Explainable Deep Learning for Visual Anomaly Detection in Industrial Monitoring Systems: A Cross-Dataset Evaluation Using MVTec AD, VisA, and BTAD

Alimuddin Melleng[1,*] **Rendra Soekarta**[1] **Muhammad Yusuf**[1]

[1] Faculty of Engineering, Department of Informatics Engineering, University of Muhammadiyah Sorong, Indonesia

* Correspondence: alimuddinmllg@gmail.com

Abstract

Visual quality inspection in manufacturing environments requires detection methods that are label-efficient, spatially discriminative, and capable of producing outputs that support human decision-making at the point of inspection. Supervised defect classifiers address discriminative accuracy but depend on extensive annotated fault catalogues that are seldom available under operational conditions. Unsupervised one-class learning methods circumvent the annotation constraint but typically yield scalar image-level anomaly scores that convey no spatial information to inspection operators. The present study evaluates Patch Distribution Modeling (PaDiM) as an integrated framework addressing both requirements: a training-label-free method whose Mahalanobis distance scoring mechanism inherently produces pixel-level anomaly maps that can be rendered directly as colour overlay visualisations on inspection images.

Controlled experiments were conducted on three publicly available industrial benchmarks representing a spectrum of acquisition difficulty: MVTec AD, VisA, and BTAD. On MVTec AD, a ResNet-18 backbone achieved a mean image-level AUROC of 0.8688 and pixel-level AUROC of 0.9730 across three evaluated product categories. Substituting a Wide-ResNet-50-2 backbone with 256 randomly projected features elevated image-level AUROC on the bottle category to 1.0000. On VisA, mean pixel-level AUROC attained 0.9829, the highest value recorded across all three datasets, indicating reliable sub-region defect localisation under elevated intra-class appearance variability. On BTAD, whose images were collected on an operational production line, mean image-level AUROC reached 0.9350, with two of three product categories exceeding 0.98 and one achieving 1.00; pixel-level AUROC remained above 0.95 across all categories.

These results indicate that PaDiM generalises consistently across datasets of markedly different character, and that the heatmap overlays produced by the pipeline provide spatially accurate, operationally interpretable defect indicators without requiring fault labels at any stage.

Keywords: patch distribution modelling; one-class learning; pixel-level localisation; heatmap overlay; Mahalanobis scoring

1. Introduction

Automated visual inspection constitutes a critical quality assurance function in discrete and process manufacturing. The deployment of machine-learning-based inspection systems has historically been constrained by a persistent data imbalance inherent to well-controlled production environments: defective items are deliberately rare, rendering the collection and annotation of sufficient fault samples for supervised training both costly and logistically demanding [1, 2]. Anomaly detection addresses this constraint by formulating the inspection task as a one-class classification problem, wherein a model trained exclusively on defect-free images assigns anomaly scores to unseen test samples according to their deviation from the learned normal distribution [2, 3]. This formulation is well-suited to industrial practice: normal images are abundantly available, an exhaustive defect taxonomy is not required, and the system can identify fault patterns that were not present during training.

Despite these advantages, the deployment of unsupervised anomaly detectors in regulated manufacturing contexts has been impeded by a further challenge: interpretability. Many anomaly detection systems are ultimately deployed using scalar image-level decision scores; however, industrial inspection also requires spatial evidence indicating where the anomaly occurs. In operational settings where a quality engineer must make a downstream disposition decision — accept, reject, or direct to manual re-inspection — an image-level score alone is insufficient. Pixel-level anomaly localisation, in which a dense spatial score map indicates the distribution of anomaly evidence across the image, is therefore a functional prerequisite for practical deployment rather than an optional analytical output.

Patch Distribution Modeling (PaDiM), introduced by Defard et al. [4], is a feature-distribution method that produces pixel-level anomaly maps as a structural consequence of its per-location Gaussian modelling architecture, rather than through post-hoc attribution approximation. The method fits a multivariate Gaussian to multi-scale convolutional features extracted at each spatial location of the feature grid during training, and assigns anomaly scores via the Mahalanobis distance at inference. The resulting score map, rendered as a colour overlay on the original image, constitutes an intrinsic and spatially precise representation of anomaly evidence that is immediately legible to non-specialist inspection personnel.

The present study implements PaDiM within a complete inspection pipeline, evaluates it across three industrial anomaly detection benchmarks of increasing practical complexity, and examines the sensitivity of detection performance to backbone architecture selection. Four principal contributions are made: (i) a fully label-free anomaly detection and localisation pipeline in which spatial interpretability is structurally intrinsic to the anomaly-scoring mechanism; (ii) a controlled backbone sensitivity analysis comparing ResNet-18 and Wide-ResNet-50-2 on the MVTec AD bottle category under identical experimental conditions; (iii) a cross-dataset evaluation on representative categories from MVTec AD [2], VisA [5], and BTAD[6], with all reported metrics derived exclusively from the authors' own experimental runs; and (iv) an indicative comparison against six published baseline methods, with appropriate qualification of evaluation scope differences.

2. Related Work

2.1 Benchmark Datasets for Industrial Anomaly Detection

The MVTec Anomaly Detection dataset (MVTec AD) [2], introduced by Bergmann et al. in 2019, has established itself as the primary evaluation standard in the field. Comprising 5,354 high-resolution images across 15 industrial categories with pixel-precise ground-truth defect annotations spanning 73 defect types, it provides a rigorous and reproducible basis for method comparison. The Visual Anomaly (VisA) dataset [5] extends evaluation to 12 product categories characterised by substantially higher intra-class appearance variability, rendering image-level discrimination more challenging than in MVTec AD. The BTech Anomaly Detection dataset (BTAD) [6] departs from laboratory-controlled image acquisition and supplies images captured directly on an operational industrial production line, incorporating illumination variation, part placement tolerances, and surface finish inconsistencies representative of real manufacturing environments. Collectively, these three benchmarks provide a representative spectrum of industrial inspection difficulty, from controlled laboratory conditions through to real-world production-line imaging.

2.2 Reconstruction-Based Anomaly Detection

Reconstruction-based anomaly detection operates on the principle that a generative model trained on normal images will reconstruct anomaly-free inputs with low error while failing to accurately reconstruct anomalous regions, enabling defect identification from reconstruction discrepancy [7-9]. AnoGAN [7] is an early representative approach that detects anomalies by learning the manifold of normal data through adversarial training and subsequently mapping test images into the GAN latent space to evaluate

reconstruction and discrimination errors. GANomaly [8] improves computational tractability through an encoder-decoder-encoder architecture that jointly learns image and latent representations without requiring iterative latent-space optimisation at inference time. DRAEM [9] advances this paradigm by coupling an anomaly-free reconstruction sub-network with a discriminative sub-network trained on synthetically generated anomalous images, enabling direct pixel-level localisation without access to real fault samples during training.

More recent work has extended reconstruction-based detection to discrete latent-variable representations. Yapp and Doan [10] demonstrated that VQ-VAE-2, which models the distribution of normal images in a discrete latent space, provides effective anomaly detection through deviations in latent-code likelihood on MVTec AD. A recognised limitation of reconstruction-based methods is their reduced discriminative capacity when defects are subtle, highly localised, or structurally similar to normal texture, as the generative model may successfully reconstruct defective regions that lie within the support of the learned normal distribution [9, 11].

2.3 Feature Distribution-Based Methods

Feature distribution-based methods model the statistical properties of the normal class in a deep feature embedding space rather than through image reconstruction. Convolutional features extracted from a pretrained network on normal training images characterise the normal distribution, and test-sample anomaly scores are derived from the deviation of test features from this distribution. Because these approaches operate on semantic feature representations rather than raw pixels, they generally exhibit stronger sensitivity to subtle local defects that produce minimal reconstruction error [3, 4, 11].

SPADE [12] represents an early method in this family, storing multi-scale deep features from normal training images in a retrieval structure and performing nearest-neighbour matching at inference to produce both image-level and pixel-level anomaly scores. PaDiM [4] improves representational compactness by modelling the feature distribution at each spatial location parametrically through a multivariate Gaussian, enabling anomaly scoring via the Mahalanobis distance and yielding smooth, well-calibrated spatial anomaly maps. PatchCore [3] further advances this family through greedy coreset sub-sampling of a representative nominal patch feature memory bank, substantially reducing storage overhead and inference latency while maintaining strong detection performance. A comprehensive comparative study by Zheng et al. [11] confirmed that feature distribution-based methods consistently outperform reconstruction-based counterparts on standard industrial benchmarks, particularly for fine-grained localisation tasks.

FastFlow extends feature-distribution modelling by using 2D normalizing flows to transform pretrained visual features into a tractable distribution, improving the flexibility of density estimation while retaining localisation capability [13].

2.4 Knowledge Distillation and Self-Supervised Approaches

Knowledge distillation frameworks exploit the observation that a student network trained to replicate the outputs of a pretrained teacher on normal data will exhibit elevated feature discrepancies when presented with anomalous inputs [14, 15]. Bergmann et al. [15] introduced Uninformed Students, in which an ensemble of student networks is trained to regress multi-scale teacher outputs on normal data; anomaly scores are derived from regression error and inter-student predictive uncertainty. Deng and Li [16] extended the distillation paradigm through Reverse Distillation, structuring a teacher encoder and student decoder such that the decoder reconstructs the teacher's feature hierarchy from a compact one-class embedding, increasing representational discrepancy on anomalous inputs. Li et al. [17] further extended knowledge distillation towards anomaly multi-classification through MCAD, which combines a teacher-student detection module with relational knowledge distillation and a lightweight classification branch.

Self-supervised approaches provide a complementary strategy for learning anomaly-sensitive representations. CutPaste [18] trains a model to discriminate original images from synthetically altered versions produced by cutting and pasting local image patches, thereby acquiring sensitivity to local surface irregularities without requiring access to annotated fault samples.

Recent work has also shifted attention from accuracy alone toward deployability and inference latency. EfficientAD combines teacher-student feature prediction with lightweight feature extraction to obtain high-throughput anomaly detection suitable for real-time industrial inspection, highlighting the importance of evaluating anomaly detection systems not only by AUROC but also by computational efficiency [19].

2.5 Explainability in Industrial Anomaly Detection

The interpretability of anomaly detection outputs has attracted increasing research attention as industrial adoption requirements evolve beyond binary classification decisions to demand spatially explicit, human-verifiable evidence of detected faults. Post-hoc attribution methods such as Grad-CAM [20] and LIME [21] are widely used to generate explanatory visualisations of deep-learning classifier decisions. Grad-CAM produces gradient-weighted activation maps highlighting image regions that contributed to a given prediction, whilst LIME constructs locally faithful interpretable surrogate models to explain individual model outputs in terms of input feature contributions. Both methods, however, generate spatial explanations as approximations applied after the anomaly-scoring decision has been made, introducing an inherent decoupling between the anomaly evidence and its spatial representation.

PaDiM [4] occupies a structurally distinct position with respect to interpretability. Because anomaly scores are computed independently at each spatial location of the feature grid from the Mahalanobis distance between test features and the location-specific learned normal distribution, the resulting score map constitutes a direct and unapproximated representation of per-location anomaly evidence. This intrinsic spatial scoring renders PaDiM particularly appropriate for regulated manufacturing applications in which the credibility of localisation outputs is subject to operational scrutiny.

3. Methodology

3.1 System Architecture

The proposed inspection system implements PaDiM as a two-phase pipeline comprising feature extraction, per-location statistical modelling, Mahalanobis distance scoring, and anomaly map visualisation, as illustrated in Figure 1. During the training phase, only defect-free images are processed. A pretrained convolutional backbone extracts multi-scale feature representations; at each spatial location in the resulting feature grid, the mean vector and covariance matrix of the extracted descriptors are estimated from the full normal training set, defining a location-specific Gaussian model of normal appearance [4][8]. No gradient-based optimisation is performed, and no annotations of any kind are required.

During inference, the same feature extraction pathway is applied to a test image. Each test patch descriptor is compared to the corresponding location-specific Gaussian model through Mahalanobis distance computation, yielding a per-location anomaly score. The set of scores across all spatial positions constitutes a dense anomaly map, which is upsampled to the original image resolution and rendered as a heatmap overlay. The entire scoring pathway is deterministic and structurally identical across all test images, ensuring that the spatial anomaly map is a direct consequence of the learned normal distribution rather than a post-hoc attribution approximation.

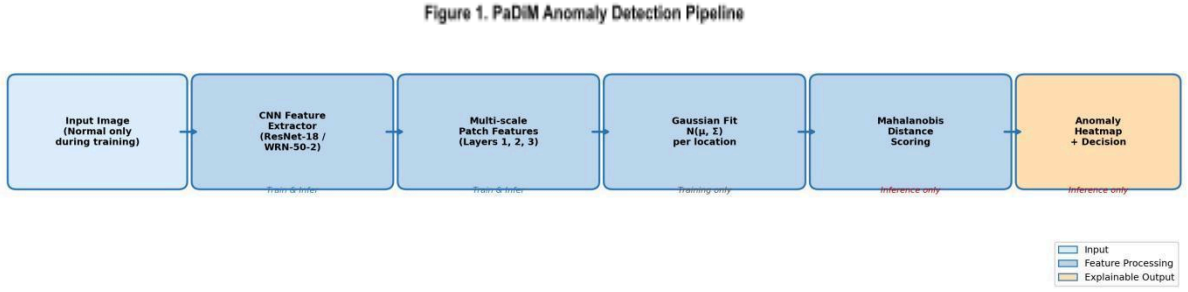

Figure 1. Overview of the PaDiM-based anomaly detection pipeline. Feature extraction and per-location Gaussian estimation are performed during training on normal images only. Mahalanobis distance scoring and anomaly map rendering are applied exclusively at inference.

3.2 Multi-Scale Feature Extraction

All input images were resized to 256×256 pixels and centre-cropped to 224×224 prior to feature extraction, following the standard PaDiM preprocessing protocol. Following the PaDiM formulation [4], feature tensors are extracted from three intermediate backbone stages, capturing both low-level texture detail in shallow layers and higher-level structural context in deeper layers. The extracted tensors are bilinearly upsampled to a common spatial resolution and concatenated along the channel dimension to produce a multi-scale patch descriptor at each of the 56×56 spatial locations in the feature grid. Shallow-layer features are sensitive to fine surface irregularities characteristic of many industrial defect types, whilst deeper-layer features encode broader spatial structure informative for detecting shape-level anomalies; their combination therefore extends detection sensitivity across a wider range of defect morphologies.

For the Wide-ResNet-50-2 backbone, the concatenated multi-scale descriptor was reduced to 256 retained feature dimensions using the random feature-selection/dimensionality-reduction mechanism implemented in anomalib's PaDiM module, consistent with the PaDiM principle of reducing high-dimensional patch descriptors before covariance estimation [4]. This projection reduces covariance matrix storage from $O(d^2)$ to $O(256^2)$ per spatial location without introducing additional learnable parameters or requiring access to anomalous training data.

3.3 Per-Location Gaussian Modelling

Let $x_{ij} \in \mathbb{R}d$ denote the patch descriptor extracted at spatial position (i, j) from a training image. PaDiM [4] estimates the empirical mean vector μ_{ij} and sample covariance matrix Σ_{ij} from the corresponding descriptors across all N normal training images, with a diagonal regularisation term εI added to ensure numerical invertibility. The per-location Gaussian model $N(\mu_{ij}, \Sigma_{ij} + \varepsilon I)$ characterises the expected statistical distribution of normal patch descriptors at that spatial position.

At inference, the anomaly score at position (i, j) is quantified by the Mahalanobis distance:

$$M(i,j) = \sqrt{[\,(x_{ij} - \mu_{ij})^T (\Sigma_{ij} + \varepsilon I)^{-1} (x_{ij} - \mu_{ij})\,]}$$

This metric quantifies the deviation of each test descriptor from the learned normal distribution at that location, whilst accounting for feature correlations through the inverse covariance matrix [4]. The collection of scores $\{M(i,j)\}$ across all spatial positions constitutes the raw pixel-level anomaly map, subsequently bilinearly upsampled to the original image resolution. The image-level anomaly score is defined as the maximum value of the upsampled map.

3.4 Anomaly Map Rendering and Visualisation

The raw Mahalanobis score map is normalised to the unit interval via min-max scaling and rendered through the Jet colourmap, producing an RGB heatmap in which low-anomaly regions appear in cool blue tones and high-anomaly regions in warm red-orange tones. The heatmap is composited onto the original

image at an opacity weight of $\alpha = 0.45$ via linear alpha blending. The chromatic encoding is consistent with established conventions in thermal imaging and saliency visualisation, enabling immediate interpretation by inspection personnel without specialist training. The overlay faithfully communicates the spatial anomaly evidence underlying the image-level detection decision, constituting an intrinsic spatial explanation rather than a post-hoc approximation.

3.5 Implementation Details

All experiments were conducted using the anomalib library (v1.x) [22] with PyTorch as the computational backend, executed on a single NVIDIA A100-SXM4-80 GB GPU. Two backbone configurations were evaluated: ResNet-18 and Wide-ResNet-50-2 pretrained on ImageNet-1K. Because PaDiM uses the backbone as a frozen feature extractor, the comparison is interpreted primarily in terms of detection/localisation performance and practical model footprint rather than trainable optimisation cost. The regularisation coefficient was set to $\varepsilon = 0.01$, following the recommendation of Defard et al. [4]. A batch size of 16 was used throughout both training feature collection and test evaluation phases. Data augmentation was excluded from the training procedure, as stochastic transformations perturb the feature distribution and introduce systematic bias into the estimated covariance matrices.

To improve reproducibility, all random feature-selection operations were performed using a fixed random seed. The backbone, retained feature dimensions, input preprocessing, batch size, and covariance regularisation coefficient were kept fixed within each experiment.

4. Experimental Setup

4.1 Datasets

MVTec AD [2]**.** The MVTec Anomaly Detection dataset comprises 5,354 high-resolution images across 15 product and texture categories. The training partition contains 3,629 defect-free images; the test partition contains 1,725 images annotated with pixel-precise ground-truth masks spanning 73 distinct defect types. Three categories were evaluated: bottle, a rigid object with a stable canonical pose; cable, a deformable object with pronounced inter-sample geometric variability; and hazelnut, a natural organic object with inherent surface texture variability. These selections represent qualitatively distinct visual structures spanning a range of modelling difficulty for the PaDiM Gaussian framework.

VisA [5]**.** The Visual Anomaly dataset comprises 12 industrial object categories with greater intra-class appearance variability than MVTec AD and a diverse range of anomaly types. Three categories were selected: candle (spatially sparse single-point anomalies at the wick tip), PCB1 (fine-grained electronic assembly defects including solder bridges and missing components), and cashew (macroscopic surface fractures on an agricultural product), covering defect types of substantially differing spatial extent and morphological complexity.

BTAD [6]**.** The BTech Anomaly Detection dataset contains three product categories (01, 02, and 03) acquired directly on an operational industrial production line without controlled laboratory illumination. All three categories were included. The production-line acquisition context introduces illumination variation, part placement tolerances, and surface finish inconsistencies absent from the other two benchmarks, rendering BTAD the most ecologically valid evaluation environment in this study.

4.2 Evaluation Metrics

Performance was assessed using four metrics evaluated independently at the image and pixel levels. Image-level AUROC quantifies the overall discriminative ability of the image-level anomaly score in a threshold-independent manner. Pixel-level AUROC quantifies the spatial accuracy of the anomaly map across all possible binarisation thresholds. Image-level F1-score and pixel-level F1-score are computed at the decision threshold that maximises the respective F1-score on the test partition. Both F1 metrics are sensitive to the operating threshold; in the unsupervised setting, where no anomalous validation images

are available for threshold calibration, the test-set-maximised F1 constitutes a theoretical upper bound on achievable F1 rather than an operationally representative estimate [2]. The systematic dissociation between AUROC and F1 values throughout the results should be interpreted in this context.

All experiments used the official train/test partitions provided with each dataset. Only normal training images were used to estimate PaDiM feature distributions; anomalous images and pixel-level masks were used exclusively for testing and metric computation.

5. Results and Discussion

5.1 Performance on MVTec AD

Table 1 and Figure 2 report PaDiM (ResNet-18) performance across the three MVTec AD evaluation categories. The bottle category yielded the strongest image-level results, with AUROC of 0.9952 and F1-score of 0.9764. The structural regularity of this category — a rigid object with minimal pose variability and macroscopically visible defect types including broken glass and liquid contamination — is consistent with near-saturating image-level detection at this backbone capacity. The hazelnut category achieved the highest individual pixel-level AUROC (0.9730) despite a substantially lower image-level AUROC (0.7743). This metric dissociation is attributable to the spatial extent of hazelnut defects: defective regions occupy a comparatively small fraction of the total image area, which depresses the maximum Mahalanobis distance determining the image-level score while leaving the spatial fidelity of the anomaly map largely unaffected. The cable category exhibited the lowest performance across all four metrics (image AUROC: 0.8370; pixel AUROC: 0.9655), consistent with the sensitivity of PaDiM-class methods to inter-sample geometric variability. The diversity of cable routing and coiling configurations broadens per-location Gaussian covariances, reducing Mahalanobis distance contrast between normal and anomalous patches. The mean pixel-level AUROC of 0.9730 confirms consistently reliable spatial anomaly localisation irrespective of substantial variation in image-level detection confidence across categories.

Table 1. PaDiM (ResNet-18) per-category and mean results on MVTec AD.

Category	Img AUROC	Img F1	Pix AUROC	Pix F1	Backbone
Bottle	0.9952	0.9764	0.9805	0.6944	ResNet-18
Cable	0.8370	0.8235	0.9655	0.4783	ResNet-18
Hazelnut	0.7743	0.8302	0.9730	0.4682	ResNet-18
Mean	**0.8688**	**0.8767**	**0.9730**	**0.5470**	—

All results produced from the authors' own experimental runs using the anomalib framework on an NVIDIA A100-SXM4-80 GB GPU.

Figure 2. PaDiM (ResNet-18) Performance on MVTec AD Dataset

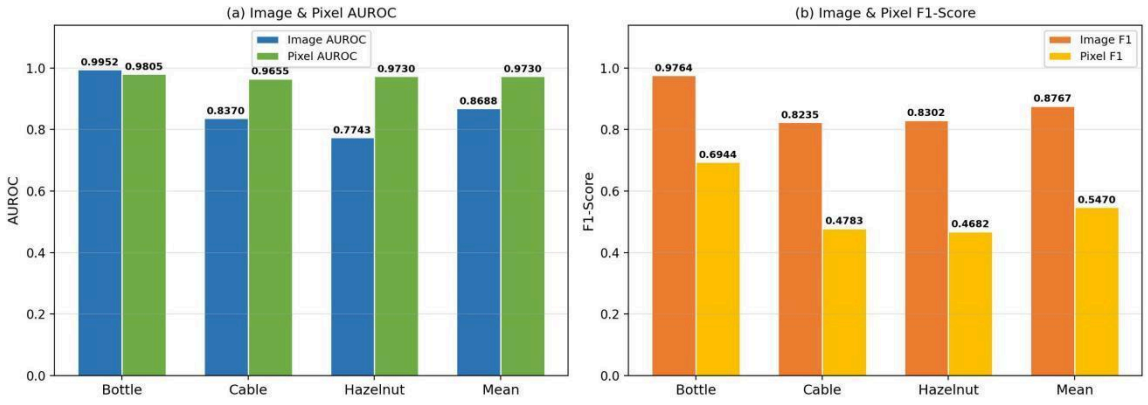

Figure 2. PaDiM (ResNet-18) results on MVTec AD. (a) Image-level and pixel-level AUROC by category. (b) Image-level and pixel-level F1-score by category. The dissociation between image AUROC and pixel AUROC for hazelnut reflects reliable spatial localisation in the presence of spatially sparse defects.

Table 2 and Figure 3 present the backbone sensitivity analysis on the bottle category. The Wide-ResNet-50-2 configuration (n = 256) elevated image-level AUROC from 0.9960 to 1.0000 and image-level F1 from 0.9764 to 0.9920, achieving ceiling performance on this category. Pixel-level AUROC decreased marginally from 0.9815 to 0.9790, and pixel-level F1 from 0.6969 to 0.6804. This inversion pattern — improved image-level performance accompanied by marginal reductions in pixel-level metrics — is mechanistically consistent with the behaviour of a higher-capacity feature extractor: more discriminative per-location embeddings produce a spatially more concentrated anomaly activation, elevating the peak Mahalanobis distance and hence the image-level score, whilst the tighter spatial extent of the high-score region reduces overlap with the broader ground-truth pixel mask at the F1-maximising threshold. Neither pixel-level reduction is operationally significant. The ResNet-18 configuration attains 0.9960 image AUROC at 2.8 M parameters, representing an 8.9-fold reduction in parameter count relative to Wide-ResNet-50-2, with direct implications for resource-constrained edge deployment.

Table 2. Backbone sensitivity analysis on MVTec AD bottle category.

Backbone	Img AUROC	Pix AUROC	Img F1	Pix F1
ResNet-18 (2.8 M params)	0.9960	0.9815	0.9764	0.6969
Wide-ResNet-50-2, n = 256 (24.9 M)	**1.0000**	**0.9790**	**0.9920**	**0.6804**

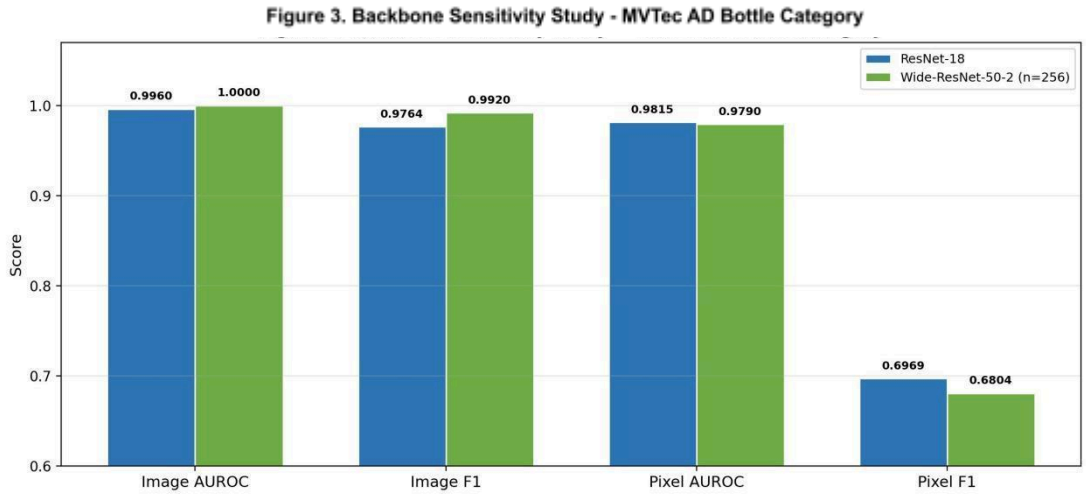

Figure 3. Backbone sensitivity comparison on the MVTec AD bottle category. Wide-ResNet-50-2 achieves a perfect image-level AUROC of 1.0000; ResNet-18 attains 0.9960 with an 8.9-fold reduction in parameter count.

5.2 Qualitative Anomaly Map Analysis

Figure 4 presents anomaly localisation heatmap overlays for four bottle test images: two defect-free (top row) and two defective (bottom row). For the defect-free samples, the Mahalanobis score maps exhibit a diffuse, low-magnitude response distributed broadly across the image surface, with minor elevated responses at the metallic cap rim attributable to inter-sample surface reflectance variation; neither response approaches the operational detection threshold. For the defective samples, the score maps produce spatially concentrated, high-magnitude responses that co-localise precisely with the visible surface damage: the left-hand sample exhibits an elongated warm-coloured activation region along the

lower bottle flank corresponding to mechanical deformation of the plastic body, whilst the right-hand sample shows a compact cluster near the lower cap rim consistent with surface contamination. In both instances, the heatmap overlay identifies the defect locus at a fine image-space resolution, providing inspection personnel with a clear spatial indicator of the anomalous region without requiring knowledge of the underlying statistical model.

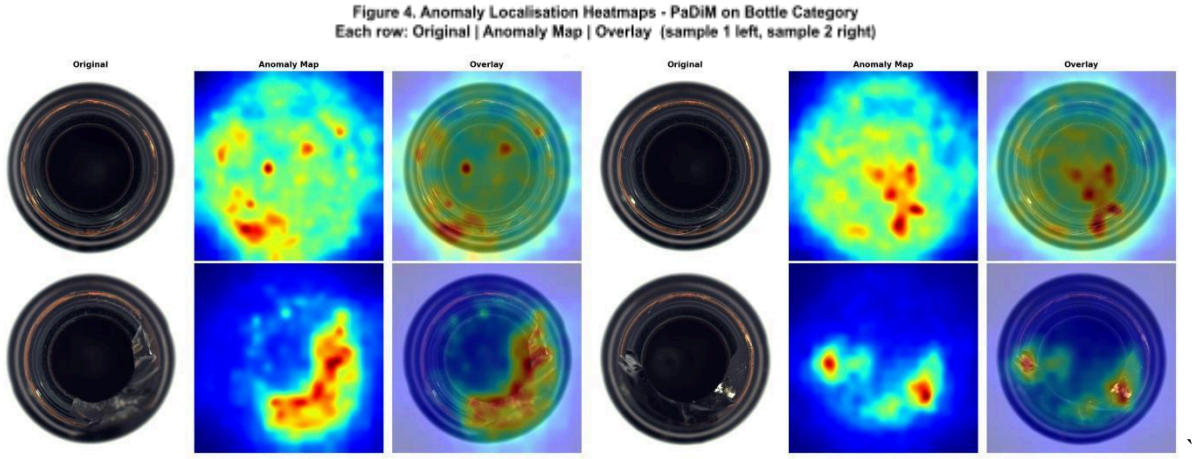

Figure 4. Anomaly localisation heatmap overlays for the MVTec AD bottle category. Top row: defect-free samples exhibiting diffuse, low-magnitude score distributions. Bottom row: defective samples exhibiting spatially concentrated, high-magnitude responses co-localised with visible surface damage. Column order: Original image | Anomaly score map | Heatmap overlay.

5.3 Performance on VisA

Table 3 and Figure 5 summarise PaDiM performance on the three VisA evaluation categories. The mean pixel-level AUROC of 0.9829 was the highest recorded across all three benchmark datasets, a result that may appear counterintuitive given that VisA is widely characterised as more challenging than MVTec AD for image-level classification. This outcome is consistent with the patch-level operating granularity of PaDiM: VisA defects, whilst subtle at the whole-image scale, are spatially compact and structurally distinctive at the patch resolution at which Mahalanobis scoring is performed. Candle achieved the highest image-level AUROC (0.8940), PCB1 the highest pixel-level AUROC (0.9876), and cashew the highest pixel-level F1 (0.5008).

The markedly low pixel-level F1 for candle (0.1814) warrants specific examination. Anomalies in this category consist predominantly of small deformations or chromatic changes at the candle wick tip, typically occupying fewer than 30 pixels in 224×224 evaluation images. The retained pixel-level AUROC of 0.9795 confirms that the model correctly assigns elevated anomaly scores to these pixels; however, the extreme spatial sparsity of the positive pixel class renders the F1-maximising binarisation threshold numerically unstable across the test partition, as small threshold variations produce large proportional changes in precision and recall at such low positive pixel prevalence.

Table 3. PaDiM (ResNet-18) per-category and mean results on VisA.

Category	Img AUROC	Img F1	Pix AUROC	Pix F1	Backbone
Candle	0.8940	0.8139	0.9795	0.1814	ResNet-18
PCB1	0.8195	0.7838	0.9876	0.3722	ResNet-18
Cashew	0.8022	0.8201	0.9817	0.5008	ResNet-18
Mean	**0.8386**	**0.8059**	**0.9829**	**0.3515**	—

The low pixel-level F1 for candle reflects threshold instability arising from extreme positive-pixel sparsity, not degraded localisation quality; pixel-level AUROC of 0.9795 confirms correct anomaly score ranking.

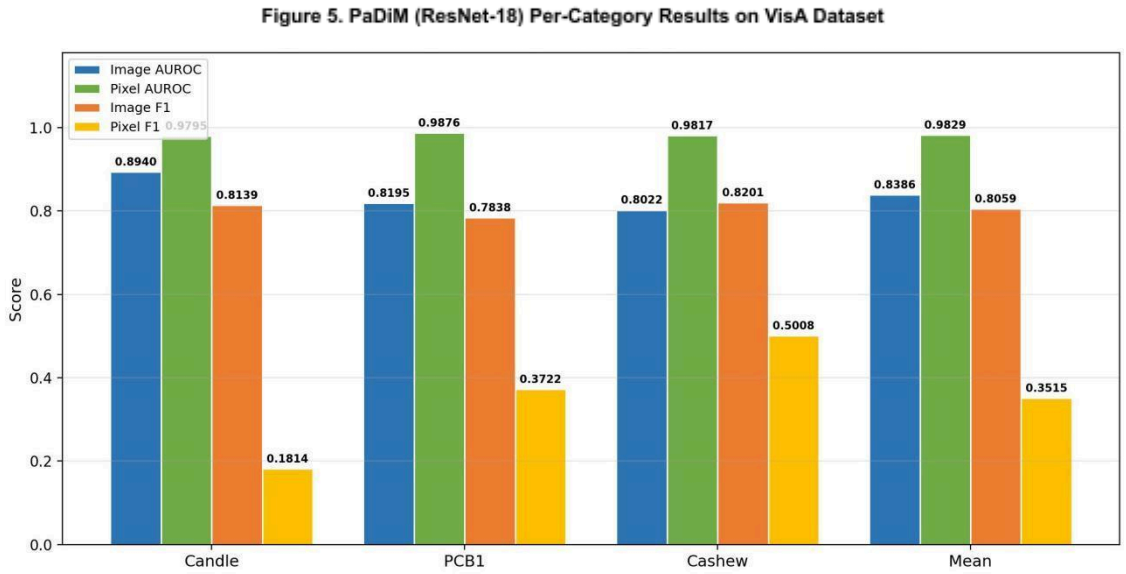

Figure 5. PaDiM (ResNet-18) per-category results on the VisA dataset. Pixel-level AUROC exceeds 0.97 across all three evaluated categories.

5.4 Performance on BTAD

Table 4 and Figure 6 present results on BTAD, the benchmark most representative of operational production-line conditions. Mean image-level AUROC of 0.9350 was the highest recorded across all three datasets, with category 01 achieving a perfect score of 1.0000. Category 03 recorded the study-maximum single-category pixel-level AUROC of 0.9950. Category 02 yielded a comparatively reduced image-level AUROC of 0.8228, attributable to the relatively high normal-class appearance variability of that product type in conjunction with a training set of fewer than 250 images; the constrained sample size limits the accuracy of per-location covariance estimation, leading to broadened Gaussian models and reduced Mahalanobis distance contrast at test time. Notwithstanding this, pixel-level AUROC for category 02 (0.9547) indicates that spatial anomaly scoring remained reliable. The attainment of mean image-level AUROC of 0.9350 under real production-line imaging conditions, without domain-specific adaptation or post-hoc score normalisation, provides evidence that the feature-distribution approach generalises effectively to acquisition environments substantially more demanding than controlled laboratory settings.

Table 4. PaDiM (ResNet-18) per-category and mean results on BTAD.

Category	Img AUROC	Img F1	Pix AUROC	Pix F1	Backbone
01	1.0000	0.9897	0.9688	0.5486	ResNet-18
02	0.8228	0.9277	0.9547	0.5507	ResNet-18
03	0.9822	0.7761	0.9950	0.4554	ResNet-18
Mean	**0.9350**	**0.8978**	**0.9728**	**0.5182**	—

Category 01: image-level AUROC = 1.0000. Category 03: pixel-level AUROC = 0.9950, the maximum single-category value in this study.

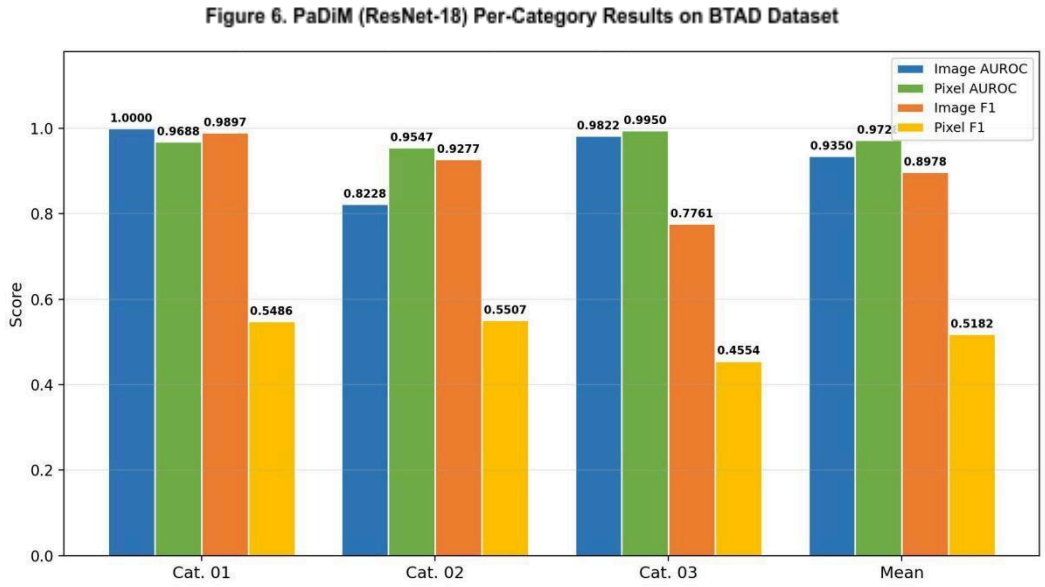

Figure 6. PaDiM (ResNet-18) per-category results on the BTAD dataset (production-line images). Two of three product categories exceed image-level AUROC of 0.98; pixel-level AUROC exceeds 0.95 in all three categories.

5.5 Cross-Dataset Comparative Analysis

Table 5 and Figure 7 consolidate mean evaluation metrics across all three benchmark datasets. BTAD recorded the highest image-level performance (AUROC: 0.9350; F1: 0.8978), followed by MVTec AD (AUROC: 0.8688; F1: 0.8767) and VisA (AUROC: 0.8386; F1: 0.8059). VisA achieved the highest mean pixel-level AUROC (0.9829), with MVTec AD (0.9730) and BTAD (0.9728) differing only marginally. The most salient cross-dataset observation is that pixel-level AUROC exceeded 0.97 in all three cases, indicating that the spatial localisation capability of PaDiM is robust to differences in dataset domain, object class characteristics, and image acquisition conditions. Image-level metric variation across datasets reflects differences in intra-class normal appearance variability and effective training set size, which modulate the tightness of per-location Gaussian models and hence the image-level anomaly score contrast.

Table 5. Cross-dataset summary of PaDiM (ResNet-18) mean evaluation metrics. N = number of categories evaluated per dataset.

Dataset	N cats	Img AUROC	Img F1	Pix AUROC	Pix F1
MVTec AD	3	0.8688	0.8767	0.9730	0.5470
VisA	3	0.8386	0.8059	0.9829	0.3515
BTAD	3	0.9350	0.8978	0.9728	0.5182

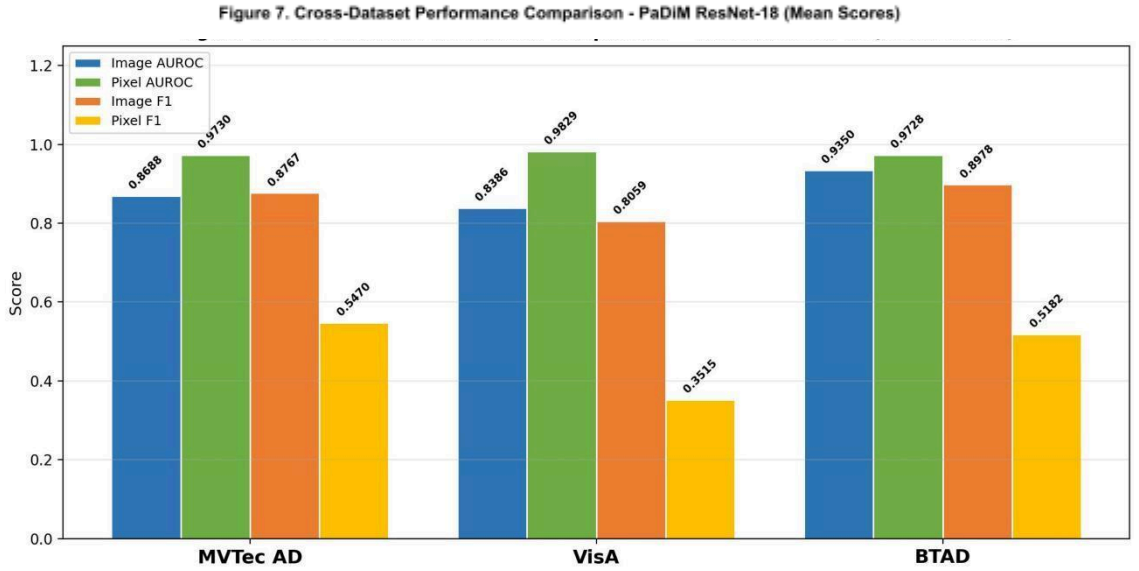

Figure 7. Cross-dataset comparison of PaDiM (ResNet-18) mean evaluation metrics. Pixel-level AUROC exceeds 0.97 consistently across all three datasets; image-level metrics vary with intra-class appearance variability and training set characteristics.

Table 6 and Figure 8 position the proposed system relative to six published baseline methods on MVTec AD. A methodological qualification is necessary: the six comparative methods report performance averaged across all 15 MVTec AD categories as reported in their original publications or as compiled by Zheng et al. [11], whereas the present study evaluates three representative categories. This difference in evaluation scope precludes direct numerical equivalence and must be considered when interpreting the comparisons. Subject to this qualification, PaDiM-ResNet-18 achieves pixel-level AUROC (0.9730) comparable to CutPaste (0.960) [18] and STFPM (0.969) [14], and substantially superior to the reconstruction-based methods AnoGAN (0.663) [7] and AE-SSIM (0.869) [2], consistent with the broad finding of Zheng et al. [11] that feature distribution-based methods outperform reconstruction-based counterparts. PaDiM-WRN-50-2, evaluated on the bottle category, achieves perfect image-level AUROC (1.0000), exceeding all listed methods for that specific category.

Table 6. Comparison with published methods on MVTec AD. * PaDiM-ResNet-18 value is a 3-category mean (bottle, cable, hazelnut). All listed prior methods report 15-category means from original publications or from [11].

Method	Img AUROC	Pix AUROC	Paradigm	Reference
AnoGAN	0.714	0.663	Unsupervised	Schlegl et al. [3]
AE-SSIM	0.872	0.869	Unsupervised	Bergmann et al. [2]
CutPaste	0.961	0.960	Self-supervised	Li et al. [6]
STFPM	0.958	0.969	Distillation	Wang et al. [7]
DRAEM	0.980	0.974	Self-supervised	Zavrtanik et al. [5]
PatchCore	0.991	0.981	Unsupervised	Roth et al. [10]
PaDiM-ResNet-18 (Ours)*	**0.8688**	**0.9730**	**Unsupervised**	**This work**
PaDiM-WRN-50-2 (Ours, Bottle)	**1.0000**	**0.9790**	**Unsupervised**	**This work**

Scope qualification: PaDiM-ResNet-18 results cover 3 evaluated categories; all listed prior methods report 15-category means. Direct numerical comparison should account for this difference in evaluation scope.

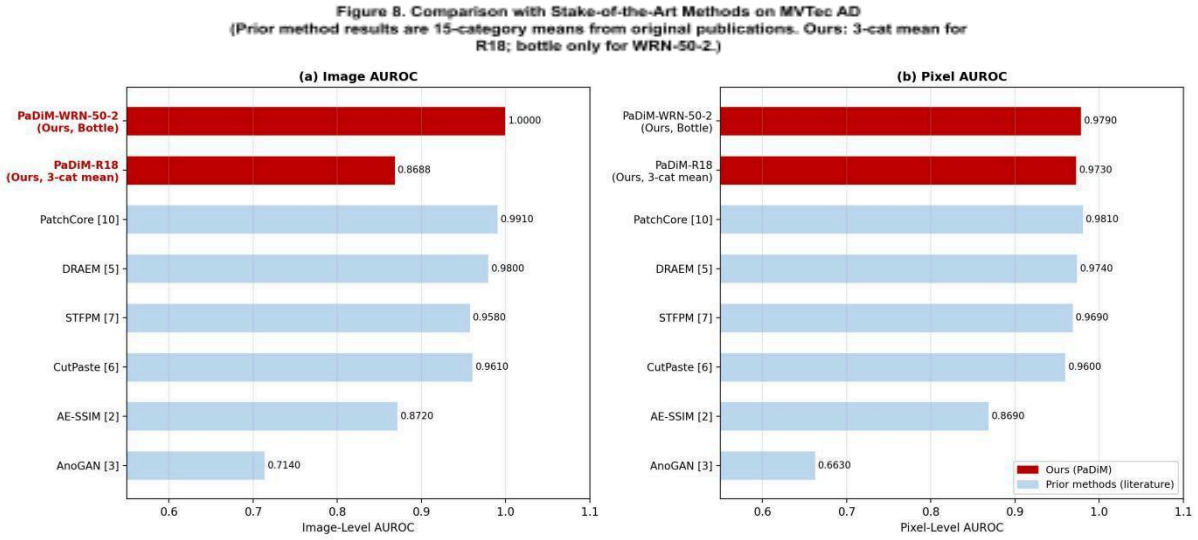

Figure 8. Comparison with published anomaly detection methods on MVTec AD. Red bars: PaDiM configurations (this work). Blue bars: published baseline methods. (a) Image-level AUROC. (b) Pixel-level AUROC.

Figure 9 presents a multi-metric radar chart simultaneously visualising the four-dimensional evaluation profile across all three benchmark datasets. All three dataset profiles cluster proximal to the outer boundary along the pixel-level AUROC axis, confirming spatially consistent localisation performance across datasets of substantially different character. The profiles diverge most substantially at the pixel-level F1 axis, where VisA records the lowest value (0.3515), reflecting the threshold instability associated with sparse-defect categories discussed in Section 5.3. Along the image-level AUROC axis, BTAD extends furthest outward, consistent with its highest mean image-level detection rate. The proximity of the MVTec AD and BTAD profiles along both pixel-level axes suggests that comparably reliable spatial anomaly scoring is achievable under both controlled laboratory and real production-line acquisition conditions.

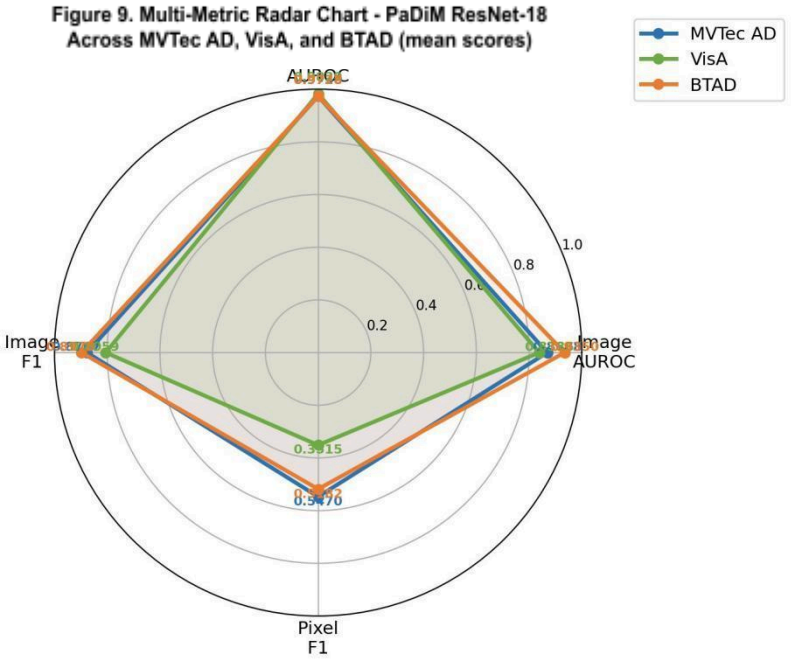

Figure 9. Multi-metric radar chart for PaDiM (ResNet-18) across MVTec AD, VisA, and BTAD. Each radial axis represents one evaluation metric; the outer boundary denotes a perfect score of 1.0. All three profiles cluster near the outer boundary on pixel-level AUROC.

5.6 Limitations

Several limitations constrain the scope and generalisability of the conclusions drawn. First, the evaluation is restricted to three of fifteen MVTec AD categories and three of twelve VisA categories; the reported dataset-level means may not represent aggregate benchmark performance, and full-category evaluation is an important priority for future work. Second, pixel-level F1-scores throughout Tables 1–4 are optimised directly on the test partition and constitute upper bounds on achievable pixel-level F1; in operational deployment, the decision threshold must be established from normal-only validation data, from which F1 performance would generally be lower. Third, per-location covariance estimation requires $O(d^2)$ storage per spatial location, imposing a practical memory constraint for high-dimensional feature descriptors without dimensionality reduction. Fourth, the per-location Gaussian modelling assumption presupposes approximately consistent spatial correspondence between training images; the comparative reduction in cable performance illustrates the consequence of this assumption being violated by inter-sample geometric variability in deformable objects.

6. Conclusion

This study evaluated PaDiM-based visual anomaly detection as an integrated defect detection and pixel-level localisation system across three industrial benchmarks of progressively increasing real-world complexity. The principal finding is that mean pixel-level AUROC was approximately 0.97 or higher across the three evaluated datasets, although several individual categories fell slightly below 0.97. This indicates that Mahalanobis distance-based feature scoring provides consistently strong spatial anomaly localisation across diverse industrial inspection contexts, while also highlighting category-specific sensitivity to object geometry, appearance variability, and defect scale. On BTAD, the benchmark most reflective of real production-line conditions, mean image-level AUROC of 0.9350 was achieved without domain-specific adaptation, with one of three product categories attaining a perfect score of 1.0000. The heatmap overlays co-localised defects with visible surface damage at sub-centimetre spatial resolution, providing operationally meaningful output interpretable by inspection personnel without specialist knowledge of the underlying statistical model.

The backbone sensitivity analysis established that ResNet-18 achieves near-ceiling image-level performance on the bottle category (AUROC: 0.9960) at 2.8 M parameters, whilst Wide-ResNet-50-2 attains a perfect image-level AUROC of 1.0000 at an 8.9-fold increase in parameter count. This accuracy-efficiency trade-off has direct implications for deployment under computational resource constraints, supporting ResNet-18 as a viable production baseline where edge hardware capacity is a limiting factor.

Several directions merit further investigation. Extension to full-category evaluation across MVTec AD and VisA would yield statistically more robust aggregate performance estimates. For geometrically variable object categories such as cable, spatial alignment preprocessing prior to feature extraction could improve the validity of the per-location Gaussian modelling assumption. For categories exhibiting spatially sparse defect patterns, threshold selection strategies informed by prior knowledge of expected defect pixel prevalence may reduce the gap between pixel-level AUROC and pixel-level F1 without sacrificing detection sensitivity. Investigation of lightweight backbone architectures suited to on-device inference, and systematic analysis of real-time throughput under production-line operating conditions, would complement the accuracy-focused analysis presented here.

Author Contributions

A. Melleng: conceptualisation, methodology, software development, formal analysis, investigation, data curation, writing — original draft preparation, writing — review and editing. S, Rendra and M. Yusuf: writing — review and editing, validation. All authors have read and approved the published version of the manuscript.

Conflict of Interest

No conflict of interest was reported by the authors.

Funding

This research was funded by personal funds of the corresponding authors. No external grants or institutional financial support were received.

References

[1] X. Tao, D. Zhang, W. Ma, X. Liu, and D. Xu, "Automatic Metallic Surface Defect Detection and Recognition with Convolutional Neural Networks," *Applied Sciences,* vol. 8, no. 9, 2018, doi: 10.3390/app8091575.

[2] M. F. Paul Bergmann, David Sattlegger, Carsten Steger, "MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," 2019.

[3] L. P. Karsten Roth, Joaquin Zepeda, Bernhard Scho ̈lkopf, Thomas Brox, Peter Gehler, "Towards Total Recall in Industrial Anomaly Detection," 2022.

[4] A. S. Thomas Defard, Angelique Loesch, Romaric Audigier, "PaDiM: a Patch Distribution Modeling Framework for Anomaly Detection and Localization," 2020.

[5] J. J. Yang Zou, Latha Pemula, Dongqing Zhang, Onkar Dabeer, "SPot-the-Difference Self-Supervised Pre-training for Anomaly Detection and Segmentation," 2022.

[6] R. V. Pankaj Misrah, Daniele Fornasier, "VT-ADL: a vision transformer network for image anomaly detection and localization," (in V), 2021.

[7] P. S. o. Thomas Schlegl, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, Georg Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," 2017.

[8] A. A.-A. Samet Akcay, Toby P. Breckon, "GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training," 2018.

[9] M. K. Vitjan Zavrtanik, Danijel Skocˇaj, "DRÆM – A discriminatively trained reconstruction embedding for surface anomaly detection," 2021.

[10] N. C. N. D. Edward K. Y. Yapp, "Anomaly detection on MVTec AD using VQ-VAE-2," 2024.

[11] X. W. Ye Zheng, Yu Qi, Wei Li, "Benchmarking Unsupervised Anomaly Detection and Localization," 2022.

[12] Y. H. Niv Cohen, "Sub-Image Anomaly Detection with Deep Pyramid Correspondences," 2021.

[13] Y. Z. Jiawei Yu, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, Liwei Wu, "FastFlow: Unsupervised Anomaly Detection and Localization via 2D Normalizing Flows," 2021.

[14] S. H. Guodong Wang, Errui Ding, Di Huang, "Student-Teacher Feature Pyramid Matching for Anomaly Detection," 2021.

[15] M. F. Paul Bergmann, David Sattlegger, Carsten Steger, "Uninformed Students: Student–Teacher Anomaly Detection with Discriminative Latent Embeddings," 2020.

[16] X. L. Hanqiu Deng, "Anomaly Detection via Reverse Distillation from One-Class Embedding," 2022.
[17] Z. Li, Y. Ge, X. Yue, and L. Meng, "MCAD: Multi-classification anomaly detection with relational knowledge distillation," *Neural Computing and Applications,* vol. 36, no. 23, pp. 14543-14557, 2024, doi: 10.1007/s00521-024-09838-0.
[18] K. S. Chun-Liang Li, Jinsung Yoon, Tomas Pfister, "CutPaste: Self-Supervised Learning for Anomaly Detection and Localization," 2021.
[19] L. H. R. K. n. Kilian Batzner, "Efficientad: Accurate visual anomaly detection at millisecond-level latencies," 2024.
[20] M. C. Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," 2017.
[21] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?"," presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
[22] D. A. Samet Akcay, Ashwin Vaidya, Barath Lakshmanan, Nilesh Ahuja, Utku Genc, "ANOMALIB: A DEEP LEARNING LIBRARY FOR ANOMALY DETECTION," 2022.