Shama Ashraf

Visual Anomaly Detection in Production Line

Metal Sleeve Dataset under Real Conditions

School of Technology and Innovations
Master's Program in Smart Energy
Robotics

Vaasa 2026

UNIVERSITY OF VAASA
School of Technology and Innovations
Author: Shama Ashraf
Title of the thesis: Visual Anomaly Detection in Production Line
Degree: Master
Degree Programme: Master's Program in Smart Energy Robotics
Supervisor: Jani Boutellier and Masud Fahim
Year: 2026 **Pages:** 66

ABSTRACT:

Modern manufacturing increasingly relies on automated visual inspection to detect surface defects, yet real industrial environments remain difficult for anomaly detection systems. Production imagery rarely behaves like benchmark data: surfaces are reflective, lighting is uneven, defects can be subtle, and samples vary naturally from one another. This thesis examines deep learning approaches to visual anomaly detection in industrial surface inspection, with a particular focus on a refined SimpleNet pipeline evaluated on a custom dataset of reflective steel sleeves. The study addresses a key limitation in current industrial anomaly detection research: many published results are obtained on controlled benchmark datasets such as MVTec AD, which do not fully represent the challenges of real production data. To examine this gap, the sleeve dataset was constructed in an MVTec-style format, with manually produced pixel-level masks for anomalous samples. Refined SimpleNet was trained on only defect-free images and evaluated alongside INP-Former, a transformer-based baseline. The work also includes hyperparameter tuning, augmentation experiments, qualitative error analysis, and a data leakage test. The results show that SimpleNet performs well on the MVTec AD categories, indicating that its architecture is dependable. However, the sleeve dataset performs worse, suggesting the challenge lies more in the data than in the model's architecture. INP-Former provides stronger defect localisation but weaker image-level classification. The augmentation study shows that strong geometric transformations improve image-level discrimination, while offline augmentation is more useful for localisation. The data leakage experiment further demonstrates how easily performance can be inflated on small custom datasets if the train-test split is not carefully controlled. Overall, the thesis shows that unsupervised anomaly detection can be useful for custom industrial inspection tasks, but only when dataset design, image capture conditions and evaluation discipline are treated as central parts of the method. The study concludes with practical recommendations for building new MVTec-style datasets for reflective metallic parts and highlights the need for larger, more realistic industrial datasets.

KEYWORDS: Visual Anomaly Detection, Deep Learning, SimpleNet, Industrial Inspection, Surface Defect Detection, INP-Former, MVTec AD, Reflective metal inspection, tiny defects, Data augmentation, Data leakage

Contents

List of Abbreviations

AD	Anomaly Detection
AI	Artificial Intelligence
AUPRO	Area Under the Per-Region Overlap Curve
AUROC	Area Under the Receiver Operating Characteristic curve
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
I-AUROC	Image-level AUROC
INP	Intrinsic Normal Prototype
MVTec AD	MVTec Anomaly Detection (dataset)
OK / NOK	Acceptable / Not Acceptable sample
P-AUROC	Pixel-level AUROC
PRO	Per-Region Overlap
ROC	Receiver Operating Characteristic
VAE	Variational Autoencoder
WRN-50	Wide Residual Network with 50 layers

List of Figures

List of Tables

1. Introduction

1.1 Background

Quality inspection still consumes a surprising amount of human attention in modern factories. Even in highly automated lines and almost any modern factory, you will still find a person standing at the end of a conveyor belt, picking up each part and deciding whether it is good enough to ship. This surprises a lot of people. We tend to picture manufacturing as fully automated, but quality inspection has proven stubbornly resistant to that assumption. Industrial anomaly detection, the problem of teaching a machine to spot something wrong with a product, has become one of the more interesting corners of applied computer vision for exactly this reason.

Catching surface defects early matters for obvious reasons. Scrap rates go down, customers stop receiving bad products, and you need fewer people staring at parts all day. Machine vision systems have been around for decades and have genuinely replaced human inspectors in certain narrow tasks, especially where you just need to check a dimension or confirm something is present. The trouble is that rule-based systems are fragile. Whenever the product changes, the lighting changes or a new defect type appears, the rules must be rewritten from scratch.

The progress in deep learning over the last decade has shifted the conversation. Convolutional neural networks can learn defect-relevant features directly from images and have been shown to outperform hand-crafted feature pipelines on a wide range of industrial inspection benchmarks (Weimer, Scholz-Reiter, and Shpitalni, 2016). However, the most successful supervised models still require thousands of labelled defective images, which do not exist for most products. A well-running production line is, by design, supposed to produce as few defective parts as possible, and the few defective parts that are produced often look nothing like each other. This is why the unsupervised setting has become so attractive in the industrial anomaly detection community: the model is

trained only on images of good parts and is then expected to flag anything abnormal during inference.

The release of the MVTec Anomaly Detection dataset in 2019 gave researchers a proper benchmark for the first time (Bergmann et al., 2019). Fifteen industrial categories, pixel-level defect masks, and standardised splits made meaningful comparisons between methods possible in a way that had not existed before. Progress on this benchmark has been fast. PatchCore (Roth et al., 2022), PaDiM (Defard et al., 2021), and SimpleNet (Liu et al., 2023) have driven image-level detection accuracy high enough that several MVTec categories now appear effectively saturated. EfficientAD has since demonstrated that strong accuracy need not come at the cost of inference speed (Batzner et al., 2024). As a result, headline benchmark numbers are increasingly difficult to interpret on their own terms. MVTec performance has plateaued to the point where improvements measured in fractions of a per cent say very little about whether a system will actually work in a production line. A growing number of researchers have made exactly this argument: the gap between performing well on MVTec and on a real production line remains uncomfortably large (Wang et al., 2024).

The practical case for automated visual inspection is not hard to make. These systems can continuously detect surface defects, misalignments, and other problems without slowing the line or requiring a break. Moreover, the improvements in deep learning over the past several years have made them genuinely more capable; models now learn meaningful representations directly from image data rather than relying on someone to specify in advance what a defect should look like, which is a significant step forward from where rule-based systems left things. However, deploying such models in real industrial environments remains challenging because the types and appearances of defects are unknown and varied, and the number of available defect samples is limited. The manufacturing sector, particularly in automotive, electronics and heavy industry, has adopted AI-assisted inspection technologies to maintain competitiveness, but limited labelled data and strict privacy regulations hinder rapid deployment

1.2 Motivation and industrial context

This thesis was written in the context of an automated inspection task on a real industrial sleeve dataset. The sleeve is a small metal part with a steel-polished, partially reflective surface. The kinds of issues that need to be found are mostly small surface problems: minor scratches, tiny dents that only show as a slight change in colour tone, light discolouration, and small uneven areas at the angled edges.

None of the issues found here is serious enough to cause any major structural problems. The main shape of the sleeve remains undamaged no matter what is seen on the outside, and most of the problems noticed are just small flaws in certain areas, not anything that would affect how the part works. What makes it hard to spot is exactly this quietness. These problems are usually very small, often taking up only a small part of the image, and they can look almost the same as regular surface shine or the kinds of texture differences that happen naturally during production. A human inspector or a model can easily look straight at one and see nothing unusual. However, from the factory's perspective, any one of them is sufficient grounds to reject the part. The threshold for what counts as acceptable is strict, even when the defect itself is visually unassuming.

From an academic point of view, the sleeve is interesting because it breaks several assumptions that benchmark-oriented anomaly detection methods quietly rely on. First, the surface is reflective, so photographing the same part twice can produce very different highlight patterns. Second, the defects are tiny: a typical dent or scratch covers approximately 0.08% of the image area, which is well below what most patch-based methods are tuned for. Third, the dataset was acquired with an inexpensive webcam under a ring light, which produces uneven illumination across the field of view and is far from the studio-style imaging used to build MVTec AD. Finally, the number of available samples is small, especially the number of normal test images, which makes the image-level evaluation statistically fragile.

1.3 Research Gap

Despite the strong performance of recent unsupervised industrial anomaly detection methods on benchmark datasets such as MVTec AD, there remains a gap between benchmark-level results and practical deployment under real production-line conditions. Many existing methods are evaluated on well-controlled MVTec datasets with relatively stable imaging conditions. At the same time, less attention has been paid to small custom datasets involving reflective metallic surfaces, subtle surface defects, illumination drift, and a limited number of normal and defective samples. In particular, it remains unclear how reliably high-performing methods such as SimpleNet and newer transformer-based approaches transfer from standard benchmark categories to a real-world sleeve inspection task. This thesis addresses this gap by evaluating these methods on a custom dataset of reflective steel sleeves and by analysing the effects of augmentation, dataset composition, data leakage, and image-level decision reliability.

1.4 Problem statement

Unsupervised anomaly detection methods achieve near-perfect image-level AUROC on standard MVTec AD categories. However, their performance often drops when applied to custom industrial datasets that contain reflective metallic parts, tiny defects, illumination drift, and noisy normal samples. This drop is especially visible at the image level. Pixel-level localisation may still identify local abnormal regions, but converting pixel-level anomaly information into a reliable image-level decision remains challenging. Therefore, the problem is not only related to model architecture. It also depends on the choice of data augmentation, the stability of the discriminator, the separation between training and testing data, and the imaging conditions under which the dataset is collected.

1.5 Research Questions

The work is built around five questions that arose directly from trying to make anomaly detection work on a real part rather than on a benchmark dataset.

1. The first is how well a refined version of SimpleNet performs on the sleeve data. SimpleNet reports near-perfect scores on standard MVTec categories, so the question is whether any of those transfers to a custom dataset built around a reflective metal surface with subtle, localised defects.

2. The second concerns the image-level AUROC and how much the test set composition influences it. Unlike MVTec AD, where images are captured under controlled studio conditions with consistent lighting and standardised splits, the sleeve dataset was collected in a real production setting using an inexpensive webcam under ring lighting. Normal samples in this dataset shift in appearance from one capture to the next in ways that simply do not occur in MVTec AD, so the image-level score is quite sensitive to which images are included in training versus testing. Getting a handle on how much that split can inflate the number is important before drawing any conclusions.

3. The third question brings in INP-Former, a newer transformer-based method, and puts it head-to-head with SimpleNet using the same sleeve images. The question, therefore, is whether substituting a transformer-based model produces any meaningful gain or whether the inherent challenges of the sleeve dataset limit both architectures to a comparable degree.

4. That naturally leads into the fourth question, which gets at something more fundamental. Most people assume that, for a model that precisely highlights the right pixels in its anomaly map at the image level, localisation quality and classification reliability go hand in hand. The sleeve experiments directly challenge this assumption. How tightly does pixel-level precision connect to the final binary decision, and what happens to that relationship when the defects are small, subtle, and sitting on a surface that reflects light unpredictably?

5. The fifth and final question is the most practical. Based on everything the experiments show, what would you recommend to someone trying to build a new MVTec-style anomaly detection class for a reflective metal part from scratch?

1.6 Significance of the Study

This study contributes to the field of industrial anomaly detection by evaluating the transferability of high-performing benchmark methods to a real industrial sleeve inspection dataset. First, it provides an empirical analysis of SimpleNet on reflective metallic parts with small and subtle surface defects, showing where the method holds up and where it falls short under real production conditions. Second, it examines how different augmentation strategies affect model performance in a setting with very few samples, with a particular focus on whether a given strategy helps at the image level, the pixel level, or both. Third, it runs a deliberate data-leakage experiment to measure how far the image-level AUROC can shift when the train-test split is not handled carefully, putting a concrete number on something that is often only warned about in passing. Fourth, it conducts a defect-wise failure analysis covering scratches, dents, colour anomalies, and bevel irregularities, providing a clearer picture of which defect types each method struggles with and why. Finally, the thesis provides practical recommendations for developing a new MVTec-style anomaly-detection class for reflective metal components.

2. Literature Review

This chapter situates the thesis in the existing literature on industrial visual anomaly detection. The chapter is organised in the order in which the field has historically developed. It begins with the classical machine-vision approaches that preceded deep learning, then moves through the three main families of deep anomaly detection methods (reconstruction-based, generative, and embedding-based), and finally turns to the two methods used in the experimental part of this thesis: SimpleNet and INP-Former. The chapter closes with a survey of the most relevant benchmark datasets and a short discussion of data augmentation and data leakage in the anomaly detection setting.

2.1 Industrial anomaly detection: a brief history

Automated optical inspection originated as an engineering problem long before it became a machine learning one. The earliest systems relied on classical image-processing techniques such as thresholding, morphological operations, edge detection, and template matching to flag deviations from a reference image (Malamas et al., 2003). These methodologies identify defects well when image quality is consistent and the defects are large and clear, which is why they are still used for tasks like checking whether a label is present or whether a screw is in the right position. However, they have trouble with small defects, rough surfaces, and parts that vary in appearance from one run to the next. The first type of learning-based methods used manually created features like Local Binary Patterns (Ojala, Pietikäinen, and Mäenpää, 2002), Histograms of Oriented Gradients (Dalal & Triggs, 2005), and Gabor filter banks, along with classifiers such as support vector machines.

Weimer, Scholz-Reiter, and Shpitalni (2016) were among the first to fully replace those manual features with features learned automatically through deep learning. Their work showed that a relatively simple CNN trained on a few thousand labelled defect images could beat carefully engineered feature pipelines on industrial inspection benchmarks. From that point onwards, the centre of gravity of the field shifted to deep learning, and

the question turned from “how do we design good features for defect X” to “how do we train a good model when we do not have any examples of defect X”.

The lack of defective training samples is the dominant constraint in industrial anomaly detection. Production lines are tuned to minimise defects, so, by construction, defective parts are rare; defect types are also open-ended, in the sense that a new failure mode can appear in production at any time. Both properties make supervised classification impractical for most industrial applications. This is the reason why the unsupervised setting, in which the model only sees good images during training, has become the dominant formulation in the field (Bergmann et al., 2019; Roth et al., 2022).

2.2 Deep anomaly detection methods

Most modern unsupervised industrial anomaly detection methods fall into one of three broad families. The taxonomy is not perfectly clean; several recent methods sit in between two families, but it is a useful map of the field. The main categories of anomaly detection methods are reconstruction-based, synthesis-based, and embedding-based approaches, as shown in Figure 2.1. Overview of anomaly detection methods categorised into three main approaches: (1) Reconstruction-Based methods, including Autoencoders & GANs, which detect anomalies by measuring reconstruction error, and DAGAN, which employs adversarial training with an encoder-decoder-discriminator architecture to generate and detect anomalies; (2) Synthesizing-Based methods, represented by DRAM, which synthesizes fake anomalies from normal samples to train a detection model capable of distinguishing normal from anomalous regions; and (3) Embedding-Based methods, represented by PaDiM, which leverages a pre-trained CNN to extract patch-level features, computes statistical representations (mean and covariance) stored in a memory bank, and classifies test samples as normal or anomalous based on their distance from the learned distribution

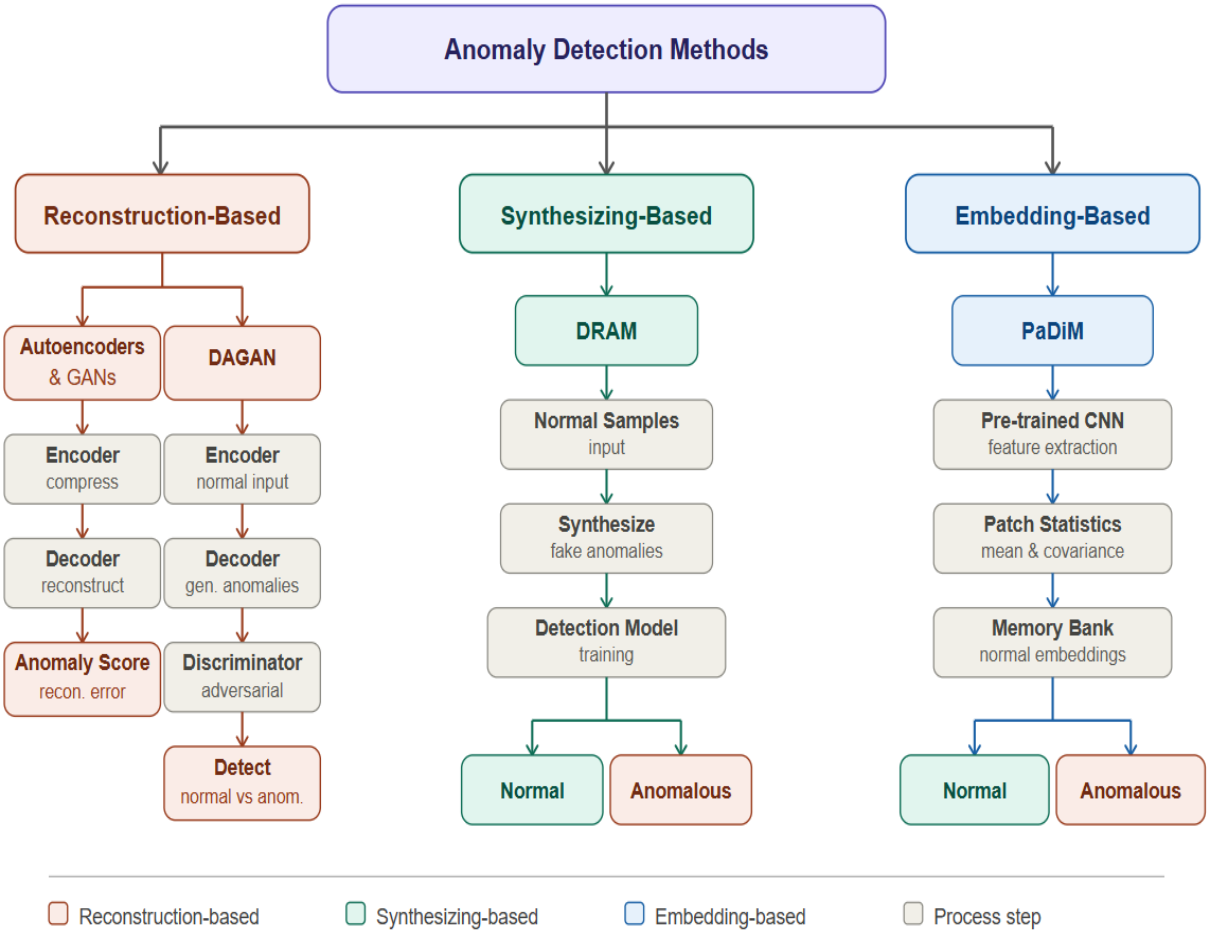

Figure 1. Taxonomy of unsupervised industrial anomaly detection methods, adapted from Tao et al. (2022).

2.2.1 Reconstruction-based methods

Reconstruction-based methods train a network, usually an autoencoder, to reconstruct images of normal parts. At test time, defective regions are assumed to be harder to reconstruct, so the pixel-wise reconstruction error is used as an anomaly score. Autoencoders (Bergmann et al., 2018) and variational autoencoders (Kingma & Welling, 2014) were among the first models tried in this setting. They have an attractive property: by design, they produce a per-pixel score, which means defect localisation comes essentially for free. The well-known weakness of reconstruction-based methods is that strong autoencoders sometimes reconstruct defects too well. If the latent space is large or the

decoder is too expressive, the network can interpolate over the defect without producing a clear error signal. Variants have been proposed to handle this problem. Memory-augmented autoencoders (Gong et al., 2019) confine the latent to be proximate to a memory bank of normal patterns, resulting in harder reconstruction of anomalous patterns. DRAEM (Zavrtanik, Kristan, and Skocaj, 2021) instead goes the route of discrimination, learning to reconstruct and segment only non-anomalous patterns, trained on generated data where fake defects are imposed on normal patterns. DRAEM was, for some time, the best-performing method on MVTec AD.

2.2.2 Generative and adversarial methods

Generative adversarial networks introduce a different idea to the table. Instead of reconstructing, they learn the distribution of normal images and use the discriminator score or a latent-space inversion as the anomaly score. AnoGAN (Schlegl et al., 2017) was the original example, with later refinements including f-AnoGAN (Schlegl et al., 2019), GANomaly (Akcay, Atapour-Abarghouei, and Breckon, 2018), and Skip-GANomaly (Akcay, Atapour-Abarghouei, and Breckon, 2019). These methods can produce very expressive models of the normal distribution, but they inherit the well-known training instabilities of GANs: mode collapse, oscillating losses and discriminator overconfidence. These problems will be revisited in Section 2.3, as the SimpleNet discriminator can exhibit similar overconfidence behaviour (Liu et al., 2023; see also Section 5.2).

2.2.3 Embedding-based methods

Rather than reconstructing images, embedding-based methods, also known as feature distance methods, operate directly in feature space. A pretrained backbone (commonly ResNet or WideResNet trained on ImageNet) is used to produce dense feature maps, and the distribution of these features over normal samples is then characterised. The anomaly score for a test image is defined as its distance from this learned distribution of normal features.

Among the earliest widely adopted methods in this family is SPADE (Cohen & Hoshen, 2020), which stores the normal feature vectors observed at each spatial position and, during inference, computes the k-nearest-neighbour distance between a test pixel and the stored features at the matching location. PaDiM (Defard et al., 2021) extended this design by substituting the stored feature set with a parametric description, modelling the distribution at each position as a multivariate Gaussian. PatchCore (Roth et al., 2022) took a different route to reducing memory cost, subsampling the patch feature bank into a compact corset and then querying that reduced set with nearest-neighbour search at inference.

Other embedding-based approaches use normalising flows to explicitly model the feature distribution. FastFlow (Yu et al., 2021) and CFlow-AD (Gudovskiy, Ishizaka, and Kozuka, 2022) belong to this family. Reverse Distillation (Deng & Li, 2022) approaches the problem from yet another angle by training a student network to mimic the teacher only on normal images, thereby making the teacher–student disagreement at inference time the anomaly score. A related and recent development is EfficientAD (Batzner et al., 2024), which combines a fast teacher–student arrangement with feature reconstruction, achieving very competitive accuracy at remarkable throughput. A common feature of these methods is that they avoid the instability inherent in GAN training while preserving the locality of reconstruction methods. They are also computationally cheaper during training, although memory bank approaches can be expensive at inference time.

2.3 SimpleNet

SimpleNet was introduced by Liu et al. (2023) at CVPR 2023 with the explicit goal of being a lightweight architecture that achieves competitive accuracy on MVTec AD while maintaining low inference cost. The discriminator alone produces both pixel-level anomaly maps and image-level scores at inference. The model is built from four components. The first component is a frozen pre-trained feature extractor, typically a WideResNet-50 trained on ImageNet, which produces multi-scale feature maps. By default, SimpleNet uses the outputs of the second and third layers. This choice is a balance between the

greater detail from earlier layers and the more meaningful information from deeper layers. The feature maps from these two layers are resized to the same size and then combined by summing their channels. First, the noise standard deviation, σ, has a significant effect on Simplenet's outcome. A low standard deviation makes the discriminator's job easier and may result in consistently high scores. The discriminator finds it difficult to determine what is typical when the standard deviation is too high, because the artificial features differ greatly from the real ones. Second, the discriminator may stop improving during training, especially when working with small datasets such as the sleeve dataset.

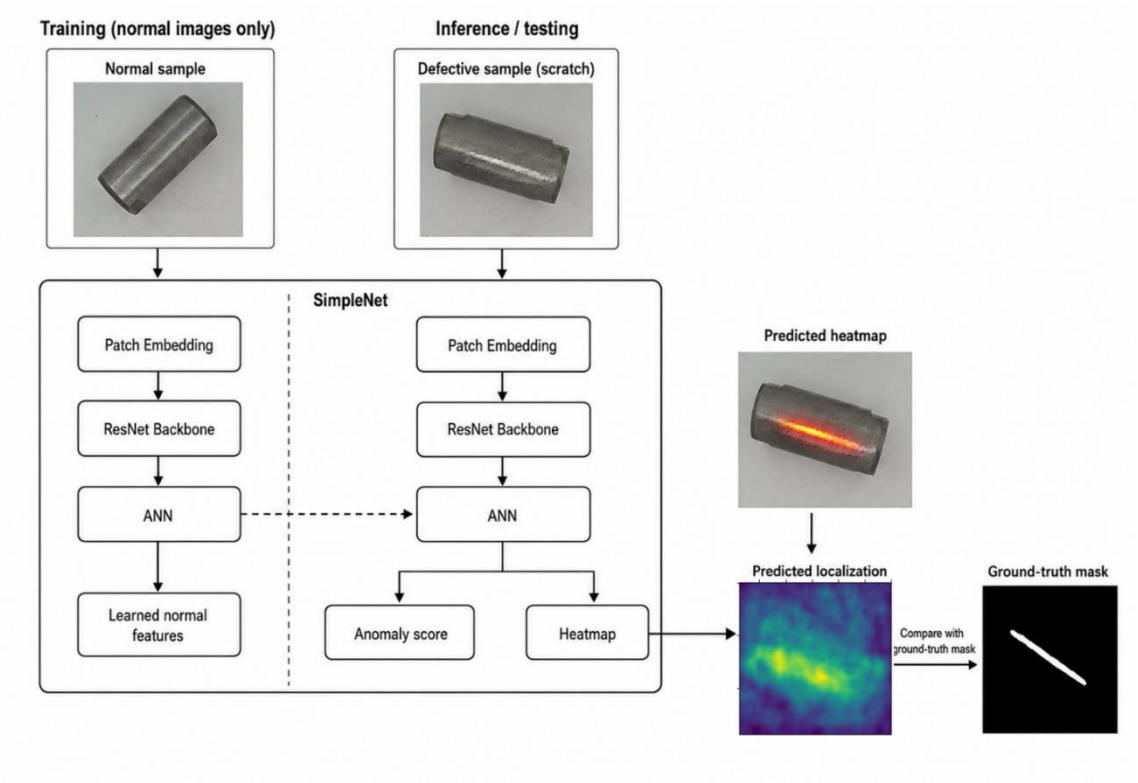

Figure 2. Flowchart outlining the SimpleNet deep learning framework: extraction of normal features via a ResNet backbone during training, and subsequent generation of anomaly scores, heatmaps, and localised predictions during inference

The second component is a feature adapter. It is a shallow linear projection that maps the ImageNet features into a domain-adapted embedding space. The motivation for the adapter is that ImageNet features are general-purpose and often suboptimal for industrial textures. The adapter has very few parameters and is trained jointly with the discriminator.

The third component is the anomaly feature generator. Instead of relying on external defect data, SimpleNet synthesises anomalous features in feature space by adding isotropic Gaussian noise of standard deviation σ to the adapted normal features. This is a simple yet surprisingly effective form of data augmentation in the feature space, allowing the model to be trained without ever seeing a real defect.

The fourth component is a binary discriminator. It is a small multi-layer perceptron that learns to distinguish real adapted features from the synthesised “fake” features. At inference time, the feature generator is discarded; the discriminator alone produces a per-patch anomaly score, and the spatial arrangement of these scores forms the anomaly map. An image-level score is obtained by taking the maximum of the smoothed anomaly map. However, alternative aggregations (such as the average of the top-k patch scores) have been proposed.

Liu et al. (2023) report a mean image-level AUROC of 99.6% and a mean pixel-level AUROC of 98.1% on MVTec AD, with an inference speed of around 77 frames per second on a single consumer-grade GPU. This combination of accuracy and speed is what makes SimpleNet attractive for industrial deployment. The simplicity of the architecture also makes it easy to modify, which is part of the reason it was chosen for the experiments in this thesis. The behaviour of SimpleNet, however, is not universally well understood. Two observations are particularly relevant for the present work. In the saturated regime, both real and fake feature probabilities approach 1.0, and the loss approaches zero, but the discriminator no longer provides a useful score at inference. These two issues are explored in detail in the experimental part of this thesis.

2.4 INP-Former: intrinsic normal prototypes

INP-Former (Luo et al., 2025) is a more recent transformer-based anomaly detection framework that addresses a specific weakness of embedding-based methods: their sensitivity to misalignment between training and test distributions. Classical embedding

methods compare test features either to a global memory bank or to position-specific Gaussians estimated from training images. Both choices implicitly assume that a normal feature at position (i, j) in a test image is comparable to normal features at the same position in training images. For objects with limited pose variation, such as the centred MVTec AD objects, the assumption holds well enough. For reflective metallic parts with framing variation, it does not.

INP-Former replaces the global memory bank with what the authors call Intrinsic Normal Prototypes. These prototypes are not stored as pre-computed feature vectors from the training set; they are learned through backpropagation as a compact summary of the normal training distribution, and at inference, the INP extractor module generates them conditioned on the test image. A small INP extractor module produces a fixed number of prototype tokens as learned linear combinations of the test-image tokens, and an INP coherence loss encourages those prototypes to represent only the normal regions of the image. After that, a guided decoder uses the extracted prototypes as a constraint to reconstruct the test feature map; the anomaly score is the residual reconstruction error. To encourage the network to concentrate on the most difficult-to-reconstruct portions of the typical image, a soft-mining loss is added. By extracting the prototypes from the test image, INP-Former adapts to local variations in appearance, pose and illumination. Luo et al. (2025) report state-of-the-art results on MVTec AD, VisA, and Real-IAD across single-class, multi-class, and few-shot settings. This adaptive behaviour is exactly the property that motivated its inclusion in the present thesis. If any current method has a chance against the framing and reflection variation of the sleeve dataset, it should be one that does not rely on global pose alignment.

2.5 Benchmark datasets

2.5.1 MVTec AD

MVTec AD (Bergmann et al., 2019) is the de facto standard benchmark in industrial anomaly detection. It has 5354 high-quality colour pictures grouped into fifteen groups.

Five of these groups are surface types: carpet, grid, leather, tile, and wood. The other ten groups include a bottle, cable, capsule, hazelnut, metal nut, pill, screw, toothbrush, transistor, and zipper. Each group has a set of pictures without any problems for training and another set that has both good and bad pictures for testing. The bad pictures in the test set also have detailed masks that show exactly where the problem is, making this set useful for checking both overall image quality and specific spots.

There are 73 distinct defect types and 1888 segmentation masks in total. The dataset has been enormously influential, partly because of its quality and partly because it established a clear protocol: train only on high-quality images, evaluate both image-level and pixel-level AUROC, and report per-category numbers. Its limitations have become more visible as the field has matured. MVTec AD images are acquired under tightly controlled studio conditions with diffuse lighting and centred objects. Test sets are small (50–100 images per category), which makes the AUROC unstable when methods are close to saturation. The dataset captures only a single view of each part. Moreover, the categories represent a narrow slice of real industrial inspection problems.

2.5.2 VisA and BTAD

VisA (Zou et al., 2022) and BTAD (Mishra et al., 2021) are two other benchmark datasets that have gained traction. VisA contains 12 categories of small electronic components, such as PCBs and capsules, with around 10,000 images in total. The defects in VisA tend to be more visually subtle than those in MVTec AD, which makes the benchmark slightly harder. BTAD focuses on three industrial categories and is smaller in scale. Both datasets, however, share many of MVTec AD's structural limitations: controlled lighting, single viewpoint, limited normal variation.

2.5.3 Real-IAD

Real-IAD (Wang et al., 2024) was specifically designed to address benchmark saturation in MVTec AD. It has 150,000 high-quality images of 30 different objects, taken from many angles under conditions more like a real factory setup. The dataset also uses a method

called Fully Unsupervised Industrial Anomaly Detection (FUIAD), where the training set includes some defective images that have not been removed. This setup is realistic because in a real factory, it is not possible to have a perfectly clean training set. Real-IAD is an important step toward more realistic testing standards, though it is still a new dataset and not all methods have been tested on it yet.

The Real-IAD dataset uses a specialised multi-view gantry setup to create images that closely reflect what happens on a real production line. Five cameras work together at the same time, one from above and four from the sides at 45-degree angles, to ensure no views are blocked. The dataset includes 30 types of products made from different materials like metal, plastic, ceramic, and wood. It includes both standard items and custom samples with real industrial defects, such as scratches and cracks. The correct answers, or ground-truth data, include detailed masks that mark the exact pixels where defects are, and these masks are checked using Cascade R-CNN. Some advanced versions of the dataset also use techniques such as photometric stereo and structured light to capture detailed 3D surface shapes.

2.5.4 MVTec LOCO and MVTec AD 2

MVTec LOCO (Bergmann et al., 2022) extends the original MVTec AD with logical-anomaly defects arising from incorrect arrangements of correct parts, in addition to the structural anomalies present in the original dataset. Most embedding-based methods perform poorly on logical anomalies, spurring a small subfield focused on this failure mode. MVTec AD 2 has been announced as a successor benchmark with even more challenging imaging conditions, including reflective objects, very small defects and defects near image borders. The motivation behind MVTec AD 2 is exactly the motivation behind the present thesis: standard benchmarks have become too easy, and the methods that succeed on them do not necessarily survive in the factory.

2.5.5 Steel surface datasets

Because the present thesis concerns a steel component, it is worth pausing to consider the datasets on steel surface inspection in the literature. The NEU surface defect database (Song & Yan, 2013) is one of the oldest: it contains six defect classes (rolled-in scale, patches, crazing, pitted surface, inclusion, and scratches) on hot-rolled steel strips, in grayscale images measuring 200×200 pixels. The Severstal Steel Defect Detection dataset was released as a Kaggle competition in 2019 and contains a much larger set of segmentation-annotated images of steel coils. GC10-DET (Lv et al., 2020) was proposed to cover metallic surface defects with greater diversity than NEU.

All three datasets are useful for steel inspection research, but none of them is a good match for the present case. They focus on flat steel sheets and strips, whereas the sleeve is a three-dimensional reflective object whose appearance varies with pose. They are also acquired under controlled industrial imaging rather than the webcam-based setup used in this thesis. Several recent works have highlighted that reflectivity is a central challenge in metallic inspection. Studies on highly reflective industrial parts have proposed polarised imaging, structured light and specialised hardware to mitigate the problem (for example, Huang et al., 2021). Most academic anomaly detection methods, however, are evaluated on diffuse-surface datasets, which is part of the reason the sleeve dataset feels harder than the literature suggests it should be.

2.6 Data Augmentation in Anomaly Detection

Augmentation has been a common tool in supervised deep learning since AlexNet (Krizhevsky, Sutskever, and Hinton, 2012). In the unsupervised anomaly detection setting, its role is more subtle because there is no clear notion of correct label preservation to anchor the choice of transformations. The standard practice in MVTec AD is to use only mild geometric augmentations, since the normal class should represent what the camera will see in the factory.

Several authors have, nevertheless, reported that more aggressive augmentation can help when the training set is small or when the normal class is too narrow. CutPaste (Li et al., 2021) randomly cuts a patch from the image and pastes it back at a different location, treating the pasted region as a synthetic defect. DRAEM (Zavrtanik et al., 2021) uses a similar idea but with more elaborate synthetic defects. There is also a question of what augmentation does to the normal distribution. If the augmentation is too aggressive, the trained model will accept as normal a much wider range of images than the production line produces, and the discriminator boundary will shift outwards. This is the trade-off that Section 4.3 of this thesis investigates empirically.

2.7 Evaluation metrics

Industrial anomaly detection methods are usually evaluated with three metrics. The image-level AUROC (I-AUROC) indicates how well the anomaly score distinguishes normal from defective images. The pixel-level AUROC (P-AUROC) works similarly but evaluates each pixel individually, as if each were its own classification task. Both AUROC variants are threshold-free, which is convenient for comparison. However, they are known to be optimistic about imbalanced data, and pixel-level anomaly detection is extremely imbalanced, since defective pixels typically form a tiny fraction of any image. Bergmann et al. (2021) introduced the Per-Region Overlap metric (PRO) to address the imbalance caused by evaluating anomalies at the pixel level. PRO measures the overlap separately for each anomalous region, giving equal consideration to small and large defects. The AUPRO score represents the area under the PRO curve, typically computed only up to a predefined false-positive rate, commonly 30%.

AUPRO is widely regarded as the most reliable localisation metric on MVTec AD, especially for categories with very small defects. For the sleeve dataset, where defects cover only about 0.08% of the image, AUPRO is particularly important: a method that produces a wide, blurry anomaly response can still reach a high P-AUROC by accident, but it will be penalised by AUPRO.

2.8 Data leakage and evaluation reliability

Data leakage is particularly relevant in the unsupervised anomaly detection setting because the model is trained only on normal images, which can easily blur the boundary between training and test data, especially on small custom datasets where the total number of normal samples is limited. When the same images appear in both sets, the resulting AUROC reflects memorisation rather than genuine detection ability. This thesis includes a deliberate leakage experiment on the sleeve dataset to quantify inflation directly. Data leakage is a recurring source of inflated results in the machine learning literature, and industrial anomaly detection is not immune. The most common form of leakage in this setting occurs when normal samples used to train the model also appear in the test set, causing the model to recognise rather than generalise. A subtler form occurs when hyperparameters are tuned on the test set, which leaks information through the modelling choices even if no individual image is shared between splits.

On small custom datasets such as the sleeve dataset of this thesis, both forms are easy to commit accidentally. It is worth noting that the test set composition affects AUROC stability more broadly than leakage alone. With only ten normal test images, a single mislabelled or unusual image can move the AUROC by several percentage points. This is one of the reasons why benchmark datasets like Real-IAD have moved towards much larger test sets, and it is also one of the reasons why the present thesis takes the question of split discipline seriously.

2.9 Research gap

Pulling the threads together, three gaps in the existing literature are particularly relevant for the present thesis. First, the augmentation question has been studied extensively in the supervised setting. However, its effect on unsupervised industrial anomaly detection remains ad hoc and is rarely decomposed into image-level and pixel-level contributions. Second, the saturation behaviour of the SimpleNet discriminator is mentioned anecdotally in several follow-up works but, to the best of our knowledge, is not systematically

documented. Third, and most importantly, the gap between MVTec-level performance and real factory performance on reflective metallic parts is large. However, it is rarely characterised in concrete terms because most academic work continues to be evaluated against the same standard benchmarks.

3. Methodology

3.1 Overall research approach

The research follows an empirical, hypothesis-driven approach. Every experiment in this section ties back to one of the research questions laid out in section 1.4. With a clear hypothesis, design a test that changes one thing at a time while keeping everything else the same, run it with a fixed random seed so the results can be reproduced, and report the I-AUROC, P-AUROC and AUPRO that come out the other side. This kind of controlled setup is not about being rigid for its own sake; it is simply the most honest way to understand what drives a result rather than getting confused by too many moving parts at once.

3.2 Dataset

3.2.1 Physical setup and imaging

The sleeve is a small metal industrial component, approximately five centimetres in length, with a cylindrical body and two bevelled rectangular cut-outs near each end. The surface is partially polished and therefore produces visible specular reflections under direct lighting. Surface defects observed in the production sample include short scratches, micro dents, mild discolouration and slight irregularities at the bevelled edges.

Strict train-test separation is enforced throughout. All 340 training images belong exclusively to the training set. The 10 normal test images and 53 anomalous test images belong exclusively to the test set. No image is ever moved between sets during model development. This discipline becomes important in Section 4.4, where a deliberate violation is used to quantify the data leakage effect.

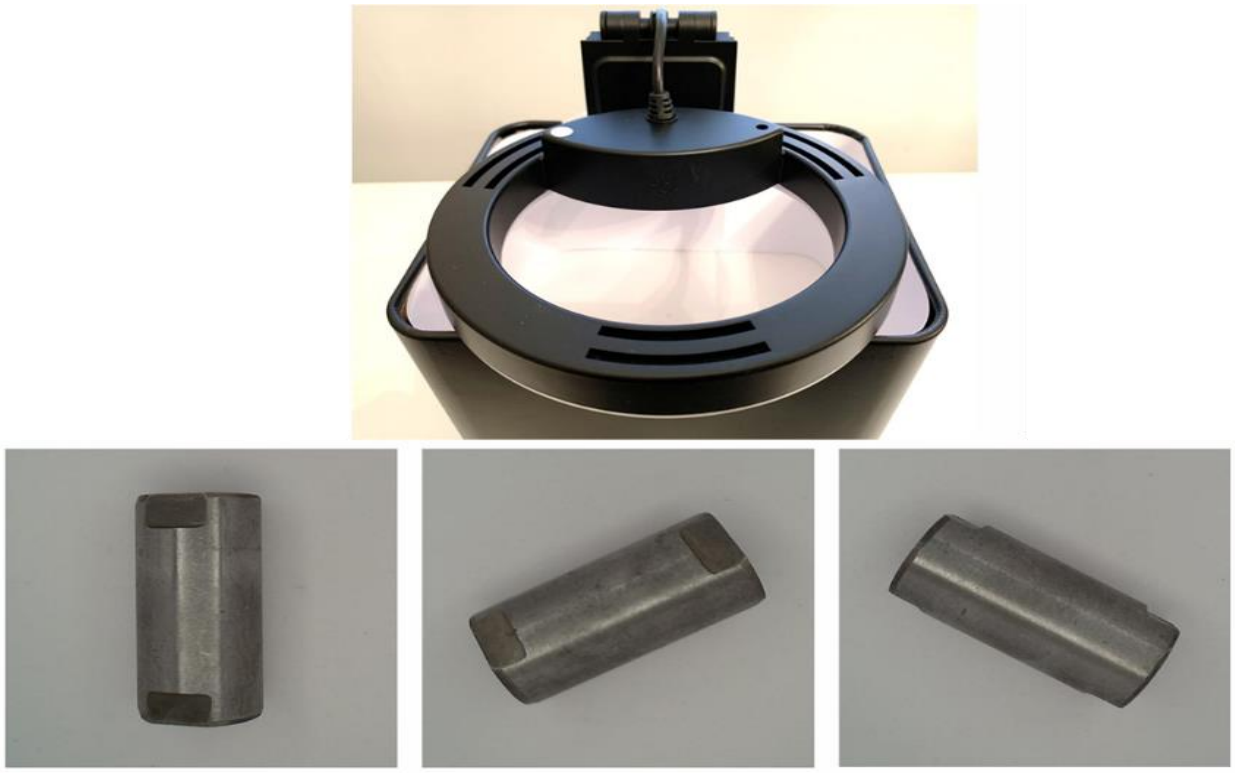

Figure 3. Imaging setup light (top) and example sleeve good images showing different poses (bottom).

The imaging setup follows common recommendations for building MVTec-style anomaly detection datasets. An inexpensive HD webcam with a short-focal-length lens is mounted on top of a ring light, and the sleeve is placed inside the ring at a fixed working distance of approximately 10 cm. The interior walls of the imaging chamber are matte white to spread the illumination. With that setup, the part fills a large fraction of the image area, which is necessary because the defects are very small in absolute pixel terms. The final image size used for training and inference is 640 by 640 pixels, which is approximately 5cm. This is either changed to 288×288 pixels within the SimpleNet procedure. The original SimpleNet paper used 256×256 pixels, but a slightly larger size worked better with this dataset.

All images are taken in the ring light, and controlled walls help, but the lighting still varies more than you would want. The webcam darkens toward the corners, and the bright circular reflection the ring light casts onto the sleeve shifts position whenever the part

is placed slightly differently. Small issues, but enough to add noise the model must deal with on top of finding actual defects. Both effects show up later in the experimental results as sources of false positives, and they are part of what makes the dataset realistic rather than benchmark clean. The dataset needs to follow the default MVTEC AD folder structure.

3.2.2 Dataset organisation

The dataset follows the standard MVTec AD folder structure as shown in Table 1.

Folder Path	Contents	Purpose
sleeve/train/good/	340 normal images	Training data (defect-free)
sleeve/test/good/	10 normal images	Normal test samples
sleeve/test/bevel1/	Anomalous images	Test defect: bevel1
sleeve/test/bevel2/	Anomalous images	Test defect: bevel2
sleeve/test/colour/	Anomalous images	Test defect: colour issue
sleeve/test/dent/	Anomalous images	Test defect: dent
sleeve/test/scratch/	Anomalous images	Test defect: scratch
sleeve/ground truth/bevel1/	Pixel masks	Ground-truth masks for bevel1
sleeve/ground truth/bevel2/	Pixel masks	Ground-truth masks for bevel2
sleeve/ground truth/colour/	Pixel masks	Ground-truth masks for colour defects
sleeve/ground truth/dent/	Pixel masks	Ground-truth masks for dent defects
sleeve/ground truth/scratch/	Pixel masks	Ground-truth masks for scratch defects

Table 1. Dataset folder structure following the MVTec AD convention.

This structure was chosen for compatibility with SimpleNet and the INP-Former experiments described in Chapter 5. The total of 53 anomalous test images is distributed across the five defect categories. Pixel-level ground-truth masks are stored as single-channel PNG images at the same resolution as the input images, with white pixels indicating defective regions and black pixels indicating normal regions.

Split	Normal images	Anomalous images	Total
Training	340	0	340
Test	10	53	63
Total	**350**	**53**	**403**

Table 2. Dataset split summary (Training / Test / Total with Normal, Anomalous and Total counts)

3.2.3 Defect taxonomy

Five defect categories were defined during dataset construction. Each has a distinct physical origin and visual signature, and each stresses the anomaly-detection model differently.

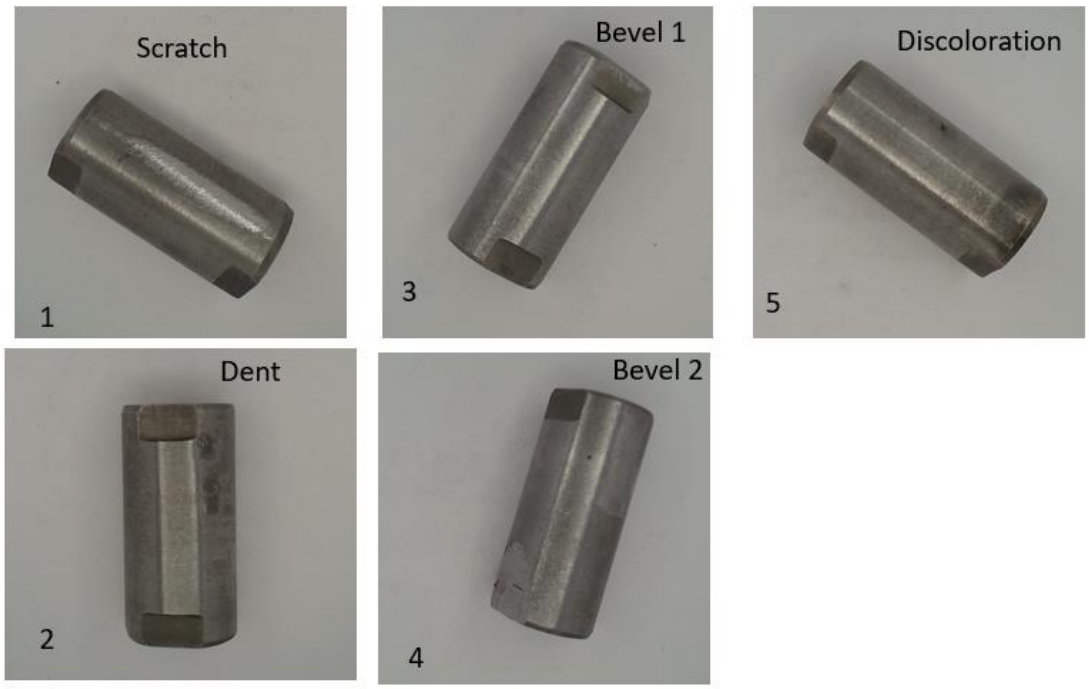

Figure 4. All defect classes in the grid: The grid shows one example from each of the dataset's four defect classes. Starting at the top left, the first image shows a scratch, a fine linear mark running across the surface. Next to it, the top canter is the first bevel variant, where the edge transition appears irregular against the surrounding steel. Bottom left shows a dent, visible as a shallow depression that disrupts the otherwise uniform surface texture. Finally, the bottom canter shows the second bevel variant, which shares the general character of the first but differs in how the irregularity presents along the edge.

Scratch. Scratches are long, thin, and anisotropic anomalies caused by mechanical contact during production or handling. In mask space, they appear as high-aspect-ratio, elongated regions, sometimes with subtle curvature and fluctuations in width. Scratch defects are particularly challenging for convolutional architectures because hierarchical downsampling tends to erode thin structures.

Dent. Dents are micro-scale deformations on the object surface. They are typically very limited in size, sparse, and often appear as disconnected regions within the same image. In some cases, a single image may contain multiple separate dent regions. These defects do not usually produce sharp structural boundaries. Instead, they are mainly observed through local variations in reflectance, indicating that dents correspond to subtle geometric perturbations expressed through photometric changes rather than major shape damage.

Colour. Colour anomalies are local discolourations of the surface caused by oxidation, contamination or mild thermal damage. They have irregular boundaries and low contrast against the metallic background. Many colour anomalies look uncomfortably similar to the sleeve's natural reflectance variation, which is why they tend to produce both false negatives and false positives.

Defect class	No. of images	Spatial character	Primary detection challenge
Scratch	10	Thin, elongated, high aspect ratio	Peak anomaly score competes with normal highlight intensity
Dent	11	Small blob(s), often 2-3-disconnected instances per image	Low contrast, only a few hundred pixels; photometric rather than structural boundary
Colour	12	Irregular patch, soft boundary	Visually overlaps with natural surface reflectance variation
Bevel 1	9	Narrow flaw near the bevelled edge	High-contrast edge background suppresses the defect signal

Defect class	No. of images	Spatial character	Primary detection challenge
Bevel 2	11	Contour deformation at the bevelled edge	Visibility is pose-dependent: near-invisible under oblique illumination and difficult to identify when viewed from below.
Total	**53**		

Table 2. Per-defect-class statistics for the sleeve test set

Bevel 1. Bevel 1 defects are small, narrow flaws that occur near the angled edges of the sleeve. The bevel area contrasts brightly, with clear shadows and reflections, so the deformity shows up against a backdrop that is already utterly noisy.

Bevel 2. Bevel 2 anomalies are larger geometric defects that change the contour of the bevelled region. Their visibility strongly depends on the sleeve's pose in the image.
In a frontal view, they are obvious; under oblique illumination, they can blend into the normal edge shading. Defects that live on the object silhouette are known to be particularly hard for SimpleNet-style classifiers, because the silhouette itself is part of the texture model.

Figure 5. Industrial inspection interface showing side-by-side comparisons of raw component captures against anomaly-mapped results, featuring colour-coded status banners

The point of keeping the categories separate, rather than lumping them into a single “anomaly” folder, is to enable per-defect failure analysis in Chapter 5. The five categories together represent four physically distinct anomaly-formation mechanisms: contact-induced surface discontinuity (scratch), local geometric deformation (dent), photometric contamination (colour), and edge geometry alteration (bevel).

3.2.4 Defect statistics

To put the dataset's difficulty on a concrete footing, the mask area per defect was measured. The average mask area across all 53 anomalous images is approximately 0.08% of the total number of pixels. This is more than an order of magnitude smaller than the average defect area in the standard MVTec AD categories, where defects often cover several per cent of the image area. The smallest individual defects in the sleeve dataset cover only a few hundred pixels at the input resolution used by SimpleNet, which is close to the spatial resolution of a single patch in the WideResNet feature maps.

3.2.5 Annotation protocol

Pixel-level masks were drawn manually using the open-source Label Studio annotation tool. Each anomalous image was labelled by one annotator and reviewed once. To keep the masks consistent across the dataset, the following rules were applied. Only the defective region itself was labelled, not its visual context. When a defect was visible only as a soft shading change (typical for dents and some colour anomalies), the mask was drawn around the most clearly perturbed region, erring on the side of being slightly smaller rather than slightly larger. When multiple disconnected defects appeared in the same image, each one was labelled separately, which is important for AUPRO.

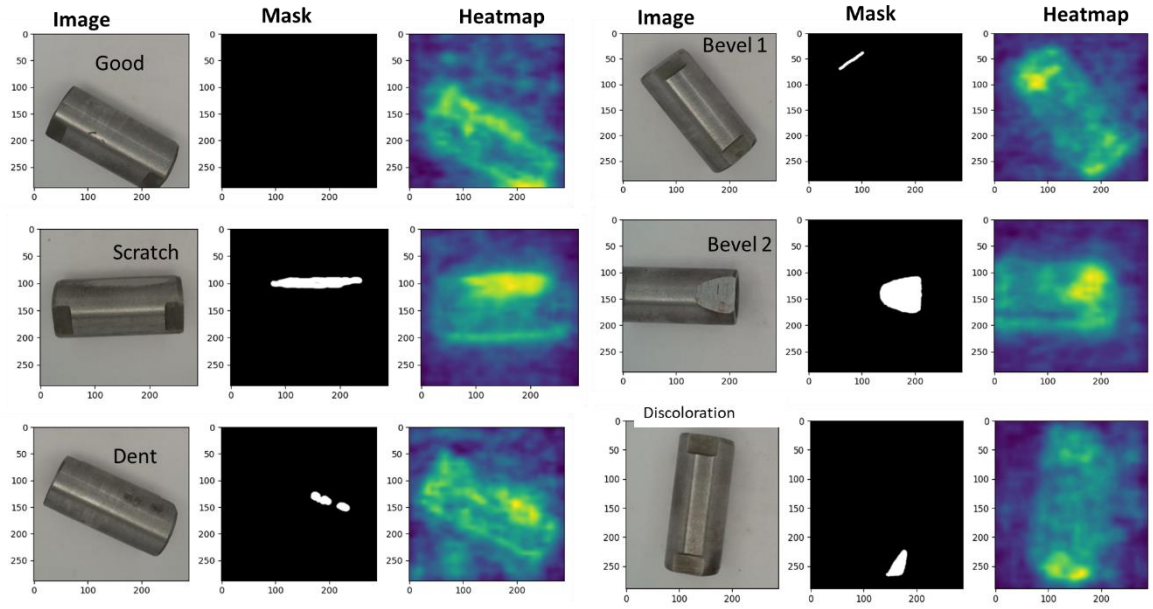

Figure 6. Representative qualitative results from the metallic sleeve surface defect dataset, illustrating visual samples, ground-truth segmentation masks, and model-generated heatmaps for different inspection conditions. The dataset contains cylindrical metallic sleeve components with both normal and defective surface states. The Good sample represents a defect-free sleeve, while the defective cases include Scratch, Dent, Bevel 1, Bevel 2, and Discolouration defects. For each category, the original RGB image is shown alongside its binary defect mask and heatmap response. The masks delineate localised defective regions, whereas the heatmaps indicate the model's attention or activation intensity across the sleeve surface. The results demonstrate that the model can focus on subtle surface irregularities, such as scratches, dents, edge bevel defects, and discolouration patterns, on reflective, cylindrical metallic surfaces.

3.3 SimpleNet pipeline

The SimpleNet setup used in this thesis is based on that of Liu et al. (2023). Features are extracted using a WideResNet-50 pretrained on ImageNet, pulling from layers 2 and 3 of the network. These mid-level layers tend to capture the structural and textural detail that matters most for spotting surface anomalies, which is why they are the standard choice over earlier or later layers. These features are upsampled to the same size using bilinear interpolation and then stacked, creating a feature vector for each position with a total of 1536 dimensions. A single linear layer is used to propagate this characteristic vector into an embedding of the same size. The discriminator is a two-layer neural network with a hidden layer of 1024 units and a scalar output. During training, the artificial anomaly generator adds random noise sampled from a Gaussian distribution

with standard deviation σ to a copy of each adapted feature vector to create a fake feature.

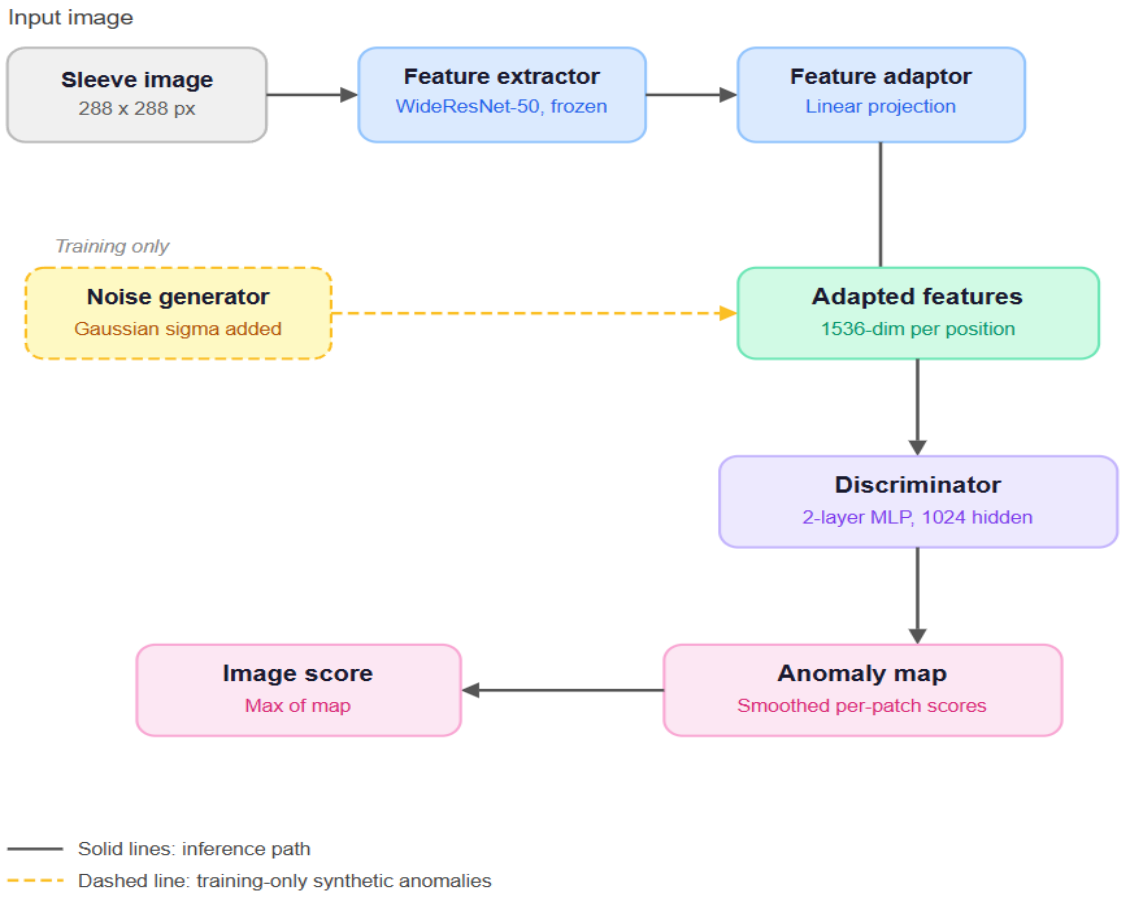

Figure 7. Block diagram of the SimpleNet anomaly detection architecture, detailing the processing pipeline.

The discriminator is trained using a loss function that pushes real features above one threshold and fake features below another. When making predictions, the generator is no longer used. Instead, the discriminator's scores for each position are arranged into an image-like map. This map is then smoothed using a Gaussian filter. This smoothed map is used both to evaluate the model at the pixel level and to create a score for the entire image.

3.3.2 Training configuration

All experiments use the same fundamental training setup unless otherwise specified. The input images are adjusted to 288×288 pixels, and then a central region of that size is extracted without random cropping. The patch size used to form the discriminator inputs is 3. Optimisation uses Adam with learning rates of 1e-4 for the adapter and 5e-4 for the discriminator, weight decay of 1e-5, batch size 4, and 40 meta-epochs of 4 inner GAN-style epochs each, for a total of 40 effective epochs. The default noise standard deviation is $\sigma = 0.015$, and the discriminator margin is 0.5.

3.3.3 Modifications introduced in this thesis

Three modifications were added to the baseline SimpleNet recipe to address the discriminator saturation behaviour described in Section 2.3 and the small-data instability that was observed early in the experimental work.

Gradient clipping. Global gradient norm clipping is applied to both the adapter and the discriminator at a threshold of 1.0 before every optimiser step. Without clipping, large gradient bursts at the start of training were occasionally observed to drive the discriminator into a saturated regime in which p_true and p_fake both approached 1.0 and the loss approached zero. Clipping does not eliminate this behaviour, but it makes it considerably less frequent and gives the model time to settle into a more useful operating point.

Reduced Gaussian smoothing sigma. The anomaly map produced by the discriminator is post-processed with a 2D Gaussian filter to suppress isolated noisy responses. The original SimpleNet code uses a smoothing sigma of 4. On the sleeve dataset, where defects are extremely small, this level of smoothing was found to wash out the response of single-blob dents and to fragment thin scratches. The sigma was reduced to 2 in all experiments unless noted otherwise. The effect on pixel-level metrics is measurable, especially for the dent and scratch categories, and is discussed in Section 5.2.

Feature distribution tracking. A lightweight tracker was added to the training loop that records the mean and the standard deviation of the discriminator's output on a small batch of training-time fake features and on a held-out batch of training-time real features after each meta-epoch. The tracker does not feed back into training; it is purely diagnostic. The trajectory of the two means over training is what allows the saturation regime to be detected in practice.

3.4 INP-Former pipeline

INP-Former was used primarily because it was released by Luo et al. (2025), with only two adjustments to ensure a fair comparison with SimpleNet. The same train-test split was applied across both methods, and the input size was kept at 288×288 to match the SimpleNet runs exactly. The only processing change was to the Gaussian smoothing sigma, which was lowered from the default of 4 to 2. Everything else, the number of prototypes, the coherence loss weighting, and the soft-mining schedule, was left at whatever the authors specified as default. Training ran for the same number of effective epochs as SimpleNet.

3.5 Evaluation metrics and protocol

Three metrics are reported for every experiment: image-level AUROC, pixel-level AUROC, and AUPRO. Image-level AUROC is computed from the per-image anomaly score, which is the maximum of the smoothed anomaly map after post-processing. Pixel-level AUROC is computed from the smoothed anomaly map by treating each pixel in each test image as an independent classification example, using the ground-truth mask as the label. AUPRO is computed as the area under the per-region overlap curve integrated up to a false-positive rate of 30%, following Bergmann et al. (2021).

For the sleeve dataset, the test split is fixed across all experiments at 10 normal and 53 anomalous images. For the MVTec AD evidence experiments in Section 4.2, the official

MVTec AD train/test split is used as is. No images are moved between the splits unless they are part of the deliberate data-leakage experiment.

3.6 Augmentation strategies

Three augmentation regimes are compared in the experimental study of Section 4.3. They were chosen to span the spectrum from no augmentation at all to heavy synthetic regeneration of the training set.

No augmentation. The training images are used as-is after resizing and centre cropping. This is the baseline.

Mild geometric augmentation: Photometric changes (contrast and brightness varying by up to 5%) are applied without any geometric transforms. This setup keeps the sleeve's orientation consistent while making the normal variations more noticeable.

Strong augmentation. Strong geometric transforms (random rotations up to 15 degrees, translations up to 5%, scaling up to 10%, horizontal flips) are combined with photometric changes (contrast and brightness up to 15%). This regime breaks pose alignment and was expected to be too aggressive a priori. However, it was kept in the comparison precisely because the question of whether aggressive augmentation helps image-level discrimination at the cost of localisation is an empirical one.

Horizontal flips are added. This regime breaks pose alignment and was expected to be too aggressive a priori. However, it was kept in the comparison precisely because the question of whether aggressive augmentation helps image-level discrimination at the cost of localisation is an empirical one.

All three regimes were trained from scratch with identical hyperparameters, except for the choice of augmentation. Each result is reported as the mean over three random seeds.

3.7 Hardware and software

The experiments in this thesis were completed on a workstation equipped with an NVIDIA RTX 4090 GPU with 24 GB of memory, an Intel Xeon CPU, and 64 GB of system RAM. The workstation operates on Ubuntu 22.04. The deep learning implementation relies on PyTorch 2.1, and CUDA. The SimpleNet configuration utilises the official release but incorporates the modifications from Section 3.3.3 as a patch. INP-Former relies on the official release from the authors' GitHub repository.

4. Experimental Results

4.1 Baseline SimpleNet on the sleeve dataset

The first experiment establishes a baseline using the default SimpleNet configuration with no augmentation, no modifications to the smoothing sigma, and no gradient clipping. The training set is the full set of 340 normal images; the test set is the fixed split of 10 normal + 53 anomalous images.

Metric	Value
I-AUROC	0.53
P-AUROC	0.91
AUPRO	0.74

Table 4. Baseline SimpleNet results on the sleeve dataset.

The clear difference between the pixel-level AUROC of 0.91 and the image-level AUROC of 0.53 is hard to ignore. The pixel-level score suggests the model is doing a reasonable job of identifying where on the surface the problem is. The image-level score tells a different story when it comes to ranking whole images and deciding which ones to flag; the model is barely better than chance. Good localisation and reliable detection are not the same thing, and this result makes that distinction concrete.

This pattern recurs across many later trials and stands as a key finding of this thesis. The AUPRO of 0.74 is lower than the pixel AU-ROC implies because AUPRO evaluates regions rather than individual pixels and penalises broad, smeared anomaly activations.

4.2 Implementation validation on MVTec AD

Before concluding on the difficulty of the sleeve dataset, possible implementation-related issues needed to be excluded first. To verify that implementation issues did not cause the observed results, the same SimpleNet pipeline was tested on three established MVTec AD categories: Leather, Zipper, and Metal Nut. Leather was selected as a comparatively easy texture category, Zipper as a medium-difficulty object category, and Metal Nut as a more challenging metallic object category due to its reflective surface properties. Across all experiments, the hyperparameters remained identical, with the dataset path being the only changed setting.

Category	**I-AUROC**	**P-AUROC**	**AUPRO**
Leather	1.000	0.992	0.970
Zipper	1.000	0.982	0.970
Metal Nut	0.999	0.983	0.875
Sleeve (Baseline)	0.530	0.910	0.740
Sleeve(Refined SimpleNet)	0.80	0.93	0.74

Table 5. Implementation validation results on three MVTec AD categories compared to the custom sleeve dataset.

For all three MVTec AD categories, the implementation achieves near-saturated numbers reported by Liu et al. (2023). The contrast between the MVTec AD rows and the sleeve row in Table 3 is one of the central messages of this thesis. The same model, same code, same hyperparameters, and same train-test protocol produce essentially perfect numbers on three benchmark categories and noticeably weaker numbers on a custom dataset built with the same MVTec-style protocol. The difficulty observed in the sleeve dataset reflects the data's inherent challenges rather than deficiencies in the implementation. The most plausible candidates, as discussed in Chapter 5, are reflectivity, defect size and the noisiness of the normal class.

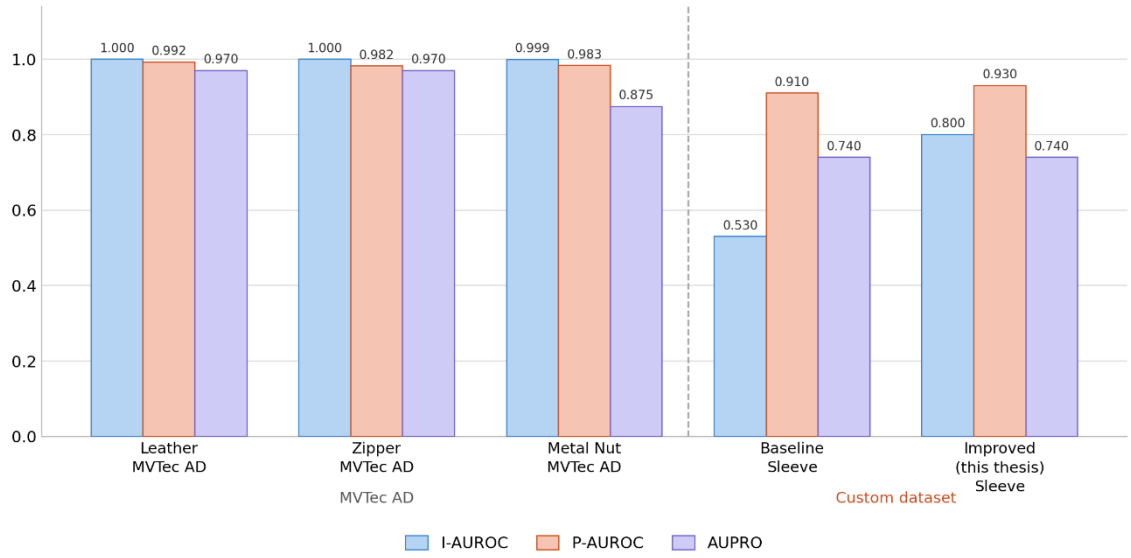

Figure 8. Bar chart comparing I-AUROC, P-AUROC, and AUPRO across the three MVTec AD validation categories and the sleeve custom dataset.

4.3 Augmentation study

The augmentation study compares the three regimes defined in Section 3.6 on the sleeve dataset. All other settings are identical to the baseline. Each row in Table 4.3 shows the mean result from three independent random seeds. The standard deviations were extremely small, normally lower than 0.01.

Augmentation	I-AUROC	P-AUROC	AUPRO
No augmentation	0.530	0.910	0.740
Mild augmentation	0.711	0.900	0.747
Strong augmentation	0.751	0.900	0.688

Table 6. Augmentation study results on the sleeve dataset (mean over three seeds).

Two observations stand out. First, all three augmentation strategies improve over the no-augmentation baseline on at least two of the three metrics. The custom dataset is small enough that augmentation is unambiguously helpful. Second, the regime that produces the best image-level number (strong augmentation, I-AUROC 0.751) also produces the worst AUPRO (0.688). The discriminator becomes better at flagging abnormal images when it is forced to deal with a more varied normal class, but the price is paid in localisation quality. The right choice depends on what the downstream system consumes: a per-image accept/reject decision benefits from stronger augmentation, while a downstream defect-localisation step benefits from mild augmentation.

4.4 Data leakage experiment

During early experimentation, a different version of the test split was constructed in which a small number of training images were moved into the test set. This was not done with any intent to inflate results; it was an oversight that became visible only when the train and test folders were audited carefully. The image-level AUROC obtained on that contaminated split was substantially higher than on the clean split. To quantify the effect properly, a small, controlled experiment was performed in which a fixed number of randomly chosen training images were deliberately reintroduced into the test set as additional “good” test samples.

The model, having essentially memorised those images during training, produces very low anomaly scores for them at test time, and the per-image AUROC ranking is correspondingly easier to achieve. The exact magnitude of the inflation depends on how many training images leak and on which random subset is chosen, but the qualitative result is robust: as soon as training images leak into the test split, the reported number ceases to measure generalisation and becomes a measure of memorisation.

The point of including this experiment is not to claim novelty; data leakage is a well-known pitfall in machine learning, but to underline how easy it is to commit on a small custom industrial dataset. With only ten normal test images, adding a handful of training images to the test set is a minor administrative slip that has an outsized effect on the headline metric. All other experiments in this thesis use the clean 340/10/53 split.

4.5 Improved SimpleNet

The final SimpleNet configuration reported in this thesis combines several of the elements introduced earlier in the chapter. It uses mild geometric augmentation (which gave the best AUPRO in the augmentation study), gradient clipping at a global norm of 1.0, the reduced Gaussian smoothing sigma, and the feature distribution tracker for diagnostic purposes.

Configuration	**I-AUROC**	**P-AUROC**	**AUPRO**
Baseline	0.530	0.910	0.740
Improved SimpleNet (this thesis)	0.800	0.930	0.740

Table 7. Improved SimpleNet results compared to the baseline.

The most notable shift is at the image level, where the gain over the baseline without augmentation comes to around 14 percentage points. That is not a marginal improvement; it suggests that the augmentation is doing real work in helping the model generalise across the normal variation in the sleeve images rather than just memorising the training set.

The AUPRO is essentially unchanged relative to the baseline, consistent with the trade-off observed in Section 4.3 between image-level and region-level metrics. The diagnostic

tracker confirmed that the discriminator did not enter the saturated regime during training for any of the three seeds, in contrast to the unmodified baseline, where saturation was observed for one of the three seeds.

4.6 Comparison with INP-Former

INP-Former was trained on the same sleeve dataset, using the same split and input resolution. Table 4.5 reports the result, together with the improved SimpleNet for reference.

Method	**I-AUROC**	**P-AUROC**	**AUPRO**
Improved SimpleNet	0.800	0.930	0.740
INP-Former	0.760	0.980	0.910

Table 8. Comparison of improved SimpleNet and INP-Former on the sleeve dataset.

The two methods land in noticeably different places when you look at localisation against classification together. INP-Former is the stronger localiser by a clear margin: P-AUROC of 0.98 against 0.93, and AUPRO of 0.91 against 0.74 for SimpleNet. The AUPRO gap is the more telling of the two numbers. At nearly 17 points, it reflects not just that INP-Former finds the right general area more often, but that its anomaly maps are more precisely shaped around the actual defect rather than bleeding into surrounding normal texture. That kind of spatial precision matters in practice, where a rough heatmap can be as frustrating to interpret as no heatmap at all.

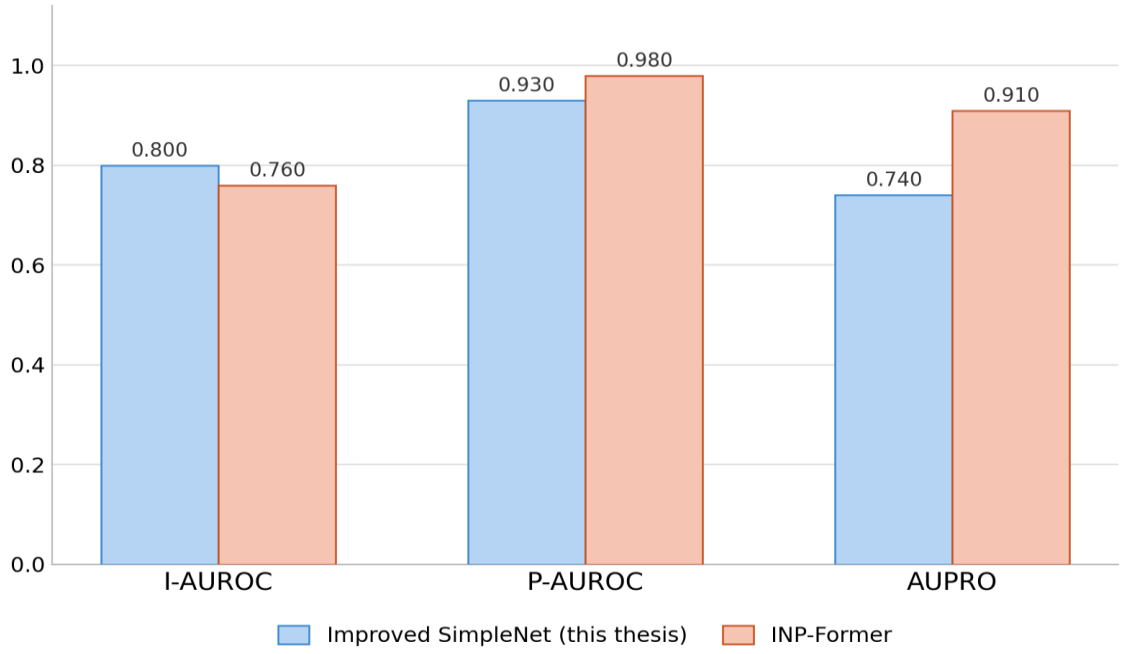

Figure 9. Quantitative comparison of anomaly detection methods on the Sleeve custom dataset. Improved SimpleNet achieves a higher I-AUROC (0.800 vs 0.760), while INP-Former outperforms on P-AUROC (0.980 vs 0.930) and AUPRO (0.910 vs 0.740).

SimpleNet has a small edge in image-level discrimination: I-AUROC 0.80 against 0.76, a difference of about 3 points. The fact that the method with substantially better localisation produces a slightly worse image-level decision is, on the surface, counterintuitive. The likely explanation is that the normal training class contains borderline samples, small dark specks, machining marks, and faint within-tolerance scratches that push the learned notion of normal outwards. Once the normal distribution is broad, even a sharp and well-localised defect response sits closer to the normal range than it would on a cleaner dataset.

5. Discussion

This chapter steps back from the experimental numbers and tries to make sense of what was observed. It opens with the most important pattern in the results, the persistent gap between image-level and pixel-level performance and works through the most likely causes. It then provides a defect-by-defect failure analysis based on misclassified images, discusses how the present results compare with the published literature, and lists the main limitations of the study.

5.1 The image-level versus pixel-level gap

The single most consistent observation across all experiments in this thesis is that pixel-level performance is substantially higher than image-level performance on the sleeve dataset. The improved SimpleNet reaches P-AUROC 0.93 but only I-AUROC 0.80. For INP-Former, the gap is even wider: P-AUROC 0.98 against I-AUROC 0.76. The same pattern recurs across the augmentation study and in the baseline configuration, suggesting it is a property of the dataset rather than of any one model.

The first ingredient is the size of the defects. On a dataset where the average defect covers only 0.08% of the image area, the anomaly map can identify the few hundred genuinely defective pixels with high local accuracy, which is what the pixel-level metrics reward, while still producing only a modest elevation in the overall per-image score. A well-localised response and a small spurious response on a normal sample can be of comparable magnitude; therefore, the per-image ranking that separates them is much more fragile than the per-pixel ranking. P-AUROC and AUPRO operate on the rich, spatially structured anomaly map; the per-image score is a single number derived from it. Information is necessarily lost in that reduction, and on a dataset with very small defects, the lost information is exactly the information that distinguishes a defective image from a normal one.

The second ingredient is the noisiness of the normal class. Inspection of the 340 training images revealed that several of them contain features that appear defect-like to a non-expert eye: small dark specks, machining marks, faint scratches within production tolerance, and occasional reflective bands caused by the ring light. The model is trained to treat all these features as normal, which pushes the learned normal distribution outwards.

When a genuine defect appears at test time, its anomaly score is closer to the inflated normal distribution than it would be on a cleaner training set, and the image-level ranking suffers accordingly. The pixel-level metrics are partially shielded from this effect because they are dominated by the millions of clearly normal pixels in each image; the image-level metric, which depends on the separation between just 53 anomalous and 10 normal test images, is not.

Of the two ingredients, the second is the more actionable. The physical part and the imaging resolution fix the size of the defects. The composition of the normal training set is a design choice that can be revisited. The same point is made from a different angle in Section 4.6, which compares SimpleNet and INP-Former: the method with substantially better localisation does not yield a meaningfully better image-level decision. If localisation were the bottleneck, INP-Former's much sharper anomaly maps should translate into a clear image-level advantage. This is consistent with the bottleneck lying not in the detector but in the data on which it was trained, specifically in the borderline samples that the model has learned to accept as normal. The same observation motivates recommendation 1 in Section 6.2 ("keep the normal class clean") and is the most concrete dataset-level lever identified by this thesis.

5.2 Discriminator saturation in SimpleNet

The diagnostic tracker introduced in Section 3.3.3 made it possible to observe a behaviour that is mentioned only in passing in the SimpleNet paper. On the sleeve dataset,

with the default hyperparameters and no gradient clipping, the discriminator occasionally enters a regime in which the predicted probabilities for both real and fake features approach 1.0 and the loss approaches zero. From a training-loss perspective, this looks like success. However, the discriminator no longer produces useful per-position scores at inference, and both the image-level and the pixel-level metrics collapse. The saturation regime seems to occur for a few specific reasons: a tiny training set, a noise standard deviation σ that is large relative to the natural feature variance, and an unclipped gradient that allows a single large update to push the discriminator past the saturation point. Lowering σ from 0.020 to 0.012 helped a bit, but the best single fix was applying gradient clipping with a global norm of 1.0.

With clipping, none of the three seeds used to produce the final SimpleNet result reached the saturated regime. Without clipping, one of the three did. This aligns with broader research on GAN training stability (Brock, Donahue, and Simonyan, 2019) and with the common issue that small adversarial setups tend to be unstable.

5.3 Defect-wise failure analysis

A purely numeric summary obscures the fact that the five defect categories of the sleeve dataset behave very differently. This section walks through each category, examining misclassified images and seeking a common cause.

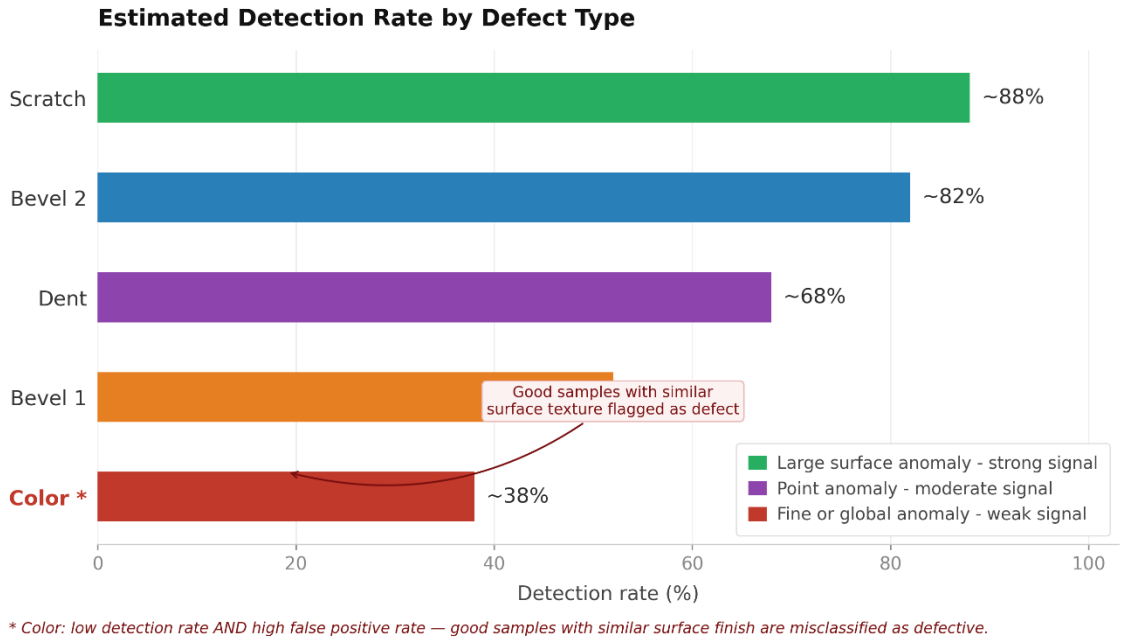

Figure 10. Estimated image-level detection rate per defect category with signal-strength classification. Horizontal bars represent the estimated detection rate (%) for each of the five defect types identified in the dataset. Bar colours encode anomaly signal strength: green denotes large-area surface anomalies that produce strong, spatially coherent activation (Scratch, ~88%; Bevel 2, ~82%); purple denotes point anomalies with moderate localised signal (Dent, ~68%; Bevel 1, ~52%); red denotes fine or global anomalies that yield a weak and diffuse signal (Colour, ~38%).

5.3.1 Scratches

Scratches are the easiest of the five defect categories to detect, at least at the pixel level. The pixel-level AUROC restricted to scratch test images is approximately 0.96 for the improved SimpleNet and approximately 0.99 for INP-Former. The misclassifications that do occur are typically for scratches that are extremely shallow, very thin, or oriented along a direction that coincides with the natural texture of the sleeve surface. These misclassifications tend to be false negatives at the image level, even when the pixel-level response is clearly present, because the anomaly map activates strongly along the scratch. Nevertheless, the peak intensity is similar to that of bright highlights on normal samples. The following recommendation is straightforward, and the existing pipeline handles scratches well.

5.3.2 Dents

Dents are the hardest category. They are tiny (often a few hundred pixels), low-contrast, sparse and often multi-instance: a single dent image may contain two or three disconnected blobs. The misclassified images in this category share a clear pattern. The false negatives tend to follow a pattern. Dent images where the defect appears as a gradual change in shading rather than a hard edge are the ones the model misses most often. The anomaly map produces a low, diffuse response that gets lost against the background reflections rather than standing out as something worth flagging. The false positives have their own logic too. Normal images that happen to catch a small reflection or carry a dark speck somewhere on the surface can look enough like a shallow dent locally for the model to treat them as suspicious. Both failure modes trace back to the same underlying difficulty: the sleeve surface is reflective enough that the boundary between a real defect and a normal surface feature is genuinely ambiguous, even to a well-trained model.

The category-level AUPRO for dents is the lowest of the five categories, and reducing the Gaussian smoothing sigma from 4 to 2 was the single change that helped most for this category, presumably because aggressive smoothing erases the already weak signal.

5.3.3 Colour anomalies

Colour anomalies are the most ambiguous category. The misclassified colour images are split into two roughly equal groups. The first group contains genuine defects whose colour is so close to the natural surface variation that the anomaly map barely activates: oxidation patches that look almost identical to normal metallic patina, for example.
The second group contains false positives caused by uneven illumination, in which a normal sleeve at a particular angle to the ring light produces a localised colour shift that the model interprets as anomalous. The colour category is the one where it is easiest to imagine human annotators disagreeing on the label, which suggests that an annotation review or a borderline-case folder could help in future iterations of the dataset.

5.3.4 Bevel 1

Bevel 1 defects are narrow structural anomalies near the bevelled edges. The local image background at the bevel is already noisy, with strong shadows and reflections. That high-contrast background makes it easy for both the model and the human eye to overlook a small defect superimposed on it. Misclassified bevel 1 images are dominated by false negatives, in which the anomaly map activates over the entire bevel region rather than focusing on the defect, leading to a low peak intensity and a low per-image score. The pixel-level AUROC for bevel 1 is reasonable, but the AUPRO is lower than for scratches and colour anomalies. Thin linear scratches (a narrow white line in the mask) produce a diffuse heatmap with no clear spatial alignment, suggesting the model lacks sufficient spatial resolution to distinguish fine linear anomalies from surface-texture variation.

5.3.5 Bevel 2

Bevel 2 defects are larger geometric anomalies that alter the contour of the bevelled region. They are visually obvious in frontal views and almost invisible in oblique views.
The misclassifications in this category are almost entirely correlated with the sleeve's orientation in the image: oblique views produce false negatives; frontal views produce correct positives. The wide gouge is similarly well-detected; the large white region in the mask corresponds to a concentrated high-activation zone in the heatmap.

Embedding-based methods carry this weakness by design. The feature extractor was pretrained on ImageNet for object recognition, not for edge or silhouette sensitivity. Hence, defects that appear right along the boundary are not well captured, and the model lacks a robust framework for interpreting them. More training data or better augmentation is unlikely to change that in any fundamental way. The architecture was not built with this kind of anomaly in mind.
Which brings up a question worth sitting with: should bevel 2 even be part of the dataset? Left in, it skews the evaluation toward something that looks like a detection problem but

is really a test of whether the model can do something outside its design. That may be intentional, and there are arguments for stress-testing a method against its own limits. However, if the intention is to measure how reliably a system catches real surface defects on this part, then a defect class that consistently falls outside what any embedding-based approach can reasonably handle probably warrants a conversation before it stays. The pragmatic answer is to keep it, because it is a real defect that the production line cares about, and to acknowledge in the evaluation that this category is intrinsically hard.

5.4 False positives and the noisy normal class

The other side of the misclassification ledger is the false positives on the ten normal test images. Visually, the false positives concentrate around three sources. The first is the bright circular highlight produced by the ring light. When the highlight falls in an unusual position relative to the part, the resulting reflection is statistically rare in the training set and gets flagged as anomalous. The second is the bevelled edge: as noted in Section 5.3.4, the high-contrast bevel produces strong anomaly responses, and a normal sample whose bevel has slightly more shadow than the average training image can be misclassified as anomalous. The third is dust and small dark specks on the surface, which are common in the training images but apparently not common enough for the model to have fully learned to ignore them.

All three failure modes have the same root cause: the normal class is not as homogeneous as the model would benefit from. A cleaner training set, either by filtering out borderline-normal samples or by acquiring new normal images under more uniform illumination, would almost certainly improve the image-level AUROC, possibly more than any algorithmic change. This is the main observation behind the recommendations in Section 6.2. A consistent source of misclassification arises from ambiguities in the annotated dataset, in which visually good samples elicit anomalous model responses because certain surface phenomena closely resemble known defect patterns. Two main failure modes were observed.

The first involves good samples that the model incorrectly flags as defective. In these cases, geometric shadows or specular reflections on the part surface produce intensity gradients that visually mimic the spatial characteristics of the bevel-2 defect class. The model has not learned to separate actual surface geometry from lighting-induced artefacts, which causes false activations near the defect decision boundary.
The second failure mode occurs in good samples that exhibit scar-like surface finishes, most likely due to tooling marks or material flow during manufacturing. These textures produce strong heatmap responses, indicating that the model treats surface discontinuities as defect-relevant features regardless of whether they represent any actual damage or functional concern. These categories are the primary reason for the low AUC scores recorded.

5.5 Comparison with the literature

The numbers reported in this thesis are difficult to compare directly with the published literature because the sleeve dataset is custom rather than a standard benchmark. The most useful comparisons are the evidence results from the MVTec AD dataset (Section 4.2) and the rankings of the distinctive methods relative to each other on the sleeve dataset (Section 4.6).

On MVTec AD, the results closely match the published numbers for SimpleNet (Liu et al., 2023), lending credibility to the implementation. On the sleeve dataset, the relative ordering SimpleNet ≈ INP-Former at the image level and INP-Former > SimpleNet at the pixel level is consistent with the strengths of each method as described in the literature: SimpleNet is tuned for fast image-level classification, while INP-Former is tuned for localisation under appearance variation. The gap between MVTec-level numbers and sleeve numbers is also broadly consistent with the motivations behind newer benchmarks such as Real-IAD (Wang et al., 2024) and the announced MVTec AD 2 successors,

both of which argue that the current generation of methods generalises poorly to realistic industrial conditions. The present thesis can be read as a small empirical confirmation of that broader claim, applied to a single industrial component.

5.6 Limitations

The dataset covers a single industrial part, so the findings may not generalise to other reflective metallic components. The annotation was performed by a single annotator with one review pass, which is acceptable for a master's thesis but falls below the standard for large benchmark datasets, which typically use multiple annotators. The MVTec AD validation was performed on three of fifteen categories, which is sufficient to rule out an implementation bug but not to claim full reproduction of the published SimpleNet results. The discriminator saturation analysis is largely qualitative, supported by a diagnostic tracker rather than by a formal statistical study. Finally, the comparison with INP-Former uses each method essentially out of the box, without per-method hyperparameter tuning on the sleeve dataset, which means the absolute numbers may underestimate what each method could achieve with more effort.

6. Conclusion and Recommendations

6.1 Summary of the work

This thesis investigated unsupervised industrial anomaly detection on a custom steel sleeve dataset using the SimpleNet framework and INP-Former as comparison baselines. The dataset was built in the standard MVTec AD layout but with deliberately realistic imaging conditions: a low-cost webcam, a ring light, a partially reflective metallic surface, and very small defects that, on average, cover only about 0.08% of the image area. The sleeve dataset contains 340 normal training images, 10 normal test images, and 53 anomalous test images across five defect categories (bevel 1, bevel 2, dent, scratch, and colour). Pixel-level masks were manually drawn for each anomalous test image.

The improved SimpleNet pipeline, which combines mild geometric augmentation, gradient clipping, a reduced Gaussian smoothing sigma, and a diagnostic feature distribution tracker, achieved I-AUROC 0.800, P-AUROC 0.930, and AUPRO 0.744 on the sleeve dataset. Validation on three standard MVTec AD categories yielded near-saturated I-AUROC scores (I-AUROC > 0.999 on Leather, Zipper, and Metal Nut), confirming that the implementation is correct and that the lower numbers on the sleeve dataset reflect the difficulty of the data rather than a bug. INP-Former reached I-AUROC 0.768, P-AUROC 0.980 and AUPRO 0.914 on the same data.

Three secondary findings supported these headline numbers. A controlled augmentation study showed that strong geometric augmentation improves image-level discrimination at the cost of region-level localisation. A small data leakage experiment showed that even a handful of training images in the test split can inflate the image-level AUROC by tens of percentage points, underscoring how easy it is to commit this mistake on small custom datasets. A qualitative misclassification analysis grouped the failure modes by defect category and identified a noisy normal class, small dark specks, machining marks, and reflective highlights as one of the main bottlenecks of the current pipeline.

6.2 Recommendations for constructing a new MVTec-style dataset

The quality of the training dataset is the single greatest determinant of model performance in industrial surface inspection. The failure patterns observed in Section 5.3, particularly the false-positive rate for the Colour defect and the near-random detection of Bevel, are largely attributable to inconsistencies in data collection rather than to limitations of the model architecture itself. The following guidelines address those root causes directly. Looking back at the experimental work, the largest single lever for improving anomaly detection performance on the sleeve component does not seem to be a change of model. It is a revision of the dataset. The following recommendations are concrete suggestions for anyone planning to build a new MVTec-style class for a reflective metallic part.

1. Keep the normal class clean. The training set should contain only samples that the production quality control would clearly accept. Samples with visible scratches, dark specks or borderline discolouration should be moved to a separate folder. The current sleeve training set contains several borderline samples (some not fully in-frame, some with marks within tolerance), and the false positives on the test set closely track the appearance of those borderline samples. The simplest improvement to the dataset would be to filter the training set in a future revision.

2. Standardise the imaging environment. The webcam-plus-ring-light setup used in this thesis is convenient and inexpensive. However, it produces a visible vignette at the corners of the frame and a bright circular highlight whose position depends on placement. Diffuse lighting, an enclosed light box, or polarised illumination would all reduce these effects. For reflective metallic components, the lighting choice tends to matter more than the camera choice. Lighting is the most critical variable in metallic surface inspection. Uncontrolled illumination is the primary source of false positives in this dataset: specular reflections and directional shadows on good parts create local brightness patterns that the model cannot distinguish from genuine surface defects. Image capturing should not be performed under ambient or natural light. Such variations may introduce

distributional shifts that undermine the robustness and reliability of anomaly detection—a dome or ring-diffuser LED panel to provide uniform, shadow-free illumination across the full cylindrical surface. Dome lights are especially effective on curved metallic parts because they suppress highlights regardless of viewing angle, making them especially advantageous for detecting colour- and surface-finish-related anomalies. Since the light is directed along the camera's optical axis, deviations in surface finish become more visually distinguishable from the normal surface response. At the same time, the formation of misleading shadows is reduced. The illumination colour temperature should be set to 5500–6500 K, and automatic white balance should be turned off on the camera. Variations in colour temperature can introduce inconsistencies in the captured images and may directly reduce the detectability of surface-finish anomalies.

3. Part position. Part positioning must be mechanically constrained. Mount the camera on a rigid fixture and place the part in a holder that constrains its position and orientation. Framing variation makes embedding-based methods harder than they need to be, and the cost of a simple physical jig is much lower than the cost of training a model to be invariant to framing. A dedicated alignment jig should be fabricated and used to ensure repeatable positioning of every inspected part. The jig should maintain the component's orientation, height, and axial rotation during each image acquisition. This is particularly important for cylindrical components, where Rotational variation is a known cause of false positives in anomaly detection because the normal surface texture appears at a different angle than in the training images. A uniform, matte, dark background should be used during image capture. The use of highly reflective, patterned, or textured backgrounds should be avoided, as they may introduce unwanted visual information near the component's boundaries. In anomaly maps, such background interference can bleed into the part region, producing misleading responses, particularly along the edges of the inspected surface. The camera should be mounted on a vibration-isolated stand to maintain imaging stability throughout the acquisition process. Even minor vibrations, including those caused by hand contact with the workbench during capture, can introduce motion blur. This reduces the apparent spatial resolution of the image and may

obscure fine defects, thereby decreasing the sensitivity and reliability of anomaly detection.

4. Camera optics. A fixed focal-length macro lens should be used for image acquisition, rather than a zoom lens. Fixed focal-length lenses provide more stable imaging geometry, whereas zoom lenses may introduce subtle variations in distortion, magnification, and apparent feature shape depending on the selected zoom setting. Such variations can reduce the consistency of defect representation across captured images. Where practical, a telecentric lens is strongly recommended, as it produces images with no perspective distortion, meaning a scratch at the edge of the part appears the same width as one in the centre. This property is particularly important for cylindrical or curved parts, where consistent feature representation is required to obtain reliable anomaly-map responses over the full inspected surface. The imaging resolution should be selected such that the smallest expected defect is represented by at least 8–10 pixels in the captured image. This ensures that fine defects are sufficiently sampled and remain detectable by patch-level or pixel-level anomaly detection methods. For example, for the Bevel 1 scratch, which is approximately 0.2 mm wide on the inspected parts, the effective pixel pitch should be no larger than approximately 20–25 µm

5. Use enough resolution. The sleeve defects are small in absolute pixel terms. The current acquisition resolution (640×640) is acceptable. However, a higher resolution combined with a tighter crop around the part would give the network more pixels per defect and would help the localisation metrics, especially AUPRO.

6. Make the test sets large enough. With only ten normal test images, a single mislabeled or unusual image can shift the AUROC by several percentage points. A target of at least 30 normal test images and at least 60 anomalous test images, distributed reasonably evenly across defect categories, would make the reported metrics much more stable. A set of roughly 30–50 normal test images and 50–80 anomalous test images is sensible, given the size of standard MVTec AD categories.

7. Keep defect categories separate. Storing scratches, dents, colour anomalies, and the two bevel categories in separate test folders made the per-defect analysis in Section 5.3 possible. A single combined “anomaly” folder would have hidden the fact that bevel 2 defects are mostly missed because of pose, while colour anomalies are mostly missed because of contrast. Per-category reporting is more work to produce, but it is much more useful for diagnosing what to fix.

8. Annotate carefully. Pixel-level masks should be compact and conservative. Over-labelled masks that include normal context around the defect inflate the apparent localisation performance and degrade AUPRO when the model correctly focuses on the defect itself. Free annotation tools such as Label Studio are sufficient for this dataset scale.

9. Document borderline cases. Some normal samples will inevitably contain marks that look defective. Document these openly. They account for a large fraction of the false positives and provide valuable information for anyone trying to reproduce the work.

10. Report defect size statistics. The mean mask area, the percentage of defective pixels, and the distribution of mask sizes per category should be reported alongside the metrics. They explain why a given dataset is easy or hard, and they justify the choice of evaluation metric. AUPRO matters more than P-AUROC when defects are very small.

11. Enforce strict split discipline. Train, test and (when relevant) validation splits should be created once, documented, and never modified during model development. The data leakage experiment in Section 4.4 is a reminder of how easy it is to inflate apparent performance through an accidental violation of this discipline.

6.3 Future work

Several directions follow naturally from the present thesis. The most direct approach is to rebuild the sleeve dataset using the recommendations above and measure how much of the remaining gap to MVTec-level numbers can be recovered through dataset engineering alone. A second direction is to investigate alternative image-level aggregation functions that are less sensitive to single-pixel maxima: the average of the top-k anomaly scores, an attention-weighted aggregation, or a small, learned head on top of the anomaly map. A third direction is to combine SimpleNet's image-level strength with INP-Former's localisation strength in a hybrid pipeline. A fourth, more ambitious direction is to extend the dataset with multi-view captures of the same part, which would directly address the bevel 2 failure mode and would align the work with the multi-view philosophy of Real-IAD. More broadly, the experience of working on this dataset reinforces a view that has been gaining traction in the industrial anomaly detection community: model architecture improvements have brought the field to the point of saturation on benchmarks, and the next round of progress is likely to come from better datasets, better acquisition protocols, and a more honest accounting of when and why methods fail. The present thesis is offered as one small contribution in that direction.

References

1. Akcay, S., AtapourAbarghouei, A., and Breckon, T. P. (2018). GANomaly: Semi-supervised anomaly detection via adversarial training. In Asian Conference on Computer Vision (pp. 622-637). Springer.
2. Akcay, S., Atapour-Abarghouei, A., and Breckon, T. P. (2019). Skip GANomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In 2019, International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
3. Batzner, K., Heckler, L., and Konig, R. (2024). EfficientAD: Accurate visual anomaly detection at millisecond-level latencies. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 128-138).
4. Bergmann, P., Lowe, S., Fauser, M., Sattlegger, D., and Steger, C. (2018). Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP) (Vol. 5, pp. 372-380).
5. Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. (2019). MVTec AD: A comprehensive real-world dataset for unsupervised anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 9592-9600).
6. Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., and Steger, C. (2021). The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. International Journal of Computer Vision, 129(4), 1038–1059.
7. Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., and Steger, C. (2022). Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localisation. International Journal of Computer Vision, 130(4), 947-969.
8. Brock, A., Donahue, J., and Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis, in the International Conference on Learning Representations (ICLR).

9. Cohen, N., and Hoshen, Y. (2020). Subimage anomaly detection with deep pyramid correspondences. arXiv preprint arXiv:2005.02357.
10. Dalal, N., and Triggs, B. (2005). Histograms of oriented gradients for human detection. In 2005, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) (Vol. 1, pp. 886-893). IEEE.
11. Defard, T., Setkov, A., Loesch, A., and Audigier, R. (2021). PaDiM: A patch distribution modelling framework for anomaly detection and localisation. In International Conference on Pattern Recognition (ICPR) (pp. 475-489). Springer.
12. Deng, H., and Li, X. (2022). Anomaly detection via reverse distillation from one-class embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 9737-9746).
13. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., and van den Hengel, A. (2019). Memorising normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 1705-1714).
14. Gudovskiy, D., Ishizaka, S., and Kozuka, K. (2022). CFLOW-AD: Real-time unsupervised anomaly detection with localisation via conditional normalising flows. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 98-107).
15. Huang, X., Bai, S., Lu, Y., and Wang, R. (2021). Surface defect detection of highly reflective metallic parts using polarisation imaging. IEEE Transactions on Instrumentation and Measurement, 70, 1-11.
16. Kingma, D. P., and Welling, M. (2014). Autoencoding variational Bayes. In the International Conference on Learning Representations (ICLR).
17. Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 1097-1105.

18. Li, C. L., Sohn, K., Yoon, J., and Pfister, T. (2021). CutPaste: Self-supervised learning for anomaly detection and localisation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 9664-9674).
19. Liu, Z., Zhou, Y., Xu, Y., and Wang, Z. (2023). SimpleNet: A simple network for image anomaly detection and localisation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20402-20411).
20. Luo, W., Cao, Y., Yao, H., Zhang, X., Lou, J., Cheng, Y., Shen, W., and Yu, W. (2025). Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
21. Lv, X., Duan, F., Jiang, J. J., Fu, X., and Gan, L. (2020). Deep metallic surface defect detection: The new benchmark and detection network. Sensors, 20(6), 1562.
22. Malamas, E. N., Petrakis, E. G., Zervakis, M., Petit, L., and Legat, J. D. (2003). A survey on industrial vision systems, applications and tools. Image and Vision Computing, 21(2), 171-188.
23. Mishra, P., Verk, R., Fornasier, D., Piciarelli, C., and Foresti, G. L. (2021). VT-ADL: A vision transformer network for image anomaly detection and localisation. In 2021, IEEE 30th International Symposium on Industrial Electronics (ISIE) (pp. 01-06). IEEE.
24. Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution grayscale- and rotation-invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 971-987.
25. Roth, K., Pemula, L., Zepeda, J., Scholkopf, B., Brox, T., and Gehler, P. (2022). Towards total recall in industrial anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 14318-14328).
26. Schlegl, T., Seebock, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In International Conference on Information Processing in Medical Imaging (pp. 146-157). Springer.

27. Schlegl, T., Seebock, P., Waldstein, S. M., Langs, G., and Schmidt Erfurth, U. (2019). f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Medical Image Analysis, 54, 30-44.
28. Song, K., and Yan, Y. (2013). A noise-robust method based on completed local binary patterns for hot-rolled steel strip surface defects. Applied Surface Science, 285, 858-864.
29. Tao, X., Gong, X., Zhang, X., Yan, S., and Adak, C. (2022). Deep learning for unsupervised anomaly localisation in industrial images: A survey. IEEE Transactions on Instrumentation and Measurement, 71, 1–21.
30. Wang, C., Zhu, H., Peng, J., Jiang, Y., Liu, T., Shu, R., Wang, L., Liu, C., Yu, Y., Lin, L., Liu, Y., and Zhang, L. (2024). Real-IAD: A real-world multiview dataset for benchmarking versatile industrial anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
31. Weimer, D., Scholz-Reiter, B., and Shpitalni, M. (2016). Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. CIRP Annals, 65(1), 417-420.
32. Yu, J., Zheng, Y., Wang, X., Li, W., Wu, Y., Zhao, R., and Wu, L. (2021). FastFlow: Unsupervised anomaly detection and localisation via 2D normalising flows. arXiv preprint arXiv:2111.07677.
33. Zagoruyko, S., and Komodakis, N. (2016). Wide residual networks. In Proceedings of the British Machine Vision Conference (BMVC).
34. Zavrtanik, V., Kristan, M., and Skocaj, D. (2021). DRAEM: A discriminatively trained reconstruction embedding for surface anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 8330-8339).
35. Zou, Y., Jeong, J., Pemula, L., Zhang, D., and Dabeer, O. (2022). Spot-the-difference self-supervised pretraining for anomaly detection and localisation. In European Conference on Computer Vision (ECCV) (pp. 392-408). Springer.