Life of Research · Published 21 November 2026 · DOI 10.5281/zenodo.24408406

A Near Neighbour Is Not a Native Feature: What Nearest-Neighbour Novelty Detection Owes to the Space It Measures In, Why the Case For and Against Pretrained Features Turns on Benchmark and Baseline More Than on Where the Features Came From, and the Comparison That Would Separate Scoring Rule From Representation

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. A novelty detector is trained on normal data and asked to flag whatever is unlike it. Since 2020 two positions have competed over how such a detector should be built. One holds that the decisive ingredient is a feature space borrowed from a network pretrained on a large external corpus, after which a nearest-neighbour distance is enough, and that self-supervised or reconstruction-based detectors trained only on the normal data are outclassed. The other holds that detectors trained in-domain, such as autoencoders whose novelty score combines the reconstructions of an input's nearest latent neighbours with the distance to them, are competitive without any pretrained network and transfer to domains, such as radio-telescope spectrograms, that look nothing like natural photographs. In this analysis the author reviews and synthesises published results to show that the two positions are answering a question neither benchmark design isolates. A published comparison of two detectors almost always changes at least three things at once: the space in which distance is measured, the rule that turns distance into a score, and the kind of novelty the benchmark plants, while a fourth factor, the evaluation (metric, baseline implementation, hyperparameter selection), decides which of the three gets the credit. On one industrial benchmark, published image-level results for detectors that all use ImageNet features and a distance or density score range from 75.0 to 99.1 AUROC, a spread several times larger than the 3.3-point margin by which the nearest-latent-neighbour autoencoder beats its own re-implementation of the pretrained baseline. On one natural-image benchmark, supervised ImageNet features with a simple density score are reported at 80.0, 86.2 and 92.5 AUROC by three groups. Off the pretraining distribution the verdict flips with the dataset: in one seven-dataset medical study a plain autoencoder beats frozen ImageNet features on two datasets and loses by up to 36 points elsewhere, and an autoencoder whose residual is measured in pretrained feature space beats both on most. The paper consolidates nineteen published results in a table and a figure, states a procedure for crediting a novelty-detection gain to a representation's origin rather than to the scoring rule or the evaluation, and names the studies that would settle what remains open.

Keywords: novelty detection; anomaly detection; nearest neighbours; autoencoders; pretrained representations; transfer learning; evaluation methodology; model selection

1. Introduction

Novelty detection asks a model to learn what normal data look like and to raise a flag when something unlike it arrives. The problem has many names, among them anomaly detection, outlier detection, out-of-distribution detection and one-class classification, and a survey that maps them onto one framework notes how much the communities have worked in isolation (Yang, J. et al., 2021). The applications are concrete: defective parts on a production line (Bergmann et al., 2019), lesions in brain scans (Baur et al., 2020), interference in radio-telescope data (Mesarcik et al., 2022). In each case the anomalies of interest are rare, varied, and unknown at training time, so the detector has to be built from normal examples alone.

For a long time the obvious tool was reconstruction. An autoencoder trained on normal data should reconstruct normal inputs well and novel inputs badly, so the reconstruction error serves as a novelty score. Deep generative models offered a second tool, the likelihood. Both tools ran into the same difficulty. Autoencoders often reconstruct anomalies well too (Gong et al., 2019; Perera, Nallapati and Xiang, 2019), and generative models trained on one image dataset can assign higher likelihood to a different one (Nalisnick et al., 2018). A third family trained classifiers on self-supervised pretext tasks, such as predicting which geometric transformation had been applied (Golan and El-Yaniv, 2018; Hendrycks et al., 2019b), or contrastive objectives on shifted views (Tack et al., 2020), and scored novelty from the trained network's behaviour.

In 2020 a short paper argued that much of this machinery could be replaced by something old. Bergman, Cohen and Hoshen (2020) took a ResNet pretrained on ImageNet classification, embedded the normal training images, and scored a test image by its mean distance to its k nearest training embeddings. On one-class CIFAR-10 their method, DN2, averaged 92.5 AUROC, above every self-supervised method they

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

compared against. A line of work followed: patch-level nearest neighbours for localisation (Cohen and Hoshen, 2020), Gaussian models of pretrained features (Rippel, Mertens and Merhof, 2020; Defard et al., 2020), adaptation of pretrained features to the normal class (Reiss et al., 2020; Reiss and Hoshen, 2021), and a memory bank of patch features that reached 99.1 image-level AUROC on the MVTec AD industrial benchmark (Roth et al., 2021). By 2022 the same group stated as a position that advances in representation learning are the main explanation for recent gains, and that a simple nearest-neighbour score on a good representation essentially solves the most commonly reported image benchmark (Reiss et al., 2022).

The anchor for this paper is a counter-position from the same period. Mesarcik et al. (2021) proposed the nearest-latent-neighbours (NLN) algorithm. It trains an autoencoder on normal data, finds the test input's nearest neighbours in the autoencoder's latent space, decodes those neighbours, and scores the input by a weighted sum of two terms: the error between the input and the neighbours' reconstructions, and the mean latent distance to the neighbours. The abstract states that the method is memory and time efficient, needs little data augmentation, and does not rely on pretrained networks. The authors later carried NLN to radio astronomy, where natural-image features are argued to be ill-suited (Mesarcik et al., 2022; Mesarcik et al., 2023). The NLN paper is careful about scope: it limits its comparison to autoencoders and explains why it did not benchmark against pretrained or self-supervised methods on AUROC alone. Yet its tables do include the pretrained nearest-neighbour baseline, and its claims, read together with the pretrained-feature literature, pose a question the field has not answered cleanly: when one novelty detector beats another, is it because of where its features came from, or because of something else?

In this analysis the author reviews and synthesises published results to show that the question, as usually posed, cannot be answered from the comparisons the literature reports, because those comparisons change several things at once. Four are separable. The first is the representation: the space in which a distance is measured, which may be raw input, a latent space learned on the normal data, a space borrowed from an external pretrained network, or a borrowed space adapted to the normal data. The second is the scoring rule: reconstruction residual, nearest-neighbour distance, Gaussian or Mahalanobis density, kernel density, a one-class boundary, or a student-teacher discrepancy. The third is the referent: what the benchmark counts as novel, whether a held-out semantic class, a local surface defect, a low-level change of texture or intensity, or a signal artefact. The fourth is the evaluation: the metric, how the baseline was implemented, and how hyperparameters were chosen. The pretrained-feature position and the NLN position each make a claim about the first factor. The published evidence shows that the other three move results by as much or more.

The argument proceeds in steps. Section 2 states the positions. Section 3 defines the four factors. Section 4 re-reads the NLN paper's own tables. Section 5 shows that, on a single benchmark, detectors from the same family differ by more than the families differ from each other. Section 6 shows that the representation and the scoring rule are each a large lever when the other is held fixed. Section 7 asks what reconstruction actually measures. Sections 8 and 9 follow the comparison off the pretraining distribution, into medical images, radio spectrograms, 3D surfaces and tabular data. Section 10 describes the hybrid that has quietly overtaken both positions: reconstruction scored in a borrowed feature space. Section 11 turns to evaluation and to the unsolved problem of choosing a detector without anomaly labels. Section 12 collects the numbers in Table 1 and Figure 1, Section 13 states a procedure (Algorithm 1) for crediting a gain to a representation's origin, Section 14 states the strongest objections, Section 15 says what is not known, and Section 16 proposes the studies that would settle it.

The contribution is the decomposition, the re-reading of the anchor and of the pretrained-feature literature from their own numbers, the within-family-spread diagnostic of Section 5, and the procedure of Section 13. The paper does not rank detectors. Every value in its table and figure is a published one, credited on its row or in its caption; arithmetic on published values (differences, ranges) is labelled as the author's.

2. The positions

2.1 Borrowed features suffice

The pretrained-feature position has a strong and a weak form. The weak form is an empirical observation: on the image benchmarks the field reports, a nearest-neighbour or Gaussian score on ImageNet features beats detectors trained from scratch on the normal data. Bergman, Cohen and Hoshen (2020) report this for CIFAR-10, CIFAR-100, Fashion-MNIST, flowers, birds and cats-versus-dogs, and also for an aerial dataset (DIOR, 92.2 against 83.2 for a rotation-prediction baseline) and a small white-blood-cell microscopy set (WBC, 82.9 against 60.5). They note that the method also wins on Fashion-MNIST, whose images are visually far from ImageNet, and conclude that DN2-like methods should be broadly applicable. Reiss et al. (2020) report that combining pretrained features with simple scoring beats far more complex methods, and that naive fine-tuning of those features on one class tends to collapse them. Xiao, Yan and Amit (2021) go further and ask whether in-domain representation learning is needed at all; replacing task-specific feature extractors with one network pretrained on ImageNet by a self-supervised loss, they report results competitive with or better than earlier two-stage methods and suggest that the benefit of domain-specific training may be marginal.

The strong form is a claim about cause: that the representation, not the scoring rule, is what makes detectors work. Reiss et al. (2022) state it directly. They

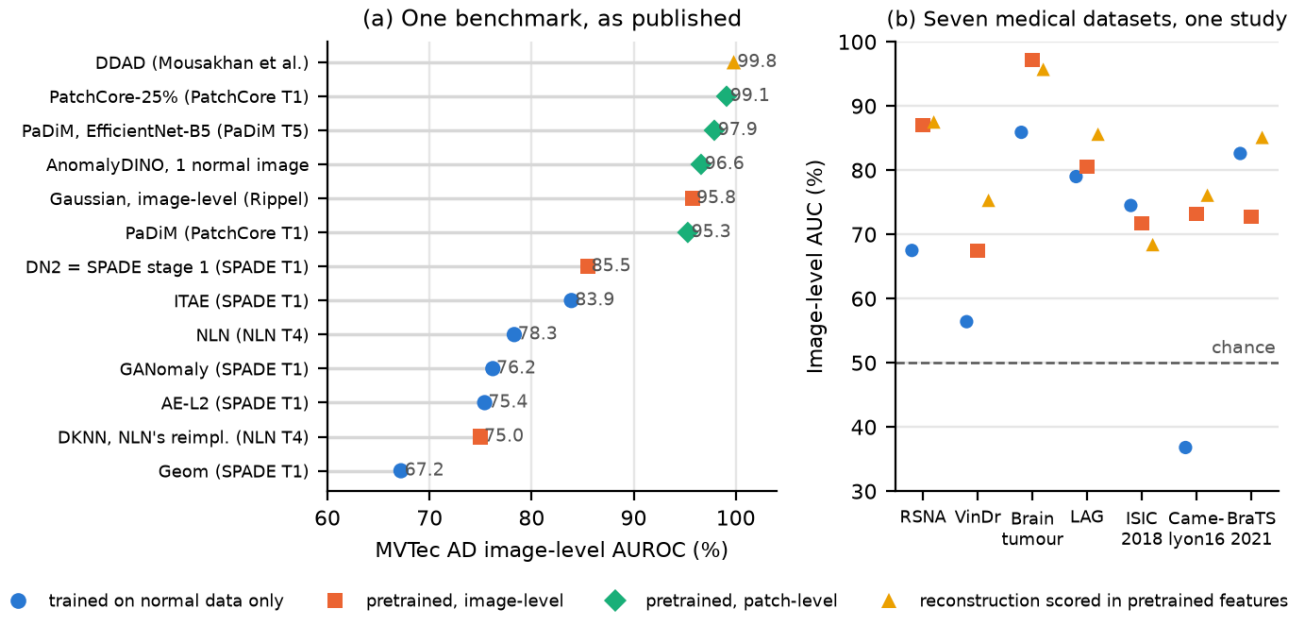


Figure 1. (a) Image-level AUROC on MVTec AD as published: detectors trained on normal data only, as listed by Cohen and Hoshen (2020, Table 1) and NLN (Mesarcik et al., 2021, Table 4); image-level pretrained detectors (NLN's nearest-neighbour re-implementation, Table 4; SPADE's first stage, Cohen and Hoshen 2020, Table 1; the image-level Gaussian of Rippel et al., 2020); patch-level pretrained detectors (PaDiM and PatchCore as listed by Roth et al., 2021, Table 1; PaDiM on EfficientNet-B5, Defard et al., 2020, Table V; AnomalyDINO, Damm et al., 2024, one normal image); and DDAD (Mousakhan et al., 2023), which scores diffusion reconstructions in adapted pretrained features. (b) Image-level AUC on seven medical datasets from MedIAnomaly (Cai et al., 2024, Table 6) for an L2 autoencoder, a frozen ImageNet ResNet18 with a Gaussian density estimator, and an autoencoder with an ImageNet perceptual residual. Colour and marker shape both encode the family. All values as published.

argue that the most successful recent approaches share a two-stage structure, a representation followed by a density estimate as simple as k nearest neighbours, and that advances in representation learning are the main explanatory factor. Notably, the same paper is also the most candid list of where that recipe runs out: masked-autoencoder representations perform much worse for anomaly detection than contrastive ones; scenes with many objects and fine-grained anomalies remain hard; handcrafted 3D descriptors beat learned ones on surface data; and on tabular data no learned representation beat the raw features.

2.2 In-domain detectors suffice

The in-domain position is older and more varied. Self-supervised detectors trained only on the normal data, by transformation prediction (Golan and El-Yaniv, 2018), rotation heads (Hendrycks et al., 2019b) or contrastive learning on shifted instances (Tack et al., 2020), were the state of the art before DN2. Sohn et al. (2020) argued for a two-stage design in which a self-supervised representation is learned from the one-class data and then a statistical one-class classifier is fitted on top, and they reported that their distribution-augmented contrastive representation, learned from as few as 100 CIFAR-10 images, matched an ImageNet-pretrained ResNet-50. Reconstruction-based detectors remained the default in medical imaging (Baur et al., 2020) and elsewhere.

NLN belongs to this position but changes the score. Its premise is that an autoencoder's failure mode, reconstructing novel inputs too well, can be countered by reconstructing the input's latent neighbours instead of the input. If a novel digit is encoded near training examples of other digits, its neighbours decode to those other digits, and the error against the input is large. The latent distance term then adds the information that a novel input sits far from the training codes. The paper reports gains of up to 17 percent AUROC in the multi-class setting and 8 percent in the single-class setting over the same autoencoders scored by reconstruction alone (Mesarcik et al., 2021, Tables 1-2).

2.3 The anchor, and the position the two share

The two positions agree on more than they admit. Both score novelty by distance to stored training examples. DN2 measures it in a borrowed space; NLN measures part of it in a learned latent space and part of it in pixel space through decoded neighbours. The two share the nearest-neighbour distance term but differ in both representation and scoring rule, since NLN adds a pixel-space residual on decoded neighbours, which its authors call "the critical difference" between DN2-style methods and their own (Mesarcik et al., 2021). The pair is therefore an instance of the confound this paper analyses, not a clean test of representation alone, and it shows why the question has proved hard: the evidence offered for each side changes other things besides the representation.

3. Four factors under one comparison

This section fixes vocabulary. The four factors are stipulated here, and the rest of the paper uses them throughout.

Representation (R). The space in which a distance or density is computed. Four kinds recur: raw input (pixels, spectrogram bins, table columns); a space learned on the normal data, such as an autoencoder's latent code or a self-supervised embedding; a space borrowed from a network trained on an external corpus, such as ImageNet classification, self-supervised ImageNet training, or a vision-language model; and a borrowed space adapted to the normal data. A representation also has a granularity, image-level or patch-level, and a depth, the layer from which it is read.

Scoring rule (S). The function from representation to score: residual between input and reconstruction; distance to the k nearest stored examples; Mahalanobis distance under a fitted Gaussian; kernel density; a one-class boundary; the discrepancy between a student network and a teacher; or a pretext-task classifier's confidence. NLN's score is a weighted sum of two rules applied in two representations.

Referent (N). What the benchmark designates as novel. Yang, J. et al. (2021) separate sensory anomaly detection, where the shift is in the input's appearance, from semantic anomaly detection, where it is in the label. Ahmed and Courville (2019) argued that the practically relevant out-distributions are semantic and that many benchmarks pairing whole datasets do not test that. Industrial benchmarks plant small local defects on an otherwise normal object (Bergmann et al., 2019). Radio-telescope benchmarks plant interference or instrument faults in time-frequency data (Mesarcik et al., 2022; Mesarcik et al., 2023). Tabular benchmarks plant local, global, dependency or clustered anomalies (Han et al., 2022). A representation can be well suited to one referent and blind to another.

Evaluation (E). The metric (image AUROC, pixel AUROC, AUPRC, region overlap, F1 at a threshold); how each baseline was obtained (the authors' code, a re-implementation, or a number copied from a paper); and how hyperparameters, including layer and backbone choices, were selected. The last matters more in novelty detection than in supervised learning, because there is no labelled validation set of the anomalies that will be met.

The decomposition is not new in outline. Ruff et al. (2020), reviewing deep and shallow anomaly detection together, organised methods by the model of normality they fit and the assumptions they make implicitly, and the NLN paper's own background section lists three steps, one or more of which a detector requires: a model of the normal distribution, a measure of fit, and a decision rule. What this paper adds is the insistence that the referent and the evaluation are factors on the same footing, because the evidence below shows them moving results as much as the representation does. The scoring rule also has theory behind it that the representation lacks. Gu, Akoglu and Rinaldo (2019) analyse nearest-neighbour detectors through the distance-to-measure, derive finite-sample guarantees under a contamination model, and relate the methods' empirical performance to the dimensionality of the problem. No comparable guarantee exists for the claim that a borrowed space preserves the distances that matter for a given referent; that claim can only be checked empirically, dataset by dataset.

The four factors give a simple rule for reading a comparison. A claim that detector A beats detector B because of A's representation is supported only if A and B share S, N and E, or if the comparison has been repeated across enough values of S, N and E for the sign to hold. Few published comparisons meet that bar, and the ones in this paper's evidence base are examined in turn.

4. The anchor re-read: what the NLN paper's tables show

4.1 Where NLN wins and loses against the pretrained baseline

The NLN paper reports two comparisons against the deep nearest-neighbour baseline, which it labels DKNN. In the multi-class setting, where all classes but one are normal (its Table 3), NLN scores 0.921 on MNIST against 0.791 for DKNN, and 0.763 on Fashion-MNIST against 0.746, but 0.560 on CIFAR-10 against 0.714. In the single-class setting (its Table 4), NLN scores 0.974 on MNIST against 0.917, 0.941 on Fashion-MNIST against 0.938, and 0.783 on MVTec AD against 0.750, but 0.658 on CIFAR-10 against 0.890.

Read as a whole, the pattern is regular. NLN wins clearly on MNIST, narrowly on Fashion-MNIST and MVTec AD, and loses by 15 to 23 points on CIFAR-10, the only dataset in the set where novelty is defined by object class under large intra-class variation in appearance. The authors give the reason themselves: images within one CIFAR-10 class can differ more in pixel terms than images of different classes, so pixel-level error cannot separate them. That explanation is a statement about the referent. A score computed partly in pixel space detects the kind of novelty that shows up in pixels, and on the dataset where novelty is semantic it is close to chance in the multi-class setting and far below the pretrained baseline in both.

4.2 Which baseline was beaten

The second observation concerns evaluation. The NLN paper's footnote to Table 4 says that the authors used DKNN's original implementation everywhere except MVTec AD, where they used their own TensorFlow re-implementation. On MVTec AD the re-implemented baseline scores 0.750. The nearest-neighbour baseline's own originators report a different number for the same thing. Cohen and Hoshen (2020) state that the first stage of their SPADE method is identical to DN2, and report it

at 85.5 average image-level AUROC on MVTec AD (their Table 1); the PatchCore paper lists the same 85.5 for SPADE (Roth et al., 2021, Table 1). NLN's 0.783 sits between the re-implemented baseline and the originators' figure. The comparison that supports the claim of a gain on MVTec AD is therefore a comparison with a weaker version of the baseline than its authors report. On the datasets where the original code was used, the numbers are close to the originators' but not identical: 0.890 against DN2's reported 92.5 on CIFAR-10, and 0.938 against 94.4 on Fashion-MNIST (Bergman, Cohen and Hoshen, 2020, Tables 1-2).

4.3 What the ablation credits

The paper's ablation (its Table 6) scores the same feature-consistent autoencoder four ways in the multi-class setting. On MNIST: reconstruction error alone, 0.778; the neighbours' reconstruction error, 0.822; a feature-consistency term that compares the input's code with codes re-extracted from the neighbours' reconstructions, 0.913; the full score, 0.921. On CIFAR-10 the four values are 0.511, 0.513, 0.551 and 0.553. Two readings follow. The largest single step on MNIST comes when the comparison moves from pixel space into a learned code space, not when the input's reconstruction is replaced by the neighbours'. And on CIFAR-10 no variant leaves the region of chance. Both readings are consistent with the thesis that where a distance is measured matters more than the neighbour trick itself, though the ablation was not designed to test that and the inference is the author's.

4.4 How the configuration was chosen

The paper reports, for each dataset and setting, the configuration it found optimal: the number of neighbours K, the latent dimension L, and the weight alpha between the two terms (for instance K = 1, L = 32, alpha = 0.5 on multi-class CIFAR-10, and K = 1, L = 128, alpha = 0.8 for single-class MVTec AD), and plots the sensitivity of AUROC to K and L. The text does not describe a validation split of novel classes separate from the test classes on which these were chosen. The radio follow-up chose its hyperparameters by grid search on the average of the maximum AUROC and AUPRC (Mesarcik et al., 2022). Neither practice is unusual in the field; Section 11 returns to why it matters.

4.5 What the efficiency claim compares

The efficiency case is real and specific. The NLN encoder and decoder take 7.41 ms and 9.63 ms per MVTec AD image on the authors' GPU, against 43.3 ms for a ResNet-50 forward pass; the latent space has 128 dimensions against 2048; the MVTec model has 1.79 million parameters against 25.58 million. These are honest engineering advantages. They are not evidence about novelty-detection accuracy, and later pretrained detectors narrowed them: EfficientAD reports a small feature extractor whose per-image cost on a current GPU is below one millisecond (Batzner, Heckler and Konig, 2023), and coreset subsampling lets PatchCore keep 1 percent of its memory bank at 99.0 image AUROC (Roth et al., 2021, Table 1).

4.6 Summary of the re-reading

NLN's evidence supports a narrower claim than its framing. Against autoencoders scored by reconstruction alone, adding latent neighbours helps on every dataset tested. Against the pretrained baseline, NLN wins where novelty is visible at the pixel level and the baseline was run as re-implemented, and loses heavily where novelty is semantic. That is a statement about the interaction of representation and referent, plus a caveat about evaluation, rather than a statement that pretrained features are unnecessary.

5. One benchmark, one family, a wide spread

If the origin of the representation were the decisive variable, detectors that share it should cluster. They do not.

On MVTec AD, image-level AUROC as reported for detectors that all score ImageNet features by a distance or a density runs as follows: 75.0 for DKNN as re-implemented by the NLN authors; 85.5 for the image-level nearest-neighbour stage of SPADE (Cohen and Hoshen, 2020); 95.3 for PaDiM's patch-level Gaussians (Defard et al., 2020, as listed by Roth et al., 2021); 95.8 for image-level Gaussians with Mahalanobis distance on average-pooled ImageNet classifier features (Rippel, Mertens and Merhof, 2020); 97.9 for PaDiM on an EfficientNet-B5 backbone (Defard et al., 2020, Table V, listed by Roth et al., 2021); and 99.1 for PatchCore with a 25 percent coreset (Roth et al., 2021, Table 1), rising to 99.6 with an ensemble of backbones at higher resolution (its Table 4). The range within the family, by the author's arithmetic, is 24.1 points for single models. AnomalyDINO, a patch-level nearest-neighbour detector on self-supervised DINOv2 features, reports 96.6 with a single normal reference image (Damm et al., 2024). The detectors trained only on the normal data, as listed by Cohen and Hoshen (2020, Table 1), span 67.2 for geometric-transformation prediction, 75.4 for an L2 autoencoder, 76.2 for GANomaly and 83.9 for ITAE, with NLN's 78.3 inside that range. Figure 1(a) plots these values.

The NLN paper's margin over its pretrained baseline on this benchmark is 3.3 points. The within-family spread of the pretrained detectors is more than seven times that, and the steps within it are not about where the features came from. The largest step in the list, 10.5 points, is an implementation difference: the move from the re-implemented baseline to its originators' figure. The next, 9.8 points from SPADE's first stage to PaDiM, changes three things at once, granularity (image to patch), density model (k nearest neighbours to Gaussian) and layers, and so cannot be credited to granularity alone; image-level Gaussians on average-pooled features from every level of an EfficientNet already reach 95.8 (Rippel, Mertens and Merhof, 2020), above patch-level PaDiM. The 2.6-point step from PaDiM to PaDiM on EfficientNet-

B5 is a backbone change. Heckler, Konig and Bergmann (2023) studied this directly and found the investigated methods highly sensitive to the particular choice of feature space, to the point where choosing a single layer well beat expensive ensembles.

The same structure appears on a natural-image benchmark. For one-class CIFAR-10 with supervised ImageNet ResNet features and a simple density score, Sohn et al. (2020, Table 2) report 80.0 with a one-class SVM on ResNet-50 features; Xiao, Yan and Amit (2021, Table 3) list 86.2 for supervised ImageNet features in their comparison; and Bergman, Cohen and Hoshen (2020, Table 1) report 92.5 with k nearest neighbours. Each paper's conclusion matches its own number. Sohn et al. present the ImageNet baseline to highlight the importance of in-domain representation learning, and their in-domain contrastive representation reaches 92.5 on CIFAR-10. Bergman et al. present the same kind of baseline as beating in-domain methods. A spread of 12.5 points on one dataset, for one kind of feature, is larger than most of the margins either camp reports. Among the variables that differ across the three are the scoring rule and the backbone; Bergman et al. themselves report that deeper ResNets give better DN2 results.

The lesson is not that pretrained features do not matter. It is that "uses pretrained features" names a family whose members differ among themselves by more than the family differs from its rivals, so a comparison between one member and one rival says little about the family.

6. Both levers move results

Holding one factor fixed and varying another isolates its effect. The literature contains several such contrasts, though few were designed for the purpose.

6.1 The representation, with the rule fixed

Reiss et al. (2022, Table 2) score two self-supervised ImageNet representations with the same nearest-neighbour rule. Masked-autoencoder features reach 78.1 AUROC on one-class CIFAR-10, 73.1 on CUB-200 and 83.2 on an ImageNet subset; DINO features reach 97.1, 93.9 and 99.3. Same rule, different self-supervised objective: a 19-point difference on CIFAR-10. The authors suggest that the masked objective, which reconstructs patches, encodes local information at the expense of the semantic properties that define the benchmark's novelty. That is again a statement about referent. In Sohn et al. (2020, Table 2), with the one-class SVM fixed, ImageNet ResNet-50 features reach a weighted mean of 84.0 over five datasets and the in-domain distribution-augmented contrastive representation reaches 89.9; a denoising-autoencoder representation reaches 80.4. Xiao, Yan and Amit (2021) report that among self-supervised ImageNet objectives the contrastive ones work best for outlier detection, even when other objectives score higher on ImageNet linear evaluation.

6.2 The rule, with the representation fixed

Sohn et al. (2020) report that a rotation-prediction network scored by its own classifier reaches 83.1 mean AUC, while the same network's representation scored by kernel density reaches 86.6, and they draw the conclusion that the second-stage detector matters. Reiss et al. (2022, Table 5) report the same on tabular data: GOAD scored by its own auxiliary task reaches 78.2 ROC-AUC on the ODDS benchmarks and 87.6 when its representation is scored by nearest neighbours instead. In out-of-distribution detection for classifiers, a non-parametric nearest-neighbour distance in the classifier's feature space lowers the false positive rate at 95 percent true positive rate by 24.77 percent against a strong Mahalanobis-based baseline on ImageNet-1k (Sun, Y. et al., 2022). Xiao, Yan and Amit (2021) found the opposite in their setting: once the representation was strong, results were consistent across choices of feature-space detector. The two observations fit together if the scoring rule matters most when the representation is weak or mismatched to the referent; that reconciliation is the author's hypothesis, not a measured result.

6.3 The referent, with both fixed

Fixed detectors also change rank with the referent. Reiss et al. (2022, Table 3) report that PANDA, the adaptive pretrained detector, reaches 61.5 AUROC when one vehicle category in whole MS-COCO images is held out, against 77.0 when objects are cropped; the same detectors are near ceiling on CIFAR-10. Fort, Ren and Lakshminarayanan (2021) report that large pretrained vision transformers raise near-OOD detection on CIFAR-100 versus CIFAR-10 from 85 to 96 AUROC, with a pretrained ViT fine-tuned on the labelled in-distribution classes, a referent made of natural-image object categories of the kind the pretraining corpus contains. Sohn et al. (2020) attribute the ImageNet baseline's relative strength on CIFAR-100, cats-versus-dogs and CelebA to the overlap between a subset of ImageNet classes and those datasets' classes. Where the borrowed space was trained to separate the very categories the benchmark holds out, nearest neighbours in it will separate them too.

The referent also decides what a representation should ignore, and here the pretrained-feature literature has documented its own limit. Cohen, Kahana and Hoshen (2022) point out that whether a sample differing from the norm in some attribute counts as anomalous depends on the operator, and propose a detector that learns a representation from which designated nuisance attributes have been removed, so that an operator need only name what to ignore. Reiss et al. (2022) extend the point to an argument that, without such guidance, the right representation for a given anomaly detection task is unidentifiable: two users of the same normal data can want opposite notions of novelty. A borrowed representation fixes those choices in advance, by the pretraining objective, and an in-domain one fixes them by the reconstruction or pretext objective. Neither is neutral. A related result shows how much the pretrained

position borrows from label structure: when the normal data contain several classes and no labels, Cohen, Abutbul and Hoshen (2021) recover much of the benefit of labelled out-of-distribution methods by clustering the normal data into pseudo-classes first. Hendrycks, Lee and Mazeika (2019a) had shown earlier that pretraining improves out-of-distribution detection and other robustness measures even where it adds little to classification accuracy, which is the observation the 2020 nearest-neighbour results generalised.

6.4 What the three contrasts establish

Each of the four factors has been shown, somewhere, to move results by as much as the margins in the disputed comparisons. A representation swap under one rule moved CIFAR-10 by 19 points, a rule swap on one representation moved tabular data by 9 points, a referent swap moved one detector by 15 points, and an implementation difference within one family moved MVTec AD by 10 points or more. None of these is the variable either position names.

7. What reconstruction measures

7.1 The identity problem

The case for reconstruction rests on a premise: a model trained to reproduce normal data will fail to reproduce novel data. The literature documents three ways the premise fails. The first is generalisation: autoencoders reconstruct some anomalies well, which motivated memory modules that force reconstructions toward stored normal prototypes (Gong et al., 2019) and constraints that keep the latent space from encoding out-of-class examples (Perera, Nallapati and Xiang, 2019). In feature reconstruction for many classes at once, the same failure appears as what You et al. (2022) call an identical shortcut, in which normal and anomalous inputs are both recovered well. Bercea, Rueckert and Schnabel (2022) found state-of-the-art autoencoders on a lung-pathology task either unable to constrain their latent manifold, and so able to reconstruct abnormal patterns, or unable to restore inputs accurately, producing blurred reconstructions. NLN's neighbour reconstruction is a response to the first failure, and Denouden et al. (2018) had earlier proposed adding a Mahalanobis distance in latent space for anomalies that sit far from the training codes yet near the learned manifold. NLN combines both ideas.

7.2 Low-level statistics dominate

The second failure is that pixel-level scores are dominated by low-level statistics. Nalisnick et al. (2018) showed flows, VAEs and PixelCNNs trained on CIFAR-10 assigning higher likelihood to SVHN. Serra et al. (2019) traced the effect to input complexity and corrected for it with a complexity estimate. Kirichenko, Izmailov and Wilson (2020) found that normalizing flows learn local pixel correlations and generic transformations that are not specific to the training data. Schirrmeister et al. (2020) traced the failure to two things acting together, the inductive bias of convolutional models and the shared statistics of natural images: whatever natural-image dataset such a network is trained on, it acquires much the same local feature statistics, these govern the likelihood, and differences that live at a higher level, such as object shape, are swamped. Havtorn et al. (2021) reached a related conclusion for hierarchical VAEs and built a likelihood-ratio score that requires an input to be in-distribution at every feature level. These results concern likelihoods rather than reconstruction error, but the mechanism they identify, scores computed in pixel space being governed by pixel-level regularities, is the same one the NLN authors invoke for CIFAR-10.

7.3 The bottleneck must be tuned

The third failure is practical. Reconstruction detectors require the information bottleneck to be set so that normal data pass and novel data do not, and the right setting depends on the data. Graham et al. (2022) note that reconstruction methods are less favoured because they need careful tuning of the latent size, and use diffusion models instead, treating the noise level as an externally controlled bottleneck and reading a multi-dimensional reconstruction error across noise levels; a second diffusion method, Lift, Map, Detect, similarly corrupts an input and maps it back with a diffusion model, scoring the distance (Liu, Z. et al., 2023). Zhou (2022) recasts autoencoder-based detection as querying a proxy of conditional data uncertainty and argues for compressing the latent space as far as reconstruction allows. In medical images, the MedIAnomaly benchmark found that datasets whose anomalies are local benefited from very small latent sizes and datasets with global semantic anomalies from large ones (Cai et al., 2024). The bottleneck is thus a hyperparameter whose right value depends on the referent, and the normal data alone do not obviously indicate which value is right.

7.4 What follows for the anchor

Reconstruction in pixel space is a detector of sensory novelty. It works where the novelty is a visible departure in appearance from a homogeneous normal class, which describes MNIST digits, many industrial defects, and interference in radio spectrograms. It fails where normal data vary more in appearance than the novelty does, which describes CIFAR-10 classes and, as Section 8 shows, some medical anomalies. NLN improves reconstruction against its own failure mode without changing which referent it is suited to. A pretrained semantic representation has the opposite profile. Neither profile is a general property of "pretrained" or "in-domain".

8. Off the pretraining distribution: medical images

If borrowed features owe their success to overlap with the benchmark's categories, they should weaken on images unlike their pretraining corpus. Medical imaging is the largest test, and its results are mixed in a way that is itself informative.

8.1 A seven-dataset comparison

MedIAnomaly evaluates thirty methods on seven medical datasets across five modalities under one protocol (Cai et al., 2024). Its Table 6 reports image-level AUC. A plain autoencoder with L2 residual scores 67.5 on RSNA chest X-rays, 56.4 on VinDr-CXR, 85.9 on a brain-tumour MRI set, 79.0 on LAG fundus images, 74.5 on ISIC2018 dermoscopy, 36.8 on Camelyon16 histopathology and 82.6 on BraTS2021 MRI. A frozen ImageNet ResNet18 with a Gaussian density estimator scores 87.0, 67.4, 97.2, 80.5, 71.7, 73.2 and 72.8. An autoencoder whose residual is measured with a perceptual loss computed from ImageNet features (AE-PL) scores 87.5, 75.3, 95.7, 85.6, 68.4, 76.1 and 85.1. Figure 1(b) plots the three series.

By the author's count, the plain autoencoder beats the frozen ImageNet features on two of seven datasets (ISIC2018 and BraTS2021, the latter by 9.8 points) and loses on five, by up to 36.4 points on Camelyon16, where it is below chance. The perceptual autoencoder beats the frozen features on five of seven and the plain autoencoder on six of seven. The study's authors conclude that ImageNet weights are highly effective in medical anomaly detection, that a fixed ImageNet ResNet18 in the two-stage mode achieves top-three performance on nearly all datasets, and that adapting the features with PANDA or mean-shifted contrastive training did not significantly improve on using them directly. They also conclude that without pretraining, reconstruction methods are more robust than self-supervised ones, and that an autoencoder is a good baseline that should be included in comparisons. Finally, they report that feature-reference methods popular in industrial inspection, knowledge distillation and patch feature modelling, perform worse on medical images than other uses of ImageNet weights.

8.2 A second benchmark that disagrees on the scoring rule

Lagogiannis et al. (2023) benchmarked unsupervised pathology detection on brain MRI, retinal fundus and chest X-ray data and reached a different conclusion about exactly those feature-reference methods. In their study, feature-modelling methods dominate the leaderboards, with a structural feature autoencoder and reverse distillation, both built on frozen pretrained encoders (Meissen et al., 2022; Deng and Li, 2022), producing the outlying performances on brain MRI, fundus and chest X-ray. They also tested whether replacing ImageNet-pretrained encoders with encoders pretrained on the target domain by self-supervision helps. It increased performance for most feature-modelling methods on brain MRI, but reverse distillation declined strongly; on chest X-rays the effect was mostly negative; on the fundus dataset it was consistently negative with one exception. Weight initialisation from domain-specific self-supervision gave no significant benefit for any method but one, which they relate to earlier findings that transfer offers little benefit for medical image classification (Raghu et al., 2019).

The two benchmarks agree that ImageNet-derived features carry most of the best results in medical anomaly detection, and they disagree about which scoring rule to put on top of them. Neither supports the view that in-domain training is uniformly better off the pretraining distribution; Lagogiannis et al. found domain-specific pretraining helping on one modality and hurting on two. Neither supports the view that borrowed features win uniformly; on BraTS2021 in MedIAnomaly the plain autoencoder is ahead of frozen ImageNet features by almost ten points.

8.3 The referent decides it

The medical results line up with referent once intensity is taken into account. Meissen, Kaissis and Rueckert (2021) showed that simply thresholding FLAIR brain MRI intensities produced better anomaly segmentations than several learned detectors for tumours and multiple-sclerosis lesions, because those lesions are hyperintense. MedIAnomaly's intensity baseline reaches 66.8 image AUC on BraTS2021 and 80.4 on the brain-tumour set but is at or below chance on the chest X-ray, fundus, dermoscopy and histopathology sets. Where the anomaly is a departure in intensity, a pixel-space residual sees it; frozen semantic features, trained to be invariant to many such changes, may not. Where the anomaly is a change in tissue structure with little intensity contrast, as in histopathology, the pixel residual fails and the borrowed features do not. Shvetsova et al. (2020) made the bridging move early, measuring an autoencoder's residual with a perceptual distance in pretrained feature space for chest X-rays and lymph-node metastases, and MedIAnomaly's AE-PL row shows the move paying off broadly.

9. Off the pretraining distribution: spectrograms, surfaces and tables

9.1 Radio-telescope spectrograms

Radio astronomy is the domain in which the NLN authors argue most explicitly that natural-image features do not transfer, citing differences in dynamic range and signal-to-noise (Mesarcik et al., 2023). Their two studies support that claim at one granularity and contradict it at another.

For pixel-level detection of radio-frequency interference in real LOFAR data, trained only on regions that a classical flagger marked clean, NLN reaches 0.862 AUROC, 0.622 AUPRC and 0.511 maximum F1 against expert labels; the pretrained nearest-neighbour baseline reaches 0.829 AUROC but only 0.070 AUPRC and 0.153 F1; an L2 autoencoder reaches 0.840 AUROC and 0.393 AUPRC (Mesarcik et al., 2022, Table 3). Supervised segmentation networks trained on the classical flags reach 0.80 to 0.83 AUROC and 0.55 to 0.60 AUPRC, and RFI-Net has the best F1, 0.598. On simulated HERA data NLN leads all three metrics (its Table 2). The pretrained baseline's AUROC is close to the others' while its AUPRC is near the floor, a reminder that AUROC can look adequate under heavy class imbalance when precision is

very poor.

For spectrogram-level classification of system anomalies on LOFAR, ROAD combines a supervised classifier with a self-supervised anomaly detector trained on context prediction and reconstruction (Mesarcik et al., 2023). Here the authors report that a backbone initialised from ImageNet classification and fine-tuned gives a 2 percent decrease in F2 score relative to their in-domain self-supervised pretraining, that a nearest-neighbour lookup in the learned space is less than 3 percent below their classifier head, and that the in-domain pretraining gives less than 5 percent over random initialisation. On one class, the galactic plane, the ImageNet-initialised model scores 0.839 against ROAD's 0.819 (their Table 2). So in the domain chosen to show that natural-image pretraining is ill-suited, the ImageNet-initialised and fine-tuned model is within about 2 percent of the in-domain one at the system level, while the frozen ImageNet nearest-neighbour baseline collapses at the pixel level. Granularity, fine-tuning and metric decide which statement is true.

9.2 3D surfaces

Reiss et al. (2022, Table 4), summarising a study of the MVTec 3D-AD dataset by the same group, report that on depth images ImageNet features reach 63.7 image AUROC, below a handcrafted point-feature-histogram descriptor on the point cloud (75.3) and SIFT on depth (71.4), while on RGB images ImageNet features reach 78.5, and combining RGB features with the handcrafted 3D descriptor reaches 86.5. On a modality the pretraining never saw, a descriptor with the right invariance (rotation, in that study's account) beats the borrowed features; on the RGB channel of the same objects the borrowed features remain useful.

9.3 Tables

Tabular data are the limiting case: there is no pretrained representation to borrow, so the comparison reduces to scoring rules on raw features versus learned representations. Bouman, Bukhsh and Heskes (2024) compared 32 unsupervised detectors on 52 real-world tabular datasets and found that the distance to the k-th nearest neighbour significantly outperforms most other algorithms; clustering the datasets, they found kNN best on those whose anomalies are local and extended isolation forest best on those whose anomalies are global. ADBench (Han et al., 2022) found that none of fourteen unsupervised detectors is statistically better than the others across 57 datasets, and that with synthetic anomaly types the best detector depends on the type, with LOF best for local anomalies and kNN best for global ones. The two large benchmarks thus disagree on which rule suits global anomalies, a disagreement that turns on how "global" is defined, synthetically in one and by clustering real datasets in the other. Livernoche et al. (2023) report that on ADBench a diffusion-time estimator beats every baseline in the semi-supervised setting except kNN, and that kNN remains a top performer; they also note that ADBench's image datasets are represented by pretrained ImageNet embeddings, so the benchmark has already chosen the representation for those tasks. Reiss et al. (2022) report kNN on raw tabular features at 90.2 ROC-AUC, against 89.4 for the best learned representation scored the same way.

9.4 What travels and what does not

Across domains, one thing travels well: a nearest-neighbour distance is rarely far from the best scoring rule once the representation suits the referent, whether that representation is ImageNet patches, DINO embeddings, an autoencoder's latent code, or raw table columns. What does not travel is any fixed answer to which representation suits which referent. That is the part of the debate the in-domain position is right to press and the borrowed-feature position is right to doubt, and the published evidence settles it only domain by domain.

10. The hybrid that dissolves the dichotomy

The strongest recent detectors are not on either side of the dispute. They reconstruct, and they score the reconstruction in a borrowed feature space.

Student-teacher detectors train a student network on normal data to regress the features of a teacher pretrained on natural-image patches and flag regions where the student fails (Bergmann et al., 2020). Reverse distillation inverts the arrangement, decoding a pretrained teacher's multi-scale features from a compact one-class embedding (Deng and Li, 2022). EfficientAD keeps the student-teacher structure with a far lighter extractor (Batzner, Heckler and Konig, 2023). The structural feature autoencoder of Meissen et al. (2022) maps images into multi-channel feature maps and trains an autoencoder there with a structural similarity loss. The perceptual autoencoder measures pixel reconstructions with pretrained features (Shvetsova et al., 2020). UniAD reconstructs pretrained features for many classes with one model and adds query embeddings, neighbour-masked attention and feature jittering to prevent the identity shortcut (You et al., 2022). DDAD conditions a diffusion model's denoising on the target image and compares input and reconstruction both pixel-wise and in a pretrained feature space adapted to the domain; it reports 99.8 image AUROC on MVTec AD and 98.9 on VisA (Mousakhan, Brox and Tayyub, 2023). In its MVTec ablation, adapting the feature extractor adds 8.2 points to the feature-wise comparison, and merging pixel-wise with feature-wise comparison adds 1.2 points to detection. DRAEM takes a different route: it embeds an input together with its reconstructed normal version and learns a normal-versus-anomalous boundary on that pair, so that the comparison between input and reconstruction is itself learned rather than fixed in pixels (Zavrtanik, Kristan and Skocaj, 2021).

These designs make the question "reconstruction or pretrained features?" ill-posed. Reconstruction supplies a per-input reference of what the normal version of this input would look like, which a nearest-neighbour search

supplies only approximately. The borrowed space supplies a distance in which that reference is compared, one that discounts nuisance variation in pixels. NLN's design sits on the same axis. Its neighbours' reconstructions play the role of the per-input reference, and its feature-consistency variant moves part of the comparison into a learned code space, which is where its ablation found most of the gain. The natural next step for NLN is to measure the residual in a stronger space. Whether that space should be borrowed or learned is, on the evidence of Section 8, a question about the referent.

11. Evaluation decides the verdict

11.1 Metric

The radio results show how much the metric can change a verdict: 0.829 AUROC and 0.070 AUPRC for the same detector on the same data. Industrial benchmarks have the converse problem of saturation. The MVTec AD 2 authors write that state-of-the-art models on MVTec AD and VisA now compete within less than a percentage point on segmentation AU-PRO, which prevents meaningful comparison, and that on their new scenarios, with transparent and overlapping objects, unusual lighting, variable normal data and very small defects, state-of-the-art methods stay below 60 percent average AU-PRO (Heckler-Kram et al., 2025). A benchmark at ceiling cannot separate representations; a benchmark below 60 can, but only for its referents. VisA, released as the largest industrial anomaly-detection set of its time, with twelve objects in three domains, widened the referents available (Zou et al., 2022).

The time-series literature shows how far evaluation artefacts can go. Kim et al. (2021) showed that the point-adjustment protocol used by many time-series anomaly-detection papers can make a random score look like the state of the art, and that an untrained model matched published methods even without the adjustment. Wu and Keogh (2020) argued that most exemplars in popular time-series anomaly benchmarks have flaws serious enough to make apparent progress illusory. Audibert et al. (2022) found no family of methods outperforming the others across five multivariate datasets, and Sarfraz et al. (2024) argued that state-of-the-art deep models in that setting effectively learn linear mappings. These results do not concern images, but they bound how much a reported margin can be trusted before the evaluation is audited.

11.2 Baseline implementation

Section 4.2 showed a re-implemented baseline scoring ten points below its originators' figure. Section 5 showed three groups reporting supervised ImageNet features on one-class CIFAR-10 at 80.0, 86.2 and 92.5. Baur et al. (2020) made the general point for brain MRI: before their comparative study, differences in datasets, resolutions and architectures prevented valid comparison of autoencoder methods. The remedy is cheap. A comparison should report the baseline's published number beside its own run, and say which was used.

11.3 Selection without anomaly labels

The deepest evaluation problem is selection. Novelty detection is defined by the absence of labelled anomalies, yet every detector in this paper has choices (layer, backbone, latent size, number of neighbours, mixing weight) that move results by several points. PaDiM on EfficientNet-B5 gains 2.6 points over PaDiM on WideResNet-50 on MVTec AD (Defard et al., 2020, Table V, author's arithmetic), a backbone swap chosen by the authors, which Roth et al. (2021, Table 1) describe as problem-specific backbone selection. Heckler, Konig and Bergmann (2023) found high sensitivity to feature layer. Ding, Zhao and Akoglu (2022) trained more than 35,000 deep outlier-detection models and concluded that hyperparameter selection cannot be avoided; they proposed an ensemble over configurations to sidestep it. Ma et al. (2021) tested seven internal selection strategies, which use only unlabelled data and model outputs, across 39 tasks and 297 candidate models, and found none practically useful: they selected models no better than a state-of-the-art detector with a random configuration. Meta-learning approaches transfer performance information from historical datasets to a new one (Zhao, Rossi and Akoglu, 2021; Zhao, Zhang and Akoglu, 2022), which works to the extent that the new task resembles the old ones. Fung et al. (2023) generate synthetic anomalies from a small set of normal images and report that selection on them often matches selection on a ground-truth validation set, for both image detectors and prompts for CLIP-based detection. Some authors relax the problem openly instead: Shvetsova et al. (2020) proposed letting a handful of anomalies, deliberately narrow in kind, seed the hyperparameter search, and said so as a change to the problem statement. In the two NLN papers re-read in Section 4, no selection split separate from the test anomalies is described; the author has not audited the remaining sources on this point, and the statement is limited to those two.

For the dispute at hand, the selection problem has a sharp consequence. Whether to use borrowed or learned features for a new domain is itself a hyperparameter, and the field has no reliable label-free way to choose it. A detector reported with its best configuration on the test anomalies measures an upper bound that a practitioner without labels cannot reach, and the gap between that bound and what is reachable is not reported. The comparisons in this paper should all be read as upper bounds in that sense.

12. The measured values

Table 1 collects the published numbers the argument rests on, one row per result, with the source on each row. Values are as reported, in the source's metric. Differences and ranges in the last column are the author's arithmetic.

Figure 1 shows the paper's argument in two pictures. In panel (a), the pretrained detectors spread across most of the axis, and the image-level ones alone run from 75.0 to

Table 1. Published novelty-detection results that separate representation, scoring rule, referent and evaluation, with the source on each row.

Source	Data / referent	Representation, rule	Reported value	Reading
Mesarcik 2021, T4	MVTec AD, defects	AE latent + neighbour recon. (NLN)	0.783 AUROC	+3.3 over re-impl. DKNN
Mesarcik 2021, T4	MVTec AD	ImageNet kNN, re-implemented	0.750	baseline as re-run
Cohen 2020, T1	MVTec AD	ImageNet kNN = DN2 stage	85.5	originators' figure
Roth 2021, T1	MVTec AD	ImageNet patches, coreset kNN	99.1	family range 24.1
Mesarcik 2021, T3-4	CIFAR-10, classes	NLN vs ImageNet kNN	0.560 vs 0.714; 0.658 vs 0.890	semantic referent
Mesarcik 2021, T6	MNIST multi-class	recon / nbr recon / code / full	.778/.822/.913/.921	code space gives most
Sohn 2020, T2	CIFAR-10 one-class	ImageNet ResNet-50, OC-SVM	80.0	in-domain DA: 92.5
Xiao 2021, T3	CIFAR-10 one-class	supervised ImageNet	86.2	SSL ImageNet: 93.8
Bergman 2020, T1	CIFAR-10 one-class	ImageNet ResNet, kNN	92.5	spread 12.5 with rows above
Reiss 2022, T2	CIFAR-10 one-class	MAE vs DINO, kNN	78.1 vs 97.1	representation lever
Sohn 2020, T2	5 datasets, mean	RotNet own head vs KDE	83.1 vs 86.6	rule lever
Reiss 2022, T5	ODDS tabular	GOAD own score vs kNN; raw kNN	78.2 vs 87.6; 90.2	rule lever; raw best
Reiss 2022, T3	MS-COCO vehicles	PANDA, image vs object crops	61.5 vs 77.0	referent lever
Cai 2024, T6	BraTS2021 MRI	AE L2 / ImageNet R18 / AE-PL	82.6 / 72.8 / 85.1	AE beats frozen IN
Cai 2024, T6	Camelyon16	AE L2 / ImageNet R18 / AE-PL	36.8 / 73.2 / 76.1	AE below chance
Mesarcik 2022, T3	LOFAR RFI, pixels	NLN vs ImageNet kNN	AUPRC 0.622 vs 0.070	AUROC .862 vs .829
Mesarcik 2023	LOFAR, system	in-domain SSL vs ImageNet FT	F2 gap 2%	near parity
Reiss 2022, T4	MVTec 3D, depth / point cloud	ImageNet (depth) vs handcrafted 3D (point cloud)	63.7 vs 75.3 I-AUROC	borrowed loses
Mousakhan 2023	MVTec AD	diffusion recon., adapted features	99.8	hybrid; adaptation +8.2

95.8, a range that granularity cannot explain and that is several times the step between NLN and its re-implemented pretrained baseline. In panel (b), no single family is on top across datasets; the plain autoencoder (circles) moves from ahead of the frozen features on BraTS2021 to below chance on Camelyon16, and the hybrid (triangles) is the most consistently high of the three.

13. Crediting a gain to a representation's origin

The re-reading suggests a procedure. A claim that detector A beats detector B because A's representation is borrowed (or learned in-domain) should survive the checks that Algorithm 1 states as a decision structure.

```
Algorithm 1. Crediting a gain to representation
origin.
Input: detectors A, B; benchmarks with referents
  N_1..N_m; metrics M; candidate configs C.
Step 1. Write each detector as (R, S): space and
  scoring rule. List which of R, S differ.
Step 2. If both differ, add the crossed pair
  (R_A, S_B) and (R_B, S_A).
Step 3. Run each baseline from its authors' code;
  print its published number beside the run.
Step 4. Choose configs on data disjoint from the
  test novelties (held-out classes or synthetic
  anomalies); say which, per detector.
Step 5. Report AUROC and AUPRC (or PRO) per class,
  with spread over >= 3 seeds.
Step 6. W = range over S-variants and layers for
  R_A; G = score(A) - score(B).
  If |G| < W: credit S or the implementation,
  not R's origin.
Step 7. Repeat on >= 1 referent off the pretraining
  distribution and >= 1 semantic referent.
Step 8. If sign(G) differs across referents: scope
  the claim to the referent class.
Output: the narrowest claim that survives.
```

Step 1 forces the comparison to name what changed. Step 2 is the test almost no published comparison runs: it would show, for NLN against DN2, whether nearest neighbours in an ImageNet space with neighbour reconstructions decoded by a domain autoencoder beat either parent. Step 3 is the check the NLN MVTec comparison fails, with a re-implemented baseline ten points below its originators' figure. Step 4 addresses Section 11.3; it does not require solving label-free selection, only saying how selection was done, so that upper bounds are not read as reachable results. Step 5 is the check the radio results show to be necessary. Step 6 is the within-family-spread diagnostic of Section 5: if the choice among scoring rules and layers for one representation moves results more than the gap between the two detectors, the gap cannot be credited to the representation's origin. The MVTec comparison fails it by a factor of seven. Steps 7 and 8 are the referent checks of Sections 6.3, 8 and 9.

Applied to the evidence base, the procedure credits the following. In-domain pixel or latent representations, scored by distance to stored neighbours, are competitive or ahead where the referent is a visible departure in appearance from a homogeneous normal class (MNIST, interference in spectrograms, hyperintense brain lesions). Borrowed semantic representations are far ahead where the referent is a held-out object category or a structural change with little intensity contrast (CIFAR-10,

Camelyon16), and they win most medical and industrial comparisons as reported. Within the borrowed family, implementation, granularity, density model, layer and backbone, changed jointly, move results more than the corpus. Reconstruction scored in a borrowed space gives the best results in Lagogiannis et al. and, in MedIAnomaly, beats the frozen ImageNet features on five of seven datasets in the author's three-row comparison. No published comparison passes all eight steps.

14. The strongest case against this paper

The pretrained-feature position is simply right in practice. A practitioner reading MedIAnomaly and Lagogiannis et al. might conclude that ImageNet-derived features carry the best result on most datasets in both studies, that the exceptions are few, and that the four-factor decomposition is a scholastic refinement of a clear answer. The reply is partial. The pretrained position wins most reported comparisons, and this paper does not dispute it. But in Lagogiannis et al. the best results come from feature reconstruction and distillation on pretrained encoders, and in MedIAnomaly an autoencoder with an ImageNet perceptual residual beats the frozen ResNet18 on five of seven datasets, though MedIAnomaly's authors also credit the frozen ResNet18 with "top-3 performance on nearly all datasets" (Cai et al., 2024); and the two benchmarks disagree about which scoring rule to put on top. A practitioner who took "use a frozen ImageNet backbone with nearest neighbours" from the 2020 results would have been behind the perceptual autoencoder on five of the seven MedIAnomaly datasets in the author's three-row comparison.

The within-family spread reflects progress over time, not noise. PatchCore came after SPADE; the spread on MVTec AD is a record of improvement, so treating it as variance is unfair. That is correct as history and does not affect the argument. The question is what a single comparison between one pretrained detector and one in-domain detector can establish about pretrained features in general. Because the pretrained family improved by about 14 points through changes in granularity, layer and density model, a comparison against any one member says little about the family, whichever member was chosen.

NLN never claimed to beat pretrained detectors. The NLN paper limited its comparison to autoencoders and said why. This paper's re-reading uses the DKNN rows the NLN paper chose to include and the abstract's claim of independence from pretrained networks. The point is not that NLN overclaimed against PatchCore, which it did not compare with, but that the evidence it offers for competitiveness without pretraining rests on benchmarks whose novelty is pixel-visible and on one re-implemented baseline. The authors' later ROAD result, an ImageNet-initialised model within about 2 percent (F2) of in-domain pretraining on spectrograms, is the strongest evidence that the independence claim needs scoping, and it is the authors' own.

The referent explanation is unfalsifiable as stated. If every reversal can be attributed to the referent, the explanation explains nothing. The reply is that it makes specific predictions. Pixel-space residuals should beat frozen semantic features when anomalies differ in intensity and normal data are homogeneous, and lose when anomalies are structural and normal data vary in appearance; adding a perceptual distance to an autoencoder should help most where pixel residuals fail. The MedIAnomaly rows are consistent with both predictions, including the intensity baseline's pattern, but they were not designed as a test, and Section 16 proposes one.

15. What is not known

No published study runs the crossed comparison of Step 2 for NLN: the same neighbour-reconstruction score with latent neighbours found in a borrowed space, or the DN2 score in the NLN autoencoder's latent space. Without it, the contributions of the representation and of the neighbour reconstruction to NLN's results cannot be separated.

No published study measures how often the representation choice that wins with test labels is the one a practitioner would pick without them. Ma et al. (2021) show that internal criteria fail for classical detectors; whether synthetic-anomaly selection (Fung et al., 2023) picks correctly between borrowed and learned representations in domains far from natural images has not been tested.

The overlap between ImageNet's categories and the benchmarks' held-out classes has been named as an explanation (Sohn et al., 2020) but not measured as a variable. No located study varies the overlap while holding the dataset's difficulty fixed.

The two medical benchmarks disagree about whether feature-reference methods suit medical images, and the cause of the disagreement, whether datasets, preprocessing, metric or implementation, is not known.

Whether vision-language or large self-supervised features change the picture off the pretraining distribution is open. Zero-shot CLIP-based detection reaches 91.8 image AUROC on MVTec AD (Jeong et al., 2023), and about 256 random natural images used as negatives give state-of-the-art results on an ImageNet-based benchmark (Liznerski et al., 2022), but these results are on natural-image or industrial referents. One 2025 study argues that ImageNet-pretrained features are suboptimal for industrial anomaly detection because of both objective mismatch and distribution shift (Yao et al., 2025); another, on supervised defect detection, reports a domain gap for ImageNet pretraining (Liu, C. et al., 2025) and also reports that naive in-domain self-supervised pretraining is often ineffective for subtle defects, which leaves the question exactly where this paper finds it.

The scoring-rule reconciliation of Section 6.2, that the rule matters most when the representation is weak, is a hypothesis drawn from two papers with different setups. It has not been tested.

16. Studies that would settle it

A crossed design on three referents. Take two representations (an in-domain autoencoder latent and a frozen borrowed space) and three scoring rules (pixel residual of neighbour reconstructions, latent kNN distance, perceptual residual), and evaluate all six combinations on a semantic referent (CIFAR-10 one-class), a sensory referent (MVTec AD or a brain-MRI set with hyperintense lesions), and an off-distribution referent (LOFAR interference). The prediction from Section 14 is an interaction: the in-domain latent with pixel residual should lead on the sensory and spectral referents and trail on the semantic one, and the perceptual residual should close most of the gap everywhere.

An overlap dose-response. Construct one-class tasks from a dataset whose classes can be matched to ImageNet classes with varying semantic distance, and measure the advantage of borrowed over in-domain features as a function of overlap. A flat curve would refute the overlap explanation; a steep one would say how much of the borrowed-feature advantage on CIFAR-style benchmarks is category knowledge rather than a detection principle.

Selection-aware reporting. Re-run the MedIAnomaly comparison with every configuration selected on synthetic anomalies (Fung et al., 2023) instead of on test labels, and report both. The difference between the two tables is the price of not having labels, and its sign across detector families would say whether the borrowed-feature advantage survives realistic selection.

Metric audit for imbalanced referents. Report AUPRC beside AUROC for every pixel-level comparison in radio and medical segmentation, as the LOFAR study did. The prediction is that frozen borrowed features lose more under AUPRC than under AUROC wherever anomalous pixels are rare.

None of these requires new datasets. All of them require running more combinations than a paper proposing a new method usually reports.

17. Conclusion

The question whether novelty detectors need pretrained features has been argued with comparisons that change the representation, the scoring rule, the referent and the evaluation at once. Once those are separated, the published record says three things. A nearest-neighbour distance is a strong scoring rule in almost every domain, provided the space it is computed in suits the kind of novelty the task cares about. Which space suits which novelty is decided by the referent, not by where the features came from: borrowed semantic features win where novelty is categorical or structural and normal data vary in appearance, in-domain pixel or latent spaces hold their own where novelty is a visible departure from homogeneous normal data, and reconstruction scored in a borrowed space leads in one medical benchmark and beats frozen borrowed features on most datasets of another. And the margins on which either position rests are smaller than the spread produced by implementation and selection choices that a practitioner without anomaly labels cannot make reliably. NLN's neighbour reconstruction is a sound remedy for an autoencoder's tendency to reproduce what it should flag. Its independence from pretrained networks is an engineering advantage and a fit to pixel-visible novelty, not evidence that borrowed features are unnecessary. The pretrained nearest-neighbour detector's success is real and large on the benchmarks the field reports, and is partly a fit between a semantic space and semantic referents. A comparison that ran Algorithm 1 would say which.

References

Ahmed, F., & Courville, A. (2019). Detecting semantic anomalies. AAAI 2020. arXiv:1908.04388. doi:10.48550/arXiv.1908.04388.

Audibert, J., Michiardi, P., Guyard, F., Marti, S., & Zuluaga, M. A. (2022). Do Deep Neural Networks Contribute to Multivariate Time Series Anomaly Detection? Pattern Recognition 132. arXiv:2204.01637. doi:10.48550/arXiv.2204.01637.

Batzner, K., Heckler, L., & Konig, R. (2023). EfficientAD: Accurate Visual Anomaly Detection at Millisecond-Level Latencies. WACV 2024. arXiv:2303.14535. doi:10.48550/arXiv.2303.14535.

Baur, C., Denner, S., Wiestler, B., Albarqouni, S., & Navab, N. (2020). Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study. arXiv:2004.03271. doi:10.48550/arXiv.2004.03271.

Bercea, C. I., Rueckert, D., & Schnabel, J. A. (2022). What do we learn? Debunking the Myth of Unsupervised Outlier Detection. arXiv:2206.03698. doi:10.48550/arXiv.2206.03698.

Bergman, L., Cohen, N., & Hoshen, Y. (2020). Deep Nearest Neighbor Anomaly Detection. arXiv:2002.10445. doi:10.48550/arXiv.2002.10445.

Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTec AD - A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. CVPR 2019. doi:10.1109/CVPR.2019.00982.

Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2020). Uninformed Students: Student-Teacher Anomaly Detection with Discriminative Latent Embeddings. CVPR 2020. arXiv:1911.02357. doi:10.48550/arXiv.1911.02357.

Bouman, R., Bukhsh, Z., & Heskes, T. (2024). Unsupervised anomaly detection algorithms on real-world data: how many do we need? Journal of Machine Learning Research 25(105). arXiv:2305.00735. doi:10.48550/arXiv.2305.00735.

Cai, Y., Zhang, W., Chen, H., & Cheng, K.-T. (2024). MedIAnomaly: A comparative study of anomaly detection in medical images. Medical Image Analysis (2025). arXiv:2404.04518. doi:10.48550/arXiv.2404.04518.

Cohen, N., & Hoshen, Y. (2020). Sub-Image Anomaly Detection with Deep Pyramid Correspondences. arXiv:2005.02357. doi:10.48550/arXiv.2005.02357.

Cohen, N., Abutbul, R., & Hoshen, Y. (2021). Out-of-Distribution Detection Without Class Labels. ECCV Workshops 2022. arXiv:2112.07662. doi:10.48550/arXiv.2112.07662.

Cohen, N., Kahana, J., & Hoshen, Y. (2022). Red PANDA: Disambiguating Anomaly Detection by Removing Nuisance Factors. arXiv:2207.03478. doi:10.48550/arXiv.2207.03478.

Damm, S., Laszkiewicz, M., Lederer, J., & Fischer, A. (2024). AnomalyDINO: Boosting Patch-based Few-shot Anomaly Detection with DINOv2. WACV 2025. arXiv:2405.14529. doi:10.48550/arXiv.2405.14529.

Defard, T., Setkov, A., Loesch, A., & Audigier, R. (2020). PaDiM: a Patch Distribution Modeling Framework for Anomaly Detection and Localization. ICPR Workshops 2020. arXiv:2011.08785. doi:10.48550/arXiv.2011.08785.

Deng, H., & Li, X. (2022). Anomaly Detection via Reverse Distillation from One-Class Embedding. CVPR 2022. arXiv:2201.10703. doi:10.48550/arXiv.2201.10703.

Denouden, T., Salay, R., Czarnecki, K., Abdelzad, V., Phan, B., & Vernekar, S. (2018). Improving Reconstruction Autoencoder Out-of-distribution Detection with Mahalanobis Distance. arXiv:1812.02765. doi:10.48550/arXiv.1812.02765.

Ding, X., Zhao, L., & Akoglu, L. (2022). Hyperparameter Sensitivity in Deep Outlier Detection: Analysis and a Scalable Hyper-Ensemble Solution. arXiv:2206.07647. doi:10.48550/arXiv.2206.07647.

Fort, S., Ren, J., & Lakshminarayanan, B. (2021). Exploring the Limits of Out-of-Distribution Detection. arXiv:2106.03004. doi:10.48550/arXiv.2106.03004.

Fung, C., Qiu, C., Li, A., & Rudolph, M. (2023). Model Selection of Anomaly Detectors in the Absence of Labeled Validation Data. arXiv:2310.10461. doi:10.48550/arXiv.2310.10461.

Golan, I., & El-Yaniv, R. (2018). Deep Anomaly Detection Using Geometric Transformations. arXiv:1805.10917. doi:10.48550/arXiv.1805.10917.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., & van den Hengel, A. (2019). Memorizing Normality to Detect Anomaly: Memory-augmented Deep Autoencoder for Unsupervised Anomaly Detection. ICCV 2019. arXiv:1904.02639. doi:10.48550/arXiv.1904.02639.

Graham, M. S., Pinaya, W. H. L., Tudosiu, P.-D., Nachev, P., Ourselin, S., & Cardoso, M. J. (2022). Denoising diffusion models for out-of-distribution detection. arXiv:2211.07740. doi:10.48550/arXiv.2211.07740.

Gu, X., Akoglu, L., & Rinaldo, A. (2019). Statistical Analysis of Nearest Neighbor Methods for Anomaly Detection. arXiv:1907.03813. doi:10.48550/arXiv.1907.03813.

Han, S., Hu, X., Huang, H., Jiang, M., & Zhao, Y. (2022). ADBench: Anomaly Detection Benchmark. NeurIPS 2022. arXiv:2206.09426. doi:10.48550/arXiv.2206.09426.

Havtorn, J. D., Frellsen, J., Hauberg, S., & Maaloe, L. (2021). Hierarchical VAEs Know What They Don't Know. ICML 2021. arXiv:2102.08248. doi:10.48550/arXiv.2102.08248.

Heckler, L., Konig, R., & Bergmann, P. (2023). Exploring the Importance of Pretrained Feature Extractors for Unsupervised Anomaly Detection and Localization. CVPR Workshops 2023. doi:10.1109/CVPRW59228.2023.00293.

Heckler-Kram, L., Neudeck, J.-H., Scheler, U., Konig, R., & Steger, C. (2025). The MVTec AD 2 Dataset: Advanced Scenarios for Unsupervised Anomaly Detection. arXiv:2503.21622. doi:10.48550/arXiv.2503.21622.

Hendrycks, D., Lee, K., & Mazeika, M. (2019a). Using Pre-Training Can Improve Model Robustness and Uncertainty. ICML 2019. arXiv:1901.09960. doi:10.48550/arXiv.1901.09960.

Hendrycks, D., Mazeika, M., Kadavath, S., & Song, D. (2019b). Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. NeurIPS 2019. arXiv:1906.12340. doi:10.48550/arXiv.1906.12340.

Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., & Dabeer, O. (2023). WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation. CVPR 2023. arXiv:2303.14814. doi:10.48550/arXiv.2303.14814.

Kim, S., Choi, K., Choi, H.-S., Lee, B., & Yoon, S. (2021). Towards a Rigorous Evaluation of Time-series Anomaly Detection. arXiv:2109.05257. doi:10.48550/arXiv.2109.05257.

Kirichenko, P., Izmailov, P., & Wilson, A. G. (2020). Why Normalizing Flows Fail to Detect Out-of-Distribution Data. arXiv:2006.08545. doi:10.48550/arXiv.2006.08545.

Lagogiannis, I., Meissen, F., Kaissis, G., & Rueckert, D. (2023). Unsupervised Pathology Detection: A Deep Dive Into the State of the Art. IEEE Transactions on Medical Imaging. arXiv:2303.00609. doi:10.48550/arXiv.2303.00609.

Liu, C., Li, H., Du, J., Hou, Y., Sun, Q., Jin, L., & Xu, K. (2025). Advancing Metallic Surface Defect Detection via Anomaly-Guided Pretraining on a Large Industrial Dataset. Pattern Recognition 179 (2026). arXiv:2509.18919. doi:10.48550/arXiv.2509.18919.

Liu, Z., Zhou, J. P., Wang, Y., & Weinberger, K. Q. (2023). Unsupervised Out-of-Distribution Detection with Diffusion Inpainting. ICML 2023. arXiv:2302.10326. doi:10.48550/arXiv.2302.10326.

Livernoche, V., Jain, V., Hezaveh, Y., & Ravanbakhsh, S. (2023). On Diffusion Modeling for Anomaly Detection. ICLR 2024. arXiv:2305.18593. doi:10.48550/arXiv.2305.18593.

Liznerski, P., Ruff, L., Vandermeulen, R. A., Franks, B. J., Muller, K.-R., & Kloft, M. (2022). Exposing Outlier Exposure: What Can Be Learned From Few, One, and Zero Outlier Images. Transactions on Machine Learning Research. arXiv:2205.11474. doi:10.48550/arXiv.2205.11474.

Ma, M. Q., Zhao, Y., Zhang, X., & Akoglu, L. (2021). A Large-scale Study on Unsupervised Outlier Model Selection: Do Internal Strategies Suffice? arXiv:2104.01422. doi:10.48550/arXiv.2104.01422.

Meissen, F., Kaissis, G., & Rueckert, D. (2021). Challenging Current Semi-Supervised Anomaly Segmentation Methods for Brain MRI. MICCAI BrainLes Workshop 2021. arXiv:2109.06023. doi:10.48550/arXiv.2109.06023.

Meissen, F., Paetzold, J., Kaissis, G., & Rueckert, D. (2022). Unsupervised Anomaly Localization with Structural Feature-Autoencoders. MICCAI BrainLes Workshop. arXiv:2208.10992. doi:10.48550/arXiv.2208.10992.

Mesarcik, M., Ranguelova, E., Boonstra, A.-J., & van Nieuwpoort, R. V. (2021). Improving Novelty Detection using the Reconstructions of Nearest Neighbours. arXiv:2111.06150. doi:10.48550/arXiv.2111.06150.

Mesarcik, M., Boonstra, A.-J., Ranguelova, E., & van Nieuwpoort, R. V. (2022). Learning to detect RFI in radio astronomy without seeing it. arXiv:2207.00351. doi:10.48550/arXiv.2207.00351.

Mesarcik, M., Boonstra, A.-J., Iacobelli, M., Ranguelova, E., de Laat, C., & van Nieuwpoort, R. (2023). The ROAD to discovery: machine learning-driven anomaly detection in radio astronomy spectrograms. Astronomy and Astrophysics 680, A74. arXiv:2307.01054. doi:10.48550/arXiv.2307.01054.

Mousakhan, A., Brox, T., & Tayyub, J. (2023). Anomaly Detection with Conditioned Denoising Diffusion Models. GCPR 2024. arXiv:2305.15956. doi:10.48550/arXiv.2305.15956.

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., & Lakshminarayanan, B. (2018). Do Deep Generative Models Know What They Don't Know? ICLR 2019. arXiv:1810.09136. doi:10.48550/arXiv.1810.09136.

Perera, P., Nallapati, R., & Xiang, B. (2019). OCGAN: One-class Novelty Detection Using GANs with Constrained Latent Representations. CVPR 2019. arXiv:1903.08550. doi:10.48550/arXiv.1903.08550.

Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding Transfer Learning for Medical Imaging. NeurIPS 2019. arXiv:1902.07208. doi:10.48550/arXiv.1902.07208.

Reiss, T., Cohen, N., Bergman, L., & Hoshen, Y. (2020). PANDA: Adapting Pretrained Features for Anomaly Detection and Segmentation. CVPR 2021. arXiv:2010.05903. doi:10.48550/arXiv.2010.05903.

Reiss, T., & Hoshen, Y. (2021). Mean-Shifted Contrastive Loss for Anomaly Detection. AAAI 2023. arXiv:2106.03844. doi:10.48550/arXiv.2106.03844.

Reiss, T., Cohen, N., Horwitz, E., Abutbul, R., & Hoshen, Y. (2022). Anomaly Detection Requires Better Representations. ECCV Workshops 2022. arXiv:2210.10773. doi:10.48550/arXiv.2210.10773.

Rippel, O., Mertens, P., & Merhof, D. (2020). Modeling the Distribution of Normal Data in Pre-Trained Deep Features for Anomaly Detection. ICPR 2020. arXiv:2005.14140. doi:10.48550/arXiv.2005.14140.

Roth, K., Pemula, L., Zepeda, J., Scholkopf, B., Brox, T., & Gehler, P. (2021). Towards Total Recall in Industrial Anomaly Detection. CVPR 2022. arXiv:2106.08265. doi:10.48550/arXiv.2106.08265.

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., & Muller, K.-R. (2020). A Unifying Review of Deep and Shallow Anomaly Detection. Proceedings of the IEEE (2021). arXiv:2009.11732. doi:10.48550/arXiv.2009.11732.

Sarfraz, M. S., Chen, M.-Y., Layer, L., Peng, K., & Koulakis, M. (2024). Position: Quo Vadis, Unsupervised Time Series Anomaly Detection? ICML 2024. arXiv:2405.02678. doi:10.48550/arXiv.2405.02678.

Schirrmeister, R. T., Zhou, Y., Ball, T., & Zhang, D. (2020). Understanding Anomaly Detection with Deep Invertible Networks through Hierarchies of Distributions and Features. NeurIPS 2020. arXiv:2006.10848. doi:10.48550/arXiv.2006.10848.

Serra, J., Alvarez, D., Gomez, V., Slizovskaia, O., Nunez, J. F., & Luque, J. (2019). Input complexity and out-of-distribution detection with likelihood-based generative models. ICLR 2020. arXiv:1909.11480. doi:10.48550/arXiv.1909.11480.

Shvetsova, N., Bakker, B., Fedulova, I., Schulz, H., & Dylov, D. V. (2020). Anomaly Detection in Medical Imaging with Deep Perceptual Autoencoders. IEEE Access 9. arXiv:2006.13265. doi:10.48550/arXiv.2006.13265.

Sohn, K., Li, C.-L., Yoon, J., Jin, M., & Pfister, T. (2020). Learning and Evaluating Representations for Deep One-class Classification. ICLR 2021. arXiv:2011.02578. doi:10.48550/arXiv.2011.02578.

Sun, Y., Ming, Y., Zhu, X., & Li, Y. (2022). Out-of-Distribution Detection with Deep Nearest Neighbors. ICML 2022. arXiv:2204.06507. doi:10.48550/arXiv.2204.06507.

Tack, J., Mo, S., Jeong, J., & Shin, J. (2020). CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances. NeurIPS 2020. arXiv:2007.08176. doi:10.48550/arXiv.2007.08176.

Wu, R., & Keogh, E. J. (2020). Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress. arXiv:2009.13807. doi:10.48550/arXiv.2009.13807.

Xiao, Z., Yan, Q., & Amit, Y. (2021). Do We Really Need to Learn Representations from In-domain Data for Outlier Detection? arXiv:2105.09270. doi:10.48550/arXiv.2105.09270.

Yang, J., Zhou, K., Li, Y., & Liu, Z. (2021). Generalized Out-of-Distribution Detection: A Survey. arXiv:2110.11334. doi:10.48550/arXiv.2110.11334.

Yao, X., Luo, Y., Qian, Z., & Zhang, C. (2025). ADPretrain: Advancing Industrial Anomaly Detection via Anomaly Representation Pretraining. NeurIPS 2025. arXiv:2511.05245. doi:10.48550/arXiv.2511.05245.

You, Z., Cui, L., Shen, Y., Yang, K., Lu, X., Zheng, Y., & Le, X. (2022). A Unified Model for Multi-class Anomaly Detection. NeurIPS 2022. arXiv:2206.03687. doi:10.48550/arXiv.2206.03687.

Zavrtanik, V., Kristan, M., & Skocaj, D. (2021). DRAEM - A discriminatively trained reconstruction embedding for surface anomaly detection. ICCV 2021. arXiv:2108.07610. doi:10.48550/arXiv.2108.07610.

Zhao, Y., Rossi, R. A., & Akoglu, L. (2021). Automating Outlier Detection via Meta-Learning. arXiv:2009.10606. doi:10.48550/arXiv.2009.10606.

Zhao, Y., Zhang, S., & Akoglu, L. (2022). Toward Unsupervised Outlier Model Selection. ICDM 2022. arXiv:2211.01834. doi:10.48550/arXiv.2211.01834.

Zhou, Y. (2022). Rethinking Reconstruction Autoencoder-Based Out-of-Distribution Detection. CVPR 2022. arXiv:2203.02194. doi:10.48550/arXiv.2203.02194.

Zou, Y., Jeong, J., Pemula, L., Zhang, D., & Dabeer, O. (2022). SPot-the-Difference Self-Supervised Pre-training for Anomaly Detection and Segmentation. ECCV 2022. arXiv:2207.14315. doi:10.48550/arXiv.2207.14315.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.24408406