CutPaste: Self-supervised learning for anomaly detection and localization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2021; pp. 9664–9674. https://doi.org/10.1109/CVPR46437.2021.00954.
16. Zavrtanik, V.; Kristan, M.; Skočaj, D. DRÆM—A discriminatively trained reconstruction embedding for surface anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 8330–8339.
17. Zhang, X.; Xu, M.; Zhou, X. RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: New York, NY, USA, 2024; pp. 16699–16708.
18. Hu, T.; Zhang, J.; Yi, R.; Du, Y.; Chen, X.; Liu, L.; Wang, Y.; Wang, C. AnomalyDiffusion: Few-shot anomaly image generation with diffusion model. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 8526–8534. https://doi.org/10.1609/aaai.v38i8.28696.
19. Bergmann, P.; Fauser, M.; Sattlegger, D.; Steger, C. MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2019; pp. 9592–9600. https://doi.org/10.1109/CVPR.2019.00982.
20. Wang, Y.O.; Chung, Y.; Wu, C.H.; De la Torre, F. Domain gap embeddings for generative dataset augmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 28684–28694. https://doi.org/10.1109/CVPR52733.2024.02710.
21. Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2017; pp. 23–30. https://doi.org/10.1109/IROS.2017.8202133.
22. Kataoka, H.; Okayasu, K.; Matsumoto, A.; Yamagata, E.; Yamada, R.; Inoue, N.; Nakamura, A.; Satoh, Y. Pre-training without natural images. *Int. J. Comput. Vis.* **2022**, *130*, 990–1007. https://doi.org/10.1007/s11263-021-01555-8.
23. Alqudah, R.; Al-Mousa, A.A.; Abu Hashyeh, Y.; Alzaibaq, O.Z. A systemic comparison between using augmented data and synthetic data as means of enhancing wafermap defect classification. *Comput. Ind.* **2023**, *145*, 103809. https://doi.org/10.1016/j.compind.2022.103809.
24. Peñaloza, R.J.S.; Moghadam, A.; Davoudi Kakhki, F. Vision-based wafer inspection in semiconductor manufacturing: A case study on scratch defect detection using synthetic data and YOLO models. *Int. J. Data Sci. Anal.* **2026**, *22*, 78. https://doi.org/10.1007/s41060-026-01034-8.
25. Liu, J.; Wu, Y.; Luo, X.; Wu, Z. Anomaly multi-classification in industrial scenarios: Transferring few-shot learning to a new task. *arXiv* **2024**, arXiv:2406.05645. Available online: https://arxiv.org/abs/2406.05645 (accessed on 30 September 2026).
26. Kornblith, S.; Shlens, J.; Le, Q.V. Do better ImageNet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2019; pp. 2661–2671. https://doi.org/10.1109/CVPR.2019.00277.
27. He, K.; Girshick, R.; Dollár, P. Rethinking ImageNet pre-training. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019; pp. 4918–4927. Available online: https://openaccess.thecvf.com/content_ICCV_2019/html/He_Rethinking_ImageNet_Pre-Training_ICCV_2019_paper.html (accessed on 30 September 2026).
28. Wang, Z.; Dai, Z.; Póczos, B.; Carbonell, J. Characterizing and avoiding negative transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2019; pp. 11293–11302. https://doi.org/10.1109/CVPR.2019.01155.
29. Geirhos, R.; Jacobsen, J.H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; Wichmann, F.A. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2020**, *2*, 665–673. https://doi.org/10.1038/s42256-020-00257-z.
30. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 11966–11976. https://doi.org/10.1109/CVPR52688.2022.01167.
31. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019. Available online: https://openreview.net/forum?id=Bkg6RiCqY7 (accessed on 30 September 2026).

32. Shen, R.; Bubeck, S.; Gunasekar, S. Data augmentation as feature manipulation. In *Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022;* Volume 162, pp. 19773–19808. Available online: https://proceedings.mlr.press/v162/shen22a.html (accessed on 30 September 2026).
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
34. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent: A new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.
35. Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 2818–2829. https://doi.org/10.1109/CVPR52729.2023.00276.
36. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.* **2024**. Available online: https://openreview.net/forum?id=a68SUt6zFt (accessed on 30 September 2026).
37. Cawley, G.C.; Talbot, N.L.C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107. Available online: https://www.jmlr.org/papers/v11/cawley10a.html (accessed on 30 September 2026).
38. Li, X.; Grandvalet, Y.; Davoine, F. Explicit inductive bias for transfer learning with convolutional networks. *Proc. Mach. Learn. Res.* **2018**, *80*, 2825–2834. Available online: https://proceedings.mlr.press/v80/li18a.html (accessed on 30 September 2026).
39. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. *Proc. Mach. Learn. Res.* **2017**, *70*, 1321–1330. Available online: https://proceedings.mlr.press/v70/guo17a.html (accessed on 30 September 2026).
40. Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **1979**, *7*, 1–26. https://doi.org/10.1214/aos/1176344552.
41. Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; Vinyals, O. Understanding deep learning requires rethinking generalization. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017. Available online: https://openreview.net/forum?id=Sy8gdB9xx (accessed on 30 September 2026).
42. Maennel, H.; Alabdulmohsin, I.M.; Tolstikhin, I.O.; Baldock, R.; Bousquet, O.; Gelly, S.; Keysers, D. What do neural networks learn when trained with random labels? *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 19693–19704. Available online: https://proceedings.neurips.cc/paper_files/paper/2020/hash/e4191d610537305de1d294adb121b513-Abstract.html (accessed on 30 September 2026).
43. Ringwald, T.; Stiefelhagen, R. Adaptiope: A modern benchmark for unsupervised domain adaptation. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 101–110. https://doi.org/10.1109/WACV48630.2021.00015.
44. Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; Wang, B. Moment matching for multi-source domain adaptation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1406–1415. https://doi.org/10.1109/ICCV.2019.00149.
45. Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; Li, H. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In Computer Vision—ECCV 2022; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 13695, pp. 493–510. https://doi.org/10.1007/978-3-031-19833-5_29.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.