



Interpretable Spatial Localisation for Visual Anomaly Detection in Industrial Monitoring Systems: A Cross-Dataset Evaluation Using MVTec AD, VisA, and BTAD

Alimuddin Melleng^{1*}, Soffian Mohammad², Muhammad Yusuf¹

¹Faculty of Engineering, Department of Informatics Engineering, University of Muhammadiyah Sorong, Sorong City, West Papua 98416, Indonesia

²Faculty of Engineering and Physical Sciences, School of Natural and Built Environment, Queen's University Belfast, Belfast, Northern Ireland BT7 1NN, United Kingdom

*Correspondence: alimuddinmllg@gmail.com; wwanmohammad02@qub.ac.uk; yusuf@um-sorong.ac.id

SUBMITTED: 29 June 2026; REVISED: 29 September 2026; ACCEPTED: 11 October 2026

ABSTRACT: Visual quality inspection requires anomaly detection using limited defect data and accurate spatial localisation. This study evaluates Patch Distribution Modeling (PaDiM), a one-class framework that generates pixel-level anomaly maps using Mahalanobis distance scoring, across three purposively selected categories from each of the MVTec AD and VisA datasets and all three categories of BTAD. Dataset-level results for MVTec AD and VisA are reported as subset means, whereas the BTAD mean represents the complete three-category benchmark. Using ResNet-18, the mean pixel-level AUROC values were 0.9713 for MVTec AD, 0.9835 for VisA, and 0.9764 for BTAD, while the mean image-level AUROC ranged from 0.8194 to 0.9381. For the MVTec AD bottle category, Wide-ResNet-50-2 improved all four evaluation metrics, achieving an image-level AUROC of 1.0000. Counterfactual masking was performed to assess whether regions with high anomaly scores contributed to the image-level anomaly score. Masking the highest-scoring 5% of pixels reduced the score by 10.98%, compared with a 2.08% increase for a size-matched random control and no change (0.00%) for an inverse control. These findings indicate that the highlighted regions contributed more substantially to the anomaly score than either size-matched control, although their removal did not eliminate the remaining anomaly evidence. For the VisA candle category, a high pixel-level AUROC of 0.9827 was accompanied by a low mean intersection-over-union of 0.091, demonstrating that strong pixel-level anomaly ranking does not necessarily translate into accurate spatial overlap with actual defects. Overall, this study demonstrates the potential of PaDiM for anomaly detection while emphasising the limitations of spatial localisation. Its contribution lies in providing empirically evaluated and explicitly bounded spatial localisation rather than a direct explanation of the model's decision-making process.

KEYWORDS: Patch distribution modelling; one-class learning; explanation faithfulness; counterfactual masking; Mahalanobis distance

1. Introduction

Automated visual inspection plays an important role in manufacturing quality control. Unsupervised industrial anomaly detection has gained increasing attention because labelled defect samples are often scarce, expensive to obtain, or unavailable [1, 2], whereas normal samples are generally easier to collect than the diverse defect examples required for supervised training [3, 4]. The MVTec AD dataset [5] reflects this setting by providing defect-free training images and anomalous test images. Consequently, anomaly detection commonly adopts a one-class formulation, in which a model trained exclusively on normal images identifies anomalies by measuring deviations from learned normal patterns [4, 6]. This approach eliminates the need for an exhaustive defect taxonomy and enables the detection of previously unseen anomalies. Recent benchmark studies have demonstrated the effectiveness of this approach. Xie et al. [7] evaluated 19 algorithms across seven industrial image datasets using a unified framework that considered supervision strategies, learning paradigms, and computational efficiency. Similarly, Gudovskiy et al. [8] introduced CFLOW-AD, which employs conditional normalising flows on features extracted from pretrained encoders for unsupervised anomaly detection and localisation. Their experiments on MVTec AD demonstrated efficient pixel-level localisation without requiring anomalous training samples.

Beyond detecting anomalies, industrial visual inspection also requires accurate identification of their spatial locations. Although a scalar anomaly score indicates whether an image contains an abnormality, it does not provide information about its location. In contrast, pixel-level localisation generates spatial anomaly maps that can support quality-control decisions, including product acceptance, rejection, and further inspection. However, these maps should not automatically be interpreted as explanations of model behaviour, as visually plausible attributions may not accurately represent the information used by a model during prediction [9–11]. This distinction motivates the use of the term *intrinsic spatial localisation* and the counterfactual masking experiment presented in Section 5.6. Patch Distribution Modeling (PaDiM), introduced by Defard et al. [12], generates spatial anomaly scores through location-specific Gaussian modelling. Multi-scale convolutional features extracted at each spatial location are represented by a multivariate Gaussian distribution, while anomaly scores are calculated using the Mahalanobis distance from the learned distribution of normal features. The resulting anomaly map is generated directly through the detector's scoring mechanism rather than through a separate post-hoc attribution method. This map can subsequently be visualised as a heatmap highlighting spatial deviations from normal appearance.

This study evaluates the performance of PaDiM across three industrial anomaly-detection benchmarks and investigates the influence of backbone architecture on detection and localisation performance. The main contributions of this study are: (i) the implementation and evaluation of a one-class anomaly detection and localisation pipeline incorporating intrinsic spatial scoring; (ii) the application of counterfactual masking experiments with size-matched random and inverse controls to assess the contribution of high-scoring regions to image-level anomaly scores; (iii) a controlled comparison between ResNet-18 and Wide-ResNet-50-2 using the MVTec AD bottle category to investigate backbone sensitivity; (iv) an empirical evaluation covering three purposively selected categories from each of MVTec AD [5] and VisA [13], together with all three categories of BTAD [14], based on the experimental results obtained in this study; and (v) a scope-qualified comparison with previously published baseline methods,

with evaluation coverage explicitly reported in the comparison table to ensure transparent interpretation of the findings.

2. Related Work

2.1. Benchmark Datasets for Industrial Anomaly Detection.

MVTec AD [4], introduced by Bergmann et al. in 2019, comprises 5,354 high-resolution images across 15 industrial categories, with pixel-level annotations covering 73 defect types. VisA [13] comprises 12 product categories characterised by substantial intra-class variation, whereas BTAD [14] contains real-world images of industrial products acquired using industrial imaging systems, reflecting acquisition conditions that differ from those of laboratory-based imaging. Collectively, these datasets represent diverse object structures, anomaly morphologies, and image-acquisition conditions, providing complementary benchmarks for evaluating industrial anomaly-detection methods.

2.2. Reconstruction-Based Anomaly Detection.

Reconstruction-based anomaly detection relies on the assumption that models trained on normal images can reconstruct normal appearances more accurately than anomalous regions [15–17]. AnoGAN [15] employs adversarial learning to capture the distribution of normal data and maps test images into a latent space to quantify reconstruction and discrimination errors. GANomaly [16] adopts an encoder–decoder–encoder architecture that eliminates the need for iterative latent-space optimisation during inference. DRAEM [17] integrates anomaly-free image reconstruction with a discriminative sub-network trained on synthetically generated anomalies, enabling pixel-level anomaly localisation without requiring real anomalous training samples. Yapp and Doan [18] applied VQ-VAE-2 to the MVTec AD dataset by representing defect-free images in a discrete latent space and identifying anomalies through deviations in latent-code likelihood. Despite their effectiveness, reconstruction-based approaches may exhibit limited sensitivity to subtle or localised defects, particularly when anomalous regions are reconstructed with relatively small residual errors [17]. This limitation highlights the importance of alternative approaches that characterise anomalies through feature-level deviations rather than reconstruction errors alone.

2.3. Feature Distribution-Based Methods.

Feature distribution-based methods characterise normal appearance using pretrained deep-feature representations rather than reconstructing input images. Features extracted from normal samples are used to establish reference distributions, while deviations observed in test samples are quantified to generate anomaly scores. These approaches can offer improved sensitivity to local abnormalities that produce relatively small reconstruction errors [6, 12]. PaDiM [12] models the distribution of normal features at each spatial location using a multivariate Gaussian distribution and calculates anomaly scores through the Mahalanobis distance. In contrast, PatchCore [6] constructs a memory bank of representative normal patch features and employs greedy coreset subsampling to reduce memory requirements and computational costs while maintaining strong anomaly-detection performance. Both approaches demonstrate the potential of pretrained feature representations for identifying deviations from normal industrial image patterns without requiring labelled defect samples during training.

2.4. Knowledge Distillation and Self-Supervised Approaches.

Knowledge distillation-based methods identify anomalies by measuring discrepancies between teacher and student networks trained on normal data [19, 20]. Uninformed Students [20] employs an ensemble of student networks to regress multi-scale teacher representations, generating anomaly scores from prediction errors and predictive uncertainty. Reverse Distillation [21] uses a student decoder to reconstruct teacher representations from a compact one-class embedding, allowing anomalies to be identified through discrepancies between the reconstructed and original feature representations. MCAD [22] further incorporates relational knowledge distillation and a lightweight classification branch to enhance anomaly-detection performance. Self-supervised approaches provide an alternative by learning anomaly-sensitive representations without requiring manually annotated defects. CutPaste [23] trains a model to distinguish normal images from synthetically modified images generated through cut-and-paste operations, thereby encouraging sensitivity to local visual irregularities. EfficientAD [24] integrates a lightweight feature extractor, student–teacher prediction, and an autoencoder to detect structural and logical anomalies while maintaining low inference latency. These developments indicate that practical industrial anomaly detection should consider computational efficiency and inference speed alongside detection accuracy and localisation performance.

2.5. Interpretability, Explanation Faithfulness, and Anomaly Localisation.

Interpretability and explanation faithfulness have become important considerations in industrial anomaly detection, particularly when spatial anomaly maps are used to support quality-control decisions. Grad-CAM [25] generates gradient-weighted activation maps to highlight image regions associated with a target prediction, whereas LIME [26] constructs a locally interpretable surrogate model to approximate individual predictions and can be applied to both image and non-image data. Both approaches provide post-hoc explanations, whereas PaDiM generates spatial anomaly scores directly through its underlying anomaly-scoring mechanism. However, visually plausible explanations do not necessarily establish faithfulness to the model's actual decision-making process. Adebayo et al. [9] demonstrated that certain saliency maps remain visually similar even after model parameters or training labels are randomised. Similarly, Ghorbani et al. [27] showed that small input perturbations can substantially alter attribution maps without changing the predicted class. Krishna et al. [28] further reported considerable disagreement among explanation methods when interpreting the same prediction. These findings suggest that explanations generated by individual methods should be interpreted cautiously, as their visual coherence does not guarantee that they accurately represent the underlying model behaviour.

PaDiM [12] differs structurally from conventional post-hoc attribution methods because its spatial anomaly scores are calculated directly from the Mahalanobis distance between each test feature descriptor and the corresponding location-specific distribution of normal features. The image-level anomaly score is subsequently derived from the maximum value of the same spatial anomaly map. This formulation eliminates the need for a separate attribution model to approximate the detector's output. Nevertheless, the direct relationship between spatial scores and image-level predictions does not, by itself, establish that the highlighted regions causally determine the model's decision. Jacovi and Goldberg [11] emphasised the distinction between

explanation plausibility and faithfulness, highlighting the importance of evaluating whether explanations accurately reflect model behaviour. Deletion-based evaluation provides one approach to assessing this relationship. Samek et al. [29] examined changes in model outputs following the removal of regions identified as important, while Petsiuk et al. [30] introduced deletion and insertion metrics to quantify the influence of selected image regions on predictions. However, Hooker et al. [10] demonstrated that removing image features can introduce distribution shifts that affect model behaviour independently of the importance of the removed features. To address this concern, they proposed the Remove and Retrain (ROAR) framework. Building on these methodological considerations, Section 3.5 introduces a counterfactual masking experiment for PaDiM, incorporating size-matched random and inverse controls, with the corresponding results presented in Section 5.6.

Although a Mahalanobis distance map identifies spatial locations where test-image features deviate from learned normal distributions, it does not explicitly determine which visual characteristics, such as texture, colour, geometry, or illumination, are responsible for those deviations. Furthermore, it does not explain why the frozen feature-extraction backbone represents particular visual patterns as anomalous. Accordingly, this study adopts the term *intrinsic spatial interpretability* in a restricted sense, referring specifically to the direct relationship between spatial anomaly scores and the detector's scoring mechanism. Rudin [31] emphasised the importance of distinguishing inherently interpretable models from post-hoc explanation approaches. However, this distinction should not be interpreted as evidence that PaDiM provides faithful explanations without empirical validation. Instead, the counterfactual masking experiments conducted in this study provide a bounded assessment of the contribution of high-scoring spatial regions to image-level anomaly scores.

3. Methodology

3.1. System Architecture.

The proposed system implements Patch Distribution Modeling (PaDiM) through four main stages: multi-scale feature extraction, location-specific statistical modelling, Mahalanobis distance scoring, and anomaly-map visualisation (Figure 1). During training, a pretrained backbone extracts multi-scale feature representations from defect-free images, and the feature distribution at each spatial location is characterised by its mean and covariance matrix [12]. The training process does not involve gradient-based optimisation, anomaly masks, or defect-class labels, as all images in the training partition are known to be defect-free. During inference, each test-image descriptor is compared with the corresponding location-specific Gaussian distribution using the Mahalanobis distance. The resulting anomaly scores are assembled into a dense spatial anomaly map, which is subsequently upsampled to the original input resolution and visualised as a heatmap. Because the anomaly map is generated directly through the detector's scoring mechanism, it represents intrinsic spatial localisation rather than an approximation produced by a separate post-hoc attribution method. Figure 1. Overview of the PaDiM-based anomaly detection pipeline. Multi-scale feature extraction and location-specific Gaussian estimation are performed during training using normal images only, whereas Mahalanobis distance scoring and anomaly-map visualisation are conducted during inference.

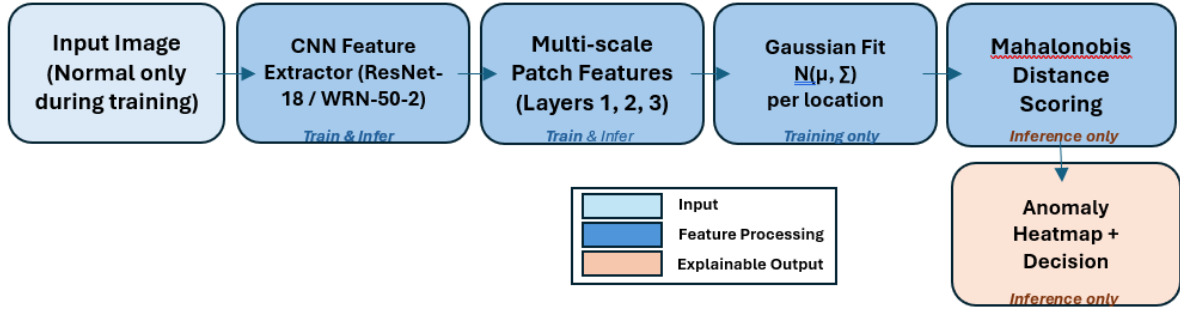


Figure 1. Overview of the PaDiM-Based Anomaly Detection Pipeline. Feature Extraction and Per-Location Gaussian Estimation are Performed During Training on Normal Images Only. Mahalanobis Distance Scoring and Anomaly-Map Rendering are Applied at Inference.

3.2. Multi-Scale Feature Extraction.

All input images were resized to 256×256 pixels before feature extraction, following the preprocessing configuration implemented in anomalib. No centre cropping was applied, ensuring that anomaly maps were generated and evaluated at a consistent resolution of 256×256 pixels. Following the original PaDiM framework [12], feature representations were extracted from three intermediate stages of the pretrained backbone to capture complementary visual characteristics, ranging from shallow texture information to deeper structural features. These feature maps were bilinearly upsampled to a common spatial resolution and concatenated to construct a multi-scale feature descriptor at each location of the feature grid. For Wide-ResNet-50-2, the concatenated feature descriptors were reduced to 256 retained dimensions using the random feature-selection mechanism implemented in anomalib, consistent with the dimensionality-reduction strategy described in PaDiM [12]. This procedure reduces covariance storage requirements from $O(d^2)$ to $O(256^2)$ per spatial location, where d represents the original feature dimensionality. The reduction is achieved without introducing additional learnable parameters or requiring anomalous training samples.

3.3. Per-Location Gaussian Modelling.

Let $x_{ij} \in \mathbb{R}^d$ denote the patch descriptor at spatial position (i, j) . PaDiM [12] estimates the empirical mean μ_{ij} and sample covariance Σ_{ij} from descriptors across all N normal training images. A diagonal regularisation term εI supports numerical invertibility, giving the per-location Gaussian $N(\mu_{ij}, \Sigma_{ij} + \varepsilon I)$.

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

In Equation (1), x_{ij} is the d -dimensional test descriptor at (i, j) ; μ_{ij} and Σ_{ij} are the mean and covariance of normal training descriptors at that location; ε is the diagonal regularisation coefficient (0.01, following Defard et al. [12]); I is the $d \times d$ identity matrix; and higher $M(i,j)$ indicates greater deviation from normality. The scores $\{M(i,j)\}$ form the raw pixel-level anomaly map, bilinearly upsampled to the input resolution. The image-level anomaly score is the maximum of this map.

3.4. Anomaly Map Rendering and Visualisation.

The raw Mahalanobis distance map was normalised to the interval $[0, 1]$ using min–max normalisation and visualised using the Jet colourmap. Lower anomaly scores were represented by cooler colours, whereas higher scores were displayed using warmer colours. The resulting heatmap was superimposed on the original input image through alpha blending, with the blending coefficient set to $\alpha=0.45$. This visualisation provides a spatial representation of the detector's anomaly scores, allowing regions exhibiting substantial deviations from learned normal features to be identified. However, the visual prominence of a region does not necessarily establish its contribution to the image-level anomaly score. Therefore, the counterfactual masking experiment described in Section 3.5 was conducted to assess whether highly scored regions made a distinctive contribution to the detector's output.

3.5. Faithfulness Assessment by Counterfactual Masking.

A deletion-based counterfactual experiment tested whether highly scored pixels contribute to the image-level anomaly score on defective MVTec AD bottle images. For each image, the highest-scoring $k\%$ of pixels, $k \in \{1, 2, 5, 10\}$, were replaced by the per-pixel mean of defect-free training images. This supplies nominal reference appearance while reducing extreme artefacts from constant-value masking [10]. The unchanged PaDiM model then recomputed the image-level score. Two controls used the same masking budget: a contiguous random region of equal pixel count and the lowest-scoring $k\%$ of pixels. Because the image-level score is the map maximum, raw top- k reduction alone is insufficient evidence of faithfulness; the key comparison is with these controls. Decision-flip rate is the proportion of defective images whose recomputed score falls below the threshold from the unmasked test partition. The protocol follows deletion-based attribution evaluation [10, 29, 30]; Section 5.6 reports the results.

3.6. Implementation Details.

All experiments were conducted using anomalib version 2.6.0 [32] and PyTorch version 2.14.0 with CUDA 13.0, running on a Windows operating system with Python 3.12 and a single NVIDIA GeForce RTX 5060 Ti GPU equipped with 16 GB of memory. ImageNet-1K-pretrained ResNet-18 and Wide-ResNet-50-2 architectures were employed as frozen feature extractors. Consequently, the backbone comparison focused on anomaly-detection performance, localisation accuracy, and feature-extraction resource requirements rather than gradient-based training costs. Covariance regularisation was applied using the coefficient specified in Section 3.3. A batch size of 16 was used for both training feature collection and testing. Data augmentation was excluded to preserve the spatial correspondence between feature locations, which is essential for PaDiM's location-specific statistical modelling. Random feature selection was performed using a fixed random seed to improve experimental reproducibility. Backbone architecture, retained feature dimensionality, preprocessing configuration, batch size, and covariance regularisation were controlled within each experimental configuration. The random state was reinitialised before each model construction because the subset of feature channels retained by PaDiM depends on the global random-number-generator state during model instantiation. To verify consistency across experimental results, the ResNet-18 configuration evaluated in Section 5.1 was compared with the

corresponding MVTec AD bottle results presented in Table 1. Both evaluations used identical configurations and produced the same performance values, confirming consistency in the reported results.

4. Experimental Setup

4.1. Datasets.

Three publicly available industrial anomaly-detection benchmarks, namely MVTec AD, VisA, and BTAD, were used to evaluate the proposed PaDiM-based framework. These datasets were selected to represent variations in object geometry, surface texture, anomaly morphology, and image-acquisition conditions. MVTec AD [4] comprises 5,354 high-resolution images across 15 object and texture categories, including 3,629 defect-free training images and 1,725 test images with pixel-precise anomaly annotations covering 73 defect types. Three categories were selected for evaluation: bottle, cable, and hazelnut. The bottle category represents objects with relatively stable spatial positioning, whereas cable exhibits substantial geometric variability. Hazelnut was selected to represent natural surface-texture variability. These categories provide different levels of spatial correspondence and normal-appearance variability, allowing the sensitivity of location-specific Gaussian modelling to be examined. VisA [13] comprises 12 industrial object categories containing diverse anomaly types and substantial variations in visual appearance. Three categories were selected: candle, PCB1, and cashew. The candle category includes localised anomalies involving features such as the wick, whereas PCB1 contains fine-grained defects associated with electronic assemblies. The cashew category represents surface abnormalities, including relatively large fractures. Together, these categories provide complementary examples of anomaly sizes, spatial distributions, and morphological characteristics.

BTAD [14] contains real-world images of three industrial product categories, designated 01, 02, and 03, acquired using industrial image-acquisition systems. All three categories were included in the evaluation. The dataset exhibits variations in illumination, object positioning, and surface appearance, providing an additional assessment of the framework under different industrial imaging conditions. Considering the available computational resources, three categories were purposively selected from each of MVTec AD and VisA rather than evaluating their complete category sets. The selection was based on differences in spatial stability, normal-appearance variability, anomaly size, and defect morphology. In contrast, all three BTAD categories were evaluated. Accordingly, the mean performance values reported for MVTec AD and VisA in Sections 5.1, 5.3, and 5.5 represent selected-category means rather than full-benchmark estimates. The BTAD mean values presented in Sections 5.4 and 5.5 represent the complete three-category benchmark. This distinction was maintained throughout the analysis to ensure transparent interpretation and comparison of the experimental findings.

4.2. Evaluation Metrics.

The performance of the PaDiM-based framework was evaluated at both image and pixel levels to assess anomaly-detection capability and spatial localisation accuracy. Image-level performance was measured using the area under the receiver operating characteristic curve (AUROC), which quantifies the ability to distinguish normal and anomalous images across different decision thresholds. Pixel-level AUROC was used to evaluate the ability of the

generated anomaly maps to discriminate between anomalous and normal pixels over a range of thresholds. Image-level and pixel-level F1-scores were also calculated to assess the balance between precision and recall. For both metrics, the decision thresholds were selected by maximising the corresponding F1-scores on the test partition. Consequently, the resulting F1-scores represent optimistic, test-set-optimised operating points rather than unbiased estimates of deployment performance. The official training and testing partitions provided by each benchmark were maintained throughout the experiments. Only defect-free training images were used to estimate the location-specific Gaussian distributions required by PaDiM. Anomalous images and their corresponding ground-truth masks were used exclusively during testing and performance evaluation, ensuring that defect annotations were not incorporated into the model-training process.

5. Results and Discussion

5.1. Performance on MVTec AD.

Table 1 presents the anomaly-detection and localisation performance of PaDiM using ResNet-18 across the three selected MVTec AD categories. The bottle category achieved the highest image-level performance, with an AUROC of 0.9968 and an F1-score of 0.9841. This strong performance may be attributed to the relatively stable spatial positioning of bottle objects, which facilitates PaDiM's location-specific statistical modelling. In contrast, the hazelnut category exhibited the lowest image-level AUROC of 0.6084, despite achieving a relatively high pixel-level AUROC of 0.9714. This discrepancy indicates that accurate pixel-level anomaly ranking does not necessarily translate into effective image-level discrimination. Extended or low-contrast anomaly responses may fail to generate sufficiently distinctive peak scores, particularly when image-level predictions rely on maximum-score aggregation. The cable category achieved an image-level AUROC of 0.8531, a pixel-level AUROC of 0.9622, and a pixel-level F1-score of 0.4943 (Table 1). Variations in cable routing, strand positioning, and geometric configurations may broaden the learned location-specific normal distributions, resulting in less concentrated anomaly responses. Across the three evaluated categories, the mean pixel-level AUROC reached 0.9713, whereas the image-level AUROC differed by 0.3884 between the bottle and hazelnut categories. These findings demonstrate considerable variation in image-level discrimination despite consistently high pixel-level ranking performance. However, these results represent the selected categories and should not be generalised to the complete MVTec AD benchmark.

Table 1. Per-Category and Mean Anomaly-Detection Performance of Padim (Resnet-18) on the Selected Mvttec AD Categories.

Category	Img AUROC	Img F1	Pix AUROC	Pix F1	Backbone
Bottle	0.9968	0.9841	0.9802	0.6924	ResNet-18
Cable	0.8531	0.8358	0.9622	0.4943	ResNet-18
Hazelnut	0.6084	0.8023	0.9714	0.4684	ResNet-18
Mean	0.8194	0.8741	0.9713	0.5517	—

All results presented in Table 1 were obtained from the authors' experimental runs using anomalib 2.6.0 and PyTorch 2.14.0 on a single NVIDIA GeForce RTX 5060 Ti GPU. To investigate backbone sensitivity, Wide-ResNet-50-2 was evaluated using the same MVTec AD bottle training partition, with 256 retained feature dimensions. Compared with ResNet-18,

Wide-ResNet-50-2 improved image-level AUROC from 0.9968 to 1.0000 and image-level F1-score from 0.9841 to 0.9920. Pixel-level AUROC also increased from 0.9802 to 0.9828, while pixel-level F1-score improved from 0.6924 to 0.7180. However, the retained feature extractor contained approximately 24.9 million parameters for Wide-ResNet-50-2, compared with 2.8 million for ResNet-18, representing an approximately 8.9-fold difference. Although the larger backbone improved all four evaluation metrics, the gains were relatively modest compared with the increase in model size. Their practical significance for deployment was not independently evaluated. The ResNet-18 results used in this comparison correspond to the same experimental configuration reported for the bottle category in Table 1.

5.2. Qualitative Anomaly Map Analysis.

Supplementary Figure S1 presents representative anomaly maps generated by PaDiM for two defect-free bottle images in the upper row and two defective images in the lower row. The defect-free samples generally exhibited diffuse, relatively low anomaly responses, although slightly elevated scores were observed around reflective surface regions. In contrast, the defective samples displayed more concentrated high-score regions that visually corresponded to the locations of surface damage. These observations suggest that the generated anomaly maps can identify spatial deviations associated with visible defects. However, the examples in Supplementary Figure S1 provide qualitative illustrations rather than quantitative evidence of localisation accuracy or explanation faithfulness. Accordingly, Section 5.6 evaluates the contribution of highly scored regions to image-level anomaly scores through counterfactual masking, while Section 5.7 examines cases in which anomaly detection or spatial localisation was unsuccessful.

5.3. Performance on VisA.

Table 2 summarises the performance of PaDiM using ResNet-18 across the three selected VisA categories. The mean pixel-level AUROC reached 0.9835, indicating strong overall discrimination between normal and anomalous pixels within the evaluated subset. Among the three categories, candle achieved the highest image-level AUROC of 0.8812, PCB1 recorded the highest pixel-level AUROC of 0.9870, and cashew achieved the highest pixel-level F1-score of 0.4738. These differences demonstrate that image-level discrimination, pixel-level anomaly ranking, and threshold-dependent localisation performance capture distinct aspects of anomaly-detection capability.

Table 2. Per-Category and Mean Anomaly-Detection Performance of Padim (Resnet-18) on the Selected Visa Categories.

Category	Img AUROC	Img F1	Pix AUROC	Pix F1	Backbone
Candle	0.8812	0.8282	0.9827	0.1897	ResNet-18
PCB1	0.8275	0.7783	0.9870	0.3946	ResNet-18
Cashew	0.8308	0.8255	0.9809	0.4738	ResNet-18
Mean	0.8465	0.8107	0.9835	0.3527	—

Despite achieving a high pixel-level AUROC of 0.9827, the candle category exhibited a substantially lower pixel-level F1-score of 0.1897 (Table 2). This discrepancy may be associated with the small spatial extent of the annotated anomalies relative to the total image

area, which creates substantial pixel-level class imbalance. Consequently, a high pixel-level AUROC does not necessarily indicate accurate spatial overlap between predicted anomaly regions and ground-truth defect masks. Among correctly detected candle anomalies, the mean intersection-over-union (IoU) between thresholded anomaly maps and ground-truth masks was only 0.091, with a median IoU of 0.034. These findings highlight the importance of evaluating spatial overlap alongside ranking-based metrics, particularly for small and localised defects. A more detailed examination of these localisation failures is presented in Section 5.7.

The relatively low pixel-level F1-score for candle was consistent with its limited spatial overlap, as indicated by the mean IoU of 0.091. This relationship is further examined through representative failure cases in Section 5.7.

5.4. Performance on BTAD.

Table 3 presents the performance of PaDiM using ResNet-18 across all three BTAD categories. The mean image-level AUROC reached 0.9381, indicating strong image-level discrimination across the complete three-category benchmark. Category 01 achieved the highest image-level AUROC of 0.9981, whereas category 03 recorded the highest single-category pixel-level AUROC in this study, reaching 0.9965. These findings indicate that PaDiM can effectively distinguish normal and anomalous visual patterns across the evaluated BTAD categories.

Table 3. Per-Category and Mean Anomaly-Detection Performance of Padim (Resnet-18) on the Complete BTAD Benchmark.

Category	Img AUROC	Img F1	Pix AUROC	Pix F1	Backbone
01	0.9981	0.9796	0.9722	0.5811	ResNet-18
02	0.8278	0.9299	0.9605	0.5897	ResNet-18
03	0.9884	0.8056	0.9965	0.5128	ResNet-18
Mean	0.9381	0.9050	0.9764	0.5612	—

In contrast, category 02 achieved a lower image-level AUROC of 0.8278, despite maintaining a relatively high pixel-level AUROC of 0.9605 (Table 3). This difference may reflect variations in normal surface appearance, illumination, or the influence of normal training-sample availability on location-specific covariance estimation. Such factors could increase the overlap between image-level anomaly-score distributions while preserving the ability to rank anomalous pixels within individual images.

Overall, the evaluated PaDiM configuration demonstrated strong image-level discrimination and pixel-level anomaly ranking across the BTAD categories without requiring domain-specific adaptation. Nevertheless, these benchmark results should not be interpreted as evidence of equivalent performance under all industrial production conditions, where acquisition variability, environmental disturbances, and operational requirements may differ substantially.

5.5. Cross-Dataset Comparative Analysis.

Supplementary Table S1 summarises the mean anomaly-detection and localisation performance across the evaluated MVTec AD, VisA, and BTAD categories. Among these evaluated category sets, BTAD achieved the highest mean image-level AUROC of 0.9381, whereas VisA recorded the highest mean pixel-level AUROC of 0.9835. The mean pixel-level

AUROC exceeded 0.97 for the selected MVTec AD categories, selected VisA categories, and complete BTAD benchmark. These findings suggest relatively consistent pixel-level anomaly ranking across the evaluated categories, although image-level discrimination exhibited greater variability. However, differences in category coverage must be considered when interpreting these comparisons. The MVTec AD and VisA results represent purposively selected subsets, whereas the BTAD results cover all three available categories. Therefore, the descriptive comparisons presented in Supplementary Table S1 do not establish that PaDiM is generally insensitive to dataset characteristics or domain differences.

Table 4 provides an indicative comparison between the present findings and previously published anomaly-detection methods evaluated on MVTec AD. To avoid misleading comparisons, the methods are grouped according to their evaluation scope. The upper group presents published results averaged across all 15 MVTec AD categories, whereas the lower group contains results for the three-category subset or the individual bottle category. Consequently, performance values should be interpreted within their respective evaluation scopes rather than directly ranked across groups. The CutPaste, STFPM, and DRAEM results were checked against their corresponding primary publications. For a more closely matched comparison, the original PaDiM-R18-Rd100 implementation reported pixel-level AUROC values of 0.981, 0.958, and 0.977 for bottle, cable, and hazelnut, respectively [12]. These values correspond to a three-category mean of 0.9720, compared with 0.9713 obtained in the present study (Table 4). The absolute difference of 0.0007 indicates closely aligned pixel-level ranking performance across the matched categories. Potential sources of this difference include variations in software implementation, preprocessing procedures, feature selection, and dimensionality reduction. However, the individual contributions of these factors were not independently examined.

Table 4. Indicative Comparison with Published Methods on Mvtec AD, Grouped by Evaluation Scope. Values from Different Scope Blocks are not Directly Comparable.

Method	Img AUROC	Pix AUROC	Paradigm	Reference / Scope
Full MVTec AD benchmark (15 categories)				
CutPaste	0.952	0.960	Self-supervised	Li et al. [23], 15 categories
STFPM	0.955	0.970	Knowledge distillation	Wang et al. [19], 15 categories
DRAEM	0.980	0.973	Reconstruction and discriminative learning	Zavrtanik et al. [17], 15 categories
PatchCore	0.991	0.981	Feature distribution	Roth et al. [6], 15 categories
PaDiM-WR50-Rd550 (original)	0.953	0.975	Feature distribution	Defard et al. [12], 15 categories
PaDiM-R18-Rd100 (original)	n/r	0.967	Feature distribution	Defard et al. [12], 15 categories
Selected categories and individual-category evaluation				
PaDiM-R18-Rd100 (matched subset)	n/r	0.9720	Feature distribution	Defard et al. [12], bottle/cable/hazelnut
PaDiM-ResNet-18 (this study)	0.8194	0.9713	Feature distribution	This study, three-category subset
PaDiM-WRN-50-2 (this study)	1.0000	0.9828	Feature distribution	This study, bottle only

Note: n/r indicates that the corresponding value was not reported.

Supplementary Figure S2 illustrates the four evaluation metrics for the selected MVTec AD and VisA categories and the complete BTAD benchmark. The performance profiles

approach the outer boundary for pixel-level AUROC across all three evaluated category sets, whereas pixel-level F1-scores exhibit greater variation. VisA recorded the lowest mean pixel-level F1-score of 0.3527, while BTAD achieved the highest mean image-level AUROC of 0.9381. These patterns reinforce the distinction between pixel-level ranking and threshold-dependent localisation accuracy. Nevertheless, Supplementary Figure S2 represents only the evaluated category sets and should not be interpreted as a comparison of complete benchmarks for MVTec AD and VisA.

5.6. Faithfulness Assessment of the Anomaly Maps.

Table 5 presents the results of the counterfactual masking experiment described in Section 3.5, conducted on 63 defective MVTec AD bottle test images using ResNet-18. The experiment examined whether replacing highly scored image regions with the corresponding training-set mean values produced greater changes in image-level anomaly scores than masking equally sized random or low-scoring regions. Masking the highest-scoring pixels consistently reduced the mean image-level anomaly score across all four masking budgets (Table 5). The mean score reduction increased from 0.73% at a masking budget of 1% to 13.35% at a budget of 10%. In comparison, the random-control condition produced mean score reductions of -1.04% , -2.46% , -2.08% , and -0.71% for masking budgets of 1%, 2%, 5%, and 10%, respectively. These negative values indicate that random masking increased the image-level anomaly score on average. Meanwhile, masking the lowest-scoring regions produced no change in the mean image-level score across all four budgets.

Table 5. Counterfactual Masking Results for the Mvtec AD Bottle Category Using Resnet-18 (N = 63 Defective Test Images). Negative Score Reductions Indicate Increases in Image-Level Anomaly Scores Following Masking.

Masking condition	k (%)	Mean score reduction (%)	Median (%)	Decision-flip rate (%)	n
Top-k Mahalanobis region masked	1	0.73	5.40	1.59	63
Size-matched random region (control)	1	-1.04	0.00	0.00	63
Inverse lowest-score region (control)	1	0.00	0.00	0.00	63
Top-k Mahalanobis region masked	2	5.99	10.34	3.17	63
Size-matched random region (control)	2	-2.46	0.00	1.59	63
Inverse lowest-score region (control)	2	0.00	0.00	0.00	63
Top-k Mahalanobis region masked	5	10.98	16.01	11.11	63
Size-matched random region (control)	5	-2.08	0.00	0.00	63
Inverse lowest-score region (control)	5	0.00	0.00	0.00	63
Top-k Mahalanobis region masked	10	13.35	17.81	6.35	63
Size-matched random region (control)	10	-0.71	-0.00	0.00	63
Inverse lowest-score region (control)	10	0.00	0.00	0.00	63

The differences between the top-scoring and random-control conditions became more pronounced as the masking budget increased. At a masking budget of 1%, the difference in mean score reduction was 1.77 percentage points, increasing to 14.06 percentage points at 10%. These results indicate that regions assigned higher Mahalanobis scores made a greater contribution to the image-level anomaly score than the two control conditions tested. However, the influence of masking remained partial. Even when 10% of the highest-scoring pixels were replaced, the mean image-level score decreased by only 13.35%. Furthermore, the decision-

flip rate increased to 11.11% at a masking budget of 5% but subsequently declined to 6.35% at 10%. This non-monotonic pattern suggests that residual anomaly evidence and potential artefacts introduced by mean-value replacement may influence the recomputed anomaly scores. As discussed by Hooker et al. [10], removing or replacing image features can alter the input distribution and affect model predictions independently of the relevance of the removed features. Although training-mean replacement may reduce extreme masking artefacts, it does not completely eliminate this potential confounding effect.

The findings in Table 5 support a restricted interpretation of the anomaly maps. Specifically, the highest-scoring pixels contributed more strongly to the maximum-pooled image-level anomaly score than the random and inverse controls. Nevertheless, removing these regions generally did not reverse the image-level anomaly decision. Although masking the lowest-scoring pixels did not affect the maximum anomaly score within the tested budgets, this does not imply that all non-highlighted pixels are irrelevant to the model's behaviour. Given that the experiment was limited to one defect category and one masking strategy, the results provide supporting evidence of score sensitivity rather than comprehensive validation of explanation faithfulness.

5.7. Failure Case Analysis.

To complement the successful qualitative examples presented in Section 5.2, a systematic failure-case analysis was conducted to identify limitations in anomaly detection and spatial localisation. Cases were selected according to predefined criteria rather than subjective visual inspection. For each category, defect-free images with the highest anomaly scores were identified as severe false-positive cases, while defective images with the lowest anomaly scores were selected as severe false-negative cases. Correctly classified anomalous images exhibiting the lowest ground-truth IoU values were examined as localisation failures. Per-image anomaly scores, predictions, and IoU values were retained and are available from the corresponding author upon reasonable request. Supplementary Table S2 summarises the classification outcomes at F1-optimal thresholds, while Figure 2 presents representative failure cases. False negatives in geometrically variable objects. The MVTec AD cable category exhibited 11 false negatives among 92 defective test images at the F1-optimal decision threshold. Missing-wire defects accounted for four missed cases among ten samples, while cable-swap defects contributed another four misses among twelve samples. The remaining false negatives involved cut-inner-insulation defects (two of fourteen) and poke-insulation defects (one of ten). In contrast, all defective images associated with bent-wire, combined, cut-outer-insulation, and missing-cable categories were correctly detected. The two lowest-scoring defective images involved missing-wire anomalies, with image-level anomaly scores of 0.459 and 0.460 and corresponding IoU values of 0.057 and 0.090. As illustrated in Figure 2(a), their anomaly responses were distributed across the conductor region rather than concentrated around a dominant peak. This behaviour is consistent with the challenges of modelling objects exhibiting substantial geometric variation through location-specific normal feature distributions, although the effects of other implementation factors were not independently isolated.

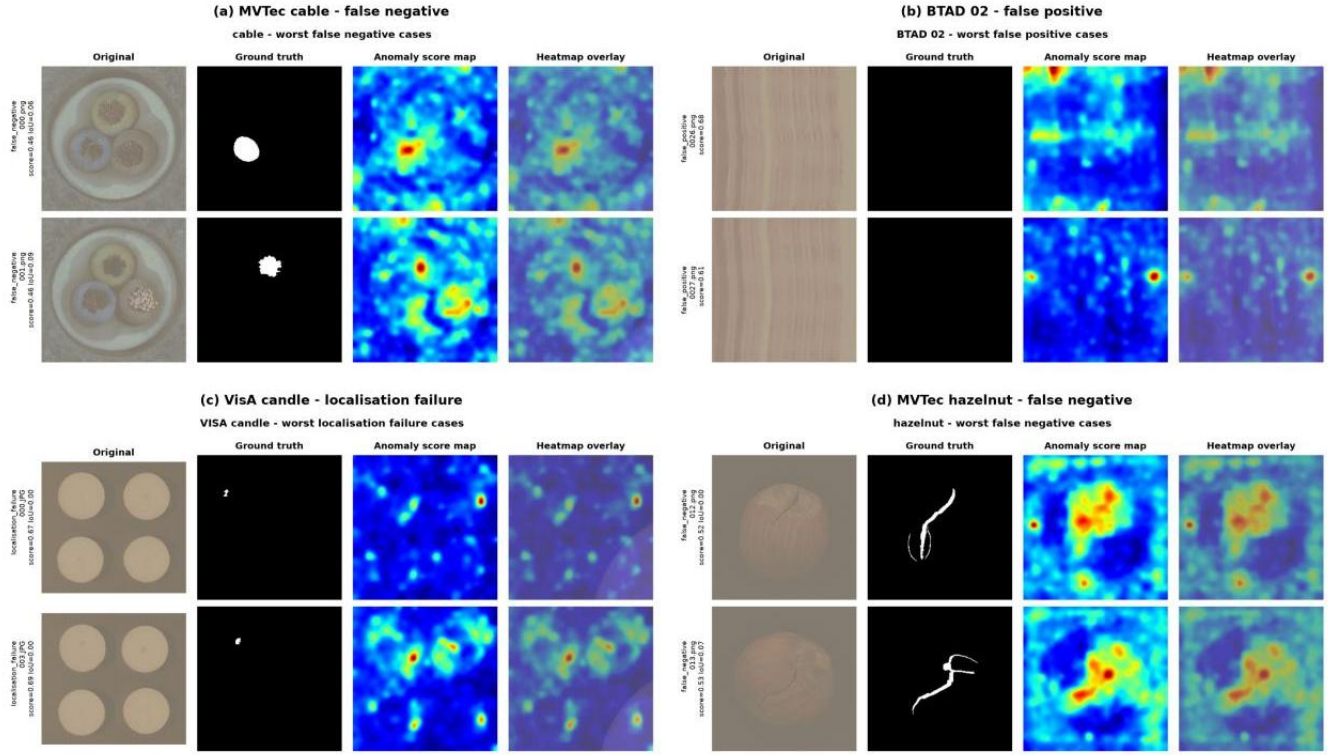


Figure 2. Representative Failure Cases, Selected by Rule Rather Than By Inspection. (A) False Negatives On Mvtec AD Cable, Where Response is Dispersed Across the Conductor Region. (B) False Positives on BTAD Category 02, Where Surface Variation Produces Compact Hotspots on Defect-Free Material. (C) Localisation Failure on Visa Candle, Where Correctly Classified Images are Localised Away from the Ground-Truth Defect (IoU = 0.000). (D) False Negatives on Mvtec AD Hazelnut, Where Diffuse Low-Contrast Response Depresses the Max-Pooled Image-Level Score. Column Order: Original Image | Ground-Truth Mask | Anomaly Score Map | Heatmap Overlay.

False positives associated with acquisition variability. At the test-set F1-optimal threshold, all 30 defect-free images in BTAD category 02 were incorrectly classified as anomalous. Similarly, 23 of 40 normal hazelnut images were classified as defective. Representative BTAD category 02 examples exhibited compact, high-scoring regions resembling genuine anomalies, despite the absence of annotated defects (Figure 2(b)). Variations in illumination, specular reflection, surface finish, and the representativeness of normal training samples may contribute to these false-positive responses. These observations demonstrate that visually coherent anomaly maps do not necessarily correspond to actual defects and highlight the importance of appropriate decision-threshold calibration.

BTAD category 02 contained 200 anomalous and 30 normal test images. Under this class imbalance, the F1-maximising threshold produced a high image-level F1-score of 0.9299 despite incorrectly classifying every normal image as anomalous. This finding illustrates the optimistic nature of test-set-optimised F1-scores discussed in Section 4.2. For practical industrial deployment, operating thresholds should therefore be calibrated independently, considering expected anomaly prevalence and the relative costs of false-positive and false-negative decisions.

Localisation failures despite correct image classification. The VisA candle category achieved a pixel-level AUROC of 0.9827 but exhibited a mean IoU of only 0.091 among correctly detected anomalous images, with a median of 0.034 and a minimum of 0.000. As illustrated in Figure 2(c), some defective images were correctly classified at the image level

even though the highest-scoring regions were located away from the annotated wick defects. These examples demonstrate that correct image-level classification can coexist with inaccurate spatial localisation. Consequently, high pixel-level AUROC should not be interpreted as sufficient evidence of accurate defect overlap, particularly for small and localised anomalies.

Partial localisation of extended and low-contrast defects. The MVTec AD hazelnut category achieved a pixel-level AUROC of 0.9714 but a substantially lower image-level AUROC of 0.6084. At the F1-optimal threshold, six of 70 defective images were missed, while 23 of 40 normal images were incorrectly classified as anomalous. The false negatives included crack defects (two of eighteen), print defects (two of seventeen), cut defects (one of seventeen), and hole defects (one of eighteen). Mean image-level anomaly scores across defect types ranged from 0.684 for print defects to 0.808 for hole defects. As shown in Figure 2(d), some defective regions generated diffuse or relatively low-contrast anomaly responses rather than distinct localised peaks. Because the image-level score was determined by the maximum value of the anomaly map, spatially extended but moderately scored anomalies could produce insufficient image-level discrimination. These findings suggest that alternative aggregation strategies incorporating the spatial extent of anomaly responses, such as high-quantile mean pooling, warrant further investigation.

Figure 2. Representative failure cases selected using predefined criteria rather than subjective visual inspection. (a) False negatives in MVTec AD cable, where anomaly responses are dispersed across the conductor region. (b) False positives in BTAD category 02, where surface variability produces compact high-scoring regions on defect-free materials. (c) Localisation failures in VisA candle, where correctly classified anomalous images exhibit high-scoring regions outside the ground-truth defects (IoU = 0.000). (d) False negatives in MVTec AD hazelnut, where diffuse, low-contrast anomaly responses produce relatively low maximum-pooled image-level scores. Columns are arranged from left to right as follows: original image, ground-truth mask, anomaly score map, and heatmap overlay.

5.8. Limitations.

Several limitations should be considered when interpreting the findings of this study. First, only three of the 15 MVTec AD categories and three of the 12 VisA categories were purposively selected for evaluation. Consequently, the reported mean values represent selected-category performance rather than complete-benchmark estimates and should not be directly compared with published full-benchmark averages. Although all three BTAD categories were evaluated, comprehensive evaluation across the remaining MVTec AD and VisA categories is necessary to establish broader generalisability. Second, the image-level and pixel-level F1-scores presented in Tables 1–3 were calculated using thresholds optimised on the test partition. These values therefore represent optimistic operating points rather than unbiased estimates of deployment performance. The failure-case analysis in Section 5.7 further demonstrated that high F1-scores can coexist with substantial false-positive rates under class-imbalanced conditions. Third, PaDiM's location-specific covariance modelling requires $O(d^2)$ storage per spatial location, which may limit scalability when high-dimensional feature descriptors are used without dimensionality reduction. Furthermore, the method assumes sufficient spatial correspondence among normal training images. The detection failures observed in the cable category illustrate the challenges associated with strong geometric variability.

The interpretation of anomaly maps also remains limited. The counterfactual masking experiment in Section 5.6 was restricted to one category and a single deletion-based perturbation strategy. Although masking the highest-scoring 10% of pixels reduced the mean image-level anomaly score by 13.35%, the reduction remained partial, and most anomaly decisions were unchanged (Table 5). The non-monotonic decision-flip rate further suggests that residual anomaly evidence and masking-induced artefacts may influence the observed responses. In addition, localisation accuracy varied considerably across categories, with mean IoU values ranging from 0.488 for MVTec AD bottle to 0.091 for VisA candle. These differences indicate that the reliability of spatial anomaly maps depends on object characteristics and defect morphology. Accordingly, the generated maps should be interpreted as spatial representations of deviations from learned normal features rather than comprehensive explanations of the model's decision-making process.

6. Conclusion

This study evaluated the performance of PaDiM for industrial anomaly detection and spatial localisation across selected MVTec AD and VisA categories and the complete BTAD benchmark. The results demonstrated consistently high pixel-level anomaly ranking performance across the evaluated categories, whereas image-level discrimination exhibited greater variability. These findings highlight the importance of assessing image-level detection and pixel-level localisation separately rather than assuming that strong performance in one necessarily indicates comparable performance in the other. The backbone comparison using the MVTec AD bottle category showed that Wide-ResNet-50-2 achieved modest improvements over ResNet-18 across all four evaluation metrics, despite requiring a substantially larger feature extractor. This finding suggests that backbone complexity should be considered alongside detection accuracy, localisation performance, and computational resource requirements when selecting anomaly-detection models for practical applications. Counterfactual masking further demonstrated that highly scored spatial regions contributed more strongly to image-level anomaly scores than the random and inverse masking controls. However, the observed effects remained partial, and most anomaly decisions were unchanged following masking. Together with the failure-case analysis, these findings support interpreting PaDiM anomaly maps as category-dependent spatial representations of deviations from learned normality rather than complete explanations of the model's decision-making process. The substantial differences observed between pixel-level AUROC and ground-truth defect overlap further emphasise the need to report overlap-based localisation metrics alongside ranking-based measures. Future research should extend the evaluation to all benchmark categories, improve spatial alignment for geometrically variable objects, investigate alternatives to maximum-score aggregation, establish independently calibrated operating thresholds, and assess explanation faithfulness across additional categories and perturbation strategies.

Data and Code Availability

All results reported in this paper were produced by the authors' own experimental runs. Evaluation notebooks, configuration and utility modules, per-image test records (score, prediction, and intersection-over-union), and counterfactual masking results are available from the corresponding author on reasonable request.

Author Contributions

A. Melleng: conceptualisation, methodology, software development, formal analysis, investigation, data curation, writing - original draft preparation, writing - review and editing. Soffian Mohammad and M. Yusuf: writing - review and editing, validation. All authors have read and approved the manuscript.

Acknowledgments

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

Conflict of Interest

The authors declare no conflict of interest.

Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.53623/amms.v3i1.1281>.

References

- [1] Tao, X.; Gong, X.; Zhang, X.; Yan, S.; Adak, C. (2022). Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement*, 71, 1–21. <https://doi.org/10.1109/TIM.2022.3196436>.
- [2] Liu, J.; Xie, G.; Wang, J.; Li, S.; Wang, C.; et al. (2024). Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21, 104–135. <https://doi.org/10.1007/s11633-023-1459-Z>.
- [3] Tao, X.; Zhang, D.; Ma, W.; Liu, X.; Xu, D. (2018). Automatic metallic surface defect detection and recognition with convolutional neural networks. *Applied Sciences*, 8, 1575. <https://doi.org/10.3390/app8091575>.
- [4] Bergmann, P.; Fauser, M.; Sattlegger, D.; Steger, C. (2019). MVTec AD — A comprehensive real-world dataset for unsupervised anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9584–9592. <https://doi.org/10.1109/CVPR.2019.00982>.
- [5] Bergmann, P.; Batzner, K.; Fauser, M.; Sattlegger, D.; Steger, C. (2021). The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129, 1038–1059. <https://doi.org/10.1007/s11263-020-01400-4>.
- [6] Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; et al. (2022). Towards total recall in industrial anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14298–14308. <https://doi.org/10.1109/CVPR52688.2022.01392>.
- [7] Xie, G.; Wang, J.; Liu, J.; Lyu, J.; Liu, Y.; et al. (2024). IM-IAD: Industrial image anomaly detection benchmark in manufacturing. *IEEE Transactions on Cybernetics*, 54, 2720–2733. <https://doi.org/10.1109/TCYB.2024.3357213>.
- [8] Gudovskiy, D.; Ishizaka, S.; Kozuka, K. (2022). CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1819–1828. <https://doi.org/10.1109/WACV51458.2022.00188>.

- [9] Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; et al. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31. <https://doi.org/10.48550/arXiv.1810.03292>.
- [10] Hooker, S.; Erhan, D.; Kindermans, P.-J.; Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, 32. <https://doi.org/10.48550/arXiv.1806.10758>.
- [11] Jacovi, A.; Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4198–4205. <https://doi.org/10.18653/v1/2020.acl-main.386>.
- [12] Defard, T.; Setkov, A.; Loesch, A.; Audigier, R. (2021). PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges; Lecture Notes in Computer Science*, Vol. 12664; Springer: Cham, Switzerland; pp. 475–489. https://doi.org/10.1007/978-3-030-68799-1_35.
- [13] Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; Dabeer, O. (2022). SPot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *Computer Vision – ECCV 2022; Lecture Notes in Computer Science*, Vol. 13690; Springer: Cham, Switzerland; pp. 392–408. https://doi.org/10.1007/978-3-031-20056-4_23.
- [14] Mishra, P.; Verk, R.; Fornasier, D.; Piciarelli, C.; Foresti, G.L. (2021). VT-ADL: A vision transformer network for image anomaly detection and localization. *Proceedings of the 30th IEEE International Symposium on Industrial Electronics (ISIE)*, 1–6. <https://doi.org/10.1109/ISIE45552.2021.9576231>.
- [15] Schlegl, T.; Seeböck, P.; Waldstein, S.M.; Schmidt-Erfurth, U.; Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Information Processing in Medical Imaging (IPMI); Lecture Notes in Computer Science*, Vol. 10265; Springer: Cham, Switzerland; pp. 146–157. https://doi.org/10.1007/978-3-319-59050-9_12.
- [16] Akçay, S.; Atapour-Abarghouei, A.; Breckon, T.P. (2019). GANomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision – ACCV 2018; Lecture Notes in Computer Science*, Vol. 11363; Springer: Cham, Switzerland; pp. 622–637. https://doi.org/10.1007/978-3-030-20893-6_39.
- [17] Zavrtanik, V.; Kristan, M.; Skočaj, D. (2021). DRÆM — A discriminatively trained reconstruction embedding for surface anomaly detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8310–8319. <https://doi.org/10.1109/ICCV48922.2021.00822>.
- [18] Yapp, E.K.Y.; Doan, N.C.N. (2024). Anomaly detection on MVTec AD using VQ-VAE-2. *Procedia CIRP*, 130, 1809–1814. <https://doi.org/10.1016/j.procir.2024.10.320>.
- [19] Wang, G.; Han, S.; Ding, E.; Huang, D. (2021). Student-teacher feature pyramid matching for anomaly detection. *Proceedings of the 32nd British Machine Vision Conference (BMVC)*, paper 1273. <https://doi.org/10.5244/C.35.349>.
- [20] Bergmann, P.; Fauser, M.; Sattlegger, D.; Steger, C. (2020). Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4182–4191. <https://doi.org/10.1109/CVPR42600.2020.00424>.
- [21] Deng, H.; Li, X. (2022). Anomaly detection via reverse distillation from one-class embedding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9727–9736. <https://doi.org/10.1109/CVPR52688.2022.00951>.
- [22] Li, Z.; Ge, Y.; Yue, X.; Meng, L. (2024). MCAD: Multi-classification anomaly detection with relational knowledge distillation. *Neural Computing and Applications*, 36, 14543–14557. <https://doi.org/10.1007/s00521-024-09838-0>.

- [23] Li, C.-L.; Sohn, K.; Yoon, J.; Pfister, T. (2021). CutPaste: Self-supervised learning for anomaly detection and localization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9659–9669. <https://doi.org/10.1109/CVPR46437.2021.00954>.
- [24] Batzner, K.; Heckler, L.; König, R. (2024). EfficientAD: Accurate visual anomaly detection at millisecond-level latencies. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 127–137. <https://doi.org/10.1109/WACV57701.2024.00020>.
- [25] Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626. <https://doi.org/10.1109/ICCV.2017.74>.
- [26] Ribeiro, M.T.; Singh, S.; Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>.
- [27] Ghorbani, A.; Abid, A.; Zou, J. (2019). Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 3681–3688. <https://doi.org/10.1609/aaai.v33i01.33013681>.
- [28] Krishna, S.; Han, T.; Gu, A.; Wu, S.; Jabbari, S.; et al. (2024). The disagreement problem in explainable machine learning: A practitioner’s perspective. *Transactions on Machine Learning Research*.
- [29] Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; Müller, K.-R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28, 2660–2673. <https://doi.org/10.1109/TNNLS.2016.2599820>.
- [30] Petsiuk, V.; Das, A.; Saenko, K. (2018). RISE: Randomized input sampling for explanation of black-box models. *Proceedings of the British Machine Vision Conference (BMVC)*, paper 1064. <https://doi.org/10.48550/arXiv.1806.07421>.
- [31] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>.
- [32] Akçay, S.; Ameln, D.; Vaidya, A.; Lakshmanan, B.; Ahuja, N.; et al. (2022). Anomalib: A deep learning library for anomaly detection. *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 1706–1710. <https://doi.org/10.1109/ICIP46576.2022.9897283>.



© 2027 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).