FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Anomaly Detection in Image Acquisition Systems for Industrial Computer Vision

Inês Ferreira de Almeida

Mestrado em Engenharia Informática e Computação

Supervisor: Professor Jaime Cardoso

Second Supervisor: Gonçalo Almeida

July 16, 2026

© Inês Ferreira de Almeida, 2026

Anomaly Detection in Image Acquisition Systems for Industrial Computer Vision

Inês Ferreira de Almeida

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Professor Carla Gonçalves

Examiner: Professor Bisheng Wang

Supervisor: Professor Jaime Cardoso

July 16, 2026

Resumo

Os sistemas de visão por computador desempenham um papel fundamental em ambientes industriais, particularmente em tarefas de inspeção visual automática e controlo de qualidade. No entanto, o seu desempenho depende fortemente da qualidade das imagens adquiridas, uma vez que perturbações como variações de iluminação, desfocagem e sujidade nas lentes podem comprometer as etapas subsequentes de processamento e tomada de decisão. Apesar da sua importância, a deteção de anomalias relacionadas com a aquisição continua a ser relativamente pouco explorada, sendo a maioria da investigação direcionada para a deteção de defeitos em produtos, em vez da deteção de anomalias que afetam o processo de aquisição de imagem. Este trabalho procura colmatar esta lacuna através da investigação da aplicação de técnicas de deteção de anomalias à monitorização de sistemas industriais de aquisição de imagem em condições reais de funcionamento. Para esse efeito, foi desenvolvido um benchmark composto por três cenários de inspeção industrial de rolhas de cortiça com diferentes sistemas de aquisição, processos e tipos de anomalias. Foram avaliados métodos representativos dos paradigmas *teacher-student*, *memory-bank*, *one-class classification*, baseados em mapas de distribuição, em *transformers* e em difusão, considerando o desempenho de deteção, a precisão da localização, a latência de inferência e os requisitos de memória GPU. Os resultados demonstram que a deteção de anomalias constitui uma solução eficaz para detetar perturbações relacionadas com a aquisição. Vários dos métodos avaliados alcançaram valores de AUROC ao nível da imagem superiores a 99,5% no cenário especificamente concebido para avaliar este tipo de anomalias. No entanto, os restantes cenários industriais revelaram diferenças substanciais ao nível da robustez e da capacidade de generalização. Para ultrapassar esta limitação, foi proposta uma framework de supervisão leve baseada num pequeno número de amostras anómalas anotadas ao nível da imagem. Níveis mínimos de supervisão melhoraram substancialmente o desempenho em diferentes métodos e cenários, permitindo que abordagens inicialmente menos eficazes atingissem ou ultrapassassem os melhores métodos totalmente não supervisionados. O FastFlow apresentou a melhoria mais significativa, aumentando o AUROC ao nível da imagem de 80,82% para 99,18% utilizando apenas 29 amostras anómalas anotadas e atingindo 99,80% com a disponibilização de amostras adicionais. A framework opera sobre representações de características extraídas de modelos de deteção de anomalias previamente treinados, eliminando a necessidade de voltar a treinar os modelos subjacentes e permitindo uma adaptação eficiente a condições industriais em evolução. Em suma, os resultados demonstram que a combinação de técnicas modernas de deteção de anomalias com supervisão leve constitui uma solução eficaz, computacionalmente eficiente e prática para a monitorização em tempo real da aquisição de imagem.

Palavras-chave: Deteção de Anomalias • Aquisição de Imagem • Visão por Computador Industrial • Aprendizagem Profunda • Supervisão Leve • Monitorização em Tempo Real

Abstract

Computer vision systems play a key role in industrial environments, particularly in automated visual inspection and quality control. However, their performance depends heavily on the quality of the acquired images, as disturbances such as illumination variations, blur, and lens dirt can compromise downstream processing and decision-making. Despite its importance, acquisition-related anomaly detection remains relatively underexplored, with most research focusing on product defects rather than anomalies affecting the image acquisition process. This work addresses this gap by investigating anomaly detection techniques for industrial image acquisition monitoring under realistic operating conditions. To this end, a benchmark comprising three industrial cork inspection scenarios with different acquisition systems, processes, and anomaly types was developed. Representative teacher-student, memory-bank, one-class classification, distribution-map-based, transformer-based, and diffusion-based methods were evaluated in terms of detection performance, localization accuracy, inference latency, and GPU memory requirements. The results demonstrate that anomaly detection is an effective solution for detecting acquisition-related disturbances, with several methods achieving image-level AUROC values above 99.5% in the scenario designed to evaluate such anomalies. However, the additional industrial scenarios revealed substantial differences in robustness and generalization. To address this, a lightweight supervision framework based on a small number of image-level labeled anomalous samples was proposed. Even minimal supervision substantially improved performance across different methods and scenarios, enabling previously underperforming approaches to reach or surpass the strongest fully unsupervised methods. FastFlow achieved the largest improvement, increasing the image-level AUROC from 80.82% to 99.18% using only 29 labeled anomalous samples and reaching 99.80% with additional samples. The framework operates on feature representations extracted from pre-trained models, avoiding retraining and enabling efficient adaptation to evolving industrial conditions. Overall, the results demonstrate that combining modern anomaly detection techniques with lightweight supervision provides an effective, computationally efficient, and practical solution for real-time image acquisition monitoring.

Keywords: Anomaly Detection • Image Acquisition • Industrial Computer Vision • Deep Learning • Lightweight Supervision • Real-Time Monitoring

The United Nations Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework for achieving a better and more sustainable future for all. It includes 17 goals (https://sdgs.un.org/goals) to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice.

This thesis contributes to several United Nations Sustainable Development Goals through the study and evaluation of anomaly detection methods for monitoring industrial image acquisition systems. By improving the reliability of automated visual inspection processes and enabling the early detection of acquisition-related disturbances, the proposed work supports technological innovation, operational efficiency, and more sustainable production practices.

The specific Sustainable Development Goals mentioned have the following names:

SDG 8 Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all.

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation.

SDG 12 Ensure sustainable consumption and production patterns.

SDG	Target	Contribution	Performance Indicators and Metrics
8	8.2	Supporting higher levels of industrial productivity through technological upgrading and innovation in automated visual inspection systems. The application and continuous improvement of anomaly detection at the acquisition stage increase the reliability of automated visual inspection systems, reducing inspection failures and errors, and minimizing the need for manual inspection.	Improved image-level anomaly detection performance (I-AUROC) across different acquisition conditions.
9	9.5	Enhancing scientific research and upgrading the technological capabilities of industrial sectors through the study and improvement of anomaly detection methods for industrial image acquisition systems.	Improved image-level anomaly detection performance (I-AUROC) and robustness across different industrial inspection scenarios.
12	12.2	Supporting the sustainable management and efficient use of natural resources by reducing inspection failures in industrial production caused by acquisition-related anomalies, thereby minimizing unnecessary product rejection and material waste.	Improved image-level anomaly detection performance (I-AUROC) across different acquisition conditions, reducing the likelihood of inspection failures and unnecessary product rejection.

Acknowledgements

I would like to express my sincere gratitude to everyone who, in one way or another, contributed to the completion of this dissertation.

First and foremost, I would like to thank my supervisor, Professor Jaime Cardoso, for his guidance, valuable feedback, and continuous support throughout this work.

I am genuinely grateful to JTA for all the opportunities, challenges, and experiences that have contributed to my professional and personal growth. I would especially like to thank Gonçalo Almeida, Ricardo Araújo, and Tiago Madeira for their support and insightful feedback throughout this thesis. I am particularly grateful to Ricardo for his guidance and for always taking the time to share his knowledge and experience throughout this journey.

To my friends, colleagues, and classmates, thank you for being part of my academic journey and for making these years so meaningful and rewarding. To my family, especially my parents and grandparents, I would like to express my deepest gratitude for their unwavering support and constant encouragement throughout my life.

Finally, I would also like to express my gratitude to all of my pets for their unconditional love, especially my cat, who was always by my side during the countless hours spent writing this dissertation.

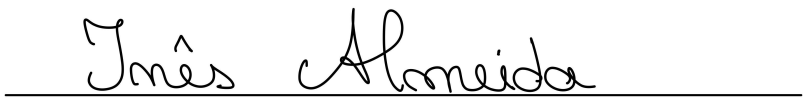
Inês Ferreira de Almeida

Declaration of honour

I declare that this dissertation is my own work and has not previously been submitted or assessed at this or any other institution. References to other authors (statements, ideas, thoughts), as well as the use of published works, strictly comply with the rules of attribution and are duly identified in the text and in the bibliographic references, in accordance with the applicable referencing standards. I am aware that the practice of plagiarism or self-plagiarism constitutes an academic offence, as established in the U.Porto Code of Ethics.

I also declare that I have used Artificial Intelligence tools such as ChatGPT to help organize ideas and improve the clarity of the writing in parts of this dissertation, and that its results have been carefully checked for errors, inaccuracies, unfounded attributions, and other forms of bias in both content and arguments.

Inês Ferreira de Almeida



"Quality means doing it right when no one is looking."

Henry Ford

Contents

1 Introduction **1**
- 1.1 Context . . . 1
- 1.2 Motivation . . . 1
- 1.3 Problem . . . 2
- 1.4 Research Questions . . . 4
- 1.5 Main Contributions . . . 4
- 1.6 Dissertation Structure . . . 6

2 Literature Review **7**
- 2.1 Literature Review Methods . . . 7
  - 2.1.1 Paper Collection . . . 8
  - 2.1.2 Paper Eligibility . . . 8
  - 2.1.3 Paper Screening . . . 9
  - 2.1.4 Paper Selection . . . 10
- 2.2 Image Quality Assessment . . . 11
  - 2.2.1 Full-Reference Image Quality Assessment . . . 11
  - 2.2.2 No-Reference Image Quality Assessment . . . 15
  - 2.2.3 Datasets . . . 18
  - 2.2.4 Metrics . . . 19
- 2.3 Anomaly Detection . . . 20
  - 2.3.1 Datasets . . . 22
  - 2.3.2 Metrics . . . 26

3 Deep Learning for Industrial Visual Anomaly Detection **28**
- 3.1 Supervised Anomaly Detection . . . 30
- 3.2 Self-Supervised Anomaly Detection . . . 33
- 3.3 Unsupervised Anomaly Detection . . . 35
  - 3.3.1 Feature embedding methods . . . 36
  - 3.3.2 Reconstruction-based methods . . . 48
- 3.4 Summary of Anomaly Detection Method Families . . . 59

4 Methodology **61**
- 4.1 Datasets . . . 61
  - 4.1.1 Acquisition-Related Anomalies Dataset . . . 62
  - 4.1.2 Foreign Object Intrusions Dataset . . . 66
  - 4.1.3 Rolling Inspection Artifacts Dataset . . . 67
- 4.2 Evaluated Models . . . 71
- 4.3 Model Evaluation . . . 71

4.4 Experimental Setup . . . 74
4.4.1 Common Anomalib Experimental Setup . . . 75
4.4.2 Model-Specific Configurations . . . 76
4.5 Improving Robustness . . . 78
4.5.1 Data Augmentations . . . 79
4.5.2 Lightweight Supervision . . . 79

5 Results **83**
5.1 Benchmark Evaluation . . . 83
5.1.1 Image-Level Detection Performance . . . 83
5.1.2 Computational Performance . . . 85
5.1.3 Localization Analysis . . . 86
5.1.4 Robustness Across Industrial Scenarios . . . 90
5.2 Lightweight Supervision Experiments . . . 95
5.2.1 Supervision on the Foreign Object Intrusions Dataset . . . 95
5.2.2 Supervision on the Rolling Inspection Artifacts Dataset . . . 97

6 Conclusion **105**
6.1 Final Remarks . . . 105
6.2 Limitations . . . 107
6.3 Future Work . . . 107

References **109**

A Chronological Summary of the considered Literature **117**

B Image Corruption Examples **124**

C Supplementary Image-Level Performance Analysis **126**
C.1 DevNet . . . 128
C.2 PatchCore . . . 130
C.3 EfficientAD-S . . . 132
C.4 EfficientAD-M . . . 134
C.5 EfficientAD-M without Autoencoder . . . 138
C.6 FastFlow . . . 142
C.7 CFA . . . 146
C.8 Dinomaly . . . 148
C.9 AST . . . 150
C.10 GLASS . . . 154
C.11 DiffusionAD . . . 158
C.12 DDAD . . . 160

List of Figures

1.1 Representative normal and anomalous samples from the three industrial application scenarios considered in the proposed benchmark. . . . 3

2.1 Schematic illustration of image quality assessment paradigms. . . . 12
2.2 Kodak test image set used as pristine reference images in several IQA datasets. . 18
2.3 Examples of synthetic distortions applied to a reference image in different IQA datasets. . . . 18
2.4 Example images from the CLIVE dataset illustrating authentic real-world distortions. 19
2.5 Example images illustrating the distinction between structural and logical anomalies. 21
2.6 Example images from all categories of the MVTec AD dataset. . . . 23
2.7 Normal and anomalous samples from representative industrial anomaly detection datasets. . . . 24

3.1 Qualitative comparison of anomaly localization results on the MVTec AD dataset. 32
3.2 Synthetic anomaly samples generated by the SDAS strategy in RealNet. . . . 35
3.3 Schematic overview of the teacher–student framework. . . . 37
3.4 EfficientAD applied to two test images from MVTec LOCO. . . . 37
3.5 Schematic overview of the one-class classification framework. . . . 39
3.6 Qualitative anomaly localization results on the MVTec AD dataset for GLASS, compared with ground truth and SimpleNet. . . . 40
3.7 Schematic overview of the distribution map framework. . . . 42
3.8 Schematic overview of the memory bank framework. . . . 43
3.9 Schematic overview of the CLIP framework. . . . 46
3.10 Schematic overview of reconstruction-based methods. . . . 49
3.11 Qualitative reconstruction results in DRAEM and DSR. . . . 50
3.12 Schematic illustration of the GAN architecture. . . . 51
3.13 Schematic comparison of local receptive field processing in convolutional networks and global context modeling through self-attention in Transformer architectures. . . . 53
3.14 Schematic illustration of the forward and reverse diffusion processes in diffusion-based models. . . . 54
3.15 Noise-to-normal reconstruction in DiffusionAD using iterative and one-step denoising. . . . 55
3.16 Schematic overview of the GPT framework. . . . 57
3.17 Example of a multi-turn interaction with AnomalyGPT. . . . 58

4.1 Standard directory structure used in the datasets, following the MVTec AD format. 62
4.2 Multi-view acquisition system used for cork inspection. . . . 63

4.3 Example of a multi-view image acquired by the inspection system, showing the body view and the two top views of the cork. 64
4.4 Representative normal and anomalous samples from the Acquisition-Related Anomalies dataset. 65
4.5 Representative examples of samples belonging to the *Boundary* anomaly type. . . 66
4.6 Representative normal and anomalous samples from the Foreign Object Intrusions dataset. 67
4.7 Representative normal and anomalous samples from the Rolling Inspection Artifacts dataset. 69
4.8 Representative samples from the unseen logo test set. 70
4.9 Representative anomalous samples and corresponding ground-truth masks used for localization evaluation. 73
4.10 Comparison of heatmaps generated using the default Anomalib normalization and the proposed visualization strategy. 76

5.1 Qualitative comparison of anomaly localization results for representative samples of each anomaly category in the Acquisition-Related Anomalies Dataset. 87
5.2 Qualitative comparison of anomaly localization results for selected samples from the Foreign Object Intrusions Dataset. 92
5.3 Qualitative comparison of anomaly localization results for selected samples from the Rolling Inspection Artifacts Dataset. 93

B.1 Example images illustrating 30 types of image corruptions. 125

C.1 Image-level ROC curves obtained with DevNet on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category. 128
C.2 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with DevNet on the Acquisition-Related Anomalies Dataset. A logarithmic scale is used for visualization purposes. 129
C.3 Image-level ROC curves obtained with PatchCore on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category. 130
C.4 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with PatchCore on the Acquisition-Related Anomalies Dataset. 131
C.5 Image-level ROC curves obtained with EfficientAD-S on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category. 132
C.6 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with EfficientAD-S on the Acquisition-Related Anomalies Dataset. . . . 133
C.7 Image-level ROC curves obtained with EfficientAD-M on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category. 134
C.8 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with EfficientAD-M on the Acquisition-Related Anomalies Dataset. . . 135
C.9 Image-level ROC curve obtained with EfficientAD-M on the Foreign Object Intrusions Dataset. 135
C.10 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with EfficientAD-M on the Foreign Object Intrusions Dataset. 136

C.11 Image-level ROC curve obtained with EfficientAD-M on the Rolling Inspection Artifacts Dataset. . . . 136
C.12 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with EfficientAD-M on the Rolling Inspection Artifacts Dataset. . . . 137
C.13 Image-level ROC curves obtained with EfficientAD-M without Autoencoder on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category. . . . 138
C.14 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with EfficientAD-M without Autoencoder on the Acquisition-Related Anomalies Dataset. . . . 139
C.15 Image-level ROC curve obtained with EfficientAD-M without Autoencoder on the Foreign Object Intrusions Dataset. . . . 139
C.16 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with EfficientAD-M without Autoencoder on the Foreign Object Intrusions Dataset. . . . 140
C.17 Image-level ROC curve obtained with EfficientAD-M without Autoencoder on the Rolling Inspection Artifacts Dataset. . . . 140
C.18 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with EfficientAD-M without Autoencoder on the Rolling Inspection Artifacts Dataset. . . . 141
C.19 Image-level ROC curves obtained with FastFlow on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category. . . . 142
C.20 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with FastFlow on the Acquisition-Related Anomalies Dataset. . . . 143
C.21 Image-level ROC curve obtained with FastFlow on the Foreign Object Intrusions Dataset. . . . 143
C.22 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with FastFlow on the Foreign Object Intrusions Dataset. . . . 144
C.23 Image-level ROC curve obtained with FastFlow on the Rolling Inspection Artifacts Dataset. . . . 144
C.24 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with FastFlow on the Rolling Inspection Artifacts Dataset. . . . 145
C.25 Image-level ROC curves obtained with CFA on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category. . . . 146
C.26 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with CFA on the Acquisition-Related Anomalies Dataset. . . . 147
C.27 Image-level ROC curves obtained with Dinomaly on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category. . . . 148
C.28 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with Dinomaly on the Acquisition-Related Anomalies Dataset. . . . 149
C.29 Image-level ROC curves obtained with AST on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category. . . . 150

C.30 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with AST on the Acquisition-Related Anomalies Dataset. A logarithmic scale is used for visualization purposes. 151
C.31 Image-level ROC curve obtained with AST on the Foreign Object Intrusions Dataset. 151
C.32 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with AST on the Foreign Object Intrusions Dataset. A logarithmic scale is used for visualization purposes. 152
C.33 Image-level ROC curve obtained with AST on the Rolling Inspection Artifacts Dataset. 152
C.34 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with AST on the Rolling Inspection Artifacts Dataset. A logarithmic scale is used for visualization purposes. 153
C.35 Image-level ROC curves obtained with GLASS on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category. 154
C.36 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with GLASS on the Acquisition-Related Anomalies Dataset. 155
C.37 Image-level ROC curve obtained with GLASS on the Foreign Object Intrusions Dataset. 155
C.38 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with GLASS on the Foreign Object Intrusions Dataset. 156
C.39 Image-level ROC curve obtained with GLASS on the Rolling Inspection Artifacts Dataset. 156
C.40 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with GLASS on the Rolling Inspection Artifacts Dataset. 157
C.41 Image-level ROC curves obtained with DiffusionAD on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category. 158
C.42 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with DiffusionAD on the Acquisition-Related Anomalies Dataset. . . . 159
C.43 Image-level ROC curves obtained with DDAD on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category. 160
C.44 Histogram of image-level anomaly scores together with the corresponding KDE, obtained with DDAD on the Acquisition-Related Anomalies Dataset. 161

List of Tables

2.1 Error-based FR-IQA metrics. . . . 13
2.2 SSIM formulation based on [13]. . . . 14
2.3 Summary of the considered FR-IQA approaches. . . . 15
2.4 Summary of the considered NR-IQA approaches. . . . 16
2.5 Overview of representative IQA datasets. Adapted from [8]. . . . 20
2.6 Evaluation metrics commonly used in IQA. . . . 21
2.7 Overview of representative industrial anomaly detection datasets. . . . 25
2.8 Metrics commonly used in industrial anomaly detection. . . . 27

3.1 Overview of deep learning-based anomaly detection paradigms according to their level of supervision. . . . 29
3.2 Summary of the considered supervised anomaly detection methods. . . . 33
3.3 Summary of the considered self-supervised anomaly detection methods. . . . 34
3.4 Summary of the considered teacher-student anomaly detection methods. . . . 38
3.5 Summary of the considered one-class classification methods. . . . 41
3.6 Summary of the considered distribution-map-based methods. . . . 42
3.7 Summary of the considered memory bank methods. . . . 44
3.8 Summary of the considered CLIP-based anomaly detection methods. . . . 48
3.9 Summary of the considered autoencoder-based anomaly detection methods. . . . 50
3.10 Summary of the considered GAN-based anomaly detection methods. . . . 52
3.11 Summary of the considered transformer-based anomaly detection methods. . . . 53
3.12 Summary of the considered diffusion-based anomaly detection methods. . . . 56
3.13 Summary of feature embedding-based anomaly detection methods and their limitations. . . . 59
3.14 Summary of reconstruction-based anomaly detection methods and their limitations. 60

4.1 Distribution of samples in the Acquisition-Related Anomalies dataset. . . . 66
4.2 Distribution of samples in the Foreign Object Intrusions dataset. . . . 67
4.3 Distribution of samples in the Rolling Inspection Artifacts dataset. . . . 70
4.4 Overview of anomaly detection methods evaluated in this work. . . . 72
4.5 Sample distribution used for the lightweight supervision experiments on the Foreign Object Intrusions dataset. . . . 81
4.6 Sample distribution used for the lightweight supervision experiments on the Rolling Inspection Artifacts dataset. . . . 82

5.1 Image-level AUROC comparison on the Acquisition-Related Anomalies Dataset. 84
5.2 Computational performance comparison of the evaluated methods. . . . 85
5.3 Image-level and pixel-level AUROC comparison across anomaly types. . . . 89
5.4 Image-level AUROC comparison on the Foreign Object Intrusions Dataset. . . . 90

5.5 Image-level AUROC comparison on the Rolling Inspection Artifacts Dataset. . . 90
5.6 Image-level AUROC comparison on the Foreign Object Intrusions Dataset using data augmentation. 94
5.7 Image-level AUROC comparison on the Rolling Inspection Artifacts Dataset using data augmentation. 94
5.8 Image-level AUROC on the Foreign Object Intrusions Dataset under different evaluation splits. 95
5.9 Image-level AUROC for supervision experiments on the Foreign Object Intrusions Dataset. 96
5.10 Image-level AUROC on the Rolling Inspection Artifacts Dataset under different evaluation splits. 97
5.11 Image-level AUROC for supervision experiments on the Rolling Inspection Artifacts dataset. 98
5.12 Image-level AUROC comparison across anomaly categories for the original unsupervised models under the initial and swapped evaluation splits. 99
5.13 Image-level AUROC comparison across anomaly categories for the supervised variants under different supervision settings. 100
5.14 Image-level AUROC comparison across anomaly categories for supervision experiments using *Roller Contamination 1* as the supervised anomaly class. 102
5.15 Image-level AUROC comparison on the unseen logo test set. 104

A.1 Chronological summary of IQA method and dataset papers considered in this thesis. 117
A.2 Chronological summary of AD method and dataset papers considered in this thesis. 119
A.3 Chronological summary of other papers considered in this thesis. 122

List of Acronyms

AD Anomaly Detection

AE Autoencoder

AFS Anomaly-aware Features Selection

AHIQ Attention-based Hybrid Image Quality Assessment Network

AI Artificial Intelligence

AST Asymmetric Student–Teacher

AU-PR Area Under the Precision–Recall Curve

AUROC Area Under the Receiver Operating Characteristic Curve

BIQI Blind Image Quality Index

BMVC British Machine Vision Conference

BRISQUE Blind/Referenceless Image Spatial Quality Evaluator

CFA Coupled-hypersphere-based Feature Adaptation

CLIP Contrastive Language–Image Pre-training

CNN Convolutional Neural Network

CSL Contextual Similarity Learning

DDAD Denoising Diffusion Anomaly Detection

Deep SAD Deep Semi-Supervised Anomaly Detection

DeepSim Deep Similarity

DevNet Deviation Network

DMOS Differential Mean Opinion Score

DSR Dual Subspace Re-projection

DTD Describable Textures Dataset

FN False Negative

FP False Positive

FPR False Positive Rate

FPS frames per second

FR-IQA Full-Reference Image Quality Assessment

FSIM Feature Similarity Index

GAN Generative Adversarial Network

GLASS Global and Local Anomaly co-Synthesis Strategy

GPT Generative Pre-trained Transformer

HVS Human Visual System

I-AUROC Image-level Area Under the Receiver Operating Characteristic Curve

IKD Informative Knowledge Distillation

IQA Image Quality Assessment

KDE Kernel Density Estimation

KRCC Kendall Rank Correlation Coefficient

MAE Mean Absolute Error

MANIQA Multi-dimension Attention Network for No-Reference Image Quality Assessment

MemSeg Memory-based Segmentation Network

MIL Multiple-Instance Learning

ML Machine Learning

MLP Multi-layer Perceptron

MOS Mean Opinion Score

MSCN Mean Subtracted Contrast Normalized

MSE Mean Squared Error

MVTec AD MVTec Anomaly Detection

MVTec LOCO AD MVTec Logical Constraints Anomaly Detection

NIQE Natural Image Quality Evaluator

NR-IQA No-Reference Image Quality Assessment

NSS Natural Scene Statistics

OCBE One-Class Bottleneck Embedding

OCC One-Class Classification

P Precision

P-AUROC Pixel-level Area Under the Receiver Operating Characteristic Curve

PaDiM Patch Distribution Modeling

PaQ-2-PiQ From Patches to Pictures

PDN Patch Description Network

PieAPP Perceptual Image-Error Assessment through Pairwise Preference

PLCC Pearson Linear Correlation Coefficient

PNI Position and Neighborhood Information

PRN Prototypical Residual Network

PRO Per-Region Overlap

PSNR Peak Signal-to-Noise Ratio

QoE Quality of Experience

R Recall

RD4AD Reverse Distillation for Anomaly Detection

ROC Receiver Operating Characteristic

RR-IQA Reduced-Reference Image Quality Assessment

RRS Reconstruction Residuals Selection

SDAS Strength-controllable Diffusion Anomaly Synthesis

SDG Sustainable Development Goal

SPAQ Smartphone Photography Attribute and Quality

SPD Spot-the-Difference

SROCC Spearman Rank Order Correlation Coefficient

sPRO Saturated Per-Region Overlap

SSIM Structural Similarity Index

SSL Self-Supervised Learning

SSPCAB Self-Supervised Predictive Convolutional Attentive Block

SSMCTB Self-Supervised Masked Convolutional Transformer

STFPM Student–Teacher Feature Pyramid Matching

SOTA State-of-the-Art

TN True Negative

TP True Positive

TPR True Positive Rate

UQI Universal Quality Index

US Uninformed Students

VisA Visual Anomaly

WinCLIP Window-based CLIP

YOLO You Only Look Once

Chapter 1

Introduction

This chapter presents the context and motivation of the thesis, defines the problem under study, introduces the research questions that guide its development, summarizes its main contributions, and outlines its structure.

1.1 Context

Computer vision systems have become increasingly relevant in industrial environments, particularly in tasks involving automatic visual inspection and quality control. The integration of Artificial Intelligence (AI) and Machine Learning (ML) methodologies has substantially advanced the automation of complex detection and classification processes, achieving levels of accuracy and operational efficiency that were previously out of reach [1].

However, the performance of these systems depends heavily on the quality of the images acquired during the capture process, since even small degradations at this stage, such as illumination variations, can compromise the reliability of computer vision algorithms used in subsequent stages.

In this context, ensuring consistent image quality at the acquisition stage becomes a fundamental requirement for the reliability of industrial computer vision systems. When the input data is compromised, even the most advanced algorithms cannot deliver trustworthy results, making image acquisition a critical component of the entire vision pipeline [2].

1.2 Motivation

Despite its importance, the image acquisition stage remains relatively unexplored. Most developments in computer vision systems have focused on later stages of the vision pipeline, often assuming that the captured images are always suitable for processing. In real-world industrial settings, however, this assumption rarely holds true [3].

Lighting variations, mechanical vibrations, dust, lens misalignment, or the occasional presence of cables or misplaced components can alter or obstruct the captured image. When unnoticed,

these disturbances may compromise not only detection accuracy but also overall process stability. In high-throughput industrial environments, such degradations can quickly propagate through the system, leading to incorrect decisions, unnecessary product rejections, or even system failures.

As a case study, this thesis focuses on the cork manufacturing industry, where visual inspection plays a key role in automating quality control processes. In this context, ensuring image quality at the acquisition stage is particularly critical, as subtle variations in texture and color can significantly impact product classification and overall production efficiency. Although developed and validated within this industrial setting, the proposed approach is designed to be extensible to other manufacturing domains that rely on consistent image quality to enable reliable automated inspection and decision-making.

This thesis was developed in collaboration with JTA[1], a data and AI engineering consultancy. This collaboration provided practical industrial insight, access to real-world production environments, and domain expertise in computer vision, ensuring that the addressed challenges, constraints, and validation scenarios reflect realistic conditions encountered in operational industrial inspection systems.

1.3 Problem

Anomalies occurring during image acquisition can undermine the reliability of industrial computer vision systems. Nevertheless, the detection of such acquisition-related anomalies remains relatively underexplored in current industrial computer vision research.

Although some approaches, such as Image Quality Assessment (IQA) methods, are designed to assess the perceptual quality of images, they primarily focus on quantifying distortions such as noise, compression artifacts, or blur [4]. As a result, these methods are not well suited for detecting acquisition anomalies caused by the presence of unexpected intruders, such as random objects or wires appearing in the field of view. Such intrusions may not significantly affect conventional image quality scores, yet they introduce foreign visual patterns that can invalidate the assumptions of downstream vision models and compromise their reliability.

In contrast, Anomaly Detection (AD) methods are inherently designed to identify irregular visual patterns and deviations from normal behavior. However, most existing research has focused on product-level inspection, targeting defects such as scratches, cracks, or surface irregularities [1], [5]. As a result, acquisition-related anomalies, including lens blur, dust accumulation, camera misalignment, or illumination disturbances, have received comparatively less attention despite their critical impact on the stability and reliability of industrial visual inspection pipelines.

Therefore, this thesis addresses the identified gap by proposing a comprehensive benchmark of anomaly detection approaches applied to the image acquisition stage of industrial computer vision systems. Representative methods from different anomaly detection paradigms are evaluated and compared to assess their suitability for acquisition-related anomaly detection under realistic industrial conditions.

[1] https://www.thedatascientists.com/

The proposed benchmark considers three industrial application scenarios, each involving distinct inspection systems, acquisition processes, and anomaly types. This diversity allows the robustness of the evaluated methods to be assessed under different industrial acquisition conditions. Representative normal and anomalous samples from each dataset are shown in Figure 1.1.

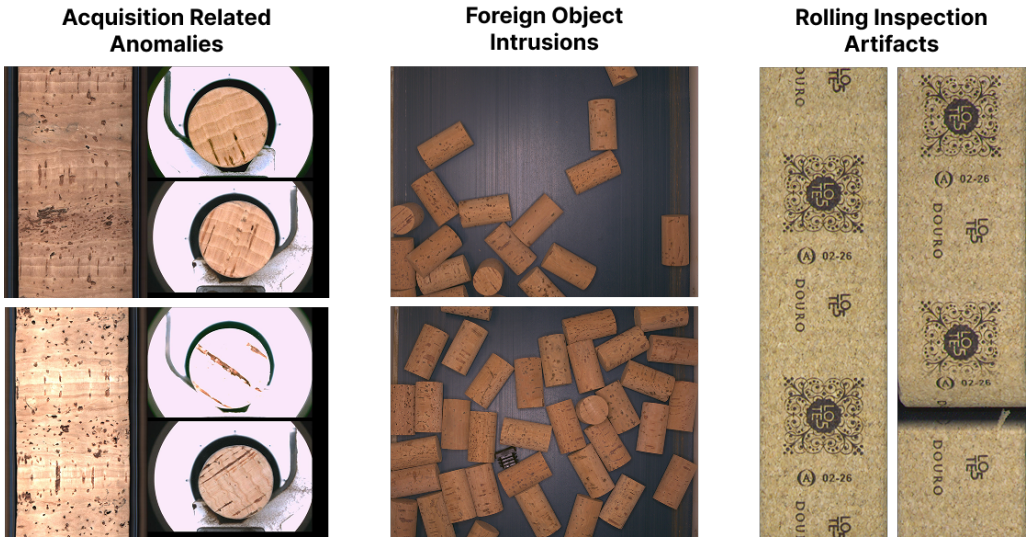


Figure 1.1: Representative normal and anomalous samples from the three industrial application scenarios considered in the proposed benchmark.

Developing effective solutions in this context presents two main challenges. First, acquisition-related anomalies are often unpredictable and may evolve over time, making it difficult to anticipate all relevant anomalous conditions in advance. Second, industrial inspection systems typically operate under strict real-time constraints, requiring anomaly detection methods that can monitor image acquisition without introducing significant computational overhead.

To ensure a meaningful and practically relevant evaluation, the proposed benchmark evaluates the considered methods in terms of anomaly detection performance, anomaly localization capability, inference latency, and GPU memory requirements. Through this evaluation, the thesis aims to identify the most suitable approaches for acquisition-related anomaly detection and investigate strategies for improving their robustness under industrial operating conditions.

Addressing this problem enables the deployment of real-time anomaly detection mechanisms capable of alerting operators to acquisition issues before image degradations propagate through the rest of the inspection pipeline. Early detection of such issues can prevent unnecessary product rejections caused by persistent acquisition faults, while also minimizing downtime and facilitating timely corrective interventions. Ultimately, this contributes to more reliable, efficient, and dependable industrial computer vision systems.

1.4 Research Questions

Based on the motivation and problem statement, this thesis is guided by the following research questions:

RQ1 How is image quality typically assessed in industrial computer vision systems, and what limitations do these approaches present at the image acquisition stage?

RQ2 Which anomaly detection approaches are most effective for identifying acquisition-related anomalies in real time?

RQ3 What strategies can be employed to improve the robustness and reliability of anomaly detection methods under industrial image acquisition conditions?

These questions define the scope of the study and guide both the analysis of existing approaches and the development of effective solutions for real-time anomaly detection at the image acquisition stage in industrial settings.

1.5 Main Contributions

The main contributions of this work are summarized as follows:

- A detailed state-of-the-art analysis of deep learning-based anomaly detection methods for industrial visual inspection. This analysis provides a structured review and comparison of the main anomaly detection paradigms and representative methods, highlighting their strengths, limitations, data requirements, and suitability for acquisition-related anomaly detection.
- The development of an industrial dataset for acquisition-related anomaly detection. Unlike existing industrial anomaly detection datasets, which primarily focus on product defects, the proposed dataset was designed to evaluate anomaly detection methods in image acquisition monitoring and includes representative acquisition-related disturbances, such as illumination variations, defocus blur, lens dirt, and partial occlusions.
- The development of a comprehensive benchmark for anomaly detection in industrial image acquisition systems. Based on the findings of the state-of-the-art analysis, representative methods employing different anomaly detection strategies are systematically evaluated and compared in terms of detection performance, localization capability, computational efficiency, and memory requirements.
- Experimental evidence demonstrating the effectiveness of modern anomaly detection methods for industrial image acquisition monitoring. Results obtained on the proposed Acquisition-Related Anomalies dataset show that methods employing different anomaly detection strategies can reliably detect acquisition-related disturbances, with several evaluated methods

achieving image-level AUROC values above 99.5% while meeting the real-time requirements of the considered industrial applications.

- An analysis of anomaly detection robustness across different industrial inspection scenarios. By extending the evaluation to three industrial datasets that cover distinct inspection systems, acquisition processes, and anomaly types, the benchmark enables the assessment of method generalization under diverse industrial conditions. The obtained results show that not all methods generalize equally well across datasets and anomaly types. While some approaches exhibit substantial performance variations depending on the considered scenario, methods such as GLASS and AST consistently achieve strong performance across multiple industrial datasets, demonstrating superior robustness under diverse operating conditions.

- A lightweight supervision approach for improving and continuously adapting anomaly detection systems using a small number of image-level labeled anomalous samples. The proposed approach enables newly observed anomalous conditions to be efficiently incorporated by training a lightweight classification head on top of feature representations extracted from pre-trained anomaly detection models. This adaptation process avoids the need to retrain the original models, resulting in a computationally efficient strategy for maintaining performance in evolving industrial environments.

- A comprehensive experimental assessment of the proposed lightweight supervision approach. The obtained results consistently demonstrate performance improvements across different methods, numbers of labeled samples, train–test splits, and industrial scenarios. In several cases, methods that achieved comparatively poor performance in the fully unsupervised setting were able to match or outperform the strongest unsupervised approaches after the incorporation of a small number of labeled anomalous samples. The most significant improvements were observed for FastFlow, with image-level AUROC increasing from 80.82% to 99.18% using only 29 labeled anomalous samples and further improving to 99.80% when additional labeled samples were available. In addition, augmentation-based supervision frequently provided further performance gains, although its effectiveness depended on the considered method and dataset.

Taken together, these contributions advance the application of anomaly detection to industrial image acquisition monitoring through the development of dedicated benchmarking resources, a comprehensive analysis of state-of-the-art methods, and the proposal of an effective supervision strategy for improving robustness and adaptability in real-world industrial environments.

The results obtained in this thesis have also been disseminated through the submission of a scientific paper to the British Machine Vision Conference (BMVC), one of the major international conferences in the field of computer vision and related areas, where it is currently under review.

1.6 Dissertation Structure

In addition to the introduction, this dissertation contains five more chapters.

Chapter 2 presents the literature review and its methodology, introducing the main concepts, datasets, and evaluation metrics related to image quality assessment and anomaly detection in industrial computer vision systems. Chapter 3 reviews deep learning-based anomaly detection methods for industrial visual inspection, discussing the main anomaly detection paradigms, representative methods, and their strengths and limitations.

Chapter 4 presents the methodology adopted in this thesis. It describes the datasets, anomaly detection methods, evaluation procedures, and experimental setup used throughout the benchmark, as well as the proposed lightweight supervision framework and the methodology adopted for its assessment. Chapter 5 presents and discusses the experimental results obtained from the benchmark evaluation and lightweight supervision experiments. The chapter analyzes the detection performance, localization capability, computational efficiency, and robustness of the evaluated methods, together with the improvements achieved through the proposed supervision framework.

Finally, Chapter 6 summarizes the main conclusions of this thesis and outlines directions for future work.

Chapter 2

Literature Review

This chapter reviews the literature that provides the conceptual foundation for this thesis, focusing on approaches relevant to the detection of acquisition-related anomalies in industrial vision systems. Its objective is to establish the key concepts, terminology, datasets, and evaluation practices that inform the methodological choices made throughout the remainder of this work.

This review focuses on two research areas relevant to this work, namely IQA and visual AD. IQA is discussed to contextualize common image degradations that arise during image acquisition and to outline how such degradations are typically characterized and assessed in the literature. Visual anomaly detection is examined as the primary framework for identifying abnormal acquisition conditions in industrial settings, as it enables the detection of deviations from expected visual patterns that may not necessarily correspond to perceptual quality degradation but can still compromise downstream inspection tasks.

The literature review methodology adopted in this work is described in Section 2.1. Section 2.2 introduces the main principles and representative methods of IQA, while Section 2.3 presents the fundamentals of visual AD, including the different anomaly types. The latter two sections also introduce representative datasets and describe the evaluation metrics commonly used to assess the corresponding methods. Together, these sections provide the necessary background for the state-of-the-art analysis and model selection presented in the following chapter.

2.1 Literature Review Methods

The present literature review adopts an exploratory approach, which is appropriate for emerging and evolving research areas such as anomaly detection in image acquisition systems for industrial computer vision. This exploratory design allowed research strategies and selection criteria to be refined iteratively as the understanding of the topic deepened.

Although the main focus of the review is on AD, a broader search also included literature on IQA, which was examined at a higher level to provide contextual insights into methods related to image acquisition quality.

2.1.1 Paper Collection

The literature search was conducted using Google Scholar, Scopus, IEEE Xplore, arXiv, and ScienceDirect. Search results were initially sorted by relevance, with priority given to peer-reviewed publications, papers published in reputable scientific venues, and works with a high number of citations.

In addition to database-driven searches, a complementary exploratory web search was conducted using Google, which led to the identification of a limited set of GitHub repositories that proved valuable for broadening the scope of the review and tracking recent developments. These resources include:

- a collection of papers on anomaly detection across multiple data modalities, including images, videos, time series, graphs, and text [6];
- Anomalib [7], an open-source deep learning library that aggregates state-of-the-art visual anomaly detection methods and provides standardized implementations and benchmarking tools for both public and private datasets;
- a comprehensive repository compiling IQA papers and datasets [8].

These complementary resources supported the identification of emerging research trends, widely used benchmarks, and representative implementations, thereby strengthening the completeness, timeliness, and overall relevance of the literature review.

2.1.2 Paper Eligibility

Studies were included if they addressed topics relevant to the detection of acquisition-related anomalies, with a primary focus on anomaly detection in computer vision and complementary contributions from image quality assessment.

For both IQA and visual anomaly detection, the following types of contributions were considered eligible:

- Methodological contributions, covering Full-Reference Image Quality Assessment (FR-IQA) and No-Reference Image Quality Assessment (NR-IQA) approaches in image quality assessment, as well as supervised, self-supervised, and unsupervised methods in visual anomaly detection;
- Dataset papers, describing representative and commonly used datasets for training and validation of IQA and visual anomaly detection methods;
- Evaluation and metric-related studies, focusing on performance metrics and evaluation protocols relevant to industrial visual inspection;
- Survey papers, providing structured overviews of methods, datasets, and open challenges in IQA and industrial anomaly detection;

- Other studies, selected for the relevance and clarity of their visual examples.

Studies were excluded if they focused on non-visual modalities, such as text, audio, time series, or purely tabular data, or if their application domain was not transferable to industrial visual inspection scenarios.

To reflect current methodological trends and recent advances in the field, works published from 2021 onward were prioritized. Earlier publications were included when identified through backward snowballing from the selected surveys, particularly when they introduced foundational concepts, widely adopted benchmarks, or baseline methods that continue to influence contemporary approaches. This strategy ensured methodological continuity while maintaining a strong emphasis on recent advances.

2.1.3 Paper Screening

The screening process was conducted in several stages to progressively refine the set of collected papers.

Initially, the titles of the retrieved articles were examined to quickly identify works potentially related to the research topic. In the second stage, the abstracts were reviewed to assess the relevance of each paper to AD in images and IQA.

For studies that appeared relevant based on the abstract, the figures and key sections such as methodology, experiments, and results were examined to confirm their contribution and applicability to the research scope. For papers considered particularly relevant, a full-text review was conducted to gain a deeper understanding of their methods, findings, and limitations.

In addition to topical relevance, indicators of community adoption and impact were considered prior to the inclusion of specific methods. These included citation counts and the level of engagement with publicly available implementations, measured by the number of stars in associated open-source repositories. While citation counts reflect academic recognition over time, repository popularity was used as a complementary signal to identify recent methods that, despite having fewer citations, demonstrated significant adoption and interest within the research community.

Furthermore, AI-based tools, specifically NotebookLM, were used to support the screening process, particularly for summarizing and reviewing longer articles. NotebookLM was selected due to its ability to operate exclusively on user-provided source documents, allowing each analysis to be constrained to a specific paper rather than relying on external or generalized knowledge. When querying the content, the tool generates responses grounded in the provided document and explicitly highlights the corresponding sections of the paper on which each response is based. This functionality facilitated efficient navigation of lengthy manuscripts and supported the rapid identification of relevant sections. Final inclusion decisions and conclusions were, however, always validated through manual inspection, ensuring that interpretation and selection remained guided by human judgment.

This process ensured that only studies with a clear contribution to the topic were included in the final selection.

2.1.4 Paper Selection

At the end of this process, a total of 83 papers were selected for inclusion in this literature review. The final set of studies spans multiple research directions within the scope of IQA and anomaly detection.

To enable a structured and systematic analysis, the reviewed literature was organized according to the nature of the proposed methods. Image quality assessment approaches were categorized into FR-IQA and NR-IQA methods. Anomaly detection methods were first grouped according to their learning paradigm, namely supervised, self-supervised, and unsupervised approaches. The unsupervised methods, which constitute the primary focus of this thesis, were further organized based on their underlying methodological principles into feature embedding–based methods and reconstruction-based methods.

Feature embedding–based methods include teacher–student frameworks, one-class classification methods, distribution-map–based methods, memory-bank methods, and contrastive language–image pre-training approaches. Reconstruction-based methods encompass autoencoders, GANs, transformer-based models, diffusion-based approaches, and GPT-based architectures. The state-of-the-art analysis presented in the following chapter adheres to this organizational structure.

In addition to methodological contributions, the review includes survey papers that provide broader context. In particular, the surveys [1] and [5], which comprehensively review deep learning–based industrial image anomaly detection, served as primary reference points. These surveys informed the organization of method categories, benchmark selection, and the identification of prevailing challenges and limitations.

Overall, the final selection comprises survey papers, methodological contributions, dataset descriptions, and experimental studies, encompassing a wide range of architectures such as Convolutional Neural Networks (CNNs), transformer-based models, and contrastive learning frameworks. Priority was given to recent state-of-the-art works, while foundational studies identified through the snowballing process were also incorporated to ensure methodological continuity.

All 83 papers meeting the established eligibility and screening criteria are listed in the References section of this document and form the basis for the analysis and discussion presented in the subsequent chapters. The References section contains a total of 97 entries, as it additionally includes links to other sources, such as official implementation repositories and websites used solely for illustrative figures. Appendix A further complements this review by providing a structured chronological overview of the considered literature, organized into three tables: one summarizing IQA methods and datasets, a second covering anomaly detection methods and datasets, and a third grouping complementary works such as survey papers and conceptual references. This organization facilitates a clear and coherent understanding of the evolution, scope, and role of the reviewed studies.

2.2 Image Quality Assessment

IQA is a fundamental task in computer vision that aims to estimate the perceptual quality of images in a way that aligns with human visual perception, commonly modeled by the Human Visual System (HVS). Its importance has increased significantly with the rapid growth of visual content delivered through connected devices and streaming platforms, where reliable image quality evaluation is essential to optimize the Quality of Experience (QoE) of end users [4].

IQA plays a critical role in a wide range of image processing and computer vision applications, including image acquisition, compression, transmission, restoration, and adaptive reconstruction [4]. Beyond general-purpose applications, IQA is also particularly relevant in medical imaging, where image quality directly impacts diagnostic accuracy and clinical decision-making.

In practice, image quality degradation can arise from a wide variety of corruption types. Common examples include noise, blur, compression artifacts, contrast changes, color distortions, and structural degradations, which may occur at different severity levels. Appendix B presents representative examples of 30 corruption types to illustrate this diversity [9].

To assess their effectiveness, IQA methods are typically evaluated according to their correlation with human subjective quality assessments, obtained through large-scale psychovisual experiments that produce quality scores such as the Mean Opinion Score (MOS) or the Differential Mean Opinion Score (DMOS).

Traditionally, IQA approaches are categorized according to the amount of information available from the original, undistorted reference image. FR-IQA methods assume complete access to the reference image and assess quality by directly comparing it with the distorted version. In contrast, NR-IQA, also known as Blind IQA, operates without any reference information, making it more challenging but also more practical in real-world scenarios. Reduced-Reference Image Quality Assessment (RR-IQA) lies between these two extremes, relying on a limited set of features extracted from the reference image. This conceptual distinction between image quality assessment paradigms is illustrated in Figure 2.1 [10].

Within the scope of this work, only a selected set of representative FR-IQA and NR-IQA methods is presented, as the objective is not to provide an exhaustive survey of the IQA literature, but to understand the core concepts and underlying approaches of these methods.

2.2.1 Full-Reference Image Quality Assessment

FR-IQA approaches estimate visual quality by directly comparing a distorted image with its corresponding pristine reference image. By exploiting the availability of the original image, FR-IQA methods aim to quantify the degradation introduced by distortions through objective measurements computed over pairs of images. These approaches are widely used in image processing and computer vision, as they provide a controlled and reproducible framework for evaluating image quality. However, their applicability is limited to scenarios where a pristine reference image is available.

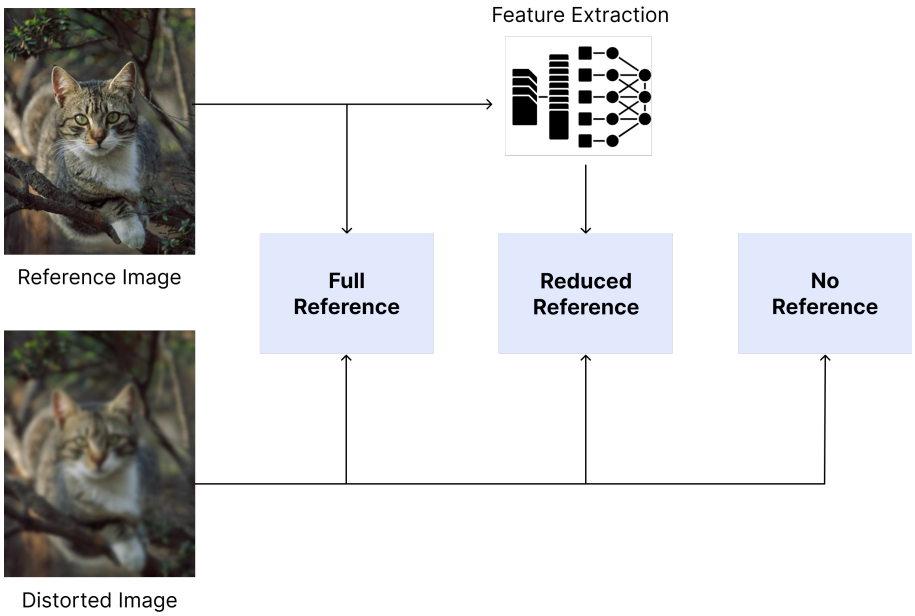


Figure 2.1: Schematic illustration of image quality assessment paradigms. Adapted from [10].

FR-IQA approaches mainly differ in how perceptual similarity between images is modeled. Early and widely adopted approaches rely on direct comparisons of pixel intensities or low-level image statistics, while more recent methods leverage perceptual modeling and learned feature representations to better align with characteristics of the HVS.

Traditional FR-IQA approaches include error-based and perceptual similarity formulations. Error-based approaches, such as the Mean Absolute Error (MAE), the Mean Squared Error (MSE) [11], and the Peak Signal-to-Noise Ratio (PSNR) [12], quantify image degradation by computing pixel-wise differences between the reference and distorted images. MAE and MSE measure absolute and squared pixel-wise discrepancies, respectively, while PSNR expresses reconstruction fidelity on a logarithmic scale derived from the MSE. However, they often correlate weakly with perceived visual quality, particularly in the presence of structural distortions, such as changes in edge continuity, loss of local spatial relationships, or geometric inconsistencies that strongly affect visual perception while producing relatively small pixel-wise errors [13]. The mathematical formulations of these error-based metrics are summarized in Table 2.1.

To better capture perceived image quality, perceptual similarity approaches have been proposed. Among these, Structural Similarity Index (SSIM) [13] is one of the most influential FR-IQA approaches. Instead of relying on pixel-wise error measurements, SSIM evaluates image quality by comparing local patterns of luminance, contrast, and structural information between the reference image and its distorted version, in accordance with the characteristics of the HVS. By explicitly modeling structural information, SSIM represents a significant advancement over traditional error-based approaches and has demonstrated improved correlation with human subjective quality judgments. The complete formulation of SSIM [13] is presented in Table 2.2 following the original definition. Since its introduction, SSIM has inspired a large number of extended and

Table 2.1: Error-based FR-IQA metrics.

Metric	Formula	Remarks / Description
Mean Absolute Error (MAE)	$\mathrm{MAE} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\lvert X(i,j)-Y(i,j)\rvert$	Measures the average absolute pixel-wise difference between a reference image and a distorted image $X(i,j)$ and $Y(i,j)$ represent the pixel intensity values of the reference and distorted images
Mean Squared Error (MSE)	$\mathrm{MSE} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(X(i,j)-Y(i,j))^2$	Measures the average squared pixel-wise difference between a reference image and a distorted image M and N denote image dimensions
Peak Signal-to-Noise Ratio (PSNR)	$\mathrm{PSNR} = 10\log_{10}\left(\frac{L^2}{\mathrm{MSE}}\right)$	PSNR quantifies how small the reconstruction error is relative to the maximum signal intensity L denotes the maximum possible pixel intensity value of the image dynamic range

derivative approaches. While these variants have demonstrated improved performance in specific scenarios, they are beyond the scope of this work and are therefore not discussed in detail.

While SSIM focuses on structural similarity derived from local luminance and contrast statistics, Feature Similarity Index (FSIM) [14] extends this paradigm by explicitly emphasizing perceptually significant low-level features that are more directly aligned with human visual perception. In particular, FSIM employs phase congruency as a primary feature to capture salient structural information and incorporates gradient magnitude as a complementary feature to account for contrast variations. Local feature similarities are aggregated through a weighted pooling strategy that assigns greater importance to perceptually salient regions, resulting in improved consistency with subjective quality assessments compared to SSIM and related structural similarity approaches.

More recent FR-IQA methods rely on learned representations to model perceptual similarity. Deep Similarity (DeepSim) [15] exploits feature representations extracted from deep convolutional neural networks pre-trained for image classification. In this approach, multi-level feature maps are extracted from both the reference and distorted images, and perceptual similarity is computed by comparing corresponding feature maps using local similarity functions inspired by structural similarity metrics. These local similarities are progressively aggregated through hierarchical pooling to produce a global quality score.

Perceptual Image-Error Assessment through Pairwise Preference (PieAPP) [16] introduces a learning-based FR-IQA model trained explicitly to align with human perceptual judgments. Instead of relying on absolute quality scores such as Mean Opinion Score (MOS) or Differential Mean Opinion Score (DMOS), PieAPP is trained using pairwise human preference data, where observers indicate which of two distorted images is perceptually closer to a reference image. The

Table 2.2: SSIM formulation based on [13].

Component / Step	**Formula**	**Remarks / Description**
Mean luminance (μ)	$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$	x_i denotes pixel intensity values within the local window
Local contrast (σ)	$\sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_x)^2}$	σ_x represents the local standard deviation
Local covariance (σ_{xy})	$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)$	σ_{xy} measures joint variation between signals
Stability constants	$C_1 = (K_1 L)^2, \quad C_2 = (K_2 L)^2, \quad C_3 = \frac{C_2}{2}$	L is the dynamic range of pixel values
Similarity combination	$S(x,y) = f(l(x,y), c(x,y), s(x,y))$	Overall similarity function combining luminance, contrast, and structure
Luminance comparison $l(x,y)$	$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$	Compares relative brightness between two image patches
Contrast comparison $c(x,y)$	$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$	Compares local contrast between two image patches
Structure comparison $s(x,y)$	$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$	Measures similarity of local structural patterns
General SSIM index	$SSIM(x,y) = l(x,y)^\alpha c(x,y)^\beta s(x,y)^\gamma$	α, β, γ control the relative importance of components
Simplified SSIM	$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_x\sigma_y + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$	Standard implementation with $\alpha = \beta = \gamma = 1$
Universal Quality Index (UQI)	$UQI = \frac{4\mu_x\mu_y\sigma_{xy}}{(\mu_x^2 + \mu_y^2)(\sigma_x^2 + \sigma_y^2)}$	Special case of SSIM without stability constants

model employs a deep neural network to estimate perceptual error scores, which are optimized using a probabilistic preference formulation. Once trained, PieAPP can predict perceptual error scores for individual distorted images relative to the reference image and has been shown to achieve high correlation with human subjective quality judgments while generalizing well to previously unseen distortion types.

Attention-based Hybrid Image Quality Assessment Network (AHIQ) [17] is a learning-based full-reference IQA model designed to address limitations of traditional quality metrics when evaluating images affected by modern synthetic distortions, particularly those introduced by Generative Adversarial Networks (GANs). Such distortions often produce visually plausible but perceptually inconsistent textures that are underestimated by pixel-wise comparison metrics and patch-based predictors that lack global contextual awareness.

To overcome this limitation, AHIQ adopts a hybrid dual-branch architecture that combines global semantic representations extracted by a Vision Transformer with local texture and spatial

details provided by a convolutional neural network. An attention-guided feature fusion mechanism leverages global semantic information to guide deformable convolution on local features, enabling the model to focus on perceptually salient regions. Patch-wise quality prediction is performed using a weighted aggregation scheme, allowing regions of higher perceptual importance to contribute more strongly to the final quality score. Experimental results demonstrate that AHIQ achieves strong performance across multiple IQA benchmarks and exhibits high alignment with human subjective quality judgments, particularly on datasets containing GAN-related distortions.

The main characteristics of the discussed methods are summarized in Table 2.3, providing a concise comparison of their core principles and key properties.

Table 2.3: Summary of the considered FR-IQA approaches.

Method	**Main Principle**	**Key Characteristics**
SSIM [13]	Structural similarity modeling	Compares local luminance, contrast, and structural patterns between reference and distorted images
FSIM [14]	Feature-based perceptual similarity	Uses phase congruency and gradient magnitude to capture perceptually salient low-level features; employs feature-based weighted pooling
DeepSim [15]	Perceptual similarity computed on deep feature representations	Extracts multi-level features from pre-trained CNNs and compares corresponding feature maps; similarities are aggregated through hierarchical pooling
PieAPP [16]	Learning perceptual error from human preferences	Trained using pairwise human comparisons to estimate perceptual error scores relative to a reference image
AHIQ [17]	Hybrid global–local quality modeling guided by attention	Combines Vision Transformer and CNN features; attention-guided fusion focuses on perceptually salient regions; weighted patch-wise quality aggregation

2.2.2 No-Reference Image Quality Assessment

NR-IQA, also known as blind IQA, aims to estimate image quality without access to a reference image. This setting is well suited to the problem addressed in this work, as reference images are not available in practice. The inspected objects exhibit significant variability in appearance, including differences in position, orientation, and surface texture. For instance, in the cork industry, which serves as the case study in this thesis, cork stoppers naturally exhibit substantial variability in appearance, including differences in surface texture, printed logos, orientation, and position within the image, making the definition and alignment of a pristine reference image impractical.

Existing NR-IQA approaches encompass both traditional methods based on natural image statistics and more recent deep learning–based models that learn quality-related patterns directly from data. Table 2.4 summarizes the NR-IQA approaches considered in this study, including Blind Image Quality Index (BIQI), Natural Image Quality Evaluator (NIQE), Blind/Referenceless

Image Spatial Quality Evaluator (BRISQUE), RankIQA, TOPIQ, and Multi-dimension Attention Network for No-Reference Image Quality Assessment (MANIQA).

Table 2.4: Summary of the considered NR-IQA approaches.

Method	**Main Principle**	**Key Characteristics**
BIQI [18]	Two-stage distortion identification and quality estimation	Extracts multi-scale statistical features to estimate distortion likelihoods and combines distortion-specific quality predictions via probability-weighted aggregation
NIQE [19]	Deviation from natural image statistics without subjective supervision	Completely opinion-unaware and distortion-unaware; models Natural Scene Statistics using a multivariate Gaussian learned from pristine image patches
BRISQUE [20]	Deviation from the statistics of natural, undistorted images	Operates in the spatial domain; uses Mean Subtracted Contrast Normalized (MSCN) coefficients; statistical features are mapped to a quality score using a support vector regressor
RankIQA [21]	Learning image quality from relative quality rankings	Trained using synthetic image pairs with known quality ordering; employs a Siamese neural network followed by fine-tuning on small IQA datasets
TOPIQ [22]	Top-down quality assessment guided by semantic understanding	Uses high-level semantic features to identify perceptually important regions and analyze local distortions while maintaining computational efficiency
MANIQA [23]	Attention-based modeling of global and local image interactions	Uses a Vision Transformer backbone; introduces attention mechanisms and a dual-branch structure to predict patch-level quality scores and their relative importance

BIQI [18] is an early NR-IQA framework designed to address the limitation that many blind IQA methods assume prior knowledge of the distortion type affecting the image. BIQI is based on Natural Scene Statistics (NSS) and follows a modular two-stage strategy. In the first stage, the method estimates the presence and relative probability of different distortion types by extracting statistical features from a wavelet-domain decomposition of the image and performing distortion classification using a support vector machine. In the second stage, distortion-specific quality scores are predicted using support vector regression models trained for each distortion category. The final quality score is obtained as a probability-weighted combination of these distortion-specific predictions. BIQI established the foundation for subsequent NSS-based blind IQA methods.

NIQE [19] extends the concept of blind IQA by adopting a completely opinion-unaware and distortion-unaware formulation. Unlike approaches that rely on training with distorted images or subjective quality scores, NIQE is based exclusively on statistical regularities observed in natural, undistorted images. A multivariate Gaussian model is learned from NSS extracted from pristine image patches. Image quality is then quantified by measuring the statistical distance between the

feature distribution of a test image and the learned natural image model. This formulation makes NIQE particularly suitable for scenarios where subjective annotations are unavailable, although its performance may degrade when applied to images that significantly deviate from natural scene statistics.

BRISQUE [20] is based on a simple idea: natural images exhibit regular statistical patterns, and image distortions disrupt these patterns. Instead of detecting specific distortions such as blur or noise, BRISQUE measures how much an image deviates from the statistics of natural, undistorted images. The method operates entirely in the spatial domain. First, the image is locally normalized by removing the local mean and normalizing the local contrast, producing coefficients known as Mean Subtracted Contrast Normalized (MSCN) coefficients. For high-quality natural images, these coefficients follow a predictable distribution, whereas distortions alter the shape of this distribution. BRISQUE captures these changes using a set of statistical features, which are mapped to a quality score using a support vector regressor. Due to its simplicity and low computational cost, BRISQUE is widely used as a baseline method for blind image quality assessment.

RankIQA [21] is a learning-based NR-IQA approach designed to address the limited availability of annotated IQA datasets. Instead of learning from absolute quality scores, RankIQA learns from relative quality rankings, which are easier to generate. The method creates synthetic image pairs with known quality ordering by applying different levels of distortion to the same image. A Siamese neural network is trained to learn which image in a pair has higher quality. The learned representation is then transferred to a standard CNN and fine-tuned using small IQA datasets. This strategy enables RankIQA to learn quality-related features effectively while reducing the risk of overfitting.

TOPIQ [22] adopts a top-down strategy inspired by the HVS, where global semantic understanding guides attention toward perceptually important regions. Instead of treating all image regions equally, TOPIQ first extracts high-level semantic features and uses them to identify regions that are more relevant for quality assessment. These regions are then examined in greater detail to detect local distortions. By combining global semantic context with local distortion information, TOPIQ is able to focus on visually important areas while maintaining computational efficiency.

MANIQA [23] is a transformer-based NR-IQA method that targets complex distortions, particularly those introduced by generative models such as GANs. The method uses a Vision Transformer backbone to capture long-range dependencies across the image. To further improve quality prediction, MANIQA introduces attention mechanisms that model both global relationships between feature channels and local interactions between image patches. A dual-branch structure predicts both patch-level quality scores and their relative importance, which are combined to produce a final image quality score. This design improves robustness and has shown strong performance on challenging NR-IQA benchmarks.

2.2.3 Datasets

The evaluation of IQA methods requires datasets that expose algorithms to a broad spectrum of image degradations. For this reason, multiple IQA datasets have been introduced, differing in size, reference image availability, and distortion types.

Several datasets are constructed by applying controlled synthetic distortions of varying types and severity levels to a set of high-quality reference images. Representative examples include LIVE-IQA [24], TID2008 [25], and TID2013 [26], which use images from the Kodak test image set as pristine references, illustrated in Figure 2.2.

Figure 2.2: Kodak test image set used as pristine reference images in several IQA datasets (from [27]).

LIVE-IQA considers five types of synthetic distortions, including JPEG and JPEG2000 compression, additive Gaussian noise, Gaussian blur, and Rayleigh fast-fading noise. TID2008 extends this framework by considering 17 distortion types, while TID2013 further increases distortion diversity by incorporating color saturation changes and more complex noise and compression artifacts. Figure 2.3 illustrates example distorted images from each dataset, presented alongside the corresponding reference image.

Figure 2.3: Examples of synthetic distortions applied to a reference image in different IQA datasets. Adapted from [24], [25], [26].

In contrast, datasets captured in real-world environments contain authentically distorted images acquired under uncontrolled conditions, where pristine reference images are unavailable.

These datasets provide subjective quality annotations derived from large-scale psychovisual experiments, enabling image quality assessment without access to reference images. One representative example is the CLIVE [28] dataset, which consists of images affected by complex mixtures of authentic distortions introduced by real imaging pipelines. Sample images from this dataset, along with their corresponding MOS values, are shown in Figure 2.4, highlighting the diversity and complexity of real-world image degradations.

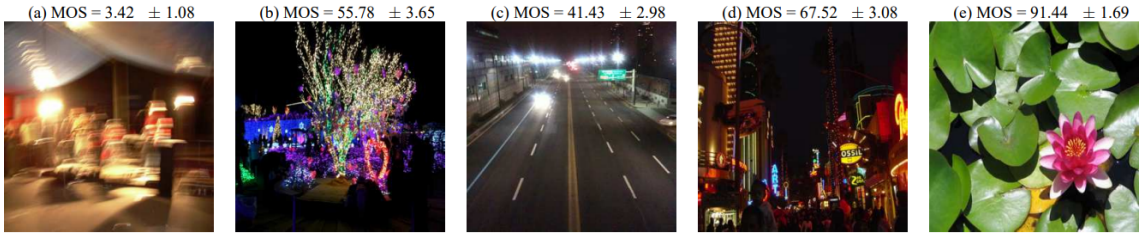


Figure 2.4: Example images from the CLIVE dataset illustrating authentic real-world distortions and the resulting variation in the quality of experience (QoE) reported by study participants. The upper label of each image indicates the corresponding MOS value together with its 95% confidence interval (from [28]).

Table 2.5 provides an overview of representative IQA datasets, detailing the number of images, the availability of reference images, and the types of distortions they contain. Collectively, these datasets cover a broad spectrum of distortion characteristics, ranging from classical synthetic degradations such as compression artifacts, blur, and additive noise to complex mixtures of authentic real-world distortions. This diversity enables a comprehensive evaluation of both FR-IQA and NR-IQA methods under controlled conditions as well as realistic in-the-wild settings.

2.2.4 Metrics

In IQA, evaluation metrics are used to quantify how well objective quality predictions produced by an IQA model agree with human subjective judgments. Since the ultimate goal of IQA is to approximate human visual perception, these metrics play a crucial role in assessing model performance and perceptual validity.

Unlike objective image quality models, which aim to estimate the perceptual quality of an image, evaluation metrics assess the relationship between predicted quality scores and subjective opinion scores, such as the MOS or the DMOS. Therefore, evaluation metrics do not measure image quality directly, but rather evaluate the effectiveness of an IQA model in predicting human perception.

Correlation-based metrics are widely adopted in the IQA literature, as they provide a direct and interpretable measure of agreement between objective predictions and subjective scores. Three commonly used correlation coefficients are considered: the Pearson Linear Correlation Coefficient (PLCC), the Spearman Rank Order Correlation Coefficient (SROCC), and the Kendall Rank

Table 2.5: Overview of representative IQA datasets. Adapted from [8].

Dataset	**Samples (Images)**	**Reference Images**	**Distortion Types**
LIVE IQA [24]	779 distorted images	29 high-quality reference images	5 synthetic distortions: JPEG2000, JPEG, White Gaussian Noise, Gaussian Blur, Rayleigh fast-fading noise
TID2008 [25]	1,700 test images	25 reference images	17 simulated distortions with 4 distortion levels each
TID2013 [26]	3,000 test images	25 reference images	24 distortion types, including color saturation change, multiplicative Gaussian noise, and noisy compression, with 5 distortion levels
CLIVE (LIVE in the Wild) [28]	1,162 images	Not applicable (authentic distortions)	Mixed authentic distortions such as low-light noise, blur, compression artifacts, and exposure errors
KonIQ-10k [29]	10,073 images	Not applicable (in-the-wild images)	Authentic distortions sampled based on content and quality diversity
From Patches to Pictures (PaQ-2-PiQ) [30]	39,810 images (119,430 patches)	Not applicable (in-the-wild images)	Authentic and mixed distortions with high diversity in content, size, and aspect ratio
Smartphone Photography Attribute and Quality (SPAQ) [31]	11,125 pictures taken by 66 smartphones	Not applicable (smartphone images)	Real camera distortions including sensor noise, defocus/motion blur, exposure issues, color shifts, and complex mixtures

Correlation Coefficient (KRCC) [15], [17], [24]. Although related, each metric captures a different aspect of IQA performance and provides complementary information. A summary of these evaluation metrics is presented in Table 2.6.

The use of correlation-based evaluation metrics requires the availability of human subjective annotations. Since the datasets considered in this work do not include subjective human quality scores, these metrics cannot be applied for quantitative performance evaluation.

2.3 Anomaly Detection

Anomaly detection aims to identify samples that deviate from patterns regarded as normal within a given system or environment. In visual settings, this typically corresponds to detecting images whose appearance or content differs from what is expected under normal operating conditions.

Table 2.6: Evaluation metrics commonly used in IQA.

Metric	Description
PLCC	Measures the linear correlation between predicted quality scores and subjective opinion scores, typically after nonlinear regression for scale alignment.
SROCC	Measures the agreement between the rank ordering of predicted and subjective scores, indicating whether an IQA model preserves relative perceptual quality ordering.
KRCC	Measures the consistency of pairwise ordering between predicted and subjective rankings, providing a robust assessment of local ranking agreement.

This capability is particularly valuable in application domains where abnormal events are rare, diverse, or difficult to characterize exhaustively in advance. In medical imaging, anomaly detection supports the identification of visual patterns associated with pathological conditions, while in security and surveillance it enables the detection of abnormal situations without relying on predefined templates. Across these domains, the common requirement is the ability to recognize deviations from expected visual behavior under limited prior assumptions.

In industrial environments, visual anomaly detection plays a critical role in quality control, operational safety, and system reliability.

Visual anomalies can be categorized as structural or logical. Structural anomalies correspond to deviations that directly affect the visual appearance of the image by altering low-level properties such as texture, edges, intensity distributions, or geometric consistency. An example is shown in Figure 2.5b, where a pushpin is visibly bent, representing a structural deviation from its expected appearance. Logical anomalies, in contrast, arise from violations of higher-level expectations regarding object presence or quantity. In industrial inspection settings, such anomalies commonly correspond to missing components or misplaced parts. Figure 2.5c depicts a representative case in which a required pushpin is missing, even though the remaining visual elements appear consistent with normal operating conditions.

(a) Normal sample (b) Structural anomaly (c) Logical anomaly

Figure 2.5: Example images illustrating the distinction between structural and logical anomalies. Adapted from [32].

The acquisition-related anomalies considered in this work, including blur, illumination variations, occlusions and intrusions, lens dirt, and object misalignment, are treated as structural anomalies from a visual standpoint. These anomalies alter the visual appearance of the image by affecting

properties such as texture, edges, intensity distributions, spatial consistency, or object boundaries. As a result, from the perspective of a vision-based model, such anomalies are effectively handled as structural anomalies.

For these reasons, visual anomaly detection provides a more appropriate framework than image quality assessment for the problem addressed in this thesis. Image quality assessment methods are primarily designed to quantify perceptual degradation and focus on distortions that affect subjective visual quality. However, several acquisition-related failures of interest may not noticeably degrade perceived image quality while still representing critical faults at the system level. In contrast, anomaly detection methods aim to identify deviations from expected visual patterns irrespective of their perceptual impact, making them better suited for monitoring the reliability of image acquisition in industrial computer vision systems.

2.3.1 Datasets

Industrial anomaly detection research commonly relies on publicly available benchmark datasets to evaluate detection methods across a wide range of inspection scenarios and product types. In these datasets, categories typically correspond to different objects or materials inspected in industrial environments, while anomaly classes describe the defects that may occur on each product, such as scratches, dents, contaminations, missing parts, or geometric deformations.

Among the available benchmarks, MVTec Anomaly Detection (MVTec AD) [33] is one of the most established and widely adopted datasets for industrial anomaly detection and is commonly regarded as a reference benchmark in the literature. As illustrated in Figure 2.6, the dataset comprises both texture-based categories, including carpet, grid, leather, tile, and wood, as well as object-based categories such as bottle, cable, capsule, hazelnut, metal nut, pill, screw, toothbrush, transistor, and zipper. The associated anomaly classes cover a broad spectrum of surface and structural defects, including scratches, dents, contamination, cracks, holes, and other structural deformations.

Due to its widespread adoption, MVTec AD is used in this thesis as the primary reference benchmark, enabling direct and consistent comparison between the anomaly detection models considered in the following chapter.

In addition to MVTec AD, several other industrial anomaly detection datasets have been proposed to address specific limitations of earlier benchmarks or to target more challenging inspection scenarios. Figure 2.7 illustrates one normal and one anomalous sample from a representative category of several of these datasets, which are discussed in more detail below.

The Visual Anomaly (VisA) dataset [36] was introduced to address the performance saturation observed on earlier benchmarks, where many methods already achieve near-ceiling detection scores. VisA includes a diverse set of challenging inspection scenarios and product categories, such as printed circuit boards (PCB), capsules, candles, macaroni, cashews, chewing gum, and fryum snacks. The dataset contains defects including surface scratches, dents, cracks, melted metal, missing or misplaced components, and other fine-grained structural anomalies. Increased

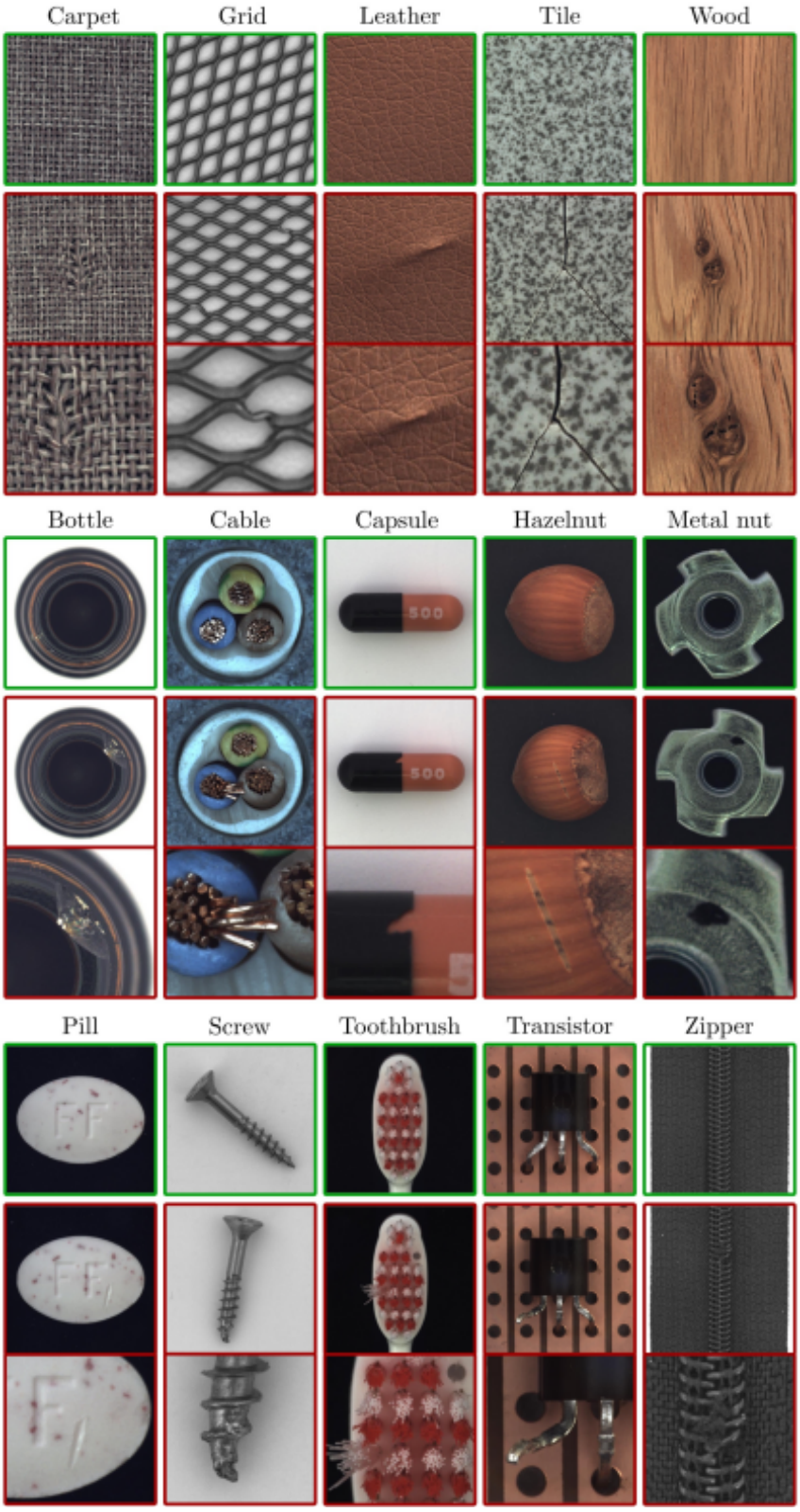


Figure 2.6: Example images for all five textures and ten object categories of the MVTec AD dataset. For each category, the top row shows an anomaly-free image. The middle row shows an anomalous example for which, in the bottom row, a close-up view that highlights the anomalous region is given (from [33]).

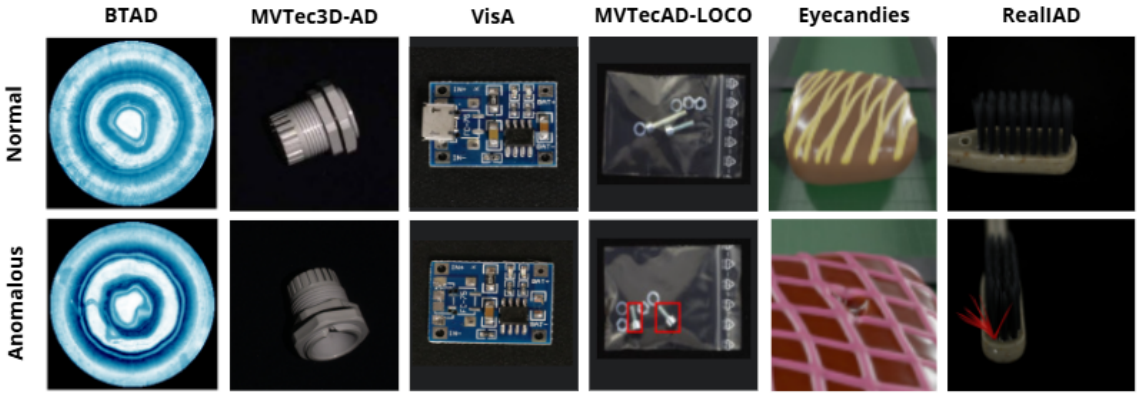


Figure 2.7: Normal and anomalous samples from representative industrial anomaly detection datasets. Adapted from [32], [34], [35], [36], [37], [38].

object complexity, the presence of multiple instances per image, and large pose variations make VisA particularly suitable for evaluating robustness and generalization.

Beyond these benchmarks, some datasets explicitly focus on specific anomaly characteristics. MVTec Logical Constraints Anomaly Detection (MVTec LOCO AD) [32] distinguishes between structural anomalies, which correspond to localized visual defects, and logical anomalies, which violate global or geometric constraints. The dataset includes categories such as breakfast boxes, screw bags, pushpins, splicing connectors, and juice bottles. Structural defects include damaged packaging, contamination, broken parts, or bent components, while logical anomalies involve incorrect filling levels, missing or extra items, swapped compartments, or incorrect labels. This dataset highlights the importance of modeling long-range dependencies and global consistency beyond local appearance cues.

Real-IAD [38] targets large-scale and realistic industrial inspection scenarios, covering a wide variety of products made from materials such as metal, plastic, wood, ceramic, and mixed materials. The dataset includes both simple components and complex assemblies, with anomalies such as holes, deformations, abrasions, scratches, damage, missing parts, foreign objects, and contaminations. In addition, Real-IAD provides multi-view coverage by capturing each object from five distinct perspectives, further increasing the realism and complexity of the anomaly detection task.

While most industrial anomaly detection datasets rely on 2D RGB imagery, certain defects are inherently geometric and are better captured using three-dimensional information. MVTec 3D-AD [35] addresses this limitation by providing aligned RGB images and 3D point cloud data. The dataset includes categories with natural shape variation, such as bagels, carrots, cookies, peaches, and potatoes, deformable objects such as foam, rope, and tires, and rigid objects like cable glands and dowels. Anomalies manifest as geometric deviations, including cracks, dents, holes, contaminations, missing material, and shape deformations, which are difficult to detect using appearance cues alone.

Finally, not all industrial anomaly detection datasets are derived from real acquisition processes. Eyecandies [37] adopts a synthetic data generation strategy to provide a scalable and controllable evaluation setting. The dataset consists of photorealistic, candy- and cookie-like objects,

Table 2.7: Overview of representative industrial anomaly detection datasets.

Dataset	Description (Categories and Defect Types)	Cat.	Normal	Anomaly	Total
MVTec AD [33]	Benchmark dataset with 5 texture categories (e.g., carpet, grid, leather) and 10 industrial object categories (e.g., bottle, cable, capsule). Defects include surface and structural anomalies such as scratches, dents, cracks, holes, and contaminations (73 defect types).	15	4096	1258	5354
BTAD [34][39]	Dataset comprising three real industrial object categories (e.g., screws, metal components). Anomalies correspond to appearance and structural deviations from normal samples, without predefined defect classes or semantic labels.	3	2250	290	2540
MVTec 3D-AD [35]	Ten object categories captured using aligned RGB images and 3D point clouds. Anomalies include geometric defects such as holes, deformations, missing material, and contaminations, which are difficult to detect using appearance cues alone (41 defect types).	10	3199	948	4147
VisA [36]	Dataset with 12 industrial object categories, including printed circuit boards (PCB), capsules, candles, and food products. Contains up to nine defect classes per object, such as scratches, dents, cracks, melted material, missing or misplaced components, and multiple anomalies per image.	12	9621	1200	10821
MVTec LOCO AD [32]	Dataset explicitly designed to distinguish between local structural anomalies (e.g., damaged packaging, contamination, broken parts) and logical anomalies (e.g., missing items, incorrect filling levels, wrong component placement). Includes categories such as breakfast boxes, screw bags, and splicing connectors (89 defect types).	5	2651	993 (432 structural + 561 logical)	3644
Eyecandies [37]	Photorealistic synthetic multimodal dataset with 10 candy- and cookie-like object categories (e.g., candy canes, gummy bears, chocolate cookies). Anomalies are procedurally generated during rendering and include color alterations, surface imperfections, scratches, bumps, and dents. Provides aligned RGB, depth, and surface normal maps.	10			∼90000
Real-IAD [38]	Large-scale real-world industrial dataset with 30 object categories covering materials such as metal, plastic, wood, and ceramic. Defects include scratches, deformations, abrasions, missing parts, foreign objects, and contaminations. Images are acquired from five different perspectives for each object.	30	99721	51329	151050

including candy canes, chocolate cookies, gummy bears, lollipops, and marshmallows. Anomalies are procedurally injected during rendering and include color alterations, surface imperfections, scratches, bumps, and dents. In addition to RGB images, depth and surface normal maps are provided to support multimodal analysis. While synthetic datasets offer advantages in scalability and precise ground-truth annotations, they may not fully capture the variability and complexity of defects observed in real industrial environments.

Table 2.7 summarizes the industrial anomaly detection datasets reviewed in this section, highlighting their object categories, scale, and types of anomalies. The exact number of normal and anomalous samples reported for the Eyecandies dataset varies across available sources and could not be determined with certainty.

2.3.2 Metrics

Evaluation metrics assess the ability of a model to distinguish between normal and anomalous samples and, when pixel-level annotations are available, to localize anomalous regions. In this work, the evaluation focuses on image-level anomaly detection, where the objective is to determine whether an input image is normal or anomalous.

Table 2.8 summarizes evaluation metrics commonly used in industrial anomaly detection, together with their mathematical definitions and typical usage.

Several of the listed metrics are derived from the basic notions of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Precision (P) measures the proportion of correctly identified anomalous samples among all samples predicted as anomalous, while Recall (R) quantifies the proportion of actual anomalies that are correctly detected. The True Positive Rate (TPR) is equivalent to recall, whereas the False Positive Rate (FPR) represents the fraction of normal samples incorrectly classified as anomalous.

The Area Under the Receiver Operating Characteristic Curve (AUROC) evaluates the trade-off between TPR and FPR across varying decision thresholds. Owing to its threshold-independent nature, AUROC is one of the most commonly used metrics for industrial anomaly detection. AUROC can be reported at two different levels: Image-level Area Under the Receiver Operating Characteristic Curve (I-AUROC), which assesses whether an entire image is correctly classified as normal or anomalous based on an aggregated anomaly score, and Pixel-level Area Under the Receiver Operating Characteristic Curve (P-AUROC), which evaluates the separability of normal and anomalous pixels.

When pixel-level ground truth is available and localization performance is of interest, additional metrics such as the Area Under the Precision–Recall Curve (AU-PR) and Per-Region Overlap (PRO) are commonly employed. AU-PR is particularly informative in highly imbalanced settings, while PRO measures the overlap between predicted anomalous regions and ground-truth defect regions. Its saturated variant (Saturated Per-Region Overlap (sPRO)) further accounts for ambiguity in region boundaries.

Table 2.8: Metrics commonly used in industrial anomaly detection. Adapted from [1].

Metric / Level	**Formula**	**Remarks / Usage**
Precision (P) ↑	$P = \frac{TP}{TP+FP}$	True positive (TP) False positive (FP)
Recall (R) ↑	$R = \frac{TP}{TP+FN}$	False negative (FN)
True positive rate (TPR) ↑	$TPR = \frac{TP}{TP+FN}$	
False positive rate (FPR) ↓	$FPR = \frac{FP}{FP+TN}$	True negative (TN)
Area under the receiver operating characteristic curve (AU-ROC) ↑	$\int_0^1 \mathrm{TPR}\, d(\mathrm{FPR})$	Classification
Area under precision-recall (AU-PR) ↑	$\int_0^1 P\, d(R)$	Localization Segmentation
Per-region overlap (PRO) ↑	$PRO = \frac{1}{N}\sum_i \sum_k \frac{P_i \cap C_{i,k}}{C_{i,k}}$	N: total GT regions P: predicted abnormal pixels C: defect regions Segmentation
Saturated per-region overlap (sPRO) ↑	$sPRO(P) = \frac{1}{m}\sum_{i=1}^{m} \min\left(\frac{A_i \cap P}{s_i}, 1\right)$	m: total GT regions P: predicted abnormal pixels A: defect regions s: saturation threshold Segmentation
F1 score ↑	$F1 = \frac{2(P \times R)}{P+R}$	Classification

Finally, the F1 score, defined as the harmonic mean of precision and recall, provides a single performance measure at a fixed decision threshold. Although useful in certain classification settings, it is less frequently emphasized in anomaly detection studies, which typically favor threshold-independent metrics such as AUROC.

Since the datasets used in this work do not provide pixel-level ground-truth annotations and localization accuracy is not a primary objective, I-AUROC is adopted as the primary evaluation metric in this thesis.

Chapter 3

Deep Learning for Industrial Visual Anomaly Detection

Anomaly Detection (AD) in visual inspection has evolved significantly over time, progressing from traditional rule-based systems to modern deep learning–based approaches [5]. Early solutions relied on handcrafted features, such as edge detectors, texture descriptors, and simple statistical measures, combined with predefined thresholds or decision rules. While these techniques could perform adequately in controlled environments, their effectiveness was limited when faced with variations such as illumination, noise, acquisition conditions, or previously unseen anomaly patterns.

Deep learning marked a fundamental shift in anomaly detection by enabling the automatic learning of hierarchical feature representations directly from raw images. In particular, CNNs significantly improved scalability and robustness to visual variability, making them well suited for complex industrial inspection scenarios.

Within this context, this chapter reviews the state of the art in deep learning–based anomaly detection for industrial visual inspection, with a particular focus on approaches that address the challenges posed by real-world image acquisition conditions. The chapter examines the main learning paradigms adopted in the literature and discusses representative methods for each paradigm in terms of their underlying assumptions, strengths and limitations.

Table 3.1 provides an overview of these anomaly detection paradigms according to their level of supervision, summarizing their data requirements as well as their main advantages and limitations [1], [5].

Based on the reviewed literature, representative anomaly detection methods were selected to establish a diverse and comprehensive benchmark spanning multiple anomaly detection paradigms. To ensure a fair and consistent comparison, model selection prioritized approaches reporting image-level AUROC values above 99% on the MVTec AD dataset, which remains the most widely adopted benchmark for industrial anomaly detection, thereby providing a common basis for performance comparison.

Table 3.1: Overview of deep learning-based anomaly detection paradigms according to their level of supervision.

Paradigm	**Required data**	**Advantages**	**Limitations**
Fully supervised anomaly detection	Labeled normal and anomalous samples from all known anomaly types, often including pixel-level annotations	High detection and localization accuracy when anomaly types are known and well represented; direct optimization for the target task	High annotation cost; limited generalization to unseen anomalies; requires representative examples of all known anomaly types
Semi-supervised anomaly detection	Large set of normal samples and a small set of labeled anomalous samples	Reduced annotation effort compared with fully supervised methods; anomaly labels improve discrimination between normal and abnormal patterns	Sensitive to class imbalance; may overfit to known anomaly types
Weakly supervised anomaly detection	Normal and anomalous samples with coarse, incomplete, or noisy annotations	Reduced annotation effort compared with fully supervised methods; can exploit imprecise annotations to improve anomaly discrimination	Needs to handle label noise; low-quality annotations can affect model performance
Self-supervised anomaly detection	Normal samples only	No anomaly annotations required; supervision is derived from pretext tasks such as pseudo-anomaly synthesis, contrastive learning, or image restoration	Performance depends on the effectiveness of the pretext task; may struggle when real anomalies differ substantially from the simulated ones
Unsupervised anomaly detection	Normal samples only	No anomaly annotations required; identifies anomalies as deviations from the learned representation of normality	Performance depends on the quality of normality modeling

3.1 Supervised Anomaly Detection

Supervised anomaly detection refers to methods that learn an explicit decision boundary between normal and anomalous samples by exploiting labeled training data. By relying on annotated examples of known anomalies, these approaches are able to directly optimize discriminative models, often achieving high detection accuracy when anomaly patterns are well defined and sufficiently represented.

Fully supervised anomaly detection methods rely on exhaustive and fine-grained annotations for all known anomaly categories. Under these conditions, CNN–based object detectors, such as You Only Look Once (YOLO) [40] and its variants, have been widely adopted in industrial inspection pipelines due to their favorable balance between detection accuracy and real-time inference performance [5].

Several studies have adapted different versions of the YOLO architecture for industrial anomaly detection. These include a lightweight YOLOv3-Tiny variant for surface anomaly inspection [41], an enhanced YOLOv5 model for detecting small-scale metal defects [42], and a more recent extension of YOLOv8 designed to capture more complex and diverse anomaly patterns [43]. These approaches demonstrate strong performance when defect types are well defined and sufficiently represented in the training data.

Despite their strong performance, fully supervised anomaly detection methods suffer from high annotation costs and limited generalization to unseen anomalies, which restrict their applicability in real-world industrial environments. In practice, industrial inspection is typically characterized by an abundance of normal data and a scarcity of anomalous samples, as defects are often rare, costly to annotate, difficult to reproduce, or unknown at deployment time.

The datasets considered in this thesis provide only image-level annotations and do not include the detailed localization labels required by many fully supervised detection frameworks. For example, YOLO-based methods typically require bounding-box annotations for training, while other approaches rely on pixel-level annotations. As a result, no fully supervised method compatible with the available annotation level and satisfying the image-level AUROC threshold established for model selection was identified during the literature review. Consequently, fully supervised methods were not considered for experimental evaluation in this thesis. This decision was further motivated by the limited ability of such methods to generalize beyond the anomaly patterns observed during training, whereas the objective of this work is to evaluate models capable of detecting acquisition-related anomalies, which can be unpredictable and may evolve or emerge over time.

These limitations have led to the development of alternative learning paradigms that reduce dependence on labeled anomalies. These include semi-supervised and weakly supervised approaches, which leverage limited abnormal data, as well as unsupervised methods that operate exclusively on normal samples, which have become the dominant paradigm in industrial anomaly detection.

Semi- and weakly supervised anomaly detection methods aim to reduce annotation costs

while maintaining strong detection performance. Representative examples include Deviation Network (DevNet) and Deep Semi-Supervised Anomaly Detection (Deep SAD), which both leverage a limited number of image-level labeled anomalies but differ fundamentally in their learning objectives.

DevNet [44] formulates anomaly detection as an end-to-end learning problem by directly optimizing anomaly scores using a deviation-based loss. By enforcing statistically significant score deviations between normal samples and the few labeled anomalies, DevNet achieves high data efficiency and strong performance when the anomaly types are well represented by the available labeled examples.

In contrast, Deep SAD [45] uses labeled anomalous samples as constraints to shape the feature representation rather than directly fitting anomaly scores. During training, normal samples are encouraged to form a compact region in latent space, while labeled anomalies are pushed away from this region. This formulation promotes learning a robust representation of normality instead of overfitting to specific anomaly instances, which generally improves generalization to unseen anomalies.

A key challenge in semi- and weakly supervised anomaly detection is the severe imbalance between normal and anomalous data, as only a limited number of labeled anomalies are typically available [5]. DRA [46] tackles the open-set nature of anomaly detection by disentangling abnormal representations into complementary components. It employs a multi-head architecture to model labeled anomalies, synthetically generated pseudo-anomalies, and latent residual abnormalities captured through feature discrepancies relative to normal references. These components are jointly optimized within a Multiple-Instance Learning (MIL) framework, encouraging sensitivity to diverse abnormal cues while reducing overfitting to limited labeled anomalies.

Prototypical Residual Network (PRN) focuses on high-precision anomaly localization under limited anomaly supervision. The method learns normality exclusively from normal samples by constructing multi-scale prototypes that capture representative feature patterns at different semantic levels. Anomalies are identified through residual feature maps, computed as deviations between input features and their nearest normal prototypes, which explicitly highlight localized abnormal regions. To handle subtle and scale-varying anomalies, PRN incorporates a multi-size self-attention mechanism. Training is performed with pixel-level supervision using losses such as Focal Loss and Smooth L1 Loss, while data imbalance is mitigated through online anomaly synthesis, including the insertion of augmented real anomaly regions into normal images and the generation of simulated anomalies using noise- and texture-based perturbations. This design enables precise anomaly localization and achieves strong performance on the MVTec AD dataset. Figure 3.1 presents a qualitative comparison of anomaly localization results on the MVTec AD dataset, comparing PRN with DRA and PatchCore, an unsupervised method described later in this chapter, and illustrating the more accurate localization behavior exhibited by PRN.

Table 3.2 summarizes the strengths and limitations of the semi- and weakly supervised anomaly detection methods discussed in this section. Although several methods have been proposed in the literature, no representative approach relying solely on image-level annotations was identified that

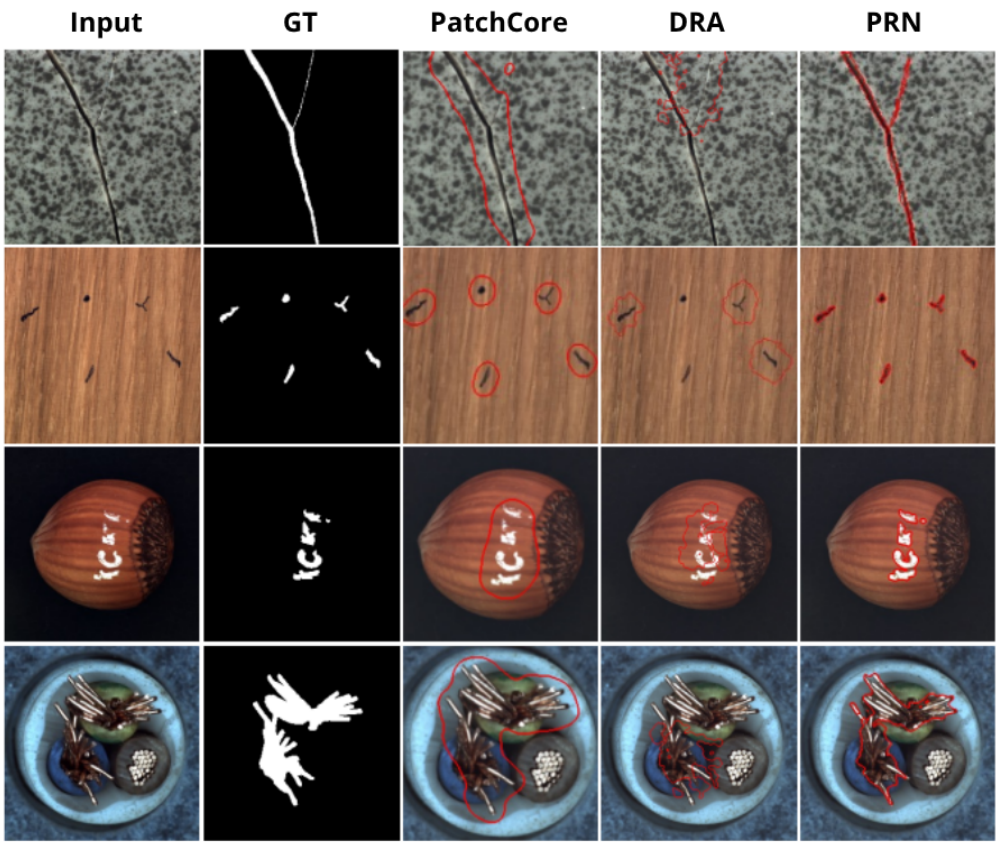


Figure 3.1: Qualitative comparison of anomaly localization results on the MVTec AD dataset for PatchCore, DRA, and PRN. Compared with the unsupervised method PatchCore and the supervised method DRA, PRN localizes anomalous regions more accurately. Adapted from [47].

satisfied the image-level AUROC threshold established for model selection. While PRN achieves image-level AUROC values above 99%, it requires pixel-level ground-truth masks during training and therefore falls outside the annotation constraints considered in this thesis. Consequently, DevNet was selected as the representative semi-supervised method for experimental evaluation, as it demonstrates stronger performance than Deep SAD while providing a more computationally efficient alternative to DRA.

Table 3.2: Summary of the considered supervised anomaly detection methods.

Method	**Strengths**	**Limitations**
YOLO [40]	Favorable trade-off between detection accuracy and real-time inference speed	Requires bounding-box annotations; high annotation cost
DevNet [44]	End-to-end anomaly scoring using few labeled anomalies; strong performance under limited anomaly supervision	Struggles when normal and anomalous samples are visually similar
Deep SAD [45]	Learns a robust representation of normality; improved generalization to unseen anomalies	Lower performance on some datasets
DRA [46]	Handles both seen and unseen anomalies; strong generalization through disentangled abnormality modeling; latent residual modeling captures subtle and hard-to-detect anomalies	Higher model complexity, making it less suitable for real-time-constrained applications
PRN [47]	Accurate pixel-level localization through direct supervision with segmentation masks; multi-scale prototype modeling; fast inference despite high localization accuracy; State-of-the-Art (SOTA) performance on MVTec AD (99.4% I-AUROC, 99.0% P-AUROC)	Requires pixel-level ground-truth masks

3.2 Self-Supervised Anomaly Detection

Self-Supervised Learning (SSL) has been widely adopted in anomaly detection, as it enables learning exclusively from normal data, avoiding the need for costly and scarce anomaly annotations. Rather than relying on explicit labels, SSL methods generate supervisory signals through pretext tasks that encourage models to learn meaningful representations of normal visual patterns. Common pretext tasks include image restoration, contrastive learning, and pseudo-anomaly generation [1], [5].

Among these strategies, pseudo-anomaly generation is particularly prevalent in the anomaly detection literature. In this setting, artificial irregularities are introduced into normal images to create surrogate abnormal samples, allowing the model to learn a decision boundary between normal and abnormal patterns in a self-supervised manner. This strategy is widely used across different paradigms and is employed by methods discussed in previous sections, such as DRA and PRN, as well as several unsupervised methods discussed later in this chapter. This highlights how

Table 3.3: Summary of the considered self-supervised anomaly detection methods.

Method	**Strengths**	**Limitations**
CutPaste [48]	Simulates defects by cutting and pasting patches within normal images	Synthetic anomalies are coarse approximations of real defects
Spot-the-Difference (SPD) [36]	Introduces subtle pseudo-anomalies through the SmoothBlend strategy	Self-supervised pre-training strategy rather than a standalone anomaly detection method
SSPCAB [49]	Self-supervised predictive block that reconstructs masked regions using contextual information; learns global structure from local features; minimal computational overhead	Architectural component rather than a standalone anomaly detection method
SSMCTB [50]	Extends masked reconstruction with enhanced channel modeling via multi-head self-attention; improves robustness to outliers through Huber loss; supports both 2D and 3D anomaly detection with consistent performance gains across domains	Architectural component rather than a standalone anomaly detection method
RealNet [51]	Generates realistic synthetic anomalies via diffusion; anomaly-aware feature selection reduces redundancy and pre-training bias; adaptive residual selection improves anomaly localization; SOTA performance on MVTec AD	Additional training cost associated with SDAS-based anomaly synthesis

self-supervised strategies often complement both supervised and unsupervised anomaly detection frameworks, rather than forming an isolated category.

Table 3.3 summarizes representative self-supervised anomaly detection methods. A well-known example is CutPaste [48], which simulates anomalies by cutting patches from normal images and pasting them at random locations. The resulting images are then used in a proxy classification task, where the model learns to distinguish between original and altered samples. Despite its conceptual simplicity, this approach has proven effective and further demonstrates how self-supervised pseudo-anomaly generation can be integrated with feature embedding methods, a paradigm discussed in the next sections.

Beyond standalone methods, self-supervised learning can also be integrated as an architectural component within larger anomaly detection models. Self-Supervised Predictive Convolutional Attentive Block (SSPCAB) [49] and its extension Self-Supervised Masked Convolutional Transformer (SSMCTB) [50] are representative examples of this trend. Both blocks introduce masked reconstruction as an intrinsic self-supervised objective, forcing the network to predict missing regions based on the surrounding context. By doing so, they encourage the learning of global structural relationships from local visual features, addressing a common limitation of CNNs. Although neither SSPCAB nor SSMCTB constitutes a standalone anomaly detection method, both have been shown to consistently improve performance when integrated into existing architectures.

Finally, RealNet [51] represents a self-supervised anomaly detection method that combines feature reconstruction with Strength-controllable Diffusion Anomaly Synthesis (SDAS) to generate realistic synthetic anomalies from normal images. Figure 3.2 illustrates example synthetic anomaly samples generated using the SDAS strategy, highlighting how diverse and realistic anomaly patterns can be synthesized from normal samples.

Figure 3.2: Synthetic anomaly samples generated by the SDAS strategy in RealNet. The SDAS strategy generates anomalous samples using only normal images (from [51]).

To further improve anomaly sensitivity, RealNet incorporates the Anomaly-aware Features Selection (AFS) module, which selects anomaly-relevant representations from pre-trained features to reduce redundancy and mitigate pre-training bias, and the Reconstruction Residuals Selection (RRS) module, which adaptively selects informative reconstruction residuals for anomaly localization. Experimental results reported on the MVTec AD dataset demonstrate image-level AUROC values above 99%, making RealNet a suitable candidate model for the problem addressed in this thesis.

3.3 Unsupervised Anomaly Detection

Unsupervised anomaly detection assumes that only normal samples are available during training, which closely reflects industrial scenarios where anomalous data are scarce, heterogeneous, and costly to annotate. Rather than relying on labeled anomalies, these methods learn a representation of normal operating conditions and identify deviations from this learned distribution at inference time as potential anomalies.

Unsupervised approaches are commonly divided into two main families. Feature embedding methods model the distribution of normal samples in a latent feature space and detect anomalies as deviations from this distribution. Reconstruction-based methods, in contrast, rely on generative models such as autoencoders or variational autoencoders to reconstruct normal inputs, with anomalies revealed through elevated reconstruction errors [1], [5].

The ability to learn exclusively from normal samples and identify anomalies without requiring prior knowledge of their characteristics has made unsupervised anomaly detection the dominant paradigm in industrial inspection. This capability is particularly relevant to the problem addressed in this thesis, where acquisition-related degradations can exhibit substantial variability and where anomaly detection systems must remain capable of identifying previously unseen anomalous conditions.

3.3.1 Feature embedding methods

Feature embedding methods aim to learn compact and discriminative representations of normal data within a latent feature space. By modeling the distribution of normal samples in this space, anomalies can be identified as deviations from the learned feature distribution. These approaches typically rely on pretrained backbone networks to extract deep visual features and employ similarity or distance-based measures to quantify deviations. Representative strategies include teacher-student architectures, one-class classification methods, distribution-map-based approaches, memory bank methods, and contrastive language image pre-training models. [1], [5]

3.3.1.1 Teacher–Student

Teacher–student methods combine transfer learning and knowledge distillation within an unsupervised anomaly detection framework. As illustrated in Figure 3.3, the teacher network is typically a large pretrained backbone that extracts rich feature representations from normal samples. The student network, in contrast, is trained exclusively on normal data to reproduce the teacher's responses, thereby learning the feature representations associated with normal operating conditions [5].

During inference, the same input image is processed by both networks, and discrepancies between the teacher and student outputs are used as indicators of abnormality. Since the student network is optimized to reproduce the teacher's responses on normal samples, normal inputs tend to produce low discrepancy values, whereas anomalous inputs typically result in larger deviations. By computing these discrepancies across the spatial feature maps, teacher–student methods naturally generate anomaly maps that support both image-level anomaly detection and anomaly localization.

Uninformed Students (US) [52] is a teacher–student framework for pixel-level anomaly detection and segmentation in high-resolution images. The method employs an ensemble of student networks, enabling anomaly detection to incorporate both teacher–student discrepancies and predictive uncertainty. In addition, US supports multi-scale anomaly detection through the use of different receptive field sizes. By varying the amount of spatial context considered during feature prediction, the model can effectively capture both local and large-scale anomalous patterns.

While conventional teacher–student methods often employ architecturally similar teacher and student networks, Asymmetric Student–Teacher (AST) [53] introduces architectural asymmetry to mitigate the undesired generalization that can arise in such settings. The method employs

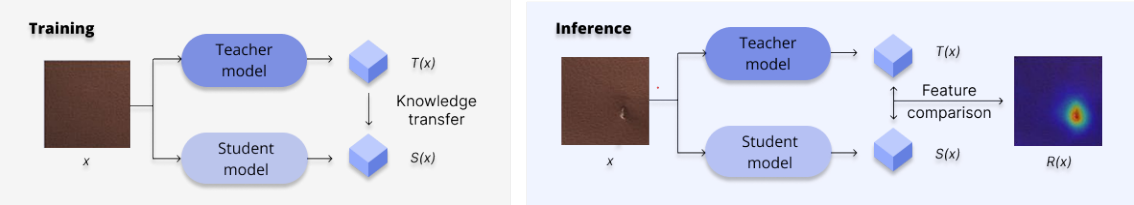


Figure 3.3: Schematic overview of the teacher–student framework. **Training phase:** The student model is iteratively trained to approximate the teacher model's feature representations, enabling knowledge transfer. **Inference phase:** During inference, the teacher model acts as a reference, and discrepancies between teacher and student features increase in the presence of anomalous inputs, enabling anomaly detection and localization. Adapted from [5].

a bijective normalizing flow as the teacher and a non-invertible CNN as the student. Due to its invertible nature, the teacher is highly sensitive to distributional shifts, making anomalous inputs more likely to produce responses that cannot be accurately reproduced by the student. This asymmetric design increases the separation between normal and anomalous samples, facilitating more reliable anomaly detection.

EfficientAD [54] is a teacher–student anomaly detection method that combines strong detection performance with low computational cost. It employs a lightweight Patch Description Network (PDN) as both teacher and student, enabling fast inference and reduced memory requirements. To mitigate undesired student generalization, a penalty term is applied using auxiliary pretraining images from an external dataset. This discourages the student from reproducing the teacher's responses on out-of-distribution samples, thereby increasing the separation between normal and anomalous inputs.

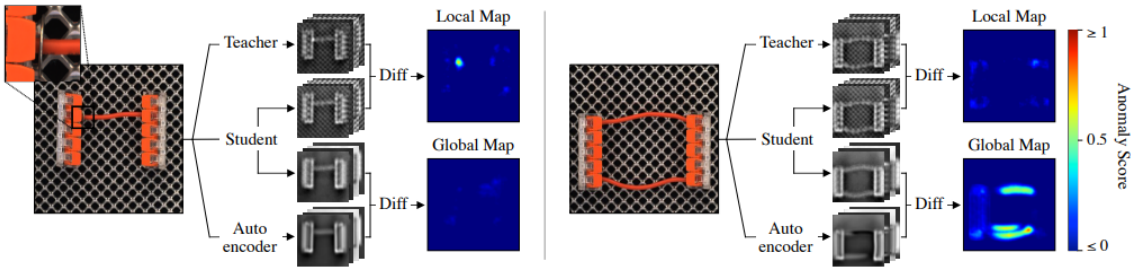


Figure 3.4: EfficientAD applied to two test images from MVTec LOCO [32]. Normal input images contain a horizontal cable connecting the two splicing connectors at an arbitrary height. The anomaly on the left is a foreign object in the form of a small metal washer at the end of the cable. It is visible in the local anomaly map because the outputs of the student and the teacher differ. The logical anomaly on the right is the presence of a second cable. The autoencoder fails to reconstruct the two cables on the right in the feature space of the teacher. The student also predicts the output of the autoencoder in addition to that of the teacher. Because its receptive field is restricted to small patches of the image, it is not influenced by the presence of the additional red cable. This causes the outputs of the autoencoder and the student to differ. "Diff" refers to computing the element-wise squared difference between two collections of output feature maps and computing its average across feature maps. To obtain pixel anomaly scores, the anomaly maps are resized to match the input image using bilinear interpolation (from [54]).

The method additionally incorporates a lightweight autoencoder. During inference, the teacher–student branch detects structural anomalies, while the autoencoder provides complementary information for identifying logical anomalies. The resulting local and global anomaly maps are normalized and fused to produce the final anomaly score. Figure 3.4 illustrates the EfficientAD architecture and the detection of structural and logical anomalies.

Table 3.4 summarizes the main strengths and limitations of representative teacher–student anomaly detection methods. Alongside the previously discussed US, AST, and EfficientAD, the table also presents Student–Teacher Feature Pyramid Matching (STFPM), Informative Knowledge Distillation (IKD), and Reverse Distillation for Anomaly Detection (RD4AD), which introduce alternative strategies for improving anomaly discrimination and localization. Among the considered approaches, AST and EfficientAD were selected as candidate models for further evaluation due to their strong reported performance on the MVTec AD benchmark, with image-level AUROC values exceeding 99%.

Table 3.4: Summary of the considered teacher-student anomaly detection methods.

Method	**Strengths**	**Limitations**
US [52]	Uncertainty-aware anomaly scoring via student ensembles; multi-scale anomaly detection through receptive field control	High training and memory overhead due to multiple ensembles; performance sensitive to receptive field selection
STFPM [55]	Feature pyramid matching across network layers for multi-scale anomaly detection; robust performance even with limited training data	Computational and memory overhead associated with dual backbone architecture
IKD [56]	Overfitting mitigation via informative knowledge distillation; context-aware learning with Contextual Similarity Learning (CSL) and hard sample mining	No adaptive selection of feature hierarchies for different defect scales; weaker performance on texture categories due to overfitting
RD4AD [57]	Reverse distillation enhances anomaly discrimination; One-Class Bottleneck Embedding (OCBE) amplifies teacher–student discrepancies; multi-scale feature distillation for local and global anomaly detection	Performance sensitive to feature layer selection and input image resolution
AST [53]	Asymmetric architecture enhances anomaly separability; Normalizing Flow teacher increases sensitivity to distribution shifts; 99.2% image-level AUROC on MVTec AD	Primarily optimized for image-level detection rather than precise pixel-level localization
EfficientAD [54]	Low-latency inference; lightweight PDN-based teacher–student architecture; dual structural and logical anomaly detection; 98.8% (EfficientAD-S) and 99.1% (EfficientAD-M) image-level AUROC on MVTec AD	Limited sensitivity to fine-grained logical anomalies

3.3.1.2 One-Class Classification

As the name suggests, One-Class Classification (OCC) focuses on modeling a single class, typically corresponding to normal operating conditions. The model learns a representation of normality, typically in the form of a decision boundary, hypersphere, or probability distribution. Samples that deviate from this representation are subsequently identified as anomalies. Figure 3.5 illustrates the one-class classification framework, showing the learning of a normality region during training and the detection of anomalous samples at inference time [1], [5].

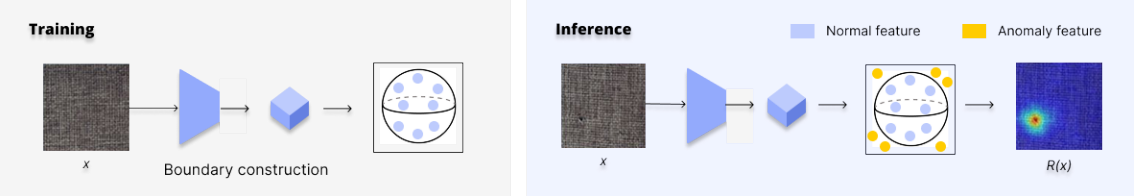


Figure 3.5: Schematic overview of the one-class classification framework. **Training phase:** The model is trained exclusively on normal samples, constructing a boundary within the feature space that encloses these normal data points. **Inference phase:** The validation data undergo the same feature extraction process to produce feature vectors, which are then compared against the previously learned boundary. Any point falling within the boundary is deemed normal; otherwise, it is classified as anomalous. Adapted from [5].

A representative OCC method is PatchSVDD [58], which extends the Deep SVDD framework to the patch level. In this approach, images are decomposed into local patches that are mapped into a latent feature space, where normal patch representations are learned. During inference, anomaly scores are computed based on the distance between a test patch and its nearest normal patch representation in the learned feature space. This patch-based formulation enables both image-level anomaly detection and fine-grained localization of anomalous regions. While the use of local patch representations improves sensitivity to subtle anomalies, the nearest-neighbor search introduces a significant computational overhead, reducing inference efficiency and limiting applicability in real-time inspection scenarios.

In contrast, SimpleNet [59] is a lightweight and efficient approach for anomaly detection and localization based on anomaly synthesis in the feature space. Its architecture consists of a pre-trained feature extractor that produces hierarchical image representations, followed by a shallow feature adaptor that projects these features into a more compact, domain-aligned space to reduce domain bias. During training, an anomalous feature generator synthesizes pseudo-anomalous samples by injecting Gaussian noise into the adapted normal features. The separation between normal and anomalous features is performed by a lightweight anomaly discriminator that acts as a normality scorer. Despite its simplicity and computational efficiency, SimpleNet exhibits limitations in terms of training stability and anomaly localization performance, particularly for small defects, which motivated the development of SuperSimpleNet [60]. SuperSimpleNet extends the original framework by leveraging a small number of labeled anomalous samples while retaining the advantages of unsupervised anomaly detection. To achieve this, the method unifies unsupervised and

supervised learning paradigms while preserving the ability to operate in scenarios where only normal samples are available. The architecture further introduces an upscaling strategy to improve the detection of small anomalies, an additional classification head to capture global contextual information and reduce false positives, and a refined anomaly synthesis process that applies Gaussian noise through binarized Perlin noise masks. These modifications produce more realistic and spatially coherent anomalous regions and result in more stable training dynamics across different runs.

Global and Local Anomaly co-Synthesis Strategy (GLASS) [61] is designed to address limitations of existing anomaly synthesis approaches, particularly their difficulty in generating and controlling weak anomalies that closely resemble normal patterns. To this end, GLASS performs anomaly co-synthesis at both global and local levels, enabling the learning of tighter and more sensitive decision boundaries. During training, normal images are processed by a frozen feature extractor and a trainable feature adaptor to mitigate domain bias, while a lightweight segmentation discriminator distinguishes normal from anomalous feature representations. Global anomaly synthesis operates in the feature space by combining Gaussian noise with gradient-guided perturbations constrained by truncated projection, producing near-distribution anomalies that refine the classification boundary. In parallel, local anomaly synthesis generates diverse and challenging anomalies at the image level through texture-based overlays guided by Perlin noise masks.

As a result, GLASS achieves state-of-the-art performance on the MVTec AD benchmark, reaching 99.9% image-level AUROC. Figure 3.6 qualitatively illustrates the anomaly localization behavior of GLASS and SimpleNet, highlighting the more precise and spatially coherent anomaly maps produced by GLASS. Its state-of-the-art image-level performance, which exceeds the model selection threshold defined in this work, together with its accurate anomaly localization capability, motivates the inclusion of GLASS as a candidate model for addressing the problem considered in this thesis. Although the method exhibits limited capability in handling logical anomalies, this limitation is not critical in the present setting, as the anomalies of interest are predominantly structural.

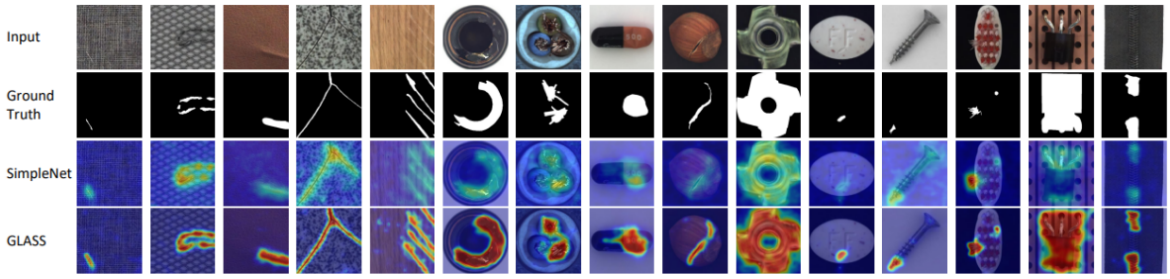


Figure 3.6: Qualitative anomaly localization results on the MVTec AD dataset for GLASS, compared with ground truth and SimpleNet. Adapted from [61].

Table 3.5 summarizes the main strengths and limitations of the OCC models presented in this section.

Table 3.5: Summary of the considered one-class classification methods.

Method	Strengths	Limitations
PatchSVDD [58]	Patch-level latent representation and nearest-neighbor distance modeling improve localization of local structural defects	Low inference efficiency due to nearest-neighbor search; performance sensitive to class-specific hyperparameter tuning
SimpleNet [59]	High inference speed enabled by a lightweight architecture; effective anomaly detection and localization through feature-space anomaly synthesis	Limited training consistency and robustness; oversimplified Gaussian-noise-based anomaly synthesis
SuperSimpleNet [60]	Unified supervised and unsupervised learning; improved training stability and global context modeling; feature upscaling enhances localization of small anomalies	Struggles with categories containing multiple objects in the unsupervised setting; performance sensitive to the choice of backbone architecture and noise-scale tuning
GLASS [61]	Effective detection of weak and near-distribution defects through global and local anomaly co-synthesis; high-throughput inference; 99.9% I-AUROC and 99.3% P-AUROC on MVTec AD	Limited capability in handling logical anomalies

3.3.1.3 Distribution map

Distribution-map-based methods model the statistical distribution of feature representations extracted from normal samples and identify anomalies as deviations from this learned distribution. As illustrated in Figure 3.7, a pre-trained backbone network maps industrial images into a latent feature space, where local patch embeddings are typically modeled using multivariate Gaussian distributions. During training, the parameters of these distributions are estimated using only normal data, resulting in an explicit probabilistic model of normality.

At inference time, feature embeddings extracted from test images are evaluated with respect to the learned normal distributions. Deviations are quantified using distribution-based discrepancy measures that capture differences between the test features and the modeled normality, commonly through the Mahalanobis distance. These discrepancy scores can be aggregated spatially to produce dense anomaly maps, enabling the localization of anomalous regions at the pixel or patch level [5].

Table 3.6 summarizes the main strengths and limitations of representative distribution-map-based methods. CFLOW-AD [62], CS-Flow [63], and FastFlow [64] leverage normalizing flows to explicitly estimate feature likelihoods in a more expressive and flexible manner. CFLOW-AD [62] employs conditional normalizing flows to model the distribution of normal features while incorporating spatial priors, allowing likelihood estimation to be conditioned on positional information and enabling accurate and memory-efficient pixel-level anomaly detection. CS-Flow [63] extends flow-based density modeling by jointly processing multi-scale feature maps through cross-scale

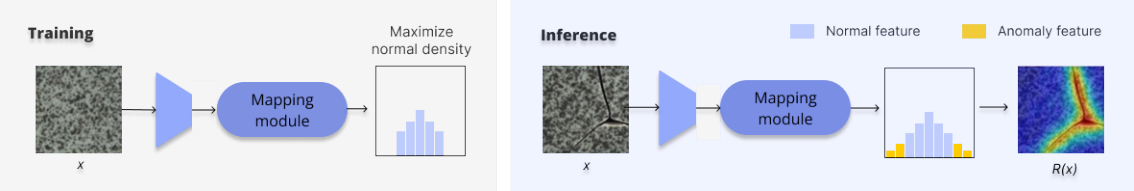


Figure 3.7: Schematic overview of the distribution map framework. **Training phase:** The model extracts features from normal samples and, via a mapping module, projects the input data into a feature distribution space with the aim of maximizing the density of normal data. **Inference phase:** The same mapping module is used to map incoming samples into the same feature space. The sample is classified as anomalous when its feature distribution differs substantially from the previously learned normal distribution. Adapted from [5].

coupling blocks, which facilitate information exchange across resolutions and improve the modeling of contextual dependencies, resulting in more consistent localization of anomalies. FastFlow [64] prioritizes inference efficiency by adopting fully convolutional two-dimensional normalizing flows that operate directly on spatial feature maps, preserving spatial relationships and enabling end-to-end inference on full-resolution images without resorting to computationally expensive sliding-window processing. Together, these methods represent complementary advances in flow-based distribution modeling, addressing spatial conditioning, multi-scale interaction, and real-time inference requirements in industrial anomaly detection.

Among the reviewed distribution-map-based methods, FastFlow is selected for evaluation in this thesis, as it satisfies the model selection criterion of achieving an image-level AUROC above 99% on the MVTec AD benchmark.

Table 3.6: Summary of the considered distribution-map-based methods.

Method	**Strengths**	**Limitations**
CFLOW-AD [62]	Conditional normalizing flows with spatial conditioning through positional encoding; high computational and memory efficiency; accurate anomaly localization	Performance sensitive to input resolution and class-specific hyperparameter tuning
CS-Flow [63]	Joint multi-scale modeling with cross-scale feature interaction; captures both local and global contextual information	Produces spatially dilated anomaly maps due to convolutional receptive fields; not optimized for pixel-precise segmentation
FastFlow [64]	Fully convolutional 2D flow modeling preserves spatial topology; high-speed end-to-end inference; compatible with both CNN and Vision Transformer backbones; 99.4% I-AUROC and 98.5% P-AUROC on MVTec AD	Reduced sensitivity to small and subtle anomalies; prone to false detections in complex background regions

3.3.1.4 Memory bank

Memory bank methods store feature embeddings extracted from normal samples and use them as references during inference. As illustrated in Figure 3.8, a feature extraction network maps input images into a latent feature space, and the resulting normal representations are stored in a memory bank that serves as an explicit memory of normality.

During inference, feature embeddings extracted from test images are compared against the stored memory entries using similarity- or distance-based measures. Feature embeddings that exhibit substantial deviations from their nearest memory representations are identified as anomalous, reflecting discrepancies with respect to previously observed normal patterns. This formulation enables both image-level anomaly detection and spatial localization, depending on the granularity at which features are stored and compared [5].

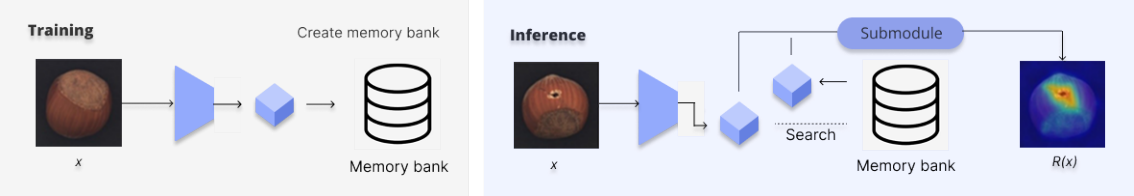


Figure 3.8: Schematic overview of the memory bank framework. **Training phase:** The model extracts features from normal samples and stores these representations in a memory bank, thereby constructing a memory of the normal data distribution. **Inference phase:** Validation data undergo the same feature extraction process, and the resulting feature vectors are queried against the memory bank for the most similar entries. Anomaly score is then computed; if the discrepancy from the stored normal features is substantial, the input is deemed anomalous. Adapted from [5].

Table 3.7 summarizes the main strengths and limitations of representative memory bank anomaly detection methods, including Patch Distribution Modeling (PaDiM), PatchCore, Memory-based Segmentation Network (MemSeg), Coupled-hypersphere-based Feature Adaptation (CFA), and Position and Neighborhood Information (PNI).

PaDiM [65] models normality by fitting a multivariate Gaussian distribution to patch-level feature embeddings at each spatial location. These embeddings are constructed by concatenating features from multiple layers of a pre-trained CNN, allowing the model to capture correlations across different feature scales. Anomaly scores are computed using the Mahalanobis distance between test features and the learned Gaussian distributions. By representing normality through learned statistical parameters rather than storing individual training samples, PaDiM achieves efficient inference and memory usage, with computational requirements that remain independent of the size of the training dataset.

PatchCore [66] builds upon the memory bank approach by maximizing the use of nominal information while remaining computationally efficient. It stores mid-level patch embeddings extracted from a pre-trained backbone and evaluates test samples by measuring their similarity to representative normal features stored in a reduced memory bank. To overcome the scalability limitations of dense memory banks, PatchCore employs greedy coreset subsampling, retaining

Table 3.7: Summary of the considered memory bank methods.

Method	**Strengths**	**Limitations**
PaDiM [65]	Captures correlations across multiple feature scales through patch-wise probabilistic modeling; inference efficiency is independent of training dataset size	Lower performance on object categories compared to texture categories
PatchCore [66]	Representative memory bank with greedy coreset subsampling; efficient inference with competitive speed; locally aware mid-level features mitigate ImageNet bias; 99.1% I-AUROC and 98.1% P-AUROC on MVTec AD	Limited global context modeling, as mid-level features capture less semantic information
MemSeg [67]	End-to-end anomaly segmentation framework; anomaly simulation improves generalization; multi-scale feature fusion and spatial attention enhance localization; 99.56% I-AUROC and 98.84% P-AUROC on MVTec AD	Limited sensitivity to subtle anomalies
CFA [68]	Mitigates domain bias through coupled-hypersphere feature adaptation; scalable memory bank independent of training dataset size; 99.3% I-AUROC and 98.2% P-AUROC on MVTec AD	Peak performance requires ensembling (CFA++), increasing computational complexity
PNI [69]	Incorporates neighborhood and positional information through conditional probability modeling; dedicated refinement network improves anomaly map quality and localization accuracy; 99.56% I-AUROC and 98.98% P-AUROC on MVTec AD	Position-aware modeling provides limited benefits for texture categories

only the most representative normal patches while discarding redundant ones. Furthermore, local neighborhood aggregation improves robustness to small spatial variations and mitigates bias introduced by high-level ImageNet pre-trained features, which tend to emphasize object-level semantics rather than fine-grained structural or texture irregularities.

Despite these strengths, PatchCore exhibits limited global context modeling capability, as mid-level features capture less semantic information and may fail to represent long-range dependencies. Nevertheless, PatchCore achieves image-level performance above 99% AUROC on the MVTec AD benchmark and is widely adopted as a reference method in the anomaly detection literature. Therefore, it is selected for inclusion in the experimental evaluation.

MemSeg [67] adopts a memory-based formulation for anomaly detection and localization that integrates anomaly simulation during training to enhance discrimination between normal and anomalous patterns without requiring pixel-level annotations. The anomaly simulation strategy combines textural perturbations drawn from external texture datasets with structural distortions generated through image rearrangement and guided by Perlin noise masks, enabling the synthesis of realistic industrial irregularities and promoting robust generalization. Multi-scale feature fusion combined with spatial attention further enhances the localization of anomalous regions. Owing to its end-to-end design, MemSeg supports real-time inference while maintaining strong anomaly detection performance on the MVTec AD dataset.

CFA [68] is designed to mitigate the domain bias of feature representations extracted from CNNs pre-trained on large-scale natural image datasets. It adapts patch-level features to the target industrial domain by enforcing compact and discriminative normal feature distributions through a coupled hypersphere formulation.

To ensure scalability, CFA employs a compressed memory bank that stores only core normal feature representations, independent of the size of the training dataset. Anomaly scores are computed based on both feature similarity and matching certainty, reducing the risk of misclassifying sparse yet normal features as anomalous. Both the standard CFA and the ensemble-based CFA++ variants achieve strong anomaly detection performance on the MVTec AD benchmark. By combining the predictions of multiple models, CFA++ further improves detection performance at the cost of increased computational complexity due to ensembling.

PNI [69] is a memory-based anomaly detection method that explicitly incorporates spatial position and neighborhood context into the estimation of normality. It models local neighborhood relationships using a Multi-layer Perceptron (MLP) trained to capture dependencies among neighboring feature embeddings. In parallel, PNI learns which feature patterns are expected at different image locations through the use of spatial histograms. This combination enables the detection of structural anomalies in which local appearance remains normal while spatial arrangement or ordering is altered.

In addition, PNI employs a dedicated refinement network trained on synthetically generated anomalies to enhance pixel-level anomaly maps and improve localization accuracy. The explicit modeling of spatial relationships and position-dependent feature distributions, together with the

refinement stage, enables highly detailed anomaly localization and accurate identification of subtle structural irregularities. As a result, PNI achieves strong image-level performance on the MVTec AD benchmark.

Among the reviewed memory bank approaches, PatchCore, MemSeg, CFA, and PNI achieve image-level AUROC values exceeding the 99% model selection threshold adopted in this work. Therefore, these methods are selected as candidate models for addressing the problem considered in this thesis.

3.3.1.5 Contrastive language–image pre-training

Contrastive Language–Image Pre-training (CLIP) learns joint image–text representations by aligning visual and textual modalities in a shared embedding space through a contrastive learning objective [70]. As illustrated in Figure 3.9, the framework consists of an image encoder and a text encoder that map visual inputs and textual descriptions into a common representation space, where cross-modal similarity scores are computed. During training, a contrastive loss promotes semantic alignment between corresponding image–text pairs while enforcing separation between non-matching pairs.

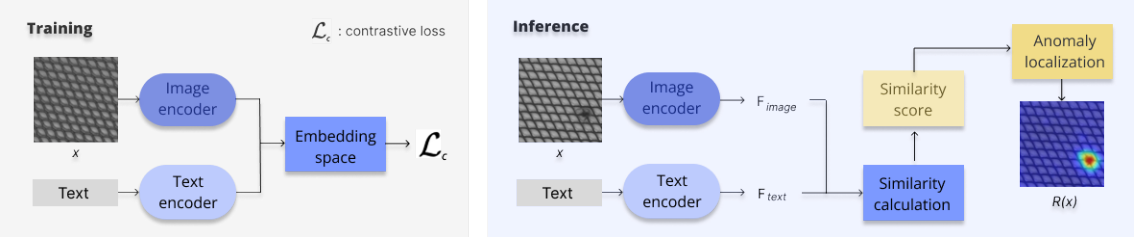


Figure 3.9: Schematic overview of the CLIP framework. **Training phase:** The CLIP model simultaneously takes an image x and its corresponding textual description as input. Through its image and text encoders, the image and text are mapped into a shared embedding space. A contrastive loss L_c is then applied to pull matched image–text pairs closer in this feature space and push mismatched pairs farther apart, thus achieving cross-modal alignment. **Inference phase:** Given a new image and text, the model's image and text encoders each generate feature representations (F_{image} and F_{text}), and a similarity score is computed. If the similarity score falls below a certain threshold, it indicates a significant discrepancy between the image and its textual description, and the input is flagged as anomalous. Adapted from [5].

Building on this multimodal representation, CLIP enables zero-shot and few-shot learning by reformulating anomaly detection as an image–text similarity problem. Rather than relying on task-specific retraining, visual features extracted from an industrial image can be directly compared against textual prompts describing expected or anomalous conditions. This capability arises from CLIP's large-scale multimodal pre-training, which allows the model to capture general visual concepts and their semantic relationships.

In the context of industrial anomaly detection, CLIP-based methods exploit this property to identify anomalies using prompts such as "normal product", "anomalous surface", or other domain-specific descriptions. Although free-form textual annotations are less readily available

in industrial environments, sources such as production standards, defect taxonomies, and quality control guidelines can provide semantically meaningful descriptions of normal and anomalous conditions. By leveraging such information, CLIP-based approaches can capture high-level semantic discrepancies between expected and observed visual patterns, improving both detection robustness and interpretability while facilitating more transparent decision-making in industrial inspection processes [5].

Window-based CLIP (WinCLIP) [70] is a representative CLIP-based approach designed for visual anomaly classification and segmentation under the zero-shot setting. It addresses the limitations of directly applying CLIP to anomaly detection by explicitly reformulating the task as a two-class problem, distinguishing between normal and anomalous states through language-guided image–text alignment. To this end, WinCLIP introduces a Compositional Prompt Ensemble, which constructs textual descriptions by combining object-specific templates with state-related attributes (e.g., normal or anomalous). This strategy enables a clearer semantic separation between acceptable variations and anomalies, improving zero-shot anomaly detection by enhancing the semantic alignment between visual observations and textual descriptions.

For pixel-level anomaly localization, WinCLIP proposes a window-based inference mechanism that extracts dense visual representations using sliding windows at multiple spatial scales. Anomaly scores from overlapping windows are aggregated to produce refined anomaly maps, allowing the detection of anomalies of varying sizes while preserving semantic consistency with the textual prompts.

WinCLIP+ [70] extends the WinCLIP framework to the few-normal-shot setting by incorporating a small number of unlabeled normal images. While the language-driven formulation of WinCLIP is effective for many anomaly types, certain deviations are more naturally characterized through visual reference than textual description. To address this limitation, WinCLIP+ introduces a reference association mechanism that stores feature representations extracted from normal samples and measures their similarity to test images. The final anomaly score is obtained by fusing the language-based prediction of WinCLIP with visual similarity scores computed against the reference memory.

This complementary fusion of language-driven and vision-driven cues enables WinCLIP+ to significantly improve both classification and localization performance in few-normal-shot scenarios. Experimental results on the MVTec AD dataset show that WinCLIP+ consistently outperforms its zero-shot counterpart, highlighting the benefit of incorporating a small amount of normal visual data while preserving the generalization capabilities of CLIP-based anomaly detection.

Table 3.8 summarizes the main strengths and limitations of WinCLIP, WinCLIP+, and other representative CLIP-based anomaly detection methods. While these approaches demonstrate the potential of vision–language models for zero-shot and few-normal-shot anomaly detection, their primary advantage lies in scenarios with limited or no normal training data, which does not reflect the conditions of the target application. Furthermore, no representative CLIP-based method was identified with reported image-level performance above the 99% AUROC threshold adopted for model selection. Therefore, these methods were not considered for the experimental evaluation

presented in this thesis.

Table 3.8: Summary of the considered CLIP-based anomaly detection methods.

Method	Strengths	Limitations
WinCLIP [70]	Zero-shot anomaly detection and segmentation; language-guided semantic anomaly definition; compositional prompting and multi-scale feature aggregation	Fails on anomalies requiring visual reference; limited sensitivity to tiny defects
WinCLIP+ [70]	Incorporates a small number of normal reference images; combines language-guided and vision-based predictions; improved robustness across defect types	Increased memory requirements; performance remains inferior to methods trained on normal data
AnomalyCLIP [71]	Zero-shot anomaly detection and segmentation; object-agnostic prompt learning; joint global and local context optimization; improved alignment between textual and visual representations	Performance remains inferior to methods trained on normal data
PromptAD [72]	Few-shot anomaly detection; automated prompt generation through semantic concatenation; explicit anomaly margin modeling for enhanced normal–anomalous feature separation	Performance remains inferior to methods trained on normal data

3.3.2 Reconstruction-based methods

Reconstruction-based methods formulate anomaly detection as a reconstruction task, in which a model is trained to learn a compact representation of normal data and to accurately reproduce it. The underlying assumption is that normal patterns can be reconstructed with high fidelity, whereas anomalous regions, which are absent during training, cannot. Consequently, anomalies manifest as reconstruction discrepancies that can be used to identify and localize abnormal regions.

As illustrated in Figure 3.10, the model is trained on normal samples by minimizing a reconstruction loss between the input image and its reconstructed output. During inference, a test image is processed by the trained reconstruction network, and the discrepancy between the input and its reconstruction is evaluated. Regions exhibiting large reconstruction errors are considered anomalous, as they deviate from the learned representation of normality [5].

Reconstruction-based methods have been explored using a wide range of architectures, including autoencoders, generative adversarial networks, transformer-based models, diffusion models, and generative pre-trained transformers [1], [5].

3.3.2.1 Autoencoders

Autoencoders (AEs) are a class of neural networks that learn compact latent representations of input data through an encoder–decoder architecture. The encoder maps the input to a lower-dimensional latent space, while the decoder reconstructs the original input from this representa-

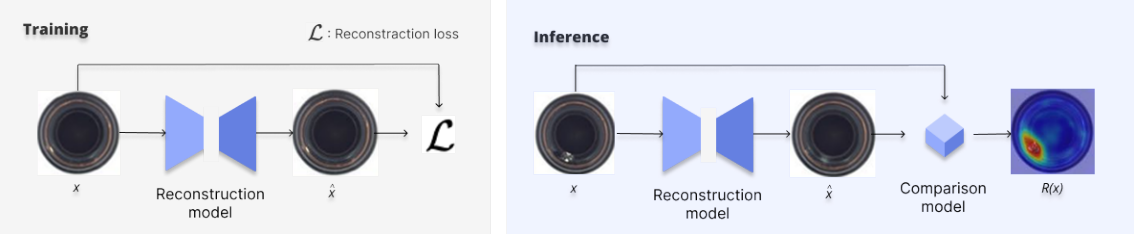


Figure 3.10: Schematic overview of reconstruction-based methods. **Training phase:** The model learns to reconstruct the primary features of normal images x. Specifically, x is passed through the reconstruction network to produce a reconstructed output $\hat{x}$, and a reconstruction loss L_r measures the discrepancy between x and $\hat{x}$. By iteratively minimizing this loss, the model refines its ability to capture the intrinsic structure and distribution of normal data. **Inference phase:** Given a new input image x, the trained reconstruction model generates a reconstruction $\hat{x}$. Both x and $\hat{x}$ are then fed into a comparison module to assess their differences. A substantial mismatch between the input and its reconstruction indicates the model's inability to faithfully reconstruct the input, thereby signaling the presence of anomalous regions within the image. Adapted from [5].

tion. By compressing the input into a compact latent representation, the model is encouraged to capture the most relevant characteristics of the data while discarding redundant information.

The popularity of AEs in industrial anomaly detection stems from their fully unsupervised training procedure, effective dimensionality reduction capabilities, and ability to model complex nonlinear patterns present in visual data. Furthermore, their relatively simple architecture and low computational requirements make them attractive for practical deployment in industrial inspection systems.

Despite these advantages, standard AEs often suffer from over-generalization. When the latent representation is overly expressive or the model capacity is too high, the network may learn to reconstruct not only normal patterns but also anomalous ones. Consequently, reconstruction errors become less discriminative, reducing anomaly detection performance. This limitation has motivated the development of several variants, including sparse AEs, denoising AEs, and variational AEs, which introduce additional constraints to improve the separation between normal and anomalous patterns [5].

Table 3.9 summarizes representative autoencoder-based methods for industrial anomaly detection. DRAEM [73] is a widely adopted approach that combines a reconstruction network with a discriminative segmentation network trained on synthetically generated anomalies. These anomalies are created at the image level by blending Perlin noise masks with textures from external datasets, enabling the model to learn both anomaly restoration and localization.

However, DRAEM relies on the realism of image-level synthetic anomalies and on auxiliary texture datasets, which can limit its sensitivity to subtle or near-distribution anomalies. To address these limitations, Dual Subspace Re-projection (DSR) [74] was proposed. Instead of synthesizing anomalies in the image space, DSR performs controlled perturbations in a quantized latent feature space and employs a dual-decoder architecture to compare general and normality-constrained reconstructions.

Although primarily designed as an unsupervised method, DSR can be extended to incorporate a small number of pixel-level anomaly annotations when available. The authors demonstrate that incorporating only 16 real anomalous samples can improve both anomaly detection and localization performance.

This design improves robustness to subtle structural deviations and enables more accurate anomaly localization. Figure 3.11 provides a qualitative comparison between the reconstruction outputs produced by DRAEM and DSR.

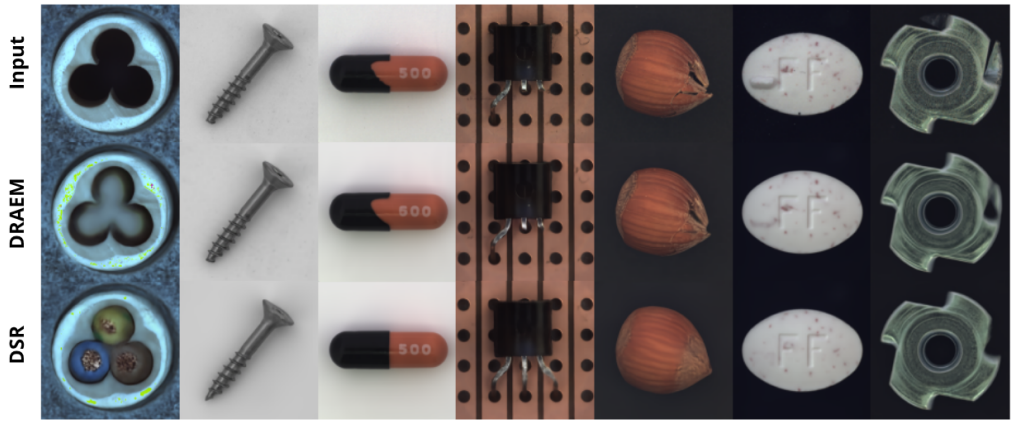


Figure 3.11: Qualitative reconstruction results in DRAEM and DSR. Compared with DRAEM, DSR produces more faithful normal reconstructions by replacing anomalous regions with their corresponding in-distribution appearance. Adapted from [74].

Nevertheless, despite these improvements, no representative Autoencoder-based method was identified with reported image-level performance above the 99% AUROC threshold adopted for model selection. Consequently, Autoencoder-based methods were not considered for the experimental evaluation conducted in this thesis.

Table 3.9: Summary of the considered autoencoder-based anomaly detection methods.

Method	**Strengths**	**Limitations**
RIAD [75]	Region-masking strategy during training and gradient similarity–based loss, enhancing reconstruction quality and anomaly sensitivity	High false-positive rates on random-pattern classes due to elevated reconstruction errors in normal regions
DRAEM [73]	Joint reconstruction and discrimination framework; direct pixel-level anomaly localization	Limited sensitivity to subtle or near-distribution anomalies due to its reliance on image-level synthetic anomaly generation
DSR [74]	Dual-decoder reconstruction separating general and object-specific appearance; feature-level synthetic anomaly generation; extensible to supervised settings benefiting from few pixel-level annotations	Lower performance than more recent state-of-the-art approaches

3.3.2.2 Generative adversarial networks

GANs constitute a class of generative models based on adversarial learning, in which two neural networks are trained simultaneously with competing objectives. As illustrated in Figure 3.12, a GAN consists of a generator, responsible for producing synthetic samples, and a discriminator, whose task is to distinguish between real data samples and those generated artificially.

During training, the generator maps random latent vectors sampled from a predefined noise distribution to synthetic images that aim to resemble samples drawn from the training data distribution. In parallel, the discriminator is presented with both real and generated samples and outputs a probability indicating whether a given input is real or generated. The discriminator is optimized through the discriminator loss L_D, which promotes correct discrimination between real and generated samples, while the generator is optimized through the generator loss L_G, which drives the generation of samples that are indistinguishable from real data according to the discriminator. Through this adversarial process, gradients propagate from the discriminator back to the generator, progressively guiding it toward producing increasingly realistic samples.

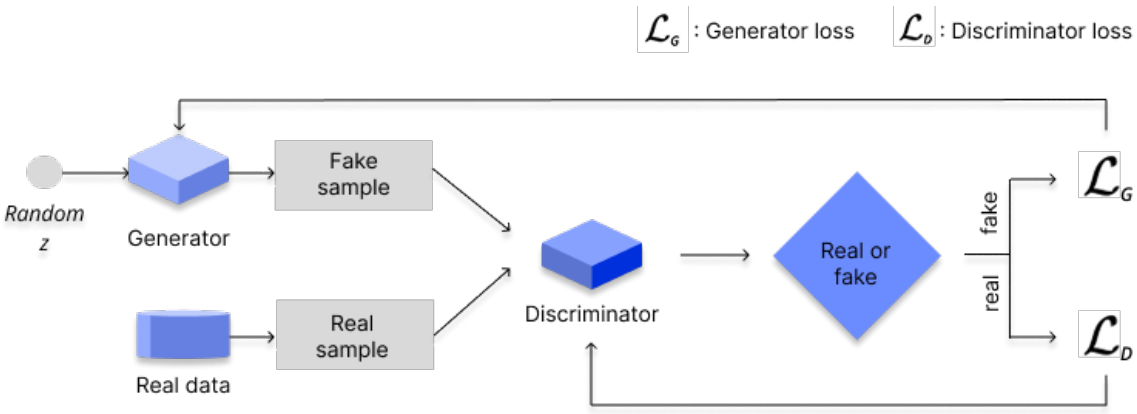


Figure 3.12: Schematic illustration of the GAN architecture. Adapted from [76].

In the context of anomaly detection, intermediate feature representations extracted from the discriminator can be exploited as a complementary signal to identify distributional mismatches between normal and anomalous inputs.

Compared to autoencoders, GAN-based methods tend to generate sharper and more visually realistic reconstructions, reducing the blurring artifacts commonly observed in Autoencoder outputs. This property may improve the detection of subtle anomalies. However, adversarial training is inherently unstable and often prone to issues such as mode collapse, which limits the diversity of generated samples and can hinder reliable modeling of normal data. Industrial inspection data often exhibit considerable variability in acquisition conditions and anomaly characteristics, which places additional demands on model robustness and generalization and can make the deployment of GAN-based methods in industrial inspection scenarios more challenging.

Table 3.10 summarizes representative GAN-based approaches for anomaly detection. Among these, AnoGAN [77] represents a pioneering method that first introduced GANs for unsupervised

anomaly detection. AnoGAN learns the manifold of normal data using a GAN trained exclusively on normal samples and identifies anomalies by searching for a latent representation that best reconstructs a given test image. The anomaly score is computed by combining reconstruction residuals with feature-level discrepancies extracted from the discriminator. While effective, this iterative latent-space optimization process is computationally expensive, resulting in slow inference.

To address this limitation, f-AnoGAN [78] replaces the iterative latent optimization with a learned encoder that directly maps input images to the latent space, enabling significantly faster inference. Additionally, f-AnoGAN adopts a more stable GAN formulation and leverages discriminator-guided feature losses to improve feature representation learning.

Table 3.10: Summary of the considered GAN-based anomaly detection methods.

Method	**Strengths**	**Limitations**
AnoGAN [77]	First GAN-based framework for unsupervised anomaly detection	Slow inference due to iterative latent optimization
f-AnoGAN [78]	Fast inference via learned latent mapping; improved detection and localization	Coarse anomaly localization; limited segmentation reliability
OCR-GAN [79]	Frequency-decoupled reconstruction improves anomaly sensitivity	High model complexity due to a large number of parameters; a lightweight version remains future work
GANomaly [80]	Encoder–decoder–encoder architecture, improving sensitivity to anomalous representations; efficient for both model training and inference	Sensitive to visually similar normal and anomalous classes

No representative GAN-based method was identified with reported image-level performance above the 99% AUROC threshold adopted for model selection. As a result, these methods were excluded from further experimental evaluation.

3.3.2.3 Transformers

Transformers have recently emerged as an alternative to convolutional architectures in reconstruction-based anomaly detection, motivated by their ability to model global context through self-attention mechanisms. Unlike convolutional networks, which rely on local receptive fields, Transformers explicitly capture long-range dependencies across the entire image, enabling a more comprehensive modeling of structural relationships and global consistency.

As illustrated in Figure 3.13, images are typically decomposed into a set of patches that are embedded and processed as a sequence of tokens. Through self-attention, the model learns interactions between distant regions of the image, allowing the reconstruction process to jointly account for both local details and global structure. This property is particularly relevant for detecting anomalies that manifest as global inconsistencies rather than strictly localized defects [5].

Table 3.11 summarizes representative Transformer-based approaches for reconstruction-based anomaly detection. Vision Transformers are often associated with high computational cost, which

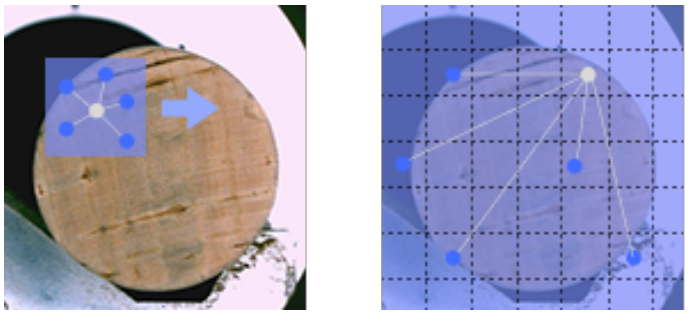


Figure 3.13: Schematic comparison of local receptive field processing in convolutional networks and global context modeling through self-attention in Transformer architectures. Adapted from [81].

can be a limiting factor in industrial scenarios requiring low-latency inference and efficient deployment. However, recent works have nevertheless demonstrated that Transformer-based anomaly detection can achieve both efficient inference and practical deployment. Among these approaches, Dinomaly [82] represents a lightweight reconstruction-based framework that leverages a pure Transformer architecture while avoiding complex designs, auxiliary modules, and specialized heuristics.

Dinomaly relies on a small set of components, including scalable foundation Transformers for feature extraction, implicit noise injection through dropout-based bottlenecks, linear attention mechanisms, and relaxed reconstruction objectives. This design addresses challenges highlighted in recent studies on unified multi-class unsupervised anomaly detection, where existing approaches often exhibit a performance gap compared to class-specific models. By targeting this limitation, Dinomaly aims to bridge the gap between unified and class-separated formulations.

Experimental results reported on the MVTec AD dataset demonstrate image-level performance above 99% AUROC. Consequently, Dinomaly was selected as a candidate model for the problem addressed in this thesis. Its inclusion enables a comparison between Transformer-based reconstruction approaches and alternative anomaly detection paradigms in the context of image acquisition-related anomalies.

Table 3.11: Summary of the considered transformer-based anomaly detection methods.

Method	**Strengths**	**Limitations**
AnoViT [83]	Transformer-based encoder–decoder architecture enabling global context modeling	High computational cost
InTra [84]	Transformer-based inpainting with self-attention; mask-based reconstruction strategy	High computational cost
Dinomaly [82]	Real-time inference; unified multi-class formulation; noise-regularized and relaxed reconstruction improves generalization; 99.56% I-AUROC and 98.4% P-AUROC on MVTec AD	Primarily designed for structural anomalies rather than logical anomalies

3.3.2.4 Diffusion models

Diffusion models have recently emerged as a powerful class of generative approaches for anomaly detection, particularly due to their ability to produce high-quality reconstructions and their improved robustness compared to traditional Autoencoder- and GAN-based methods. These models are built around a two-stage process that explicitly models data generation through gradual noise corruption and subsequent denoising.

As illustrated in Figure 3.14, the forward diffusion process progressively corrupts a normal image by adding Gaussian noise over a sequence of time steps, eventually transforming the input into pure noise. The model is then trained to learn the reverse diffusion process, which iteratively removes noise and reconstructs the original image by reversing the corruption steps. This denoising process is typically implemented using a neural network trained to predict the noise added at each diffusion step, enabling the gradual recovery of the underlying image.

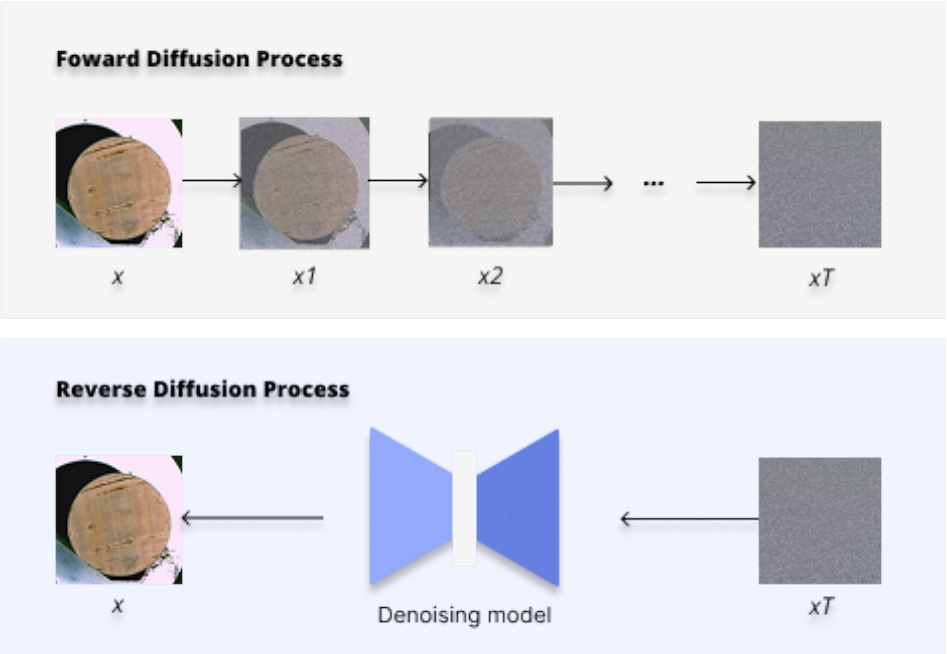


Figure 3.14: Schematic illustration of the forward and reverse diffusion processes in diffusion-based models. Adapted from [85].

However, diffusion-based methods incur high inference latency due to the inherently multi-step denoising process required during reconstruction, which limits their applicability in real-time settings. In addition, their performance can be influenced by the characteristics of the noise process employed during diffusion. Effective anomaly detection relies on sufficiently corrupting anomalous regions during the forward diffusion process so that they can be suppressed during reconstruction. When the injected noise does not adequately interact with the underlying anomaly structure, anomalous regions may not be properly removed, leading to reduced detection performance. This

dependency can introduce inductive biases that affect the ability of diffusion-based methods to generalize across different anomaly types.

Table 3.12 summarizes representative diffusion-based approaches for anomaly detection.

DiffusionAD [86] addresses one of the main limitations of diffusion-based anomaly detection, namely the high inference cost associated with iterative denoising. To improve efficiency, it replaces the conventional iterative denoising process with a one-step denoising strategy, significantly reducing inference time while maintaining reconstruction quality, as illustrated in Figure 3.15.

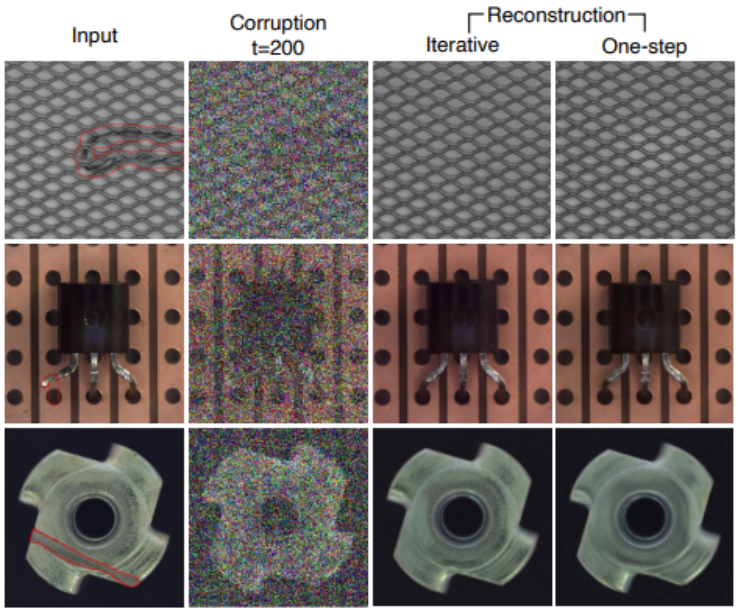


Figure 3.15: Noise-to-normal reconstruction in DiffusionAD using iterative and one-step denoising. Adapted from [86].

In addition, DiffusionAD introduces a norm-guided reconstruction mechanism that integrates reconstructions obtained from multiple noise levels. This strategy enables the model to capture both subtle and prominent anomalies by balancing semantic consistency and fine-grained detail preservation. Anomaly localization is subsequently performed by a dedicated segmentation network trained through an anomaly synthesis strategy and exploiting discrepancies between the input image and its reconstructed normal counterpart.

The reported results demonstrate that reducing the denoising process to a single step does not compromise anomaly detection performance, with DiffusionAD achieving image-level AUROC values above 99% on the MVTec AD dataset. This demonstrates that the high computational cost traditionally associated with diffusion models can be substantially reduced while preserving detection accuracy. Consequently, DiffusionAD was selected as a candidate model for the problem addressed in this thesis.

Denoising Diffusion Anomaly Detection (DDAD) [87] introduces a conditional diffusion framework that reconstructs images while preserving the identity and structural characteristics of the

Table 3.12: Summary of the considered diffusion-based anomaly detection methods.

Method	**Strengths**	**Limitations**
AnoDDPM [88]	Multi-scale simplex noise improves detection of large anomalies	High inference latency
DiffusionAD [86]	Noise-to-norm diffusion framework; one-step denoising significantly improves inference efficiency; 99.7% I-AUROC and 98.7% P-AUROC on MVTec AD	Limited generalization to complex anomaly patterns; requires foreground masks during training
DDAD [87]	Conditional diffusion framework preserving object identity; partial diffusion strategy reduces inference cost; 99.8% I-AUROC and 98.8% P-AUROC on MVTec AD	Requires multiple denoising iterations during inference, limiting suitability for highly time-constrained applications
DiAD [89]	Preserves image categories and pixel-wise structural integrity in multi-class settings; multi-space reconstruction combining pixel, semantic, and feature representations	Sensitive to background impurities, leading to localization and classification errors

inspected object. By conditioning the denoising process on the input image, the model generates defect-free reconstructions that can be directly compared with the original image for anomaly detection.

To reduce the number of denoising iterations required during inference, DDAD adopts a partial diffusion strategy in which the reverse process starts from an intermediate noise level rather than from pure noise. This design reduces inference time while preserving sufficient structural information for accurate reconstruction. Anomaly detection and localization are performed by comparing the input image and its reconstruction at both the pixel and feature levels.

Experimental results reported on the MVTec AD dataset demonstrate image-level performance above 99% AUROC. Based on these results, DDAD was selected for further evaluation in this thesis.

3.3.2.5 Generative Pre-trained Transformers

Generative Pre-trained Transformers (GPTs) are large-scale models originally developed for language understanding and generation, which can be extended with vision–language capabilities to jointly process images and text. In the context of anomaly detection, GPT-based approaches exploit this multimodal reasoning capacity to associate visual content with semantic representations, enabling anomaly detection to be formulated at a semantic level rather than relying solely on low-level visual cues.

As illustrated in Figure 3.16, GPT-based anomaly detection frameworks typically leverage vision–language alignment to associate visual observations with semantic descriptions. During inference, anomalies can be identified when the inferred semantics deviate from those expected for normal samples [5].

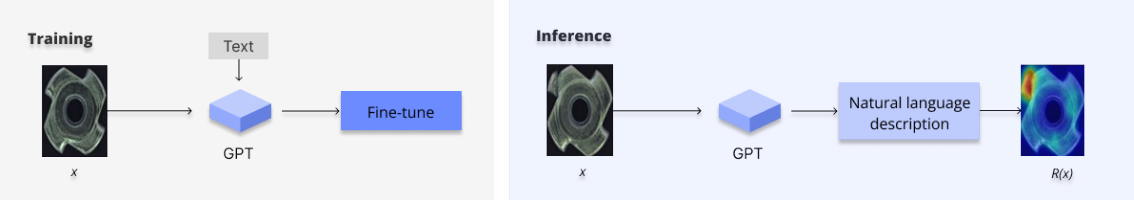


Figure 3.16: Schematic overview of the GPT framework. **Training phase:** The input image x and its corresponding textual description are fed into GPT, which is fine-tuned to generate natural language outputs aligned with the image content. Through iterative training, the model progressively learns and captures the semantic characteristics of normal images, building a representation of expected visual–semantic correspondence. **Inference phase:** During inference, given a new input image x, GPT generates a corresponding natural language description based on its learned knowledge. If the generated description deviates significantly from the previously learned semantics of normal samples, the input is identified as anomalous. Adapted from [5].

A key advantage of GPT-based methods lies in their interpretability, as detected anomalies can be accompanied by natural language descriptions that facilitate human understanding and decision-making. However, this semantic expressiveness comes at the cost of increased computational complexity and a strong dependence on reliable vision–language alignment, which can limit scalability and real-time applicability in industrial inspection scenarios.

AnomalyGPT [90] is a representative example of GPT-based industrial anomaly detection. Unlike conventional anomaly detection approaches that typically output only anomaly scores or heatmaps, AnomalyGPT formulates anomaly detection as a conversational task. This enables the model to determine the presence of anomalies, localize them, and describe their characteristics through natural language interaction, without relying on manually defined decision thresholds.

As illustrated in Figure 3.17, the model can respond to queries such as “Is there any anomaly in the image?” and “Tell me some details about the defect,” providing both anomaly localization and semantic descriptions. This example highlights one of the key advantages of AnomalyGPT: its ability to support multi-turn interactions while offering interpretable explanations of detected defects.

Nevertheless, no representative GPT-based anomaly detection method was identified with reported image-level performance above the 99% AUROC threshold adopted for model selection. Consequently, GPT-based methods were not considered for further evaluation in this thesis.

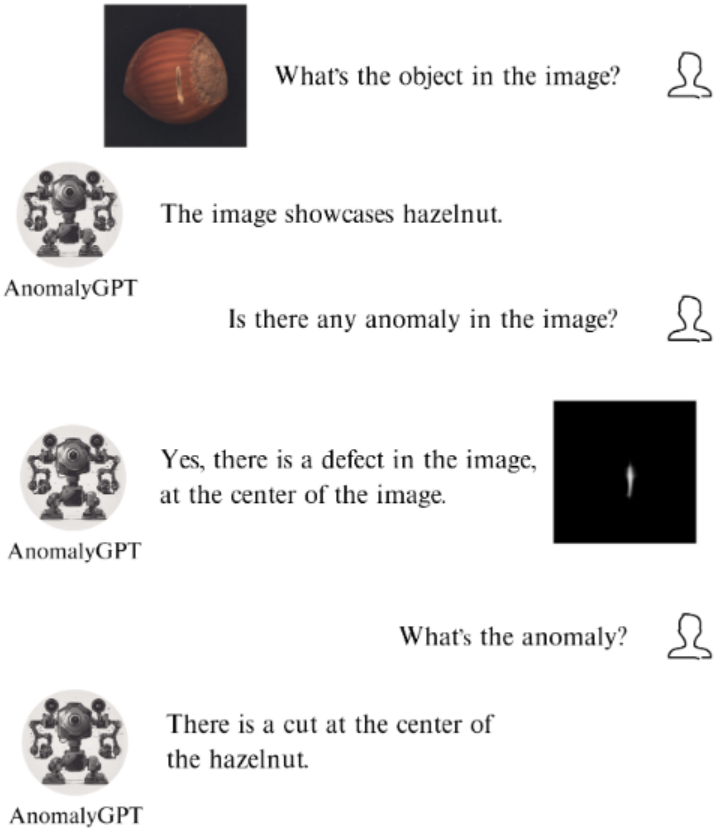


Figure 3.17: Example of a multi-turn interaction with AnomalyGPT (from [91]).

3.4 Summary of Anomaly Detection Method Families

The literature review indicates that unsupervised anomaly detection constitutes the most suitable paradigm for the problem addressed in this work. In industrial image acquisition systems, anomalous conditions are often diverse, difficult to anticipate, and may evolve over time. Consequently, collecting representative examples of all possible anomaly types is generally impractical. By learning exclusively from normal data, unsupervised methods can detect previously unseen anomalies while avoiding the need for costly anomaly annotation procedures.

Tables 3.13 and 3.14 summarize the two main families of unsupervised anomaly detection methods, namely feature embedding-based and reconstruction-based approaches, together with their principal limitations. Although these methods differ substantially in how normality is modeled and how anomalies are detected, they share the common objective of learning representations of normal operating conditions and identifying deviations from them at inference time.

Table 3.13: Summary of feature embedding-based anomaly detection methods and their limitations.

Method type	**Description**	**Limitations**
Feature embedding methods	Learn discriminative feature representations of normal data and detect anomalies as deviations in the learned feature space	Sensitive to feature extractor bias; may exhibit limited generalization across datasets with differing visual statistics
Teacher-Student	Learn normal feature representations by training a student network to reproduce the outputs of a pre-trained teacher network	Student may overgeneralize by reproducing teacher representations for anomalous inputs, reducing anomaly sensitivity
One-Class Classification	Learn a compact decision boundary around normal samples in the feature space	Risk of trivial solutions and feature collapse; often rely on pseudo-anomaly generation or auxiliary objectives
Distribution map	Model the statistical distribution of normal feature representations and detect deviations from the learned distribution	Performance depends on accurate density estimation and may degrade in complex high-dimensional feature spaces
Memory bank	Store representative normal feature embeddings and detect anomalies through feature similarity comparisons	Inference complexity and memory requirements increase with the size of the stored feature bank
CLIP	Leverage vision-language representations for zero- or few-shot anomaly detection through image-text alignment	May focus on object semantics rather than subtle visual irregularities; no representative method with image-level AUROC above 99% on MVTec AD was identified

Many recent unsupervised methods incorporate self-supervised learning strategies, such as

Table 3.14: Summary of reconstruction-based anomaly detection methods and their limitations.

Method type	**Description**	**Limitations**
Reconstruction-based methods	Learn to reconstruct normal samples and identify anomalies through reconstruction discrepancies	May reconstruct anomalous regions due to overgeneralization, reducing anomaly sensitivity
Autoencoders	Compress and reconstruct normal samples using an encoder-decoder architecture	Prone to overgeneralization, resulting in accurate reconstruction of anomalous samples; no representative method with image-level AUROC above 99% on MVTec AD was identified
GANs	Generate normal reconstructions through adversarial training between a generator and a discriminator	Training instability and mode collapse; no representative method with image-level AUROC above 99% on MVTec AD was identified
Transformers	Reconstruct normal samples using self-attention mechanisms to capture global context	High computational cost and increased inference latency
Diffusion models	Learn normal data distributions through progressive noise corruption and denoising processes	Iterative denoising results in high computational cost and slow inference; robustness may be influenced by assumptions regarding the noise distribution and corruption process
GPTs	Perform anomaly detection through semantic reasoning over visual and textual information	No representative method with image-level AUROC above 99% on MVTec AD was identified; high computational complexity and limited suitability for real-time industrial deployment

synthetic anomaly generation mechanisms, to improve representation learning from normal samples. In addition, limited anomaly supervision is increasingly being explored to improve performance in challenging scenarios. This is particularly relevant in industrial environments, where anomalous samples may gradually become available during system operation and can be leveraged to further improve detection performance while preserving the ability to identify previously unseen anomalies.

Based on this analysis, representative methods from the most relevant anomaly detection families identified in the literature were selected for experimental evaluation in the proposed benchmark. The selection focused on approaches that demonstrated strong performance on MVTec AD, with image-level AUROC values above 99%. No representative CLIP-, autoencoder-, GAN-, or GPT-based method meeting this criterion was identified in the literature. Consequently, these families were not considered for experimental evaluation. The methodology adopted for this evaluation is presented in the next chapter.

Chapter 4

Methodology

This chapter presents the methodology adopted for the development and evaluation of the proposed benchmark. The considered datasets, evaluated models, performance metrics, and experimental protocol are described. In addition, the strategies investigated to improve robustness in the most challenging inspection scenarios are presented, with particular emphasis on the proposed lightweight supervision framework and its evaluation methodology.

4.1 Datasets

To evaluate the selected anomaly detection methods under realistic industrial conditions, three datasets from the cork manufacturing domain were considered in this work.

The first dataset was specifically collected for this study. Unlike conventional industrial anomaly detection benchmarks, which primarily focus on product-level defects such as scratches, cracks, dents, or contaminations, this dataset contains acquisition-related anomalies commonly encountered in industrial visual inspection systems, including defocus blur, illumination variations, partial occlusions, intrusions, lens dirt, and object misalignment. The second and third datasets represent more challenging industrial inspection scenarios involving distinct acquisition setups, visual characteristics, image formation processes, and anomaly categories. Their inclusion enables a broader assessment of the evaluated methods under diverse industrial conditions and reduces the influence of dataset-specific characteristics on the overall conclusions.

All datasets were organized according to the directory structure adopted by the MVTec AD benchmark [33], one of the most widely used datasets in industrial anomaly detection. Using this standardized format enables the direct use of existing data-loading pipelines and anomaly detection frameworks with minimal modifications, thereby facilitating reproducible experiments. The adopted directory structure is illustrated in Figure 4.1. Unlike the original MVTec AD benchmark, which contains multiple object and texture categories, the datasets considered in this work comprise a single category, denoted as *cork*, within which different anomaly types are defined.

Each dataset follows the standard MVTec AD organization, consisting of three main directories: *train*, *test*, and *ground_truth*. The *train* directory contains only normal samples stored in a

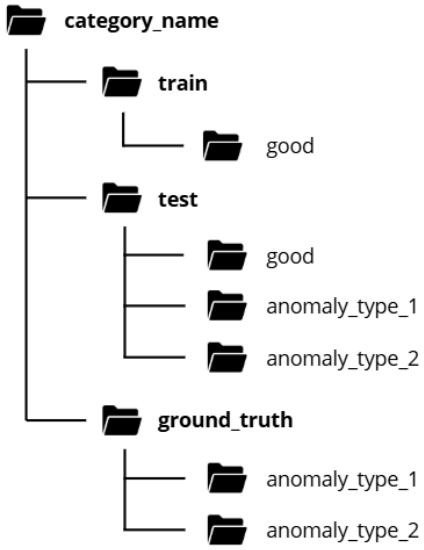


Figure 4.1: Standard directory structure used in the datasets, following the MVTec AD format.

subdirectory named *good*. The *test* directory contains both normal and anomalous samples, with the latter organized into separate subdirectories according to anomaly type. The *ground_truth* directory stores pixel-level segmentation masks corresponding to anomalous samples. Since the primary objective of this work is image-level anomaly detection, pixel-level annotations were not available for the considered datasets. Nevertheless, to maintain compatibility with frameworks that expect the standard MVTec AD structure, dummy masks were generated for all anomalous samples. These masks consist of black images with the same dimensions and filenames as their corresponding images.

The following subsections provide a detailed description of the three datasets considered in this work.

4.1.1 Acquisition-Related Anomalies Dataset

To represent common image acquisition disturbances encountered in industrial visual inspection systems, a dedicated dataset was collected using a prototype cork inspection system developed in partnership with JTA and deployed in an operational industrial environment. The dataset contains acquisition-related anomalies such as defocus blur, illumination variations, and partial occlusions, which may compromise image quality and affect the reliability of subsequent inspection stages.

The acquisition system employs three synchronized cameras. As shown in Figure 4.2, one camera is dedicated to acquiring images of the cork body, while two additional cameras capture opposite top views. During operation, the cork is initially rotated to enable the acquisition of its lateral surface. It is then displaced along the inspection line by a mechanical actuator and passes through the top-view acquisition stage. The acquisition of the top views is triggered by a sensor that detects the presence of the cork at the inspection position. The resulting multi-view image generated by the system is presented in Figure 4.3. By combining information from the body view

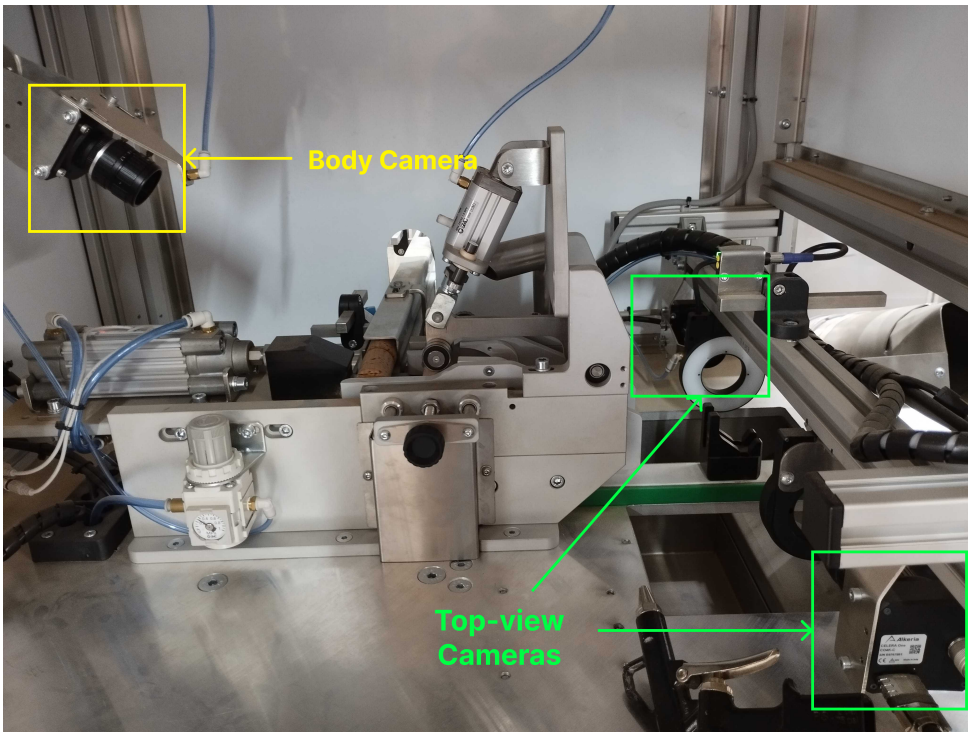


Figure 4.2: Multi-view acquisition system used for cork inspection. The system employs three synchronized cameras: one body camera (highlighted in yellow) and two top-view cameras (highlighted in green) positioned to acquire opposite views of the cork tops.

and the two top views, the system achieves more comprehensive coverage of the cork surface. However, the use of multiple cameras also introduces additional sources of image degradation. Acquisition disturbances may affect one or more views simultaneously, resulting in anomalies that can be present in a single view or across multiple views.

Representative examples of normal samples and the anomaly categories included in the dataset are shown in Figure 4.4. The dataset combines disturbances deliberately introduced during image acquisition with anomalies naturally observed during system operation, providing a representative set of acquisition-related failures encountered in industrial visual inspection systems. The anomaly categories were identified through a detailed analysis of the acquisition process and validated in collaboration with domain experts from JTA to ensure their relevance to industrial inspection scenarios. The annotation process was performed according to these definitions, with ambiguous cases discussed with domain experts from JTA whenever necessary. This process also ensures that each anomalous sample contains only a single anomaly type, allowing the effect of each acquisition-related disturbance to be evaluated independently, although the disturbance may be present in one or more camera views simultaneously.

The dataset includes anomalies related to defocus blur, illumination variations, lens dirt, occlusions and cork misalignment. *Defocus Blur* was generated by intentionally adjusting the focus of the cameras, resulting in degradations affecting individual views or multiple views simultaneously. *Illumination Variations* were introduced by independently adjusting the lens aperture

Figure 4.3: Example of a multi-view image acquired by the inspection system, showing the body view and the two top views of the cork.

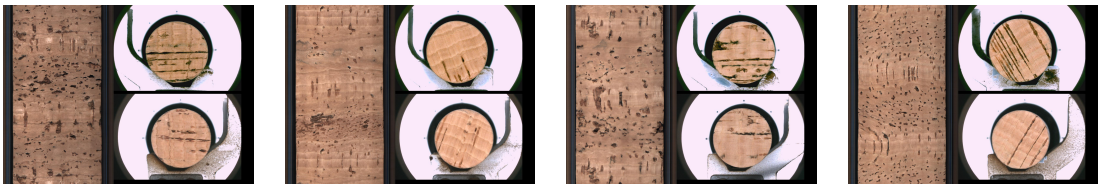
of each camera, resulting in darker or brighter acquisitions that simulate changes in illumination conditions.

Lens Dirt was simulated by adding dust particles to one of the top-view camera lenses, reproducing a common source of image degradation in industrial environments. *Occlusions* were introduced by positioning foreign objects that could realistically be encountered during industrial operation, such as cables, tools, or additional corks, within the camera field of view. In some cases, moving mechanical components of the inspection system temporarily entered the imaging area during acquisition. Both scenarios result in the partial obstruction of the captured scene.

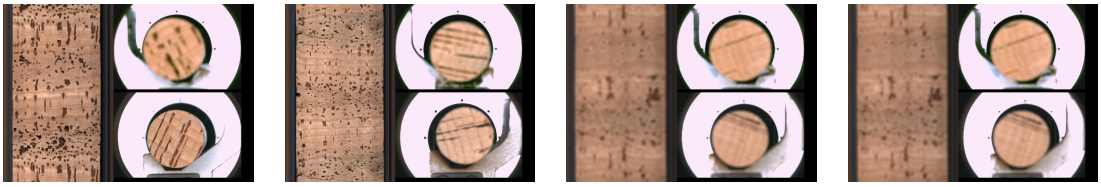
The dataset also contains anomalies naturally observed during system operation. *Misaligned Cork* anomalies occur when a cork is incorrectly positioned on the conveyor belt, leading to abnormal spatial alignment across one or more views. *Missing Cork* anomalies correspond to situations in which no cork is present in the acquisition area at the expected capture instant, resulting in images containing only background elements.

During the analysis of the collected data, it was observed that some samples exhibit only minor deviations from normal operating conditions, such as slight illumination variations or small occlusions that do not affect relevant regions of the cork surface. As these samples lie close to the boundary between normal and anomalous acquisitions, they may be difficult to distinguish from normal samples, even for high-performing anomaly detection methods. To account for such cases, an additional anomaly type, referred to as *Boundary*, was introduced. The definition of this category was established in collaboration with domain experts from JTA, who identified these samples as representing borderline cases between normal and anomalous acquisition conditions. The deviations from normal operation are generally minor and would not typically be considered problematic in a production environment, as their expected impact on the inspection process is limited. Nevertheless, these samples were treated as a separate anomaly category to enable a dedicated evaluation of the performance of the evaluated models under such acquisition conditions.

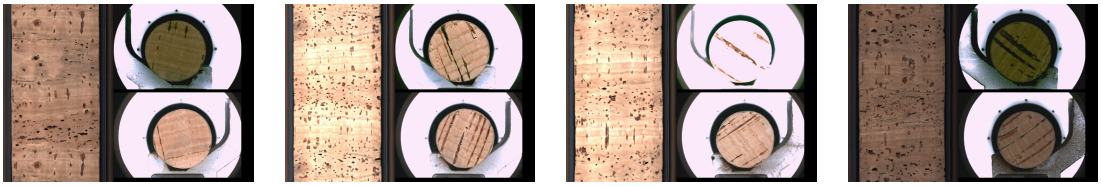
Representative examples of samples belonging to the *Boundary* anomaly type are shown in Figure 4.5. The first example presents a mild illumination variation affecting one of the top views. The second and third examples correspond to borderline cases of occlusions. In the second exam-



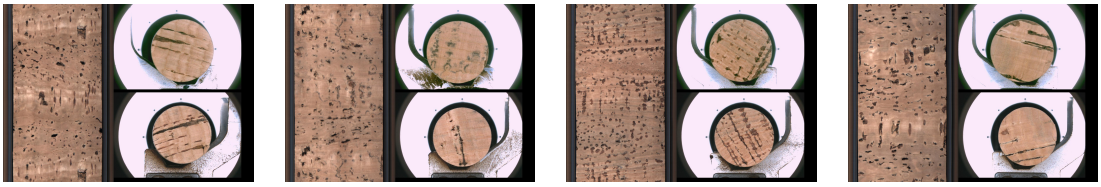
(a) Normal samples



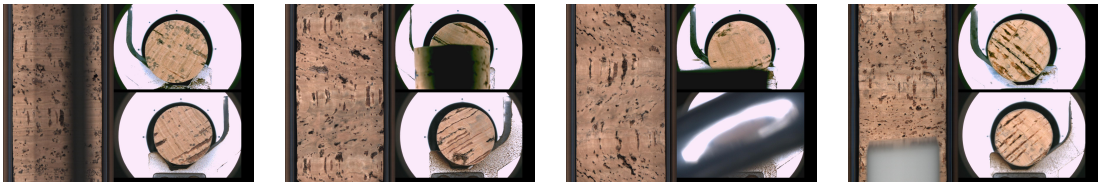
(b) Defocus Blur



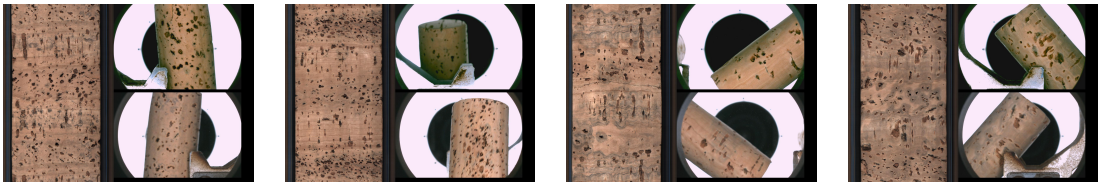
(c) Illumination Variations



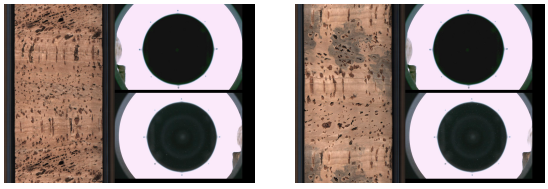
(d) Lens Dirt



(e) Occlusions



(f) Misaligned Cork



(g) Missing Cork

Figure 4.4: Representative normal and anomalous samples from the Acquisition-Related Anomalies dataset.

ple, the shadow cast by an unexpected cork appears in one of the top views without affecting the visible surface of the inspected cork. The third example shows a faint cable shadow projected onto the cork body. In both cases, the occlusions have a limited visual impact and do not significantly obstruct the relevant regions of the cork.

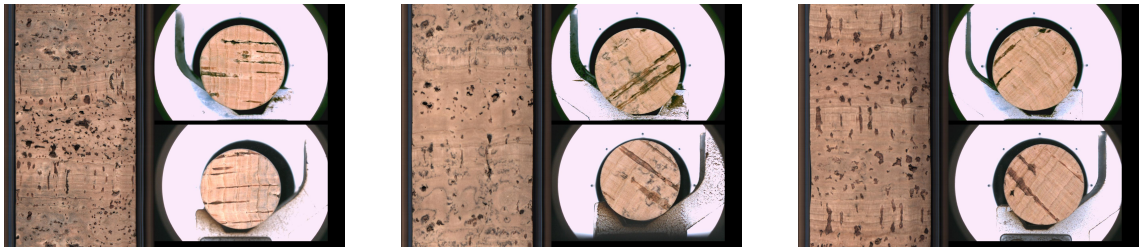


Figure 4.5: Representative examples of samples belonging to the *Boundary* anomaly type.

Table 4.1 summarizes the distribution of samples across the different anomaly types included in this dataset. The lower number of samples in the *Misaligned Cork* and *Missing Cork* categories reflects their natural occurrence during system operation, which limited their availability during data collection.

Table 4.1: Distribution of samples in the Acquisition-Related Anomalies dataset. All anomalous samples belong to the test set.

Type	**Number of Samples**
Normal (train)	515
Normal (test)	178
Defocus Blur	107
Illumination Variations	201
Lens Dirt	32
Occlusions	256
Boundary	16
Misaligned Cork	4
Missing Cork	2
Total	**1311**

4.1.2 Foreign Object Intrusions Dataset

A second application scenario was considered involving industrial cork feeding systems. During transport through the feeding system, foreign objects may occasionally appear among the corks, resulting in intrusion anomalies. Examples include colored wires, small wooden fragments, and other unwanted objects. Representative normal and anomalous samples are shown in Figure 4.6.

The dataset was collected by JTA as part of a proof-of-concept study conducted for an industrial client. Unlike the previous dataset, only a single anomaly category is considered, namely *Intrusions*. However, multiple types of intrusion objects are represented in the dataset, and one or more intrusion objects may be present simultaneously within the same sample. The dataset

was provided with image-level annotations indicating only the presence or absence of an intrusion anomaly in each image.

Figure 4.6: Representative normal and anomalous samples from the Foreign Object Intrusions dataset.

Compared to the Acquisition-Related Anomalies Dataset, this dataset presents substantially greater variability among normal samples. The number of visible corks, their orientation, and their spatial arrangement vary considerably across acquisitions. In addition, corks may contain different printed logos and markings, further increasing the visual diversity of the scene.

As a result, anomaly detection models must account for a broader range of normal appearances, making the distinction between normal and anomalous conditions more challenging. Due to confidentiality constraints, examples containing customer-specific logos are not shown in this thesis.

Table 4.2 summarizes the distribution of normal and anomalous samples in the Foreign Object Intrusions dataset.

Table 4.2: Distribution of samples in the Foreign Object Intrusions dataset. All anomalous samples belong to the test set.

Type	**Number of Samples**
Normal (train)	1477
Normal (test)	237
Intrusions	284
Total	**1998**

4.1.3 Rolling Inspection Artifacts Dataset

A third application scenario was based on a rolling inspection system in which cork images are reconstructed from multiple partial acquisitions captured while the cork rotates along a conveyor. This acquisition strategy enables complete surface inspection from a single reconstructed image but also introduces additional sources of error.

Incorrect cork positioning, dirt accumulation on the conveyor rollers, and inaccuracies in cork detection or image stitching may introduce reconstruction artifacts that result in visual distortions not present on the actual cork surface, potentially leading to incorrect inspection decisions. Representative normal and anomalous samples are shown in Figure 4.7.

The dataset consists of real production data acquired from an operational rolling inspection system. Analysis of the collected data revealed several recurring reconstruction artifacts, which were subsequently organized into anomaly categories. The definition of these categories and the annotation of the samples were carried out in consultation with domain experts from JTA to ensure consistency and relevance to the target industrial application.

Roller Contamination 1 occurs when dirt accumulated on the conveyor rollers interferes with cork detection, causing the cork boundary to be incorrectly estimated. Consequently, part of the conveyor belt becomes visible in the reconstructed image, together with a curved portion of the cork surface that should not appear after the unwrapping process. *Roller Contamination 2* is caused by the same underlying mechanism but affects the lateral boundaries of the reconstruction. In this case, portions of the conveyor belt are incorrectly included along the sides of the reconstructed image, appearing as dark rectangular regions. Some samples in this category also exhibit characteristics of *Roller Contamination 1*, as both effects may occur simultaneously.

Swirl represents a more severe consequence of roller contamination. In these cases, errors introduced during cork detection propagate throughout the reconstruction process, producing heavily distorted reconstructions with repeated and irregular image patterns that result in a characteristic swirl-like appearance.

Duplication artifacts occur when the same region of the cork surface is incorporated multiple times into the reconstruction. As a result, portions of the cork appear duplicated, creating repeated visual patterns that do not correspond to the actual cork surface. *Discontinuity* artifacts occur when adjacent image strips are incorrectly aligned during reconstruction, producing visible discontinuities in the cork surface and causing features such as printed logos to appear broken or incomplete.

Cropped artifacts occur when the cork deviates from its expected position during acquisition. Consequently, only part of the cork surface is captured, resulting in incomplete reconstructions with missing image regions. Finally, the *Other* category includes miscellaneous reconstruction artifacts that cannot be assigned to any of the previously defined anomaly types.

Although the anomaly categories were defined to be mutually exclusive, some overlap exists between *Roller Contamination 1*, *Roller Contamination 2*, *Duplication*, and *Discontinuity*. In particular, samples affected by *Roller Contamination 1* or *Roller Contamination 2* may also exhibit local duplication or discontinuity artifacts as a consequence of the reconstruction errors introduced by the contamination. However, such samples were assigned to the corresponding roller contamination category, as this represents the dominant failure mode and underlying cause of the observed artifacts. Furthermore, as previously discussed, some samples belonging to *Roller Contamination 2* may also exhibit characteristics associated with *Roller Contamination 1*. Apart from these cases, the remaining anomaly categories are mutually exclusive.

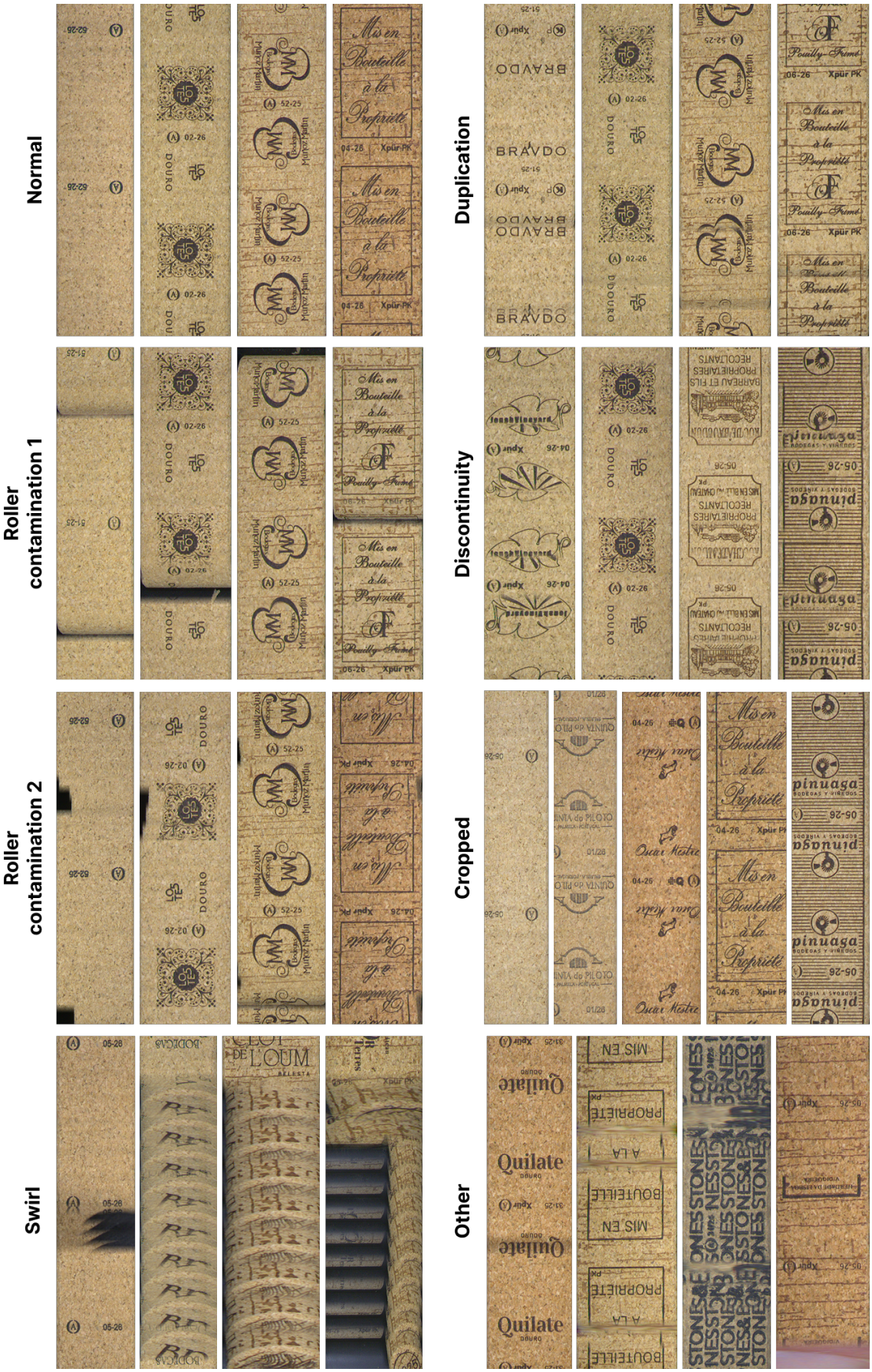


Figure 4.7: Representative normal and anomalous samples from the Rolling Inspection Artifacts dataset.

Table 4.3 presents the distribution of samples in the Rolling Inspection Artifacts dataset.

Table 4.3: Distribution of samples in the Rolling Inspection Artifacts dataset. All anomalous samples belong to the test set.

Type	**Number of Samples**
Normal (train)	1110
Normal (test)	266
Roller contamination 1	113
Roller contamination 2	41
Swirl	42
Duplication	49
Discontinuity	47
Cropped	20
Other	16
Total	**1704**

The detection of *Roller Contamination 1*, *Roller Contamination 2*, and *Swirl* is of particular interest, as these artifacts are directly associated with dirt accumulation on the conveyor rollers. Domain experts from JTA identified these anomaly categories as the most relevant for industrial operation, as their detection enables timely maintenance and cleaning interventions, helping to preserve reconstruction quality and prevent the propagation of reconstruction errors. These anomaly categories are also characterized by stronger visual deviations than *Duplication* and *Discontinuity*, making them easier to detect for most evaluated methods. In contrast, *Cropped* artifacts can be detected through simple image-dimension checks and therefore do not necessarily require dedicated anomaly detection methods.

To evaluate the robustness of the anomaly detection methods to variations in cork appearance, an additional test set containing corks with three logo types not present in the training data was considered. For each logo type, both normal and anomalous samples were collected, enabling the evaluation of model performance on previously unseen visual patterns while preserving the

(a) Logo 1 (b) Logo 2 (c) Logo 3

Figure 4.8: Representative samples from the unseen logo test set.

acquisition-related anomalies of interest. The additional test set comprises 143 samples distributed across the three unseen logo types, including 59 normal samples and 84 anomalous samples. Representative examples of the three logo types are shown in Figure 4.8. For each logo type, one normal sample and two anomalous samples are presented.

4.2 Evaluated Models

The candidate models presented in the previous chapter were selected for evaluation to identify the most suitable approaches for the target anomaly detection problem. Candidate methods were identified through recent anomaly-detection surveys, scientific literature searches, and the Anomalib [92] framework. The final selection aimed to include representative and widely adopted methods from different anomaly detection paradigms reporting image-level AUROC values above 99% on the MVTec AD benchmark.

Table 4.4 summarizes the main characteristics of the evaluated models.

Although DevNet does not achieve the same performance on the MVTec AD benchmark as the other selected methods, it was included as a representative semi-supervised anomaly detection approach.

Three EfficientAD variants were evaluated. EfficientAD-S and EfficientAD-M differ in the architecture of the teacher and student PDNs, with the medium variant incorporating additional convolutional layers and a larger number of feature channels. This increases model capacity and representational power at the cost of higher computational requirements. EfficientAD-M without the autoencoder was additionally evaluated to assess the contribution of the autoencoder component, which provides a complementary anomaly signal based on global contextual information that is fused with the teacher–student anomaly map during inference.

Whenever available, implementations provided by the anomalib framework were adopted, including EfficientAD, Dinomaly, CFA, FastFlow, and PatchCore. For the remaining models, the official implementations released by the original authors were used to preserve the original training procedures and architectural configurations. Models without publicly available implementations or incompatible with the adopted experimental setup were excluded from the evaluation. In particular, MemSeg was excluded due to the absence of an official public implementation, whereas PNI and RealNet were not compatible with the experimental setup adopted in this work.

4.3 Model Evaluation

The proposed benchmark is evaluated across three industrial datasets targeting different image acquisition and inspection scenarios. Performance evaluation is primarily conducted at the image level using the image-level AUROC, which serves as the main metric for comparing anomaly detection methods throughout this work. Since all considered datasets contain a single object category (cork), results are analyzed across anomaly types rather than object categories. This

Table 4.4: Overview of anomaly detection methods evaluated in this work.

Method	**Type**	**Description**
DevNet [44]	Semi-supervised	Learns anomaly detection in an end-to-end manner by directly optimizing anomaly scores with a deviation-based loss, providing high data efficiency when labeled anomalies are representative.
PatchCore [66]	Memory bank	Uses coreset sampling to retain a representative subset of normal feature embeddings, enabling efficient nearest-neighbour comparison during inference and strong performance in industrial anomaly detection tasks.
EfficientAD [54]	Teacher–student	Combines a teacher–student architecture with an autoencoder, detecting anomalies through feature discrepancies and reconstruction errors while maintaining fast inference.
FastFlow [64]	Distribution map	Uses fully convolutional two-dimensional normalizing flows operating directly on spatial feature maps, preserving spatial relationships.
CFA [68]	Memory bank	Models normal feature representations within a hypersphere-constrained latent space, improving separability between normal and anomalous patterns.
Dinomaly [82]	Transformer	Performs feature reconstruction using a lightweight transformer architecture with linear attention mechanisms and dropout-based bottlenecks to effectively model normal visual patterns.
AST [53]	Teacher–student	Detects anomalies by measuring discrepancies between feature representations produced by asymmetric networks.
GLASS [61]	One-class classification	Improves discrimination between normal and anomalous patterns through the synthesis of near-distribution anomalies at both global and local levels during training.
DDAD [87]	Diffusion model	Uses conditional diffusion and partial denoising to generate normal reconstructions while preserving object identity and structural information.
DiffusionAD [86]	Diffusion model	Performs anomaly detection through a noise-to-normal reconstruction framework using a one-step denoising strategy to improve inference efficiency.

enables a detailed assessment of model behavior under different anomaly conditions and facilitates the identification of anomaly types that are particularly difficult for each method.

The Acquisition-Related Anomalies Dataset serves as the primary dataset and is used to compare all evaluated methods. Based on the obtained image-level AUROC results, the best-performing approaches are subsequently selected for evaluation on the Foreign Object Intrusions and Rolling Inspection Artifacts datasets. This evaluation strategy enables the performance of the most promising methods to be further investigated across industrial scenarios exhibiting different visual characteristics and anomaly types.

In addition to detection performance, computational efficiency is considered a key requirement for deployment in industrial image acquisition systems. In the inspection setups considered in

this work, a minimum throughput of 5 FPS is required to ensure that anomaly detection does not become a bottleneck in the production process. Consequently, computational efficiency is assessed on the Acquisition-Related Anomalies Dataset and considered alongside detection performance when selecting the methods for further evaluation on the remaining benchmark datasets.

To assess deployment feasibility, each method is evaluated in terms of inference time, throughput, and GPU memory footprint. The mean inference time, its standard deviation, and the corresponding frames per second (FPS) are reported. All measurements are performed using a batch size of one, reflecting the intended deployment scenario in which images are processed individually as they are acquired. Prior to measurement, a warm-up phase is executed to eliminate initialization overheads and obtain representative inference times.

GPU memory footprint was measured during inference as the maximum memory usage of the corresponding process, as reported by `nvidia-smi`. Together, these metrics provide a practical assessment of the computational resources required for deployment and enable a balanced comparison between detection performance and computational efficiency.

The localization capability of the evaluated methods is also investigated through both qualitative and quantitative analyses. The qualitative evaluation is based on the visual inspection of the generated anomaly maps. For each evaluated model, representative heatmaps are presented for each anomaly category, enabling a visual comparison of localization behavior across different methods.

To enable quantitative localization analysis, pixel-level annotations were manually created for a subset of the Acquisition-Related Anomalies test set. The annotated subset contains 150 samples, including 50 randomly selected images from each of the *Defocus Blur*, *Illumination Variations*, and *Occlusions* categories. Examples of the resulting annotations are shown in Figure 4.9.

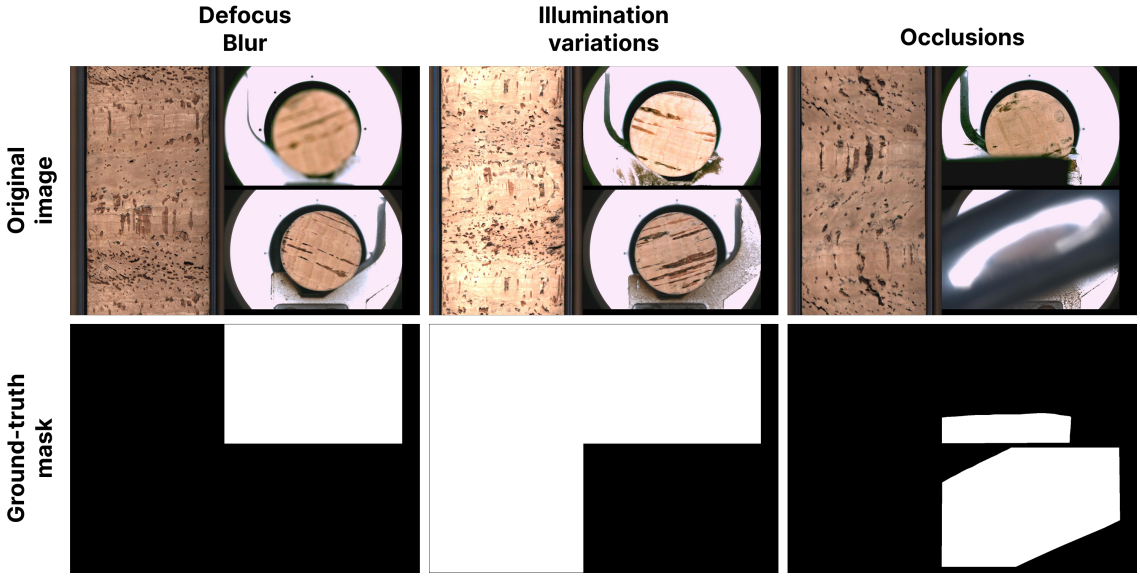


Figure 4.9: Representative anomalous samples and corresponding ground-truth masks used for localization evaluation.

For *Defocus Blur* and *Illumination Variations*, the annotated regions correspond to the complete field of view of the affected camera, as these anomalies alter the appearance of the entire image rather than a localized region. In contrast, *Occlusions* were annotated only within the regions directly affected by the anomaly, including both the occluding objects and the corresponding shadowed regions.

This annotated test set enables the computation of pixel-level AUROC in addition to image-level AUROC, providing a quantitative basis for comparing the localization performance of the evaluated methods.

4.4 Experimental Setup

This section describes the experimental setup adopted for training and evaluating the anomaly detection models considered in this work.

The original training procedures and hyperparameter configurations proposed by the respective authors were preserved whenever possible. Modifications were introduced only when required by hardware constraints, repository recommendations, or the objectives of the conducted experiments.

As described in the previous section, the evaluation process was standardized across all methods to ensure fair and consistent comparisons. Dedicated evaluation scripts were developed and adapted to generate a common set of outputs, including overall and anomaly-specific image-level AUROC scores, inference time, inference time standard deviation, throughput (FPS), GPU memory footprint, anomaly maps, and heatmap overlays. In addition, image-level ROC curves and anomaly score histograms were generated for all methods and are presented in Appendix C. This standardized evaluation pipeline enabled a comprehensive assessment of both detection performance and computational efficiency.

Since several of the evaluated methods were implemented within the Anomalib framework, a number of experimental settings and implementation details were shared across multiple models. Consequently, this section is divided into two subsections. Section 4.4.1 describes the common experimental settings shared across the models implemented within the Anomalib framework, while Section 4.4.2 presents the model-specific configurations, implementation details, and training procedures.

All experiments were conducted on a workstation equipped with an NVIDIA GPU, using driver version 581.42 and CUDA version 13.0. This hardware and software configuration ensured compatibility with the selected deep learning frameworks and stable execution of the training and evaluation pipelines.

To improve reproducibility and reduce variability across experiments, a fixed random seed of 42 was used throughout all experiments. This commonly adopted value controls stochastic components such as weight initialization, data shuffling, and mini-batch sampling, thereby improving the consistency and reproducibility of the obtained results.

4.4.1 Common Anomalib Experimental Setup

All experiments involving models implemented within the Anomalib [7] framework were conducted using Anomalib version 2.1.0 within a Conda environment running Python 3.10. The framework provides standardized utilities for dataset handling, model training, inference, and performance evaluation, facilitating consistent implementation and reproducible experimentation across different anomaly detection methods. Model training and evaluation were performed using the Anomalib Engine, which extends the PyTorch Lightning Trainer with functionality specifically designed for anomaly detection tasks. This high-level interface provides a unified workflow for training, inference, and evaluation, ensuring consistency across the evaluated models.

Each model was trained using its corresponding training script, where parameters such as batch size and number of training epochs are defined. While the default Anomalib configuration typically specifies 1000 training epochs, the number of epochs was adjusted when necessary according to the training requirements of each model. Additional modifications, such as changes to batch size or other training parameters, were introduced only when explicitly recommended in the corresponding model documentation or required by hardware constraints. To improve reproducibility, the random seed was set to 42 using the PyTorch Lightning `seed_everything()` function.

Some models require external resources that Anomalib downloads automatically. For example, EfficientAD uses pretrained teacher weights and ImageNette images to compute the student's penalty loss, discouraging the student from imitating the teacher on images outside the normal training distribution.

A common evaluation script was developed and used across all models implemented within the Anomalib framework. The script requires only the model architecture and the corresponding trained checkpoint, ensuring a consistent evaluation procedure across different methods.

Initially, the evaluation pipeline relied on the high-level `engine.predict` interface. However, preliminary experiments showed that this approach introduced non-negligible computational overhead, resulting in substantially reduced throughput and execution times that were not representative of the actual model inference cost. To obtain more accurate measurements, the evaluation script was modified to execute the standard inference pipeline directly through the `model.pre_processor`, `model.model`, and `model.post_processor` stages. Execution time was measured for the `model.model` stage, while preprocessing and postprocessing were still performed as part of the complete inference pipeline. Preliminary measurements showed that these additional stages contributed negligibly to the total execution time, indicating that the reported runtime closely approximates the complete inference cost.

The same evaluation script was also used to generate anomaly scores, anomaly maps, heatmap visualizations, and the performance metrics reported throughout the benchmark.

For heatmap visualization, a dedicated normalization strategy was adopted. In Anomalib, anomaly maps are normalized during post-processing using both the observed score range and the anomaly decision threshold. More specifically, the normalization maps the anomaly decision

threshold to a value of 0.5. As a result, regions classified as normal may still receive relatively high normalized values and appear with intermediate colors when visualized. This reduces the visual contrast between normal and anomalous areas, potentially making the heatmaps appear more active than the underlying model responses suggest.

To improve visual interpretability, the anomaly maps were normalized using only the min-max scaling component of the default Anomalib procedure, without incorporating the anomaly decision threshold. Consequently, the full color range reflects the relative magnitude of the anomaly responses, allowing the regions most strongly associated with anomalous behavior to be more clearly distinguished from the background. This normalization is applied exclusively for visualization purposes and does not affect anomaly scores, model predictions, decision thresholds, or any quantitative evaluation results.

Figure 4.10 illustrates the impact of the adopted visualization strategy using anomaly maps generated by FastFlow for a normal sample and a sample containing an *Occlusion* anomaly. In the default Anomalib visualization, regions of the normal sample are still displayed with intermediate and warm colors due to the threshold-centered normalization. In contrast, the proposed visualization suppresses these responses and more clearly highlights the regions most strongly associated with anomalous behavior.

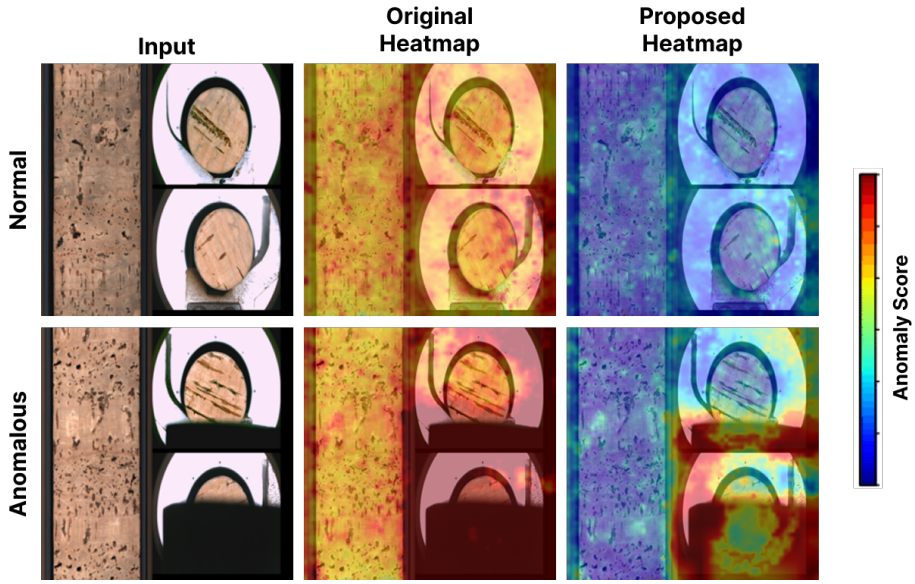


Figure 4.10: Comparison of heatmaps generated using the default Anomalib normalization and the proposed visualization strategy.

4.4.2 Model-Specific Configurations

This subsection describes the model-specific training and evaluation configurations adopted for the evaluated methods. Unless otherwise stated, the default hyperparameter settings provided by the respective implementations were used.

PatchCore, EfficientAD, CFA, Dinomaly, and FastFlow were trained and evaluated using their implementations available in the Anomalib framework [92] and therefore followed the common training and evaluation procedures described in Section 4.4.1.

For PatchCore, the training batch size was reduced from the default value of 32 to 8 due to GPU memory limitations. For EfficientAD, training was controlled through the number of optimization steps by setting `-trainer.max_steps` to 70000 and `-trainer.max_epochs` to `-1`. Both the *Small* and *Medium* variants were evaluated through the `-model.model_size` parameter. In addition, the implementation was modified to support disabling the autoencoder component. This modification removes the autoencoder reconstruction and student–autoencoder consistency losses during training and relies exclusively on the teacher–student discrepancy for anomaly scoring during inference, enabling the evaluation of a teacher–student-only variant. For CFA, the number of training epochs was set to 30.

For Dinomaly, both the training and evaluation batch sizes were set to 16, while training was controlled through `-trainer.max_steps`=5000 and `-trainer.max_epochs`=`-1`. For FastFlow, training was limited to 500 epochs with early stopping enabled. Although the default repository configuration specifies an early stopping patience of 3 epochs, preliminary experiments indicated that this setting could terminate training prematurely. Consequently, the patience value was increased to 50 epochs.

DevNet, AST, GLASS, DDAD, and DiffusionAD were trained and evaluated using the official implementations released by their respective authors [93], [94], [95], [96], [97]. For DevNet, no modifications to the default training configuration were required. The provided `localization.py` script was adapted to generate the outputs required for the benchmark.

For AST, the provided `eval.py` script was adapted to generate outputs consistent with those obtained for the remaining models. The implementation reports image-level AUROC using both the mean anomaly score and the maximum anomaly score extracted from the anomaly map. The maximum-score AUROC was retained throughout the benchmark, as it consistently provided higher and more stable performance across the evaluated datasets.

For GLASS, the custom dataset configuration recommended in the repository issue tracker was adopted. Training requires the Describable Textures Dataset (DTD), which is used to generate synthetic anomalies through texture-based perturbations. Furthermore, training was limited to 640 epochs and early stopping with a patience of 50 epochs was introduced due to the prolonged training times observed during preliminary experiments. The evaluation script uses the internal `_predict` function, which directly outputs image-level anomaly scores and the corresponding anomaly maps.

The default Dinomaly and GLASS pre-processing pipelines include image resizing followed by center cropping. Since relevant information may appear close to the image boundaries in the evaluated cork inspection datasets, experiments were conducted both with and without center cropping. When center cropping was disabled, the entire image was resized directly to the network input resolution, preserving all image content at the cost of increased geometric distortion. For the Acquisition-Related Anomalies Dataset, both methods achieved slightly better performance

using the default center-cropping configuration, indicating that the removed border regions did not contain information critical for detecting the evaluated acquisition-related anomalies.

Additional experiments with GLASS on the Foreign Object Intrusions and Rolling Inspection Artifacts datasets showed the opposite behavior. In the Foreign Object Intrusions Dataset, small foreign objects may appear near the image boundaries and can therefore be partially or completely removed by cropping. Similarly, the elongated aspect ratio of the reconstructed images in the Rolling Inspection Artifacts Dataset causes center cropping to discard a substantial portion of the available visual information. In these cases, preserving the complete image proved more beneficial than reducing the geometric distortions introduced by resizing.

Based on these observations, the results reported for the Acquisition-Related Anomalies Dataset were obtained using the default center-cropping configuration, whereas the results reported for the Foreign Object Intrusions and Rolling Inspection Artifacts datasets were obtained without center cropping.

Due to the computational requirements of diffusion-based models, DDAD was trained using three GPUs through `DataParallel`, taking advantage of the multi-GPU support already available in the implementation. Despite this, the default training batch size of 32 could not be used due to GPU memory limitations. Consequently, the DDAD-S configuration with a batch size of 16 was adopted. Following the recommendations of the authors, training was performed for 3000 epochs. Model checkpoints were evaluated on the validation set every 500 epochs from epoch 1000 onwards, with the best performance being obtained at epoch 1500. The results reported in this thesis therefore correspond to the checkpoint obtained at this epoch. A dedicated evaluation script was developed based on the original DDAD inference pipeline to generate outputs consistent with the remainder of the benchmark. In the original implementation, anomaly maps are visualized directly from the generated anomaly responses without the use of a dedicated colormap. To ensure consistency with the visualization procedure adopted for the other evaluated methods, anomaly maps were normalized using global min–max scaling and visualized using the Jet colormap.

For DiffusionAD, the default batch size of 16 was reduced to 2 due to GPU memory limitations. Additional experiments using `DataParallel`, following the same strategy adopted for DDAD, were conducted across three GPUs. However, this increased the maximum feasible batch size only to 4, which remained substantially below the default configuration. Additionally, in the original DiffusionAD implementation, anomaly maps are normalized independently for each image prior to visualization. To ensure consistency with the evaluation pipeline adopted throughout the benchmark, this procedure was replaced by global min–max normalization.

4.5 Improving Robustness

Several evaluated models achieved image-level AUROC values above 99% on the Acquisition-Related Anomalies Dataset. However, some of these top-performing methods demonstrated lower detection performance on the Foreign Object Intrusions and Rolling Inspection Artifacts datasets,

highlighting the additional challenges posed by more complex industrial acquisition conditions. To improve robustness under these scenarios, additional improvement strategies were investigated.

4.5.1 Data Augmentations

Initially, data augmentation techniques were explored to increase the variability of the normal training data and improve robustness to appearance variations. Care was taken to avoid unrealistic transformations that could introduce artificial visual patterns or significantly distort the underlying normal data distribution. For the Foreign Object Intrusions Dataset, random horizontal and vertical flips ($p = 0.5$) together with small random rotations ($\pm 3^\circ$) were applied during training, whereas the Rolling Inspection Artifacts Dataset used only random horizontal and vertical flips.

However, these augmentations generally did not improve performance and, in some cases, degraded anomaly detection capability. While data augmentation is often beneficial in supervised learning settings, anomaly detection methods trained exclusively on normal samples rely on learning a precise representation of the normal data distribution. Increasing the variability of the training data may therefore broaden the learned notion of normality, reducing the separability between normal and anomalous samples.

Since the evaluated methods typically exhibited lower discrimination between normal and anomalous samples on the more challenging datasets, the incorporation of labeled anomalous samples was investigated. Although unsupervised training remains attractive due to the scarcity of anomalous samples, real anomalous examples often become available during system operation as new anomalies are observed and manually verified. Such samples contain information that is not captured by synthetic anomalies or normality modeling and may therefore help improve the discrimination between normal and anomalous conditions. Motivated by this observation, a lightweight supervision strategy based on a limited number of image-level labeled anomalous samples was investigated.

4.5.2 Lightweight Supervision

To exploit the labeled anomalous samples available for the Foreign Object Intrusions and Rolling Inspection Artifacts datasets, a lightweight supervision strategy was adopted. Instead of retraining the anomaly detection models, compact supervised heads were trained on top of frozen representations extracted from pretrained anomaly detection networks. By leveraging the visual representations learned during unsupervised training, the available labeled anomalous samples can primarily be used to learn the decision boundary between normal and anomalous conditions.

Since only the supervised head is optimized, adaptation to newly observed anomaly types can be achieved without retraining the underlying anomaly detection model, substantially reducing the computational resources and training time required to incorporate additional anomaly information.

The remainder of this subsection describes the proposed supervision strategy, the design of the supervised heads, and the methodology used to evaluate the impact of different levels of anomaly supervision on detection performance.

4.5.2.1 Implementation Details

In all evaluated models, the original anomaly detection networks were kept frozen and used to extract anomaly-related representations. A lightweight supervised head was then trained on top of these representations. This strategy preserves the anomaly detection capabilities learned during unsupervised training while enabling the incorporation of anomaly-specific information through a limited number of labeled anomalous samples.

The proposed heads share a common architecture consisting of batch normalization, a fully connected layer with 128 hidden units, ReLU activation, dropout regularization ($p = 0.3$), and a final binary classification layer. Training was performed using the AdamW optimizer with a learning rate of 10^{-3} and a weight decay of 10^{-4}. The heads were intentionally kept lightweight and architecturally simple. This design choice allows the impact of the additional anomaly supervision to be evaluated independently of classifier complexity. Although more sophisticated architectures could potentially achieve further performance gains, the objective of these experiments is to assess whether a small amount of supervision can effectively improve anomaly detection performance while maintaining low computational overhead.

For EfficientAD, the supervised head combines information derived from the teacher–student discrepancy and anomaly maps produced by the model. Global average and maximum pooled representations of the student–teacher distance tensor, the student anomaly map, and the student–autoencoder anomaly map are extracted and combined with the normalized image-level anomaly score produced by EfficientAD.

For AST, the supervised head combines information from multiple stages of the anomaly detection pipeline. Global average and maximum pooled representations are extracted from the teacher embeddings, student embeddings, and their absolute differences. These features are complemented by pooled anomaly-map representations, the original AST image-level anomaly scores computed using both mean and maximum anomaly-map aggregation, and summary statistics derived from the teacher and student flow Jacobians. Although the original implementation supports and recommends the use of pre-extracted features, all inputs to the supervised head were computed on-the-fly from the original AST architecture to ensure consistency with the intended deployment scenario.

For FastFlow, the supervised head operates on the latent variables produced by the normalizing flow blocks at multiple feature scales. Global average and maximum pooled representations are extracted from each latent tensor and combined with the normalized image-level anomaly score generated by the original FastFlow model.

For GLASS, supervision relies on the patch-level representations extracted by the model backbone. Global average and maximum pooled patch embeddings are combined with the official GLASS image-level anomaly score and the maximum patch-level discriminator score. This enables the classifier to exploit both global anomaly evidence and localized patch-level responses.

Across all models, training was monitored using validation AUROC, and the checkpoint achieving the highest validation AUROC was retained for final evaluation.

4.5.2.2 Experiments

To evaluate the proposed lightweight supervision strategy under different levels of anomaly supervision, several experimental settings were considered. For all experiments, anomalous samples originally belonging to the test set were re-partitioned into training, validation, and test subsets. A 40%/20%/40% split was adopted, with 40% of the anomalous samples used for supervised training, 20% for validation and model selection, and the remaining 40% for testing. In addition, approximately 30% of the normal samples were allocated to the validation set, while the remaining normal samples were retained for testing.

For the Rolling Inspection Artifacts Dataset, anomalous samples were partitioned independently for each anomaly category to preserve a balanced distribution of anomaly types across the training, validation, and test sets.

To investigate the influence of supervision level, additional experiments were conducted using 10%, 20%, and 40% of the total anomalous samples for training. These settings correspond to using 25%, 50%, and 100% of the anomalous samples available in the training partition, respectively. The validation and test sets remained unchanged throughout these experiments, enabling a direct comparison between different supervision levels. Tables 4.5 and 4.6 detail the sample distributions associated with the different experimental settings.

Table 4.5: Sample distribution used for the lightweight supervision experiments on the Foreign Object Intrusions dataset. For the train–test swapped experiment, intrusion samples assigned to the training and test partitions were exchanged, resulting in 115 and 113 intrusion samples in the training and test sets, respectively.

Type	**Train (10%)**	**Train (20%)**	**Train (40%)**	**Val**	**Test**
Normal	1477	1477	1477	71	166
Intrusions	29	57	113	56	115

To assess the sensitivity of the proposed approach to the particular anomalous samples selected for training, an additional train–test swapped experiment was performed. In this setting, the anomalous samples originally assigned to the training and test partitions were exchanged while preserving the validation set. This experiment provides insight into the robustness of the obtained results with respect to the selected supervision samples.

For the Rolling Inspection Artifacts Dataset, a further experiment was conducted using supervision from a single anomaly category. Specifically, only *Roller Contamination 1* samples available in the training partition were used for supervised training, while samples belonging to all remaining anomaly categories were excluded. This setting was designed to evaluate whether supervision obtained from a single anomaly category could improve the detection of other anomalies present in the dataset.

Since the standard supervised split and the train–test swapped split introduced alternative test partitions, the original unsupervised models were re-evaluated on the corresponding test sets to ensure fair comparisons between supervised and unsupervised approaches.

Table 4.6: Sample distribution used for the lightweight supervision experiments on the Rolling Inspection Artifacts dataset. For the single anomaly category experiment, only the 45 *Roller Contamination 1* samples available in the training partition were used for supervised training. For the train–test swapped experiment, anomalous samples assigned to the training and test partitions were exchanged independently for each anomaly category, resulting in 133 and 130 anomalous samples in the training and test sets, respectively.

Type	**Train (10%)**	**Train (20%)**	**Train (40%)**	**Val**	**Test**
Normal	1110	1110	1110	80	186
Roller contamination 1	11	23	45	23	45
Roller contamination 2	4	8	16	8	17
Swirl	4	8	17	8	17
Duplication	5	10	19	10	20
Discontinuity	5	9	19	9	19
Cropped	2	4	8	4	8
Other	2	3	6	3	7
Anomalous samples (Total)	33	65	130	65	133

The data augmentation strategies described previously were also applied during supervised head training. Unlike the fully unsupervised setting, these augmentations increase the variability of the limited anomalous samples available during training, exposing the supervised heads to a broader range of anomaly appearances while reducing the risk of overfitting.

Finally, an additional experiment was conducted to evaluate the ability of the proposed supervision framework to generalize to previously unseen logo patterns. For each model, the best-performing supervised configuration identified in the previous experiments was compared against its corresponding unsupervised baseline on a test set containing logo patterns that were excluded from both the unsupervised and supervised training data.

Together, these experiments provide a comprehensive evaluation of the proposed supervision framework across the two application scenarios, considering different numbers of anomalous training samples and dataset partitions.

Chapter 5

Results

This chapter presents a detailed analysis of the results obtained from the benchmark evaluation and the supervision experiments described in the previous chapter. These results provide insights into the performance, robustness, limitations, and generalization capabilities of the evaluated methods and proposed approaches.

5.1 Benchmark Evaluation

The considered anomaly detection methods are benchmarked in terms of detection performance, computational efficiency, anomaly localization capabilities, and robustness across different industrial inspection scenarios. The obtained results enable a comparative analysis of how different anomaly detection approaches handle acquisition-related anomalies and perform across different industrial acquisition systems.

5.1.1 Image-Level Detection Performance

Table 5.1 summarizes the image-level AUROC results obtained for each anomaly category, as well as the overall image-level AUROC, on the Acquisition-Related Anomalies Dataset. Overall, the evaluated methods achieve strong performance, with six methods exceeding 99.5% image-level AUROC, indicating that acquisition-related anomalies, such as defocus blur, illumination variations, lens dirt, and partial occlusions, can be reliably detected using modern anomaly detection techniques.

GLASS achieves the best overall performance, followed by EfficientAD-M w/o AE, FastFlow, DDAD, AST, and EfficientAD-M. Consistent with the evaluation strategy defined in Section 4.3, these methods satisfy the image-level AUROC selection criterion and are therefore considered for further evaluation on the remaining datasets, subject to the runtime requirements. PatchCore achieves the lowest overall performance, followed by DiffusionAD. However, both methods were trained using reduced batch sizes due to GPU memory limitations. Consequently, their results may not fully reflect the performance achievable under the configurations recommended by the original authors.

Table 5.1: Image-level AUROC comparison on the Acquisition-Related Anomalies Dataset.

Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
DevNet	99.97	99.64	72.88	97.84	73.42	100.0	100.0	96.94
PatchCore	93.56	85.57	57.04	98.34	81.29	100.0	100.0	90.79
EfficientAD-S	96.54	100.0	95.94	99.60	96.14	100.0	100.0	98.93
EfficientAD-M	98.31	100.0	98.98	99.81	98.14	100.0	100.0	99.53
EfficientAD-M w/o AE	99.96	100.0	100.0	99.90	96.14	100.0	100.0	99.85
FastFlow	100.0	100.0	100.0	99.77	97.30	100.0	100.0	99.83
CFA	95.04	97.39	84.22	99.67	89.19	100.0	100.0	97.06
Dinomaly	99.47	98.65	86.71	99.69	90.38	100.0	100.0	98.40
AST	99.88	99.96	100.0	99.60	95.58	100.0	100.0	99.69
GLASS	99.97	99.99	99.79	99.95	98.17	100.0	100.0	**99.91**
DDAD	98.97	100.0	99.84	99.97	99.19	100.0	100.0	99.78
DiffusionAD	86.30	99.76	90.41	96.57	85.71	100.0	98.60	95.26

1: Defocus Blur 2: Illumination Variations 3: Lens Dirt
4: Occlusions 5: Boundary 6: Misaligned Cork 7: Missing Cork

Performance also varies across anomaly categories. *Misaligned Cork* is detected perfectly by all evaluated methods, while *Missing Cork* is detected perfectly by all methods except DiffusionAD, which achieves 98.60% AUROC. *Illumination Variations* and *Occlusions* are detected reliably by most evaluated methods, with the majority achieving AUROC values above 99.5%. For *Occlusions*, only DevNet (97.84%), PatchCore (98.34%), and DiffusionAD (96.57%) fall below this threshold. For *Illumination Variations*, PatchCore (85.57%), CFA (97.39%), and Dinomaly (98.65%) are the only methods achieving less than 99.5% AUROC. Although the evaluated methods generally achieve strong performance on *Defocus Blur*, this category proves to be more challenging than *Illumination Variations* and *Occlusions*, with several methods obtaining AUROC values below the 99.5% threshold, namely DiffusionAD (86.30%), PatchCore (93.56%), CFA (95.04%), EfficientAD-S (96.54%), DDAD (98.97%), EfficientAD-M (98.31%), and Dinomaly (99.47%).

Lens Dirt proves particularly challenging for several approaches, producing the lowest AUROC values for PatchCore, DevNet, CFA and Dinomaly. This behavior is consistent with the subtle nature of *Lens Dirt* anomalies, which can be more difficult to distinguish from normal samples than anomalies associated with stronger structural changes in the image.

Notably, GLASS, AST, FastFlow, DDAD, EfficientAD-M, and EfficientAD-M w/o AE achieve near-perfect performance on this category, with several methods reaching 100% AUROC. This suggests that these methods are better able to distinguish subtle acquisition-related anomalies. For several of these top-performing approaches, the *Boundary* category proves to be the most challenging anomaly type. This observation is consistent with the intended design of the category, which contains borderline cases with only mild deviations from normal operating conditions. Never-

theless, GLASS, DDAD, and EfficientAD-M achieve AUROC values of 98.17%, 99.19%, and 98.14%, respectively, demonstrating that even such subtle deviations can be detected reliably.

Comparing the two EfficientAD-M variants, removing the autoencoder results in slightly better overall detection performance. These results suggest that the autoencoder provides limited benefit for the acquisition-related anomalies considered in this dataset. This observation is consistent with the original motivation of the autoencoder, which was introduced to improve the detection of logical anomalies. Most of the anomalies evaluated in this benchmark correspond to acquisition disturbances and structural image degradations, which appear to be adequately modeled by the teacher–student component alone. The only exception is the *Boundary* category, where the additional reconstruction-based information provided by the autoencoder may help distinguish subtle deviations that are visually closer to normal operating conditions.

5.1.2 Computational Performance

Table 5.2 presents the computational performance metrics obtained on the Acquisition-Related Anomalies Dataset. Most evaluated methods satisfy the real-time requirements defined for the considered industrial applications, namely a minimum throughput of 5 frames per second (FPS). The only exceptions are DDAD and DiffusionAD, which achieve 0.73 FPS and 4.28 FPS, respectively. Consequently, despite DDAD being among the best-performing methods in terms of detection accuracy, it was not selected for further evaluation on the remaining datasets. Nevertheless, the strong performance achieved by DDAD demonstrates that diffusion-based methods can effectively detect acquisition-related anomalies, even though their computational requirements currently limit their suitability for real-time deployment in the considered industrial scenario, despite the efficiency improvements introduced by recent diffusion-based approaches.

Table 5.2: Computational performance comparison of the evaluated methods.

Method	**GPU Memory (MB)**	**Inference Time (ms)**	**FPS**
DevNet	5192	14.61 ± 6.40	68.45
PatchCore	1402	38.99 ± 13.02	25.65
EfficientAD-S	548	14.10 ± 3.70	70.93
EfficientAD-M	764	28.49 ± 7.33	35.10
EfficientAD-M w/o AE	746	23.75 ± 2.15	42.11
FastFlow	462	45.73 ± 19.84	21.87
CFA	1036	28.85 ± 12.04	34.66
Dinomaly	2428	41.63 ± 24.17	24.02
AST	1440	139.85 ± 32.37	7.15
GLASS	864	39.09 ± 18.31	25.58
DDAD	1494	1363.17 ± 117.22	0.73
DiffusionAD	2096	233.82 ± 68.25	4.28

Among the methods satisfying the real-time requirement, AST exhibits the highest inference latency. Despite delivering one of the strongest detection performances, AST is substantially

slower than the remaining approaches due to the additional feature extraction performed during inference. Nevertheless, its throughput remains above the minimum real-time requirement.

FastFlow combines one of the strongest detection performances with the lowest GPU memory footprint, requiring only 462 MB of GPU memory. In contrast, DevNet exhibits the highest memory consumption, requiring 5192 MB. Among the unsupervised methods, Dinomaly is the most memory-intensive approach.

EfficientAD-S is the most computationally efficient method, achieving the highest throughput while requiring only 548 MB of GPU memory. Although EfficientAD-M incurs an increase in computational cost, it achieves stronger detection performance while remaining both fast and memory-efficient. As expected, removing the autoencoder branch from EfficientAD-M further reduces both memory consumption and inference latency, resulting in a model that is both lighter and faster than the original variant, with a slight improvement in detection performance.

Overall, the results indicate that several methods achieve comparable detection performance while exhibiting substantially different computational requirements. In particular, GLASS, DDAD, AST, FastFlow, and the EfficientAD variants achieve similarly strong detection performance despite large differences in inference speed and memory consumption, highlighting the importance of considering both detection accuracy and computational efficiency when selecting a model for industrial deployment.

5.1.3 Localization Analysis

Figure 5.1 presents representative anomaly heatmaps generated by the evaluated methods for each anomaly category in the Acquisition-Related Anomalies Dataset. Overall, the evaluated methods are generally able to localize the acquisition-related anomalies, although substantial differences can be observed in the quality and spatial extent of the generated anomaly maps.

Across all anomaly categories, FastFlow and GLASS generally produce the strongest anomaly responses. However, FastFlow frequently highlights portions of the background, a limitation also reported by the original authors of the method. A similar behavior can be observed for PatchCore, whose heatmaps often exhibit anomaly responses in visually normal regions, resulting in the greener background appearance visible in several examples. In contrast, GLASS typically provides a clearer visual separation between normal and anomalous regions, yielding heatmaps that are generally easier to interpret despite occasionally producing somewhat dilated responses. The EfficientAD variants tend to generate comparatively weak anomaly responses, although they often produce more compact and spatially localized heatmaps. An exception can be observed for the *Illumination Variations* example, where a strong and well-localized response is produced. DevNet and AST occasionally exhibit isolated high-intensity anomaly activations within visually normal regions, indicating that parts of the anomaly response may not always correspond directly to the underlying acquisition disturbance. CFA and Dinomaly generally produce stable and well-localized heatmaps, providing a good compromise between anomaly confidence and localization precision. DDAD generally localizes the anomalous regions correctly; however, in some cases, such as *Defocus Blur* and *Lens Dirt*, the anomaly responses are comparatively subtle, making

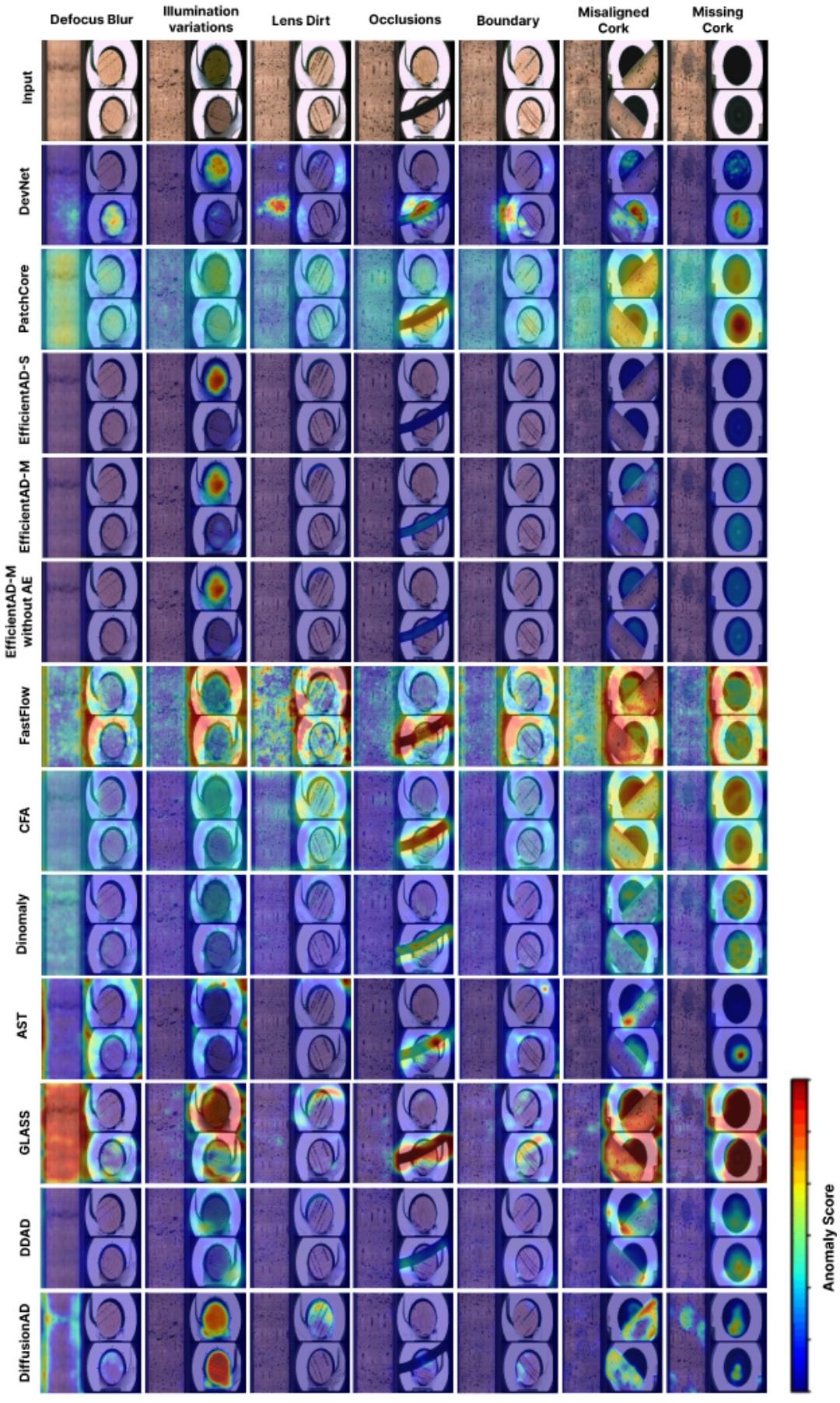


Figure 5.1: Qualitative comparison of anomaly localization results for representative samples of each anomaly category in the Acquisition-Related Anomalies Dataset.

the affected regions less visually distinguishable. DiffusionAD exhibits a less stable localization behavior, occasionally producing anomaly responses in visually normal regions, as illustrated by the *Missing Cork* example. For *Occlusions*, the model fails to highlight the complete extent of the occluding object and instead produces anomaly responses in portions of the cork surface, resulting in a less accurate localization of the anomaly.

For the *Missing Cork* category, despite the anomaly affecting both top views equally, most methods assign noticeably different anomaly response intensities to the two views. This indicates that the anomaly response magnitude may vary across views even when the underlying disturbance is visually similar. Nevertheless, for both *Illumination Variations* and *Defocus Blur*, the heatmaps generally reflect the relative severity of the anomaly across the top views. In the *Illumination Variations* example, the disturbance is more pronounced in the top-right view than in the top-left view, and most methods assign a correspondingly stronger anomaly response to that region. A similar behavior can be observed for *Defocus Blur*. However, a different pattern emerges when comparing the top views with the cork body view. In the *Defocus Blur* example, the strongest blur affects the cork body view, yet most methods assign a stronger anomaly response to the affected top view. A similar tendency can be observed for *Illumination Variations*, where the anomaly response associated with the cork body view is generally weaker than that of the affected top view. These observations suggest that several methods are less sensitive to acquisition-related disturbances affecting the cork body view than to similar disturbances occurring in the top views. GLASS constitutes a notable exception, producing a strong response in the cork body view and yielding a localization pattern that more closely matches the visual severity of the disturbance. DiffusionAD also produces noticeable responses in the cork body view, although its localization pattern is generally less consistent with the relative severity of the disturbances across the different top views.

Additional differences can be observed for the *Lens Dirt* and *Boundary* categories. In the *Lens Dirt* example, the anomaly is not clearly highlighted by DevNet and PatchCore, which is consistent with their comparatively lower quantitative performance on this anomaly category. The EfficientAD variants produce only a very subtle anomaly response, making the highlighted region barely distinguishable in the heatmap, although the contaminated area is still correctly localized. In contrast, the *Boundary* example generates virtually no response for the EfficientAD variants, indicating that the subtle deviations present in this sample are not detected by these models.

Table 5.3 presents the image-level and pixel-level AUROC results for the anomaly categories for which pixel-level annotations were created. Overall, most methods achieve strong image-level performance, with FastFlow and GLASS obtaining the highest overall image-level AUROC. However, larger differences emerge when evaluating anomaly localization performance.

GLASS achieves the highest overall pixel-level AUROC, followed by FastFlow and DDAD. A major factor contributing to the superior performance of these methods is their localization behavior on the *Defocus Blur* and *Illumination Variations* categories. Unlike the remaining methods, which primarily highlight the cork region, GLASS, FastFlow, and DDAD also assign anomaly responses to the affected background regions. Since the pixel-level annotations cover the entire

Table 5.3: Image-level and pixel-level AUROC comparison across anomaly types.

Method	1		2		3		Total	
	I	P	I	P	I	P	I	P
DevNet	99.96	75.97	99.66	81.39	95.87	91.82	98.46	80.61
PatchCore	96.17	83.46	87.07	80.10	99.07	99.32	94.10	82.00
EfficientAD-S	96.18	74.79	100.0	81.89	97.96	91.92	98.04	78.31
EfficientAD-M	98.39	73.63	100.00	82.87	99.04	91.98	99.15	77.93
EfficientAD-M w/o AE	99.96	75.57	100.0	82.41	99.47	91.92	99.81	78.95
FastFlow	100.0	95.23	100.0	87.49	99.96	97.99	**99.99**	89.18
CFA	95.98	91.30	97.22	84.83	99.97	99.51	97.72	87.76
Dinomaly	99.42	86.90	98.76	81.69	99.01	98.82	99.06	84.40
AST	99.91	92.16	99.97	80.56	99.18	96.86	99.69	84.98
GLASS	99.98	**97.65**	99.99	**97.28**	100.0	**99.14**	**99.99**	**96.03**
DDAD	99.93	92.93	100.0	96.88	100.0	99.49	99.98	89.04
DiffusionAD	87.36	59.34	99.72	69.87	86.34	75.13	94.47	64.51

1: Defocus Blur 2: Illumination Variations 3: Occlusions
I: Image-level AUROC P: Pixel-level AUROC

affected field of view, this behavior results in a substantially higher overlap with the ground-truth masks and consequently higher pixel-level AUROC values.

DiffusionAD achieves the lowest overall pixel-level AUROC and consistently ranks as the lowest-performing method across all evaluated anomaly categories. This observation is consistent with the qualitative heatmap analysis, where the model occasionally produces anomaly responses in visually normal regions and struggles to accurately localize the full extent of certain anomalies.

The EfficientAD variants achieve the second lowest overall pixel-level AUROC values, despite obtaining strong image-level detection performance. This observation is consistent with the heatmap analysis, where the anomaly responses are often very weak and barely distinguishable from the surrounding regions. While these responses are generally sufficient for image-level classification, they result in comparatively poor localization performance. This behavior is particularly evident for the *Defocus Blur* and *Occlusions* categories, where the EfficientAD variants rank among the lowest-performing methods. In contrast, the performance gap is less pronounced for *Illumination Variations*, for which the EfficientAD variants produce stronger and more clearly localized anomaly responses.

Overall, the pixel-level evaluation is consistent with the qualitative heatmap analysis. Methods that generate more spatially coherent and accurately localized anomaly maps achieve superior localization performance, whereas methods producing weaker or less precise anomaly responses obtain lower pixel-level AUROC values.

5.1.4 Robustness Across Industrial Scenarios

Tables 5.4 and 5.5 present the image-level AUROC results on the Foreign Object Intrusions Dataset and the Rolling Inspection Artifacts Dataset, respectively. Compared to the Acquisition-Related Anomalies Dataset, these datasets reveal substantially larger performance differences between the evaluated methods, providing a clearer assessment of their robustness across different anomaly types and operational conditions.

Table 5.4: Image-level AUROC comparison on the Foreign Object Intrusions Dataset.

Method	**Total**
EfficientAD-M	**99.50**
EfficientAD-M (w/o AE)	99.17
FastFlow	78.29
AST	93.51
GLASS	99.40

Table 5.5: Image-level AUROC comparison on the Rolling Inspection Artifacts Dataset.

Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
EfficientAD-M	88.96	90.79	99.96	60.30	55.58	70.96	95.94	80.78
EfficientAD-M w/o AE	86.78	91.26	99.96	68.00	71.39	58.12	92.69	82.56
FastFlow	98.19	99.62	99.71	76.05	75.28	95.88	96.42	**91.75**
AST	98.85	98.09	99.90	60.58	75.51	83.57	89.33	88.43
GLASS	96.58	99.11	99.92	68.18	72.54	70.45	93.12	87.87

1: Roller contamination 1 2: Roller contamination 2 3: Swirl
4: Duplication 5: Discontinuity 6: Cropped 7: Other

EfficientAD-M achieves the best overall performance on the Foreign Object Intrusions Dataset but the lowest performance on the Rolling Inspection Artifacts Dataset. Conversely, FastFlow achieves the best overall performance on the Rolling Inspection Artifacts Dataset while obtaining the lowest performance on the Foreign Object Intrusions Dataset. These results indicate that the performance of both methods is strongly influenced by the characteristics of the anomalies present in each dataset.

In contrast, GLASS consistently ranks among the top-performing methods across all evaluated datasets. Although it does not achieve the highest score on either the Foreign Object Intrusions Dataset or the Rolling Inspection Artifacts Dataset, it achieves the best overall performance on the Acquisition-Related Anomalies Dataset and maintains consistently strong results across all benchmarks. This behavior suggests greater robustness to variations in anomaly characteristics and acquisition conditions. AST exhibits a similar trend, maintaining competitive performance across both datasets without the large performance fluctuations observed for EfficientAD-M and FastFlow.

The performance ranges observed across the evaluated methods further highlight the distinct characteristics of the two datasets. On the Foreign Object Intrusions Dataset, the evaluated methods achieve AUROC values ranging from 78.29% to 99.50%, whereas on the Rolling Inspection Artifacts Dataset the corresponding range is 80.78% to 91.75%. While the best-performing method achieves substantially higher performance on the Foreign Object Intrusions Dataset, the larger gap between the best- and worst-performing methods indicates that detection performance is more strongly influenced by the selected anomaly detection approach. In contrast, the lower overall AUROC values observed on the Rolling Inspection Artifacts Dataset suggest that this dataset is inherently more challenging for all evaluated methods, limiting the performance attainable even by the strongest approaches.

Across all evaluated datasets, EfficientAD-M w/o AE achieves comparable or better performance than the original model. This observation reinforces the trend already observed on the Acquisition-Related Anomalies Dataset, indicating that the additional autoencoder branch provides limited benefits for the anomaly detection tasks considered in this work. Nevertheless, the original EfficientAD-M model is used in the subsequent experiments, as it represents the complete version of the method proposed by the authors.

For the Rolling Inspection Artifacts Dataset, performance also varies considerably across anomaly categories. Overall, the evaluated methods achieve very high AUROC values on *Roller Contamination 1*, *Roller Contamination 2*, and *Swirl*, which correspond to the most relevant anomaly types in this inspection scenario. *Swirl* exhibits near-perfect performance across all evaluated methods, reflecting the pronounced visual deviations associated with this anomaly type. The main exception is the EfficientAD-M variants, which achieve noticeably lower AUROC values on *Roller Contamination 1* and *Roller Contamination 2* than the remaining approaches.

Substantially lower AUROC values are observed for *Duplication*, *Discontinuity*, and *Cropped* across most methods, indicating that these anomaly types are considerably more challenging to detect. This behavior is consistent with the nature of these anomalies, which often produce more complex or less distinctive deviations from normal samples than the remaining anomaly categories. Consequently, the lower performance observed on these categories is largely responsible for the reduced overall AUROC achieved on the Rolling Inspection Artifacts Dataset.

Figures 5.2 and 5.3 show example anomaly heatmaps generated by the evaluated methods for selected anomalous samples from the Foreign Object Intrusions Dataset and the Rolling Inspection Artifacts Dataset, respectively. The examples shown for the Rolling Inspection Artifacts Dataset include the *Roller Contamination 1*, *Roller Contamination 2*, and *Swirl* categories, together with a *Discontinuity* sample, as this was one of the most challenging anomaly categories. In the selected *Discontinuity* example, a small portion of the upper logo is missing; however, none of the evaluated methods successfully highlights this missing region, illustrating the difficulty of detecting this anomaly. AST produces anomaly responses in the image; however, the strongest activations occur outside the defective region, suggesting that they correspond to false-positive detections. This interpretation is supported by the observation of similar activations on normal samples and

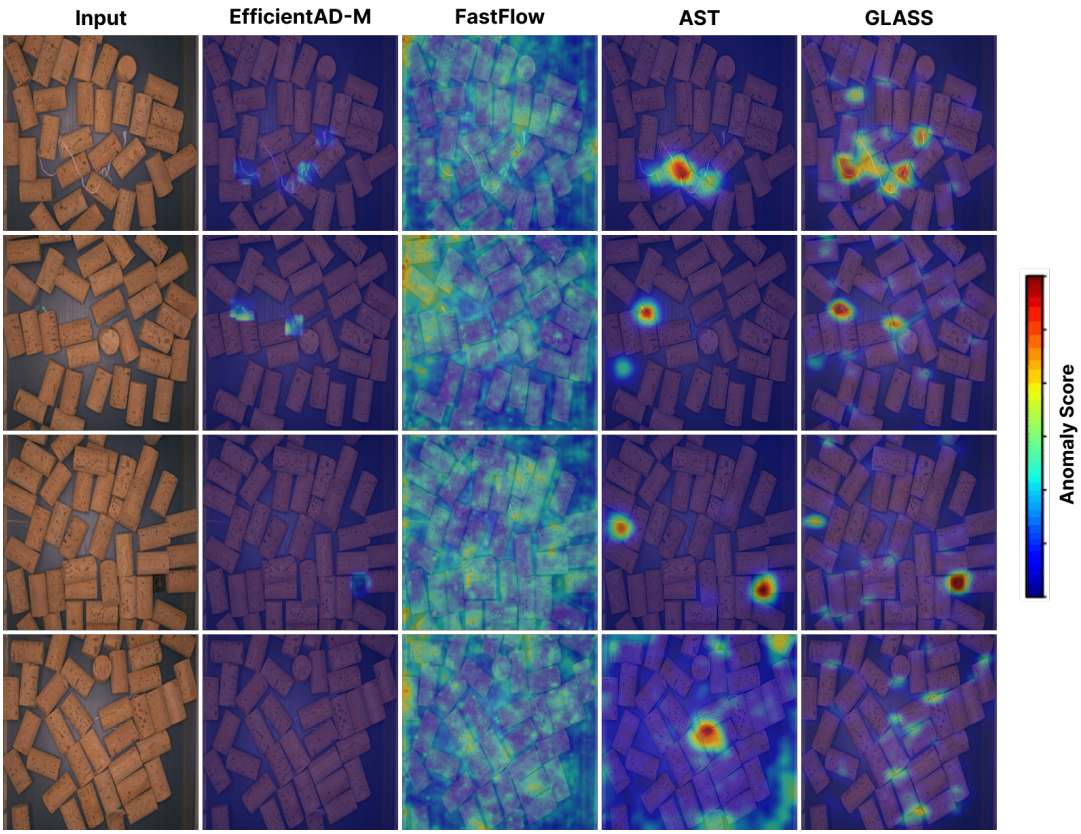


Figure 5.2: Qualitative comparison of anomaly localization results for selected samples from the Foreign Object Intrusions Dataset.

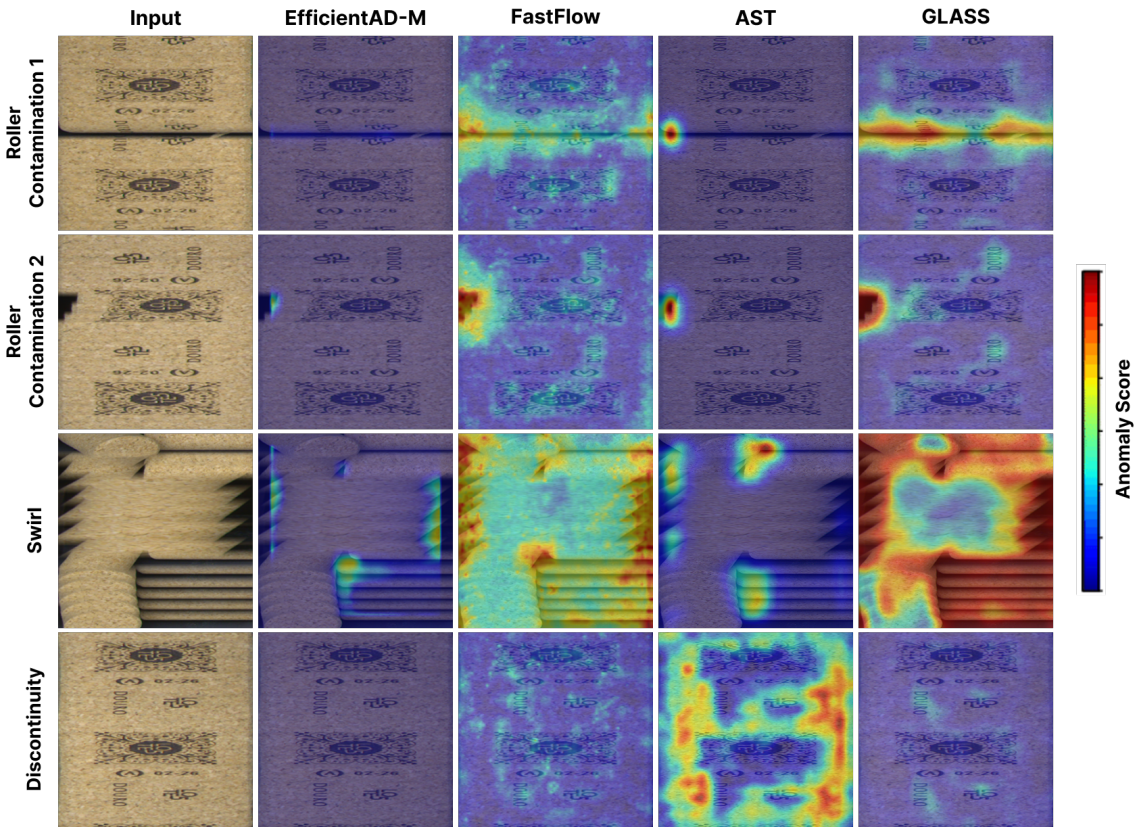


Figure 5.3: Qualitative comparison of anomaly localization results for selected samples from the Rolling Inspection Artifacts Dataset.

is consistent with the previous findings that the method occasionally generates false-positive responses.

Overall, the qualitative localization results are consistent with those obtained on the Acquisition-Related Anomalies Dataset. GLASS continues to produce the most reliable anomaly maps, combining strong anomaly responses with accurate localization across both datasets. FastFlow continues to exhibit some background activations, with mild anomaly responses occasionally appearing in visually normal regions across both datasets. Nevertheless, it provides accurate localization of the anomalous regions on the Rolling Inspection Artifacts Dataset. In contrast, its localization performance is noticeably weaker on the Foreign Object Intrusions Dataset, consistent with its substantially lower image-level detection performance. AST often produces spatially restricted anomaly responses, highlighting only small portions of the anomalous regions, as illustrated by the wire intrusion and *Roller Contamination 1* examples. Finally, EfficientAD-M continues to produce comparatively weaker anomaly responses than the remaining methods, making the anomalous regions less visually distinguishable, although the affected regions are generally correctly identified.

Tables 5.6 and 5.7 present the results obtained when the evaluated models were trained using the augmentation strategy described in Section 4.5.1.

Table 5.6: Image-level AUROC comparison on the Foreign Object Intrusions Dataset using data augmentation.

Method	**Total**
EfficientAD-M + Aug	99.19
FastFlow + Aug	76.25
AST + Aug	69.36
GLASS + Aug	**99.40**

Table 5.7: Image-level AUROC comparison on the Rolling Inspection Artifacts Dataset using data augmentation.

Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
EfficientAD-M + Aug	87.78	90.47	99.78	50.65	63.81	67.84	89.52	79.54
FastFlow + Aug	98.39	99.45	99.80	72.37	74.61	97.26	94.97	**91.17**
AST + Aug	96.26	97.30	98.34	63.77	70.80	67.11	92.27	86.18
GLASS + Aug	96.47	99.85	99.96	64.08	71.69	58.93	93.30	86.51

1: Roller contamination 1 2: Roller contamination 2 3: Swirl
4: Duplication 5: Discontinuity 6: Cropped 7: Other

The proposed augmentations do not improve performance and consistently result in lower AUROC values across all evaluated methods and datasets. The most pronounced performance

degradation is observed for AST on the Foreign Object Intrusions Dataset, where the AUROC decreases from 93.51% to 69.36%. In contrast, GLASS is only marginally affected, exhibiting a negligible reduction of only 0.0015 AUROC points. These results suggest that, rather than improving robustness, the introduced transformations may reduce the discriminability between normal and anomalous samples, making it more difficult for the models to distinguish normal observations from anomalous ones. Unlike supervised classification tasks, anomaly detection methods trained exclusively on normal samples rely on learning a precise representation of the normal data distribution. The proposed augmentations increase the variability of the normal training data, potentially broadening the learned notion of normality. As a consequence, the distinction between normal and anomalous samples may become less pronounced, reducing anomaly detection performance.

5.2 Lightweight Supervision Experiments

This section evaluates the proposed supervision framework using a limited number of image-level labeled anomalous samples. The conducted experiments investigate its behavior under different conditions and evaluation scenarios, providing insight into its effectiveness, robustness, and generalization capability on the Foreign Object Intrusions and Rolling Inspection Artifacts datasets.

5.2.1 Supervision on the Foreign Object Intrusions Dataset

Tables 5.8 and 5.9 present the results obtained on the Foreign Object Intrusions Dataset. Table 5.8 reports the performance of the original unsupervised models on the evaluation splits, while Table 5.9 summarizes the results obtained with the supervised variant and its augmentation-based extension under different supervision settings.

Table 5.8: Image-level AUROC on the Foreign Object Intrusions Dataset under different evaluation splits.

Method	Initial Split	Swapped Split
EfficientAD-M	**99.74**	**99.58**
FastFlow	80.82	78.94
AST	96.10	92.92
GLASS	99.38	99.36

Overall, the introduction of lightweight supervision improves the performance of all evaluated models. As expected, the largest gains are observed for the methods with the lowest baseline performance, as these are the methods for which the fully unsupervised setting is least effective and therefore stand to benefit the most from the introduction of supervision. FastFlow benefits the most from supervision, improving from 80.82% AUROC in the unsupervised setting to 97.93% using only 29 labeled anomalous samples. When data augmentation is incorporated, performance exceeds 99% AUROC, reaching 99.18%. With 113 labeled anomalous samples, FastFlow achieves

Table 5.9: Image-level AUROC for supervision experiments on the Foreign Object Intrusions Dataset.

Method	**10%**	**20%**	**40%**	**Swapped**
EfficientAD-M + Sup	98.30	99.11	99.86	99.89
EfficientAD-M + Sup + Aug	97.68	99.76	99.87	99.90
FastFlow + Sup	97.93	98.07	98.30	98.94
FastFlow + Sup + Aug	99.18	99.77	99.80	99.82
AST + Sup	**99.91**	99.93	**99.97**	99.75
AST + Sup + Aug	99.90	**99.94**	98.19	**99.94**
GLASS + Sup	99.51	99.90	99.88	99.89
GLASS + Sup + Aug	99.85	99.88	99.93	99.69

10%: 29 samples 20%: 57 samples 40%: 113 samples

98.30% AUROC using supervision alone and 99.80% AUROC when augmentation is incorporated. These results demonstrate that a small amount of supervision can substantially improve the performance of methods that struggle in the fully unsupervised setting, while data augmentation can provide additional benefits.

EfficientAD-M already achieves a very strong unsupervised performance of 99.74% AUROC, leaving limited room for improvement. Nevertheless, supervision enables the model to further improve its results, reaching 99.86% AUROC under the 40% supervision setting and 99.87% AUROC when augmentation is incorporated. However, performance improvements are only observed from the 20% supervision level onward, with data augmentation being required. In lower supervision settings, performance remains below the unsupervised baseline. This behavior suggests that, for highly performant models, a minimum amount of labeled data may be required before supervision becomes beneficial.

In contrast, GLASS benefits from supervision even with only 29 labeled anomalous samples despite already achieving 99.38% AUROC in the unsupervised setting. Performance increases to 99.51% AUROC without augmentation and 99.85% AUROC with augmentation. Under the 40% supervision setting with augmentation, the model achieves 99.93% AUROC. Together with the EfficientAD-M results, these observations indicate that lightweight supervision can remain beneficial even for methods that already achieve very high performance in the fully unsupervised setting, although the amount of labeled data required to obtain improvements may vary across methods.

The effect of augmentation is not uniformly beneficial across all evaluated models. While the augmented FastFlow configuration exhibits a steady improvement as the number of labeled samples increases, this behavior is not observed across all models. For example, AST achieves its best performance under the 40% supervision setting, reaching 99.97% AUROC, whereas the corresponding augmented configuration achieves only 98.19% AUROC. This observation suggests that synthetic anomalous samples do not always provide complementary information and may, in

some cases, reinforce characteristics already present in the labeled data rather than increasing the diversity of the anomaly distribution.

Among all evaluated approaches, AST emerges as the strongest supervised method, consistently achieving the highest performance levels across the evaluated supervision settings. Interestingly, although EfficientAD-M and GLASS achieve the strongest unsupervised performance, AST achieves the best results once supervision is introduced. A possible explanation is that the supervision module operates on the feature representations produced by each anomaly detection model. Consequently, the supervised variants are not trained on a common feature space, and their performance depends not only on the supervision strategy but also on the discriminative quality of the underlying feature representations.

The results obtained on the swapped split are consistent with those observed for the initial split. All evaluated methods benefit from supervision, while only minor performance variations are observed relative to the original partition, indicating that the conclusions are not strongly dependent on a particular train-test split. Some differences are nevertheless observed in the effect of augmentation. In contrast to the initial split, augmentation benefits AST under the 40% supervision setting. This behavior is not observed for GLASS, whose best performance is obtained without augmentation, further reinforcing the observation that the benefits of augmentation are not consistent across all methods and supervision levels. Despite these differences, the overall performance trends remain unchanged, indicating that the evaluated methods generalize consistently across different train-test partitions and that the reported results are not driven by a particular selection of intrusion samples.

5.2.2 Supervision on the Rolling Inspection Artifacts Dataset

Tables 5.10 and 5.11 present the results obtained on the Rolling Inspection Artifacts Dataset. Table 5.10 reports the performance of the original unsupervised models on the evaluation splits, while Table 5.11 summarizes the results obtained with the supervised variants under different supervision settings.

Table 5.10: Image-level AUROC on the Rolling Inspection Artifacts Dataset under different evaluation splits.

Method	Initial Split	Swapped Split
EfficientAD-M	81.83	78.81
FastFlow	**91.61**	**92.50**
AST	88.81	86.87
GLASS	88.60	88.10

Overall, lightweight supervision improves the performance of all evaluated models on the Rolling Inspection Artifacts Dataset. EfficientAD-M benefits the most from supervision, which is consistent with its lower unsupervised performance on this dataset. Its performance improves from 81.83% AUROC in the unsupervised setting to 88.70% AUROC under the 40% supervision

Table 5.11: Image-level AUROC for supervision experiments on the Rolling Inspection Artifacts dataset.

Method	**10%**	**20%**	**40%**	**Swapped**
EfficientAD-M + Sup	82.77	83.97	88.70	89.46
EfficientAD-M + Sup + Aug	83.32	85.41	89.71	87.02
FastFlow + Sup	90.57	92.03	91.52	91.95
FastFlow + Sup + Aug	89.99	90.63	93.38	92.66
AST + Sup	89.97	91.23	92.70	92.75
AST + Sup + Aug	**93.21**	**94.82**	95.61	93.96
GLASS + Sup	90.91	94.39	95.14	95.58
GLASS + Sup + Aug	91.60	92.72	**95.71**	**96.21**

setting and to 89.71% AUROC when augmentation is incorporated. In contrast, FastFlow already achieves the strongest performance in the unsupervised setting (91.61% AUROC), leaving less potential for improvement. Consequently, supervision provides more modest gains, with the best result reaching 93.38% AUROC under the 40% supervision setting with augmentation.

AST and GLASS achieve the strongest supervised results. For the initial split, AST achieves the best performance under the 10% and 20% supervision settings, reaching 93.21% and 94.82% AUROC, respectively. Under the 40% supervision setting, GLASS achieves the highest performance with 95.71% AUROC, improving from 88.60% AUROC in the unsupervised setting, while AST achieves a comparable result of 95.61% AUROC. These results reinforce the observations from the previous dataset, suggesting that the effectiveness of the proposed supervision framework is influenced by the quality of the underlying feature representations. The superior performance achieved by both methods suggests that their learned representations are more informative and therefore better able to exploit the additional information provided by labeled anomalous samples.

As in the experiments conducted on the Foreign Object Intrusions Dataset, the impact of augmentation is not consistently positive across all methods and supervision settings. AST benefits from augmentation at all evaluated supervision levels. In contrast, EfficientAD-M exhibits reduced performance on the swapped split, FastFlow benefits from augmentation only under the 40% supervision setting and on the swapped split, and GLASS does not consistently benefit from augmentation across all supervision levels.

The swapped split produces results that are broadly consistent with those obtained on the initial split, and the overall performance trends remain unchanged. Supervision continues to improve performance across all evaluated methods. These findings indicate that the observed gains are not dependent on a particular train–test partition and further support the robustness and generalization capability of the proposed supervision framework.

5.2.2.1 Category-Level Analysis

Tables 5.12 and 5.13 present the performance of the evaluated methods for each anomaly category of the Rolling Inspection Artifacts Dataset. Table 5.12 reports the results of the original unsu-

pervised models, while Table 5.13 presents the corresponding results for the supervised variants under different supervision settings.

Table 5.12: Image-level AUROC comparison across anomaly categories for the original unsupervised models under the initial and swapped evaluation splits.

	Method	1	2	3	4	5	6	7	Total
Initial	EfficientAD-M	90.33	88.46	100.0	68.25	50.37	77.54	98.78	81.83
	FastFlow	99.67	100.0	99.94	75.71	69.84	97.53	99.92	**91.61**
	AST	99.41	98.21	100.0	57.45	75.96	87.63	99.62	88.81
	GLASS	97.92	98.71	99.97	64.76	75.67	73.93	98.78	88.60
Swapped	EfficientAD-M	88.98	90.47	99.93	49.87	59.30	56.22	95.28	78.81
	FastFlow	97.86	99.90	99.97	76.41	81.25	95.45	93.40	**92.50**
	AST	98.63	97.36	100.0	55.08	72.31	83.36	84.76	86.87
	GLASS	97.08	99.57	99.97	68.98	72.99	65.51	94.56	88.10

1: Roller contamination 1 2: Roller contamination 2 3: Swirl
4: Duplication 5: Discontinuity 6: Cropped 7: Other

Several consistent patterns can be observed across the different supervision levels and evaluation splits. First, *Roller Contamination 1*, *Roller Contamination 2*, and *Swirl* already achieve very high AUROC values in the unsupervised setting for most methods. Consequently, supervision provides only limited improvements for these categories, as their detection performance is already close to saturation.

For *Roller Contamination 1*, all evaluated methods benefit from supervision, with improvements already observable under the 10% supervision setting. The largest gains are obtained by EfficientAD-M and GLASS, which start from lower unsupervised performance levels. In contrast, AST and FastFlow already achieve AUROC values above 99% in the unsupervised setting and therefore exhibit only marginal improvements.

The behavior of *Roller Contamination 2* differs slightly. None of the evaluated methods improves under the 10% supervision setting. At the 20% supervision level, EfficientAD-M exhibits the most noticeable improvement, whereas AST improves only marginally. Increasing the supervision level to 40% provides further improvements only for EfficientAD-M, while the remaining methods do not surpass their unsupervised performance.

In the case of *Swirl*, all methods already achieve AUROC values close to 100% in the unsupervised setting. Consequently, supervision generally maintains performance at similar levels or introduces only minor fluctuations. EfficientAD-M exhibits the largest performance variations across supervision levels and augmentation settings, whereas the remaining methods maintain consistently high performance throughout the evaluated configurations.

Duplication and *Discontinuity* exhibit very similar behavior across all evaluated methods and supervision settings. These categories correspond to some of the most challenging anomaly types in the fully unsupervised setting and consistently yield some of the lowest AUROC values across

Table 5.13: Image-level AUROC comparison across anomaly categories for the supervised variants under different supervision settings.

	Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
10%	EfficientAD-M + Sup	92.98	84.84	94.44	71.15	76.66	48.06	72.73	82.77
	EfficientAD-M + Sup + Aug	93.76	87.04	94.44	75.76	78.88	41.84	59.82	83.32
	FastFlow + Sup	98.74	96.92	99.69	75.75	73.74	90.98	90.03	90.57
	FastFlow + Sup + Aug	99.77	88.14	98.90	78.74	72.35	89.77	91.67	89.99
	AST + Sup	96.78	92.26	100.00	75.96	84.44	65.24	99.92	89.97
	AST + Sup + Aug	99.97	96.16	100.00	77.06	85.37	89.04	99.85	**93.21**
	GLASS + Sup	99.63	92.51	99.97	87.22	80.27	46.86	98.17	90.91
	GLASS + Sup + Aug	99.88	96.89	99.97	88.46	81.76	40.57	98.70	91.60
20%	EfficientAD-M + Sup	92.62	90.09	94.44	67.58	83.45	42.71	83.12	83.97
	EfficientAD-M + Sup + Aug	92.54	88.77	98.37	73.19	83.26	47.19	83.65	85.41
	FastFlow + Sup	98.96	97.86	99.57	80.77	75.11	92.68	94.23	92.03
	FastFlow + Sup + Aug	99.63	93.05	99.73	74.97	72.06	93.32	99.16	90.63
	AST + Sup	97.86	92.92	100.00	81.87	82.70	68.25	99.77	91.23
	AST + Sup + Aug	99.93	98.24	100.00	86.15	85.35	88.50	100.00	**94.82**
	GLASS + Sup	99.98	96.79	99.97	92.92	85.11	68.38	98.09	94.39
	GLASS + Sup + Aug	99.91	89.37	99.97	92.51	80.40	67.78	99.39	92.72
40%	EfficientAD-M + Sup	97.44	95.09	100.00	80.52	79.06	49.80	83.27	88.70
	EfficientAD-M + Sup + Aug	98.41	87.26	100.00	87.83	80.94	52.67	85.10	89.71
	FastFlow + Sup	98.59	93.58	99.64	81.53	74.10	94.12	95.87	91.52
	FastFlow + Sup + Aug	99.70	97.77	99.88	82.17	77.54	98.46	97.48	93.38
	AST + Sup	99.08	91.51	99.97	79.42	83.90	92.91	99.77	92.70
	AST + Sup + Aug	99.90	97.99	100.0	87.75	87.17	93.92	100.0	95.61
	GLASS + Sup	99.88	98.68	99.94	93.99	84.04	78.01	97.71	95.14
	GLASS + Sup + Aug	99.97	98.65	99.97	93.48	84.52	86.03	99.47	**95.71**
Swapped	EfficientAD-M + Sup	99.13	79.28	99.97	87.56	80.72	56.68	91.89	89.46
	EfficientAD-M + Sup + Aug	98.17	81.38	93.75	80.69	71.98	69.39	89.13	87.02
	FastFlow + Sup	98.85	95.97	99.60	84.30	83.63	69.75	87.92	91.95
	FastFlow + Sup + Aug	99.77	95.29	100.00	84.60	80.30	77.21	95.99	92.66
	AST + Sup	98.50	95.32	100.00	87.25	82.06	81.28	88.15	92.75
	AST + Sup + Aug	99.85	92.95	100.00	95.61	81.34	79.48	88.41	93.96
	GLASS + Sup	100.00	99.10	100.00	93.19	88.92	77.41	92.96	95.58
	GLASS + Sup + Aug	99.94	97.83	100.00	93.55	89.96	83.62	97.77	**96.21**

1: Roller contamination 1 2: Roller contamination 2 3: Swirl
4: Duplication 5: Discontinuity 6: Cropped 7: Other

the evaluated models. Consequently, they also benefit the most from the introduction of supervision. For both categories, performance generally improves as the number of labeled anomalous samples increases, reaching the strongest overall results under the 40% supervision setting.

The similar behavior observed for *Duplication* and *Discontinuity* suggests that supervision is particularly effective for anomaly categories characterized by subtle structural deviations from the expected image pattern. Unlike *Roller Contamination 1*, *Roller Contamination 2*, and *Swirl*, these anomalies do not introduce large black regions or severe visual distortions, making them more difficult to detect using purely unsupervised representations.

The behavior of *Cropped* is considerably more complex than that observed for *Duplication* and *Discontinuity*. While supervision ultimately provides substantial benefits for some methods, performance often decreases at lower supervision levels and does not improve consistently across all approaches. This behavior is particularly evident for EfficientAD-M and GLASS. EfficientAD-M decreases from 77.54% AUROC in the unsupervised setting to 48.06% under the 10% supervision setting and remains below its unsupervised performance even at the highest supervision level. Similarly, GLASS decreases from 73.93% to 46.86% AUROC under the 10% supervision setting before progressively recovering and ultimately surpassing its unsupervised performance, reaching 86.03% AUROC under the 40% supervision setting. FastFlow also shows limited benefits from supervision, with noticeable improvements becoming apparent only under the 40% supervision setting, particularly when augmentation is incorporated. In contrast, AST benefits from augmentation under the 10% and 20% supervision settings, while supervision alone improves performance only at the 40% supervision level.

One possible explanation for the poorer performance observed at lower supervision levels is the limited number of labeled samples available for this category. Under the 10% supervision setting, only two *Cropped* samples are included in the training set, which may prevent the supervised head from learning a representative decision boundary for this anomaly type. As additional labeled samples become available, this limitation is progressively reduced, leading to improved detection performance.

The *Other* category exhibits greater variability than most of the remaining anomaly categories. Unlike the remaining categories, *Other* groups multiple heterogeneous anomaly patterns, making it inherently more difficult to model using a limited number of labeled samples. EfficientAD-M is particularly affected, decreasing from 98.78% AUROC in the unsupervised setting to 72.73% under the 10% supervision setting and recovering only partially at higher supervision levels. FastFlow also exhibits noticeable variability across the evaluated supervision settings. In contrast, AST responds particularly well to supervision, with performance improvements already observed under the 10% supervision setting and reaching perfect performance (100% AUROC) from the 20% supervision setting onwards when augmentation is incorporated. GLASS also benefits from supervision, although improvements over the unsupervised baseline become apparent only from the 20% supervision setting onwards. These findings further support the observation that the learned representations of AST and GLASS are better able to exploit labeled anomalous samples, enabling the supervision framework to generalize more effectively across diverse anomaly patterns.

Considering all anomaly categories and supervision settings, augmentation generally contributes to improved performance, although its effectiveness depends on both the anomaly category and the underlying anomaly detection model. Its benefits become more evident at higher supervision levels, where a larger number of labeled anomalous samples enables the augmented data to better capture the variability of each anomaly category. With only a few labeled samples, augmentation may reinforce the characteristics of a limited set of anomalies, resulting in smaller gains or even performance degradations.

5.2.2.2 Single-Category Supervision

Table 5.14 presents the image-level AUROC results obtained when supervision was provided exclusively for the *Roller Contamination 1* anomaly category.

Table 5.14: Image-level AUROC comparison across anomaly categories for supervision experiments using *Roller Contamination 1* as the supervised anomaly class.

	Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
RC1	EfficientAD-M + Sup	98.83	91.35	88.77	77.16	55.64	47.06	62.34	82.06
	EfficientAD-M + Sup + Aug	98.61	79.55	87.97	77.90	55.03	48.06	75.25	81.16
	FastFlow + Sup	99.22	98.27	99.73	80.75	64.13	86.03	87.70	89.86
	FastFlow + Sup + Aug	99.33	90.22	98.99	78.18	46.79	96.46	89.53	86.57
	AST + Sup	99.59	91.76	100.00	78.84	80.48	67.91	99.92	**90.87**
	AST + Sup + Aug	99.99	96.41	100.00	86.25	67.65	63.57	99.85	90.59
	GLASS + Sup	100.00	89.53	100.00	94.30	71.82	47.66	98.47	90.58
	GLASS + Sup + Aug	100.00	90.41	100.00	93.94	74.33	40.37	89.84	90.13

1: Roller contamination 1 2: Roller contamination 2 3: Swirl
4: Duplication 5: Discontinuity 6: Cropped 7: Other

At the overall level, supervision without augmentation improves the performance of EfficientAD-M from 81.83% to 82.06% AUROC, AST from 88.81% to 90.87%, and GLASS from 88.60% to 90.58%. In contrast, FastFlow decreases from 91.61% to 89.86% AUROC. The strongest improvements are obtained by AST and GLASS, further reinforcing the observations from the previous experiments, while EfficientAD-M exhibits only a marginal gain and remains the lowest-performing supervised method. Furthermore, the introduction of augmentation reduces the overall performance of all evaluated methods under this setting. Since both the supervised samples and the generated anomalies originate exclusively from *Roller Contamination 1*, the supervision process may become overly specialized towards characteristics associated with this anomaly category, limiting its ability to generalize across the remaining anomaly types.

Analyzing the results at the category level, performance on *Roller Contamination 1* improves for EfficientAD-M, AST, and GLASS, while FastFlow exhibits a slight reduction. Given that FastFlow already achieves 99.67% AUROC in the unsupervised setting, this decrease likely reflects the limited additional information provided by the available labeled samples.

Performance on *Roller Contamination 2* generally decreases following supervision. The only exception is EfficientAD-M without augmentation, which achieves a small improvement. Notably, EfficientAD-M also exhibits the lowest unsupervised performance on this category, whereas the remaining methods already achieve very strong results. A similar behavior is observed in the previous experiments, where improvements for this anomaly type generally require larger numbers of labeled anomalous samples.

The effect on *Swirl* is limited. AST already achieves perfect detection performance in the unsupervised setting and maintains the same performance after supervision. FastFlow experiences a small reduction, whereas EfficientAD-M shows a more noticeable degradation. Similar behavior is observed in the previous experiments under lower-supervision settings. Notably, GLASS achieves 100% AUROC on the initial split, extending the perfect detection performance previously observed only on the swapped split.

Some of the largest improvements are obtained for the most challenging anomaly categories, which is consistent with the observations from the previous experiments. All evaluated methods improve on *Duplication*, with particularly large gains for AST and GLASS. For *Discontinuity*, improvements are observed for EfficientAD-M and AST, whereas FastFlow and GLASS experience performance reductions. These findings indicate that supervision on a single anomaly category can still improve the detection of previously unseen anomaly types when they share relevant visual characteristics with the supervised class, although the magnitude of this transfer varies across models.

A different behavior is observed for *Cropped*, where performance decreases for all evaluated methods. This result is consistent with the previous experiments, where *Cropped* also proves to be one of the most challenging anomaly categories under limited supervision.

Similarly, performance on the *Other* category generally decreases following supervision. The only exception is AST, which improves slightly from 99.62% to 99.92% AUROC. Since this category contains heterogeneous anomaly patterns, supervision derived from a single anomaly type appears insufficient to capture the full variability of the anomalies grouped within *Other* for the remaining models.

Overall, supervision using a single anomaly category improves the performance of three of the four evaluated methods. Interestingly, the observed improvements are not limited to *Roller Contamination 1*, but also extend to several anomaly categories that were not included in the supervised training set. This suggests that the proposed supervision framework is not solely learning characteristics specific to the supervised anomaly type. Since the supervised head is trained using a large number of normal samples, the observed improvements may result not only from learning the appearance of the supervised anomaly category, but also from a better separation between normal and anomalous samples. Consequently, the benefits of supervision appear to extend beyond the supervised anomaly category, contributing to a more general refinement of the learned representation.

5.2.2.3 Generalization to Unseen Logos

Table 5.15 presents the image-level AUROC results obtained on the unseen logo test set. In the unsupervised setting, substantial differences are observed between the evaluated methods. FastFlow achieves the highest performance (95.76% AUROC), followed closely by AST (95.04% AUROC), indicating a strong ability to generalize to logo variations not observed during training. EfficientAD-M obtains the lowest performance (65.31% AUROC), which is consistent with its comparatively weaker results on the original dataset. In contrast, although GLASS achieved performance comparable to FastFlow and AST on the original dataset, its AUROC drops to 77.44% on the unseen logo test set. These results indicate that EfficientAD-M and GLASS are more strongly affected by the distribution shift introduced by previously unseen logo types, whereas FastFlow and AST exhibit substantially stronger generalization capabilities.

Table 5.15: Image-level AUROC comparison on the unseen logo test set.

Method	**Unsupervised**	**Supervised**
EfficientAD-M	65.31	76.76
FastFlow	**95.76**	93.18
AST	95.04	92.45
GLASS	77.44	**93.42**

The supervised results also reveal distinct behaviors across the evaluated methods. The largest gain is observed for GLASS, whose AUROC increases from 77.44% to 93.42%, allowing it to achieve the highest performance among the supervised methods despite its comparatively weak performance in the unsupervised setting. Nevertheless, its performance remains below that achieved by FastFlow and AST in the unsupervised setting. EfficientAD-M also benefits considerably from supervision, improving from 65.31% to 76.76% AUROC. Despite this gain, EfficientAD-M remains the weakest supervised method, indicating that the available labeled anomalous samples are insufficient to fully compensate for the distribution shift introduced by the unseen logo types. These results are consistent with the observations from the previous experiments, where the largest benefits of supervision are generally obtained by the weakest-performing unsupervised models. The substantial gains achieved by GLASS and EfficientAD-M demonstrate that even a limited amount of supervision can significantly improve robustness to previously unseen visual patterns.

In contrast, FastFlow and AST experience slight performance reductions following supervision, decreasing from 95.76% to 93.18% and from 95.04% to 92.45% AUROC, respectively. Since supervision is derived exclusively from samples containing the training logo types, it may encourage the models to rely more strongly on logo-specific characteristics, slightly reducing their ability to generalize to previously unseen logos.

Chapter 6

Conclusion

This chapter presents the final conclusions of the thesis, summarizing the main findings and contributions of the proposed work and discussing their implications for industrial image acquisition monitoring. The chapter also reflects on the principal limitations of the study and outlines several directions for future research.

6.1 Final Remarks

The image acquisition stage plays a fundamental role in the reliability of industrial computer vision systems, as degradations introduced during image capture can compromise the performance of all subsequent processing stages. Despite its importance, acquisition monitoring remains relatively underexplored. Most developments in industrial computer vision have focused on later stages of the inspection pipeline, often assuming that the acquired images are always suitable for processing. In real-world industrial environments, however, illumination variations, blur, occlusions, lens dirt, and unexpected intrusions can significantly affect image quality and compromise the reliability of downstream vision algorithms.

Traditional acquisition monitoring approaches are typically based on IQA techniques, which evaluate image degradations such as blur, noise, contrast, or illumination changes. While effective for assessing image quality, these methods are not designed to detect acquisition anomalies that may lead to incorrect inspection decisions without necessarily affecting the perceptual quality of the image. Conversely, anomaly detection methods have demonstrated remarkable success in industrial visual inspection, but their application has primarily focused on product defect detection rather than image acquisition monitoring. Consequently, the suitability of modern anomaly detection methods for industrial image acquisition monitoring has received limited attention in the literature.

Motivated by this gap, this thesis investigated the use of anomaly detection techniques for industrial image acquisition monitoring under realistic operating conditions. The main objective was to develop effective mechanisms for the real-time monitoring of industrial image acquisition systems. To achieve this goal, a dedicated benchmark covering multiple acquisition-related

disturbances and industrial scenarios was developed, enabling the systematic evaluation and comparison of representative anomaly detection approaches. Furthermore, strategies for improving the robustness, reliability, and adaptability of anomaly detection systems under evolving industrial conditions were also investigated.

The obtained results demonstrate that anomaly detection constitutes an effective and practical solution for industrial image acquisition monitoring. Acquisition-related disturbances can be detected reliably under realistic industrial conditions, with several evaluated methods achieving image-level AUROC values above 99.5% on the proposed Acquisition-Related Anomalies dataset while satisfying the real-time requirements of the considered industrial applications. Among the evaluated methods, GLASS emerged as the most suitable approach for the considered problem, achieving the best overall performance on the Acquisition-Related Anomalies dataset while also providing the most accurate anomaly localization and maintaining practical computational requirements. Furthermore, GLASS consistently demonstrated strong performance across the additional industrial datasets, highlighting its robustness under diverse operating conditions. Nevertheless, the obtained results also showed that not all methods generalize equally well. While some approaches consistently achieved strong performance across multiple scenarios, others exhibited substantial performance variations, achieving excellent results on certain datasets while performing considerably worse on others.

To further improve the robustness and adaptability of anomaly detection systems under diverse industrial conditions, a lightweight supervision framework was proposed and experimentally validated. The obtained results demonstrated that the incorporation of a small number of image-level labeled anomalous samples can substantially improve performance across different methods and industrial scenarios. Notably, performance improvements were observed across all evaluated methods and datasets. The largest gains were observed for the methods that exhibited the weakest performance in the fully unsupervised setting. As a result, several methods that initially achieved comparatively poor unsupervised performance were able to reach performance levels comparable to, or even exceeding, those obtained by the strongest unsupervised approaches. For example, on the Foreign Object Intrusions dataset, FastFlow improved from 80.82% I-AUROC in the fully unsupervised setting to 99.18% using only 29 labeled anomalous samples and further to 99.80% when additional labeled samples were employed, surpassing the 99.74% I-AUROC achieved by EfficientAD-M in the unsupervised setting.

Furthermore, the incorporation of data augmentation strategies frequently provided additional performance improvements within the proposed supervision framework, although these benefits were not observed consistently across all methods and datasets. In contrast, applying the same augmentations during fully unsupervised training consistently degraded performance across the evaluated methods and datasets. A possible explanation is that the increased variability introduced into the normal training data broadens the learned notion of normality, thereby reducing the separability between normal and anomalous samples.

Overall, the objectives defined for this thesis were successfully achieved. The developed benchmark enabled a systematic evaluation of anomaly detection methods for industrial image

acquisition monitoring and confirmed the suitability of anomaly detection for this application. Furthermore, the proposed lightweight supervision framework demonstrated that high levels of performance and robustness can be achieved across diverse industrial scenarios using only a small number of labeled anomalous samples, enabling multiple anomaly detection methods to become viable solutions for acquisition monitoring applications. Since the proposed approach operates on feature representations extracted from pre-trained anomaly detection models, adaptation can be achieved without retraining the underlying models. This provides a computationally efficient and scalable mechanism for incorporating newly observed anomalous conditions and continuously adapting anomaly detection systems to evolving industrial environments.

6.2 Limitations

Despite the promising results obtained in this thesis, some limitations should be acknowledged.

- Although the proposed benchmark covers a wide range of acquisition-related disturbances, it does not encompass all possible acquisition anomalies that may occur in industrial environments. Examples include camera misalignment, camera vibrations, noise, and motion blur caused by conveyor-speed variations.
- The experimental evaluation was restricted to a selected set of representative anomaly detection methods. While these methods were chosen based on a comprehensive literature review and cover the anomaly detection paradigms considered most suitable for the target application, other existing approaches or future developments may achieve superior performance for industrial image acquisition monitoring.
- The proposed lightweight supervision framework demonstrated significant performance improvements. However, the potential benefits of alternative approaches were not explored. Consequently, further improvements in detection performance, robustness, and generalization may be achievable through different supervision and adaptation strategies.
- Despite including multiple industrial inspection systems and acquisition setups, the proposed benchmark and lightweight supervision framework were evaluated exclusively within the cork industry. Consequently, their generalization to other industrial domains has not been validated and requires further investigation.

6.3 Future Work

Building upon the findings and limitations identified in this thesis, several directions for future research can be explored:

- Investigate the behavior of anomaly detection methods under a broader range of acquisition-related disturbances and industrial operating conditions, providing a deeper understanding of their applicability to image acquisition monitoring.

- Evaluate recently proposed anomaly detection methods and future developments in the field. Given the rapid evolution of anomaly detection research, newly proposed approaches may provide additional insights into their suitability for industrial image acquisition monitoring and potentially outperform the methods evaluated in this work.
- Investigate more advanced supervision and adaptation mechanisms for incorporating labeled anomalous samples into anomaly detection systems.
- Validate the proposed benchmark and lightweight supervision framework across different industrial domains to assess their generalization beyond the cork industry.
- Develop acquisition monitoring systems capable not only of detecting anomalies but also of automatically identifying their root cause. Such systems could distinguish between different acquisition failures, such as illumination problems, blur, lens dirt, or foreign object intrusions, enabling more effective corrective actions in industrial environments.

References

[1] J. Liu et al., "Deep industrial image anomaly detection: A survey," *Machine Intelligence Research*, vol. 21, no. 1, pp. 104–135, 2024, ISSN: 2731-5398. DOI: 10.1007/s11633-023-1459-z.

[2] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, Jun. 2016, pp. 1–6. DOI: 10.1109/QoMEX.2016.7498955.

[3] A. Alzarooni, E. Iqbal, S. U. Khan, S. Javed, B. Moyo, and Y. Abdulrahman, "Anomaly detection for industrial applications: Challenges, solutions, and future directions," *International Journal of Computer Integrated Manufacturing*, pp. 1–26, 2025. DOI: 10.1080/0951192X.2025.2599548.

[4] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2018. DOI: 10.1109/TIP.2017.2760518.

[5] Z. Li, Y. Yan, X. Wang, Y. Ge, and L. Meng, "A survey of deep learning for industrial visual anomaly detection," *Artificial Intelligence Review*, vol. 58, no. 9, 2025, Art. no. 279. DOI: 10.1007/s10462-025-11287-7.

[6] MALA Lab. "Awesome anomaly detection foundation models." GitHub repository, Accessed: Jan. 30, 2026. [Online]. Available: https://github.com/mala-lab/Awesome-Anomaly-Detection-Foundation-Models.

[7] Open Edge Platform. "Anomalib." GitHub repository, Accessed: Jun. 30, 2026. [Online]. Available: https://github.com/open-edge-platform/anomalib.

[8] C. Chen. "Awesome image quality assessment." GitHub repository, Accessed: Jan. 2, 2026. [Online]. Available: https://github.com/chaofengc/Awesome-Image-Quality-Assessment.

[9] C. Li et al., "Image quality assessment: From human to machine preference," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 7570–7581. DOI: 10.1109/CVPR52734.2025.00709.

[10] M. A. Aslam et al., "Qualitynet: A multi-stream fusion framework with spatial and channel attention for blind image quality assessment," *Scientific Reports*, vol. 14, 2024, Art. no. 26039, ISSN: 2045-2322. DOI: 10.1038/s41598-024-77076-4.

[11] J. Søgaard, L. Krasula, M. Shahid, D. Temel, K. Brunnström, and M. Razaak, "Applicability of existing objective metrics of perceptual quality for adaptive video streaming," *Electronic Imaging*, vol. 28, no. 13, pp. 1–7, 2016. DOI: 10.2352/ISSN.2470-1173.2016.13.IQSP-206.

[12] K.-H. Thung and P. Raveendran, "A survey of image quality measures," in *2009 International Conference for Technical Postgraduates (TECHPOS)*, 2009, pp. 1–4. DOI: 10.1109/TECHPOS.2009.5412098.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. DOI: 10.1109/TIP.2003.819861.

[14] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011. DOI: 10.1109/TIP.2011.2109730.

[15] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu, "Deepsim: Deep similarity for image quality assessment," *Neurocomputing*, vol. 257, pp. 104–114, 2017, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2017.01.054.

[16] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual image-error assessment through pairwise preference," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018, pp. 1808–1817. DOI: 10.1109/CVPR.2018.00194.

[17] S. Lao et al., "Attentions help CNNs see better: Attention-based hybrid image quality assessment network," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 1139–1148. DOI: 10.1109/CVPRW56347.2022.00123.

[18] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513–516, 2010. DOI: 10.1109/LSP.2010.2043888.

[19] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013. DOI: 10.1109/LSP.2012.2227726.

[20] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/referenceless image spatial quality evaluator," in *2011 Conference Record of the Forty-Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2011, pp. 723–727. DOI: 10.1109/ACSSC.2011.6190099.

[21] X. Liu, J. van de Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1040–1049. DOI: 10.1109/ICCV.2017.118.

[22] C. Chen et al., "Topiq: A top-down approach from semantics to distortions for image quality assessment," *IEEE Transactions on Image Processing*, vol. 33, pp. 2404–2418, 2024. DOI: 10.1109/TIP.2024.3378466.

[23] S. Yang et al., "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 1190–1199. DOI: 10.1109/CVPRW56347.2022.00126.

[24] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full-reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006. DOI: 10.1109/TIP.2006.881959.

[25] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008: A database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, pp. 30–45, 2009. [Online]. Available: https://www.researchgate.net/publication/221678002_TID2008_-_A_Database_for_Evaluation_of_Full-Reference_Visual_Quality_Assessment_Metrics.

[26] N. Ponomarenko et al., "Color image database TID2013: Peculiarities and preliminary results," in *2013 Fourth European Workshop on Visual Information Processing (EUVIP)*, 2013, pp. 106–111. [Online]. Available: https://ieeexplore.ieee.org/document/6623960.

[27] D. Zhou, W. Dong, and W. Chen, "Joint demosaicking and zooming using moderate spectral correlation and consistent edge map," *Journal of Electronic Imaging*, vol. 23, no. 4, 2014, Art. no. 043010. DOI: 10.1117/1.JEI.23.4.043010.

[28] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016, ISSN: 1941-0042. DOI: 10.1109/tip.2015.2500021.

[29] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020. DOI: 10.1109/TIP.2020.2967829.

[30] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. C. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 3572–3582. DOI: 10.1109/CVPR42600.2020.00363.

[31] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3674–3683. DOI: 10.1109/CVPR42600.2020.00373.

[32] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization," *International Journal of Computer Vision*, vol. 130, pp. 947–969, 2022. DOI: 10.1007/s11263-022-01578-9.

[33] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9584–9592. DOI: 10.1109/CVPR.2019.00982.

[34] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti, "Vt-adl: A vision transformer network for image anomaly detection and localization," in *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, 2021, pp. 1–6. DOI: 10.1109/isie45552.2021.9576231.

[35] P. Bergmann, X. Jin, D. Sattlegger, and C. Steger, "The MVTec 3D-AD dataset for unsupervised 3d anomaly detection and localization," in *17th International Conference on Computer Vision Theory and Applications (VISAPP)*, 2022, pp. 202–213. DOI: 10.5220/0010865000003124.

[36] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "SPot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *Computer Vision – ECCV 2022*, 2022, pp. 392–408, ISBN: 978-3-031-20056-4. DOI: 10.1007/978-3-031-20056-4_23.

[37] L. Bonfiglioli, M. Toschi, D. Silvestri, N. Fioraio, and D. De Gregorio, "The eyecandies dataset for unsupervised multimodal anomaly detection and localization," in *Computer Vision – ACCV 2022*, 2023, pp. 459–475, ISBN: 978-3-031-26348-4. DOI: 10.1007/978-3-031-26348-4_27.

[38] C. Wang et al., "Real-IAD: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 22 883–22 892. DOI: 10.1109/CVPR52733.2024.02159.

[39] Dataset Ninja. "BTAD dataset," Accessed: Jan. 31, 2026. [Online]. Available: https://datasetninja.com/btad.

[40] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91.

[41] H. Yuan, Y. He, X. Zheng, C. Li, and A. Wu, "Surface defect detection based on improved YOLOv3-Tiny algorithm," in *2022 41st Chinese Control Conference (CCC)*, 2022, pp. 5769–5774. DOI: 10.23919/CCC55666.2022.9902500.

[42] C. Zhou et al., "Metal surface defect detection based on improved YOLOv5," *Scientific Reports*, vol. 13, no. 1, 2023, Art. no. 20803, ISSN: 2045-2322. DOI: 10.1038/s41598-023-47716-2.

[43] X. Song, S. Cao, J. Zhang, and Z. Hou, "Steel surface defect detection algorithm based on YOLOv8," *Electronics*, vol. 13, no. 5, 2024, Art. no. 988, ISSN: 2079-9292. DOI: 10.3390/electronics13050988.

[44] G. Pang, C. Shen, and A. van den Hengel, "Deep anomaly detection with deviation networks," in *25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2019, pp. 353–362. DOI: 10.1145/3292500.3330871.

[45] L. Ruff et al., "Deep semi-supervised anomaly detection," in *International Conference on Learning Representations (ICLR)*, 2020. [Online]. Available: https://openreview.net/forum?id=HkgH0TEYwH.

[46] C. Ding, G. Pang, and C. Shen, "Catching both gray and black swans: Open-set supervised anomaly detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 7378–7388. DOI: 10.1109/CVPR52688.2022.00724.

[47] H. Zhang, Z. Wu, Z. Wang, Z. Chen, and Y.-G. Jiang, "Prototypical residual networks for anomaly detection and localization," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16 281–16 291. DOI: 10.1109/CVPR52729.2023.01562.

[48] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2021, pp. 9659–9669. DOI: 10.1109/CVPR46437.2021.00954.

[49] N.-C. Ristea et al., "Self-supervised predictive convolutional attentive block for anomaly detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 566–13 576. DOI: 10.1109/CVPR52688.2022.01321.

[50] N. Madan et al., "Self-supervised masked convolutional transformer block for anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 1, pp. 525–542, 2024. DOI: 10.1109/TPAMI.2023.3322604.

[51] X. Zhang, M. Xu, and X. Zhou, "RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 16 699–16 708. DOI: 10.1109/CVPR52733.2024.01580.

[52] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4182–4191. DOI: 10.1109/CVPR42600.2020.00424.

[53] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Asymmetric student-teacher networks for industrial anomaly detection," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 2591–2601. DOI: 10.1109/WACV56688.2023.00262.

[54] K. Batzner, L. Heckler, and R. König, "EfficientAD: Accurate visual anomaly detection at millisecond-level latencies," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 127–137. DOI: 10.1109/WACV57701.2024.00020.

[55] G. Wang, S. Han, E. Ding, and D. Huang, "Student-teacher feature pyramid matching for anomaly detection," in *British Machine Vision Conference (BMVC)*, 2021. DOI: 10.5244/C.35.349.

[56] Y. Cao, Q. Wan, W. Shen, and L. Gao, "Informative knowledge distillation for image anomaly segmentation," *Knowledge-Based Systems*, vol. 248, 2022, Art. no. 108846, ISSN: 0950-7051. DOI: 10.1016/j.knosys.2022.108846.

[57] H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9727–9736. DOI: 10.1109/CVPR52688.2022.00951.

[58] J. Yi and S. Yoon, "Patch SVDD: Patch-level SVDD for anomaly detection and segmentation," in *Computer Vision – ACCV 2020*, 2021, pp. 375–390. DOI: 10.1007/978-3-030-69544-6_23.

[59] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "Simplenet: A simple network for image anomaly detection and localization," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2023, pp. 20 402–20 411. DOI: 10.1109/CVPR52729.2023.01954.

[60] B. Rolih, M. Fučka, and D. Skočaj, "SuperSimpleNet: Unifying unsupervised and supervised learning for fast and reliable surface defect detection," in *Pattern Recognition – ICPR 2024*, 2025, pp. 47–65. DOI: 10.1007/978-3-031-78192-6_4.

[61] Q. Chen, H. Luo, C. Lv, and Z. Zhang, "A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization," in *Computer Vision – ECCV 2024*, Springer Nature Switzerland, 2025, pp. 37–54. DOI: 10.1007/978-3-031-72855-6_3.

[62] D. Gudovskiy, S. Ishizaka, and K. Kozuka, "CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 1819–1828. DOI: 10.1109/WACV51458.2022.00188.

[63] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Fully convolutional cross-scale flows for image-based defect detection," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 1829–1838. DOI: 10.1109/WACV51458.2022.00189.

[64] J. Yu et al., *FastFlow: Unsupervised anomaly detection and localization via 2D normalizing flows*, 2021. arXiv: 2111.07677 [cs.CV].

[65] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A patch distribution modeling framework for anomaly detection and localization," in *Pattern Recognition. ICPR International Workshops and Challenges*, 2021, pp. 475–489. DOI: 10.1007/978-3-030-68799-1_35.

[66] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2022, pp. 14 298–14 308. DOI: 10.1109/CVPR52688.2022.01392.

[67] M. Yang, P. Wu, and H. Feng, "MemSeg: A semi-supervised method for image surface defect detection using differences and commonalities," *Engineering Applications of Artificial Intelligence*, vol. 119, 2023, Art. no. 105835, ISSN: 0952-1976. DOI: 10.1016/j.engappai.2023.105835.

[68] S. Lee, S. Lee, and B. C. Song, "CFA: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization," *IEEE Access*, vol. 10, pp. 78 446–78 454, 2022. DOI: 10.1109/ACCESS.2022.3193699.

[69] J. Bae, J.-H. Lee, and S. Kim, "PNI: Industrial anomaly detection using position and neighborhood information," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 6350–6360. DOI: 10.1109/ICCV51070.2023.00586.

[70] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "WinCLIP: Zero-/few-shot anomaly classification and segmentation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 606–19 616. DOI: 10.1109/CVPR52729.2023.01878.

[71] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. [Online]. Available: https://openreview.net/forum?id=buC4E91xZE.

[72] X. Li et al., "PromptAD: Learning prompts with only normal samples for few-shot anomaly detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 16 848–16 858. DOI: 10.1109/CVPR52733.2024.01594.

[73] V. Zavrtanik, M. Kristan, and D. Skočaj, "DRÆM: A discriminatively trained reconstruction embedding for surface anomaly detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 8310–8319. DOI: 10.1109/ICCV48922.2021.00822.

[74] V. Zavrtanik, M. Kristan, and D. Skočaj, "DSR – a dual subspace re-projection network for surface anomaly detection," in *Computer Vision – ECCV 2022*, 2022, pp. 539–554. DOI: 10.1007/978-3-031-19821-2_31.

[75] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognition*, vol. 112, 2021, Art. no. 107706, ISSN: 0031-3203. DOI: 10.1016/j.patcog.2020.107706.

[76] Packt Publishing. "Inside the generative adversarial networks (GAN) architecture." Medium post, Accessed: Jan. 24, 2026. [Online]. Available: https://medium.com/@Packt_Pub/inside-the-generative-adversarial-networks-gan-architecture-2435afbd6b3b.

[77] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Information Processing in Medical Imaging*, 2017, pp. 146–157. DOI: 10.1007/978-3-319-59050-9_12.

[78] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," *Medical Image Analysis*, vol. 54, pp. 30–44, 2019, ISSN: 1361-8415. DOI: 10.1016/j.media.2019.01.010.

[79] Y. Liang, J. Zhang, S. Zhao, R. Wu, Y. Liu, and S. Pan, "Omni-frequency channel-selection representations for unsupervised anomaly detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 4327–4340, 2023. DOI: 10.1109/TIP.2023.3293772.

[80] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in *Computer Vision – ACCV 2018*, Springer International Publishing, 2019, pp. 622–637. DOI: 10.1007/978-3-030-20893-6_39.

[81] AIToolHouse. "ViTs vs. CNNs in AI image processing," Accessed: Sep. 9, 2025. [Online]. Available: https://blog.aitoolhouse.com/vits-vs-cnns-in-ai-image-processing/.

[82] J. Guo, S. Lu, W. Zhang, F. Chen, H. Li, and H. Liao, "Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 20 405–20 415. DOI: 10.1109/CVPR52734.2025.01900.

[83] Y. Lee and P. Kang, "AnoViT: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder," *IEEE Access*, vol. 10, pp. 46 717–46 724, 2022. DOI: 10.1109/ACCESS.2022.3171559.

[84] J. Pirnay and K. Chai, "Inpainting transformer for anomaly detection," in *Image Analysis and Processing – ICIAP 2022*, 2022, pp. 394–406. DOI: 10.1007/978-3-031-06430-2_33.

[85] Steins. "Diffusion model clearly explained." Medium post, Accessed: Jan. 24, 2026. [Online]. Available: https://medium.com/@steinsfu/diffusion-model-clearly-explained-cd331bd41166.

[86] H. Zhang, Z. Wang, D. Zeng, Z. Wu, and Y.-G. Jiang, "DiffusionAD: Norm-guided one-step denoising diffusion for anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 8, pp. 7140–7152, 2025. DOI: 10.1109/TPAMI.2025.3570494.

[87] A. Mousakhan, T. Brox, and J. Tayyub, "Anomaly detection with conditioned denoising diffusion models," in *Pattern Recognition*, 2025, pp. 181–195. DOI: 10.1007/978-3-031-85181-0_12.

[88] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks, "AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 649–655. DOI: 10.1109/CVPRW56347.2022.00080.

[89] H. He et al., "A diffusion-based framework for multi-class anomaly detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, pp. 8472–8480, 2024. DOI: 10.1609/aaai.v38i8.28690.

[90] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang, "AnomalyGPT: Detecting industrial anomalies using large vision-language models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, pp. 1932–1940, 2024. DOI: 10.1609/aaai.v38i3.27963.

[91] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang. "AnomalyGPT," Accessed: Jan. 24, 2026. [Online]. Available: https://anomalygpt.github.io/.

[92] S. Akcay, D. Ameln, A. Vaidya, B. Lakshmanan, N. Ahuja, and U. Genc, "Anomalib: A deep learning library for anomaly detection," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 1706–1710. DOI: 10.1109/ICIP46576.2022.9897283.

[93] G. Pang, C. Ding, C. Shen, and A. van den Hengel. "Explainable deep few-shot anomaly detection with deviation networks." Official GitHub repository, Accessed: Jun. 30, 2026. [Online]. Available: https://github.com/mala-lab/deviation-network-image.

[94] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt. "AST." Official GitHub repository, Accessed: Jun. 30, 2026. [Online]. Available: https://github.com/marco-rudolph/AST.

[95] Q. Chen, H. Luo, C. Lv, and Z. Zhang. "GLASS." Official GitHub repository, Accessed: Jun. 30, 2026. [Online]. Available: https://github.com/cqylunlun/GLASS.

[96] A. Mousakhan, T. Brox, and J. Tayyub. "DDAD." Official GitHub repository, Accessed: Jun. 30, 2026. [Online]. Available: https://github.com/arimousa/DDAD.

[97] H. Zhang, Z. Wang, D. Zeng, Z. Wu, and Y.-G. Jiang. "DiffusionAD." Official GitHub repository, Accessed: Jun. 30, 2026. [Online]. Available: https://github.com/HuiZhang0812/DiffusionAD.

Appendix A

Chronological Summary of the considered Literature

This appendix provides a structured summary of the literature considered throughout this thesis. The reviewed works are organized in chronological order of publication to reflect the progression of research relevant to the problem addressed. For clarity and readability, the literature is divided into three tables according to the nature and role of the referenced works.

Table A.1 summarizes literature related to IQA, including FR-IQA and NR-IQA methods, as well as datasets commonly used in image quality assessment.

Table A.2 presents the anomaly detection–related literature considered in this thesis. This table includes both methods and datasets, covering a broad range of anomaly detection paradigms, such as supervised and self-supervised approaches, teacher-student frameworks, memory-based methods, and transformer-based models.

Finally, Table A.3 lists additional reference works that support the thesis but do not fall directly into the categories of image quality assessment or anomaly detection methods and datasets. These include survey papers, publications describing evaluation metrics, and studies used for conceptual explanations or illustrative purposes.

Across all tables, each entry reports the year of publication, the paper title, and additional information relevant to its role in the thesis, such as the methodological category, the commonly used name of the method or dataset, or a brief description of its purpose.

Table A.1: Chronological summary of IQA method and dataset papers considered in this thesis.

Year	Paper Title	Category	Name
2004	Image quality assessment: from error visibility to structural similarity [13]	FR-IQA method	SSIM
2006	A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms [24]	IQA dataset	LIVE IQA

Continued on next page

Table A.1 – continued from previous page

Year	Paper Title	Category	Name
2009	TID2008: A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics [25]	IQA dataset	TID2008
2010	A Two-Step Framework for Constructing Blind Image Quality Indices [18]	NR-IQA method	BIQI
2011	FSIM: A Feature Similarity Index for Image Quality Assessment [14]	FR-IQA method	FSIM
2011	Blind/Referenceless Image Spatial Quality Evaluator [20]	NR-IQA method	BRISQUE
2012	Making a "Completely Blind" Image Quality Analyzer [19]	NR-IQA method	NIQE
2013	Color Image Database TID2013: Peculiarities and Preliminary Results [26]	IQA dataset	TID2013
2016	Massive Online Crowdsourced Study of Subjective and Objective Picture Quality [28]	IQA dataset	CLIVE
2017	RankIQA: Learning from Rankings for No-reference Image Quality Assessment [21]	NR-IQA method	RankIQA
2017	DeepSim: Deep similarity for image quality assessment [15]	FR-IQA method	DeepSim
2018	PieAPP: Perceptual Image-Error Assessment through Pairwise Preference [16]	FR-IQA method	PieAPP
2019	From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality [30]	IQA dataset	PaQ-2-PiQ
2020	KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment [29]	IQA dataset	KonIQ-10k
2020	Perceptual Quality Assessment of Smartphone Photography [31]	IQA dataset	SPAQ
2022	MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment [23]	NR-IQA method	MANIQA
2022	Attentions Help CNNs See Better: Attention-based Hybrid Image Quality Assessment Network [17]	FR-IQA method	AHIQ
2024	TOPIQ: A Top-Down Approach From Semantics to Distortions for Image Quality Assessment [22]	NR-IQA method	TOPIQ

Table A.2: Chronological summary of AD method and dataset papers considered in this thesis.

Year	**Paper Title**	**Category**	**Name**
2016	You Only Look Once: Unified, Real-Time Object Detection [40]	Fully supervised AD	YOLO
2017	Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery [77]	GANs	AnoGAN
2018	GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training [80]	GANs	GANomaly
2019	f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks [78]	GANs	f-AnoGAN
2019	Deep Semi-Supervised Anomaly Detection [45]	Supervised AD	Deep SAD
2019	MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection [33]	AD Dataset	MVTec AD
2019	Deep Anomaly Detection with Deviation Network [44]	Supervised AD	DevNet
2020	Patch SVDD: Patch-level SVDD for Anomaly Detection and Segmentation [58]	One-Class Classification	Patch SVDD
2020	Uninformed Students: Student-Teacher Anomaly Detection With Discriminative Latent Embeddings [52]	Teacher-Student	US
2020	PaDiM: a Patch Distribution Modeling Framework for Anomaly Detection and Localization [65]	Memory bank	PaDiM
2021	Student-Teacher Feature Pyramid Matching for Anomaly Detection [55]	Teacher-Student	STFPM
2021	CutPaste: Self-Supervised Learning for Anomaly Detection and Localization [48]	Self-supervised AD	CutPaste
2021	VT-ADL: A Vision Transformer Network for Image Anomaly Detection and Localization [34]	AD Dataset	BTAD
2021	Reconstruction by inpainting for visual anomaly detection [75]	Autoencoders	RIAD
2021	Inpainting Transformer for Anomaly Detection [84]	Transformers	InTra
2021	Towards Total Recall in Industrial Anomaly Detection [66]	Memory bank	PatchCore

Continued on next page

Table A.2 – continued from previous page

Year	Paper Title	Category	Name
2021	CFLOW-AD: Real-Time Unsupervised Anomaly Detection with Localization via Conditional Normalizing Flows [62]	Distribution-maps	CFLOW-AD
2021	DRAEM - A discriminatively trained reconstruction embedding for surface anomaly detection [73]	Autoencoders	DRAEM
2021	Fully Convolutional Cross-Scale-Flows for Image-based Defect Detection [63]	Distribution-maps	CS-Flow
2021	FastFlow: Unsupervised Anomaly Detection and Localization via 2D Normalizing Flows [64]	Distribution-maps	FastFlow
2021	Self-Supervised Predictive Convolutional Attentive Block for Anomaly Detection [49]	Self-supervised AD	SSPCAB
2021	The MVTec 3D-AD Dataset for Unsupervised 3D Anomaly Detection and Localization [35]	AD Dataset	MVTec 3D-AD
2022	Beyond Dents and Scratches: Logical Constraints in Unsupervised Anomaly Detection and Localization [32]	AD Dataset	MVTec LOCO AD
2022	Anomaly Detection via Reverse Distillation from One-Class Embedding [57]	Teacher-Student	RD
2022	Omni-frequency Channel-selection Representations for Unsupervised Anomaly Detection [79]	GANs	OCR-GAN
2022	AnoViT: Unsupervised Anomaly Detection and Localization With Vision Transformer-Based Encoder-Decoder [83]	Transformers	AnoViT
2022	MemSeg: A semi-supervised method for image surface defect detection using differences and commonalities [67]	Memory bank	MemSeg
2022	CFA: Coupled-hypersphere-based Feature Adaptation for Target-Oriented Anomaly Localization [68]	Memory bank	CFA
2022	AnoDDPM: Anomaly Detection with Denoising Diffusion Probabilistic Models using Simplex Noise [88]	Diffusion model	AnoDDPM
2022	Catching Both Gray and Black Swans: Open-set Supervised Anomaly Detection [46]	Supervised AD	DRA

Continued on next page

Table A.2 – continued from previous page

Year	Paper Title	Category	Name
2022	Surface Defect Detection based on Improved YOLOv3-Tiny Algorithm [41]	Fully supervised AD	Improved YOLOv3-Tiny
2022	Informative knowledge distillation for image anomaly segmentation [56]	Teacher-Student	IKD
2022	SPot-the-Difference Self-Supervised Pre-training for Anomaly Detection and Segmentation [36]	AD Dataset; Self-supervised AD	VisA; SPot-the-Difference
2022	DSR - A dual subspace re-projection network for surface anomaly detection [74]	Autoencoders	DSR
2022	Self-Supervised Masked Convolutional Transformer Block for Anomaly Detection [50]	Self-supervised AD	SSMCTB
2022	The Eyecandies Dataset for Unsupervised Multimodal Anomaly Detection and Localization [37]	AD Dataset	Eyecandies
2022	Asymmetric Student-Teacher Networks for Industrial Anomaly Detection [53]	Teacher-Student	AST
2022	PNI : Industrial Anomaly Detection using Position and Neighborhood Information [69]	Memory bank	PNI
2022	Prototypical Residual Networks for Anomaly Detection and Localization [47]	Supervised AD	PRN
2023	DiffusionAD: Norm-guided One-step Denoising Diffusion for Anomaly Detection [86]	Diffusion models	DiffusionAD
2023	WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation [70]	Contrastive language-image pre-training	WinCLIP
2023	SimpleNet: A Simple Network for Image Anomaly Detection and Localization [59]	One-Class Classification	SimpleNet
2023	Anomaly Detection with Conditioned Denoising Diffusion Models [87]	Diffusion models	DDAD
2023	AnomalyGPT: Detecting Industrial Anomalies Using Large Vision-Language Models [90]	GPTs	AnomalyGPT
2023	AnomalyCLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection [71]	Contrastive language-image pre-training	AnomalyCLIP
2023	Metal surface defect detection based on improved YOLOv5 [42]	Fully supervised AD	Improved YOLOv5

Continued on next page

Table A.2 – continued from previous page

Year	Paper Title	Category	Name
2023	A Diffusion-based Framework for Multi-class Anomaly Detection [89]	Diffusion model	DiAD
2024	Steel Surface Defect Detection Algorithm Based on YOLOv8 [43]	Fully supervised AD	Improved YOLOv8
2024	EfficientAD: Accurate Visual Anomaly Detection at Millisecond-Level Latencies [54]	Teacher-Student	EfficientAD
2024	RealNet: A Feature Selection Network with Realistic Synthetic Anomaly for Anomaly Detection [51]	Self-supervised AD	RealNet
2024	Real-IAD: A Real-World Multi-View Dataset for Benchmarking Versatile Industrial Anomaly Detection [38]	AD Dataset	Real-IAD
2024	PromptAD: Learning Prompts with only Normal Samples for Few-Shot Anomaly Detection [72]	Contrastive language-image pre-training	PromptAD
2024	Dinomaly: The Less Is More Philosophy in Multi-Class Unsupervised Anomaly Detection [82]	Transformers	Dinomaly
2024	A Unified Anomaly Synthesis Strategy with Gradient Ascent for Industrial Anomaly Detection and Localization [61]	One-Class Classification	GLASS
2024	SuperSimpleNet: Unifying Unsupervised and Supervised Learning for Fast and Reliable Surface Defect Detection [60]	One-Class Classification	SuperSimpleNet

Table A.3: Chronological summary of other papers considered in this thesis.

Year	Paper Title	Relevance
2009	A survey of image quality measures [12]	Presents error-based metrics used for IQA.
2016	Understanding How Image Quality Affects Deep Neural Networks [2]	Examines the impact of image quality on deep neural network performance.
2016	Applicability of Existing Objective Metrics of Perceptual Quality for Adaptive Video Streaming [11]	Presents error-based metrics used for IQA.

Continued on next page

Table A.3 – continued from previous page

Year	**Paper Title**	**Relevance**
2017	Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment [4]	Introduces the fundamentals of IQA and its main paradigms.
2022	Anomalib: A Deep Learning Library for Anomaly Detection [92]	Introduces anomalib, a library for unsupervised anomaly detection and localization in industrial vision tasks.
2024	Deep Industrial Image Anomaly Detection: A Survey [1]	Reviews deep learning-based methods for industrial image AD.
2024	QualityNet: A multi-stream fusion framework with spatial and channel attention for blind image quality assessment [10]	Presents a figure illustrating IQA paradigms, adapted for use in this thesis.
2025	Anomaly Detection for industrial applications: challenges, solutions, and future directions [3]	Highlights the importance of anomaly detection in industrial and manufacturing environments.
2025	Image Quality Assessment: From Human to Machine Preference [9]	Provides example images of multiple image corruption types.
2025	A survey of deep learning for industrial visual anomaly detection [5]	Reviews deep learning-based methods for industrial image AD.

Appendix B

Image Corruption Examples

Figure B.1 [9] presents example images for 30 types of image corruptions. For each corruption type, the severity increases from left to right, illustrating progressively stronger degradation levels. These examples illustrate the diversity of image degradation types observed in corrupted visual data.

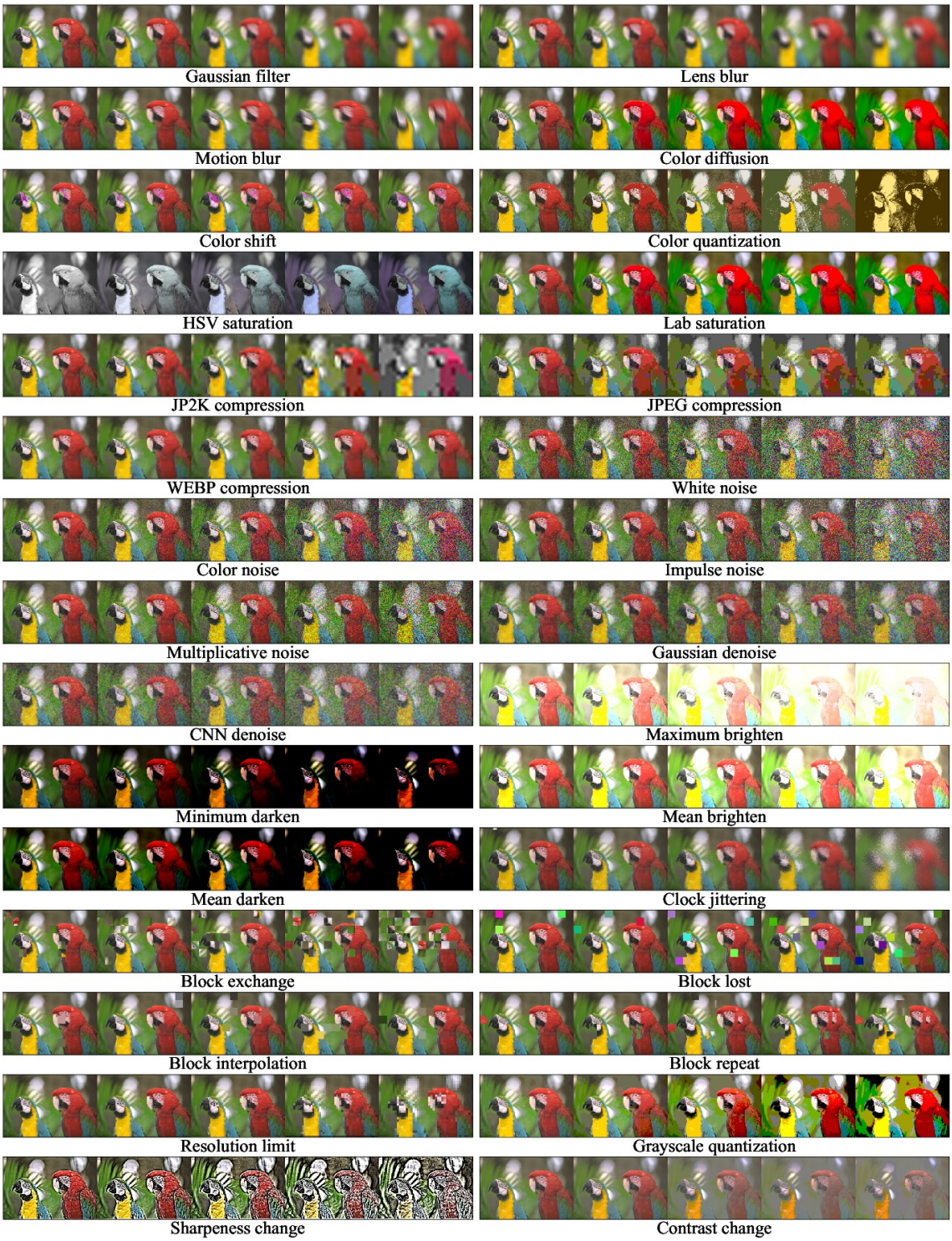


Figure B.1: Example images illustrating 30 types of image corruptions. For each corruption type, severity increases from left (Level 1) to right (Level 5). Zoom in for detail (from [9]).

Appendix C

Supplementary Image-Level Performance Analysis

This appendix complements the quantitative evaluation presented in Section 5.1 by providing additional analyses of the image-level performance of the evaluated anomaly detection models. While Section 5.1 focuses on image-level AUROC values, the visualizations presented here provide further insight into the discriminative behavior of each method.

For each evaluated model, this appendix presents image-level Receiver Operating Characteristic (ROC) curves for the Acquisition-Related Anomalies Dataset, including one curve corresponding to the overall performance and seven corresponding to the individual acquisition-related anomaly categories. Since all evaluated models were assessed on this dataset, these visualizations are provided for every method.

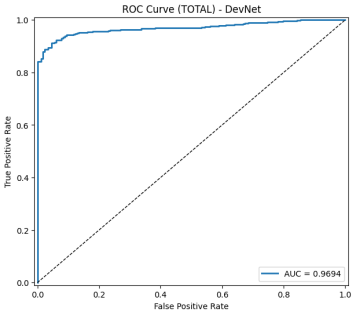
Additional image-level ROC curves are also presented for the Foreign Object Intrusions and Rolling Inspection Artifacts datasets. As shown in Chapter 5, EfficientAD-M, EfficientAD-M without Autoencoder, FastFlow, AST, and GLASS achieved the strongest image-level performance on the Acquisition-Related Anomalies Dataset while meeting the real-time requirements defined for the target industrial application. These methods were therefore further evaluated on the Foreign Object Intrusions and Rolling Inspection Artifacts datasets. Consequently, ROC curves for these datasets are provided only for these methods. Unlike the Acquisition-Related Anomalies Dataset, only the overall image-level ROC curve is presented for each model.

In addition, image-level anomaly score distributions are presented using score histograms together with Kernel Density Estimation (KDE)-based density estimates. While histograms summarize the frequency of anomaly scores, KDE provides a smooth, non-parametric estimate of the underlying score distribution, making the separation and overlap between normal and anomalous samples easier to visualize. For DevNet and AST, the histograms are displayed using a logarithmic scale for visualization purposes only, as the anomaly scores produced by these methods are not normalized. Unlike most of the evaluated methods, whose anomaly scores are normalized to the range [0,1], DevNet and AST produce anomaly scores spanning a much wider range of values. The logarithmic scale improves the visualization of the lower-score regions without modifying the

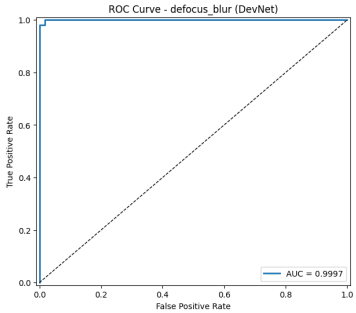
underlying anomaly scores or affecting the quantitative evaluation.

Taken together, the ROC curves, score histograms, and KDE-based density estimates complement the quantitative AUROC results by providing an even more comprehensive characterization of the image-level behavior of the evaluated models. Beyond the overall image-level detection performance summarized by the AUROC, these visualizations provide additional insight into the distribution of anomaly scores, class separability, false-positive and false-negative behavior, and the trade-offs associated with decision threshold selection.

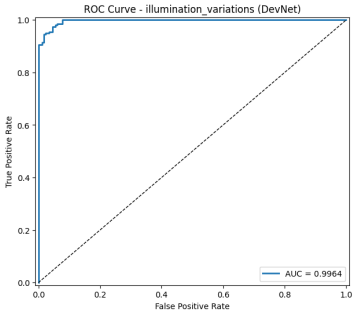
C.1 DevNet



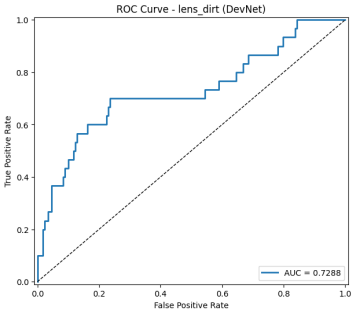
(a) Total (96.94 I-AUROC)



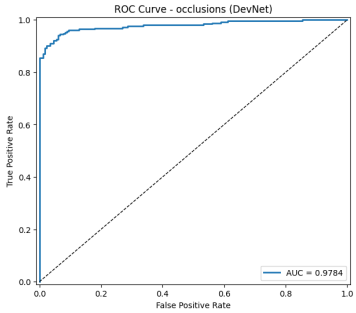
(b) Defocus Blur (99.97 I-AUROC)



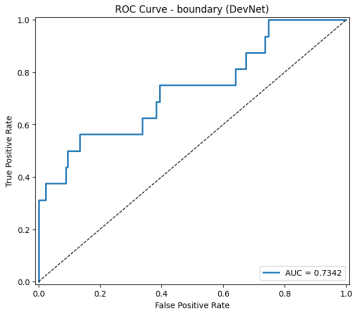
(c) Illumination variations (99.64 I-AUROC)



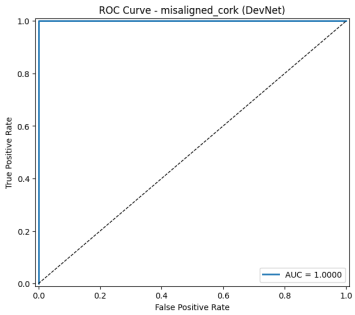
(d) Lens dirt (72.88 I-AUROC)



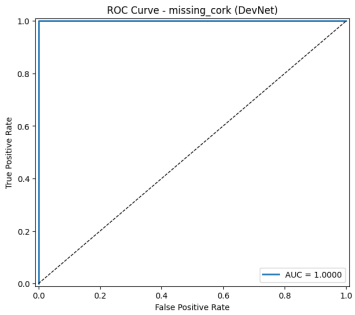
(e) Occlusions (97.84 I-AUROC)



(f) Boundary (73.42 I-AUROC)



(g) Misaligned cork (100.00 I-AUROC)



(h) Missing cork (100.00 I-AUROC)

Figure C.1: Image-level ROC curves obtained with DevNet on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category.

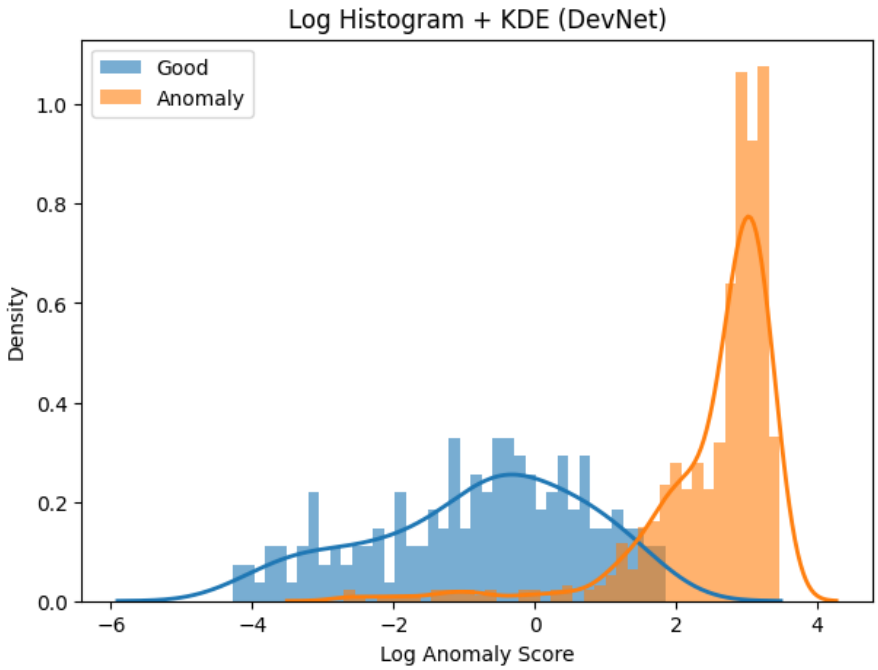
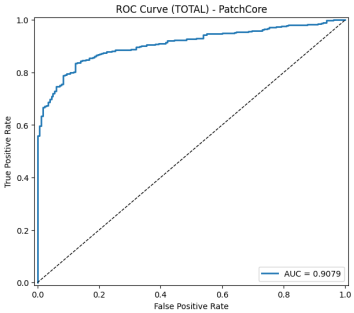
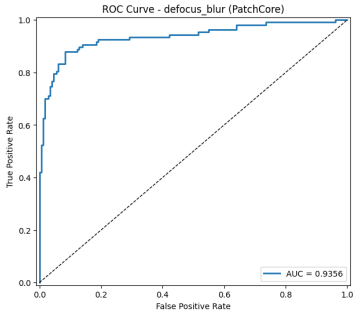


Figure C.2: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with DevNet on the Acquisition-Related Anomalies Dataset. A logarithmic scale is used for visualization purposes.

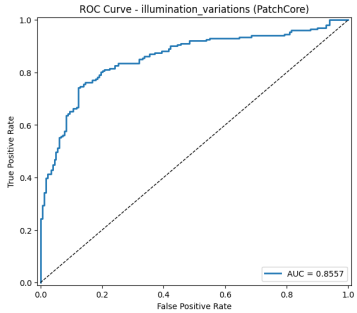
C.2 PatchCore



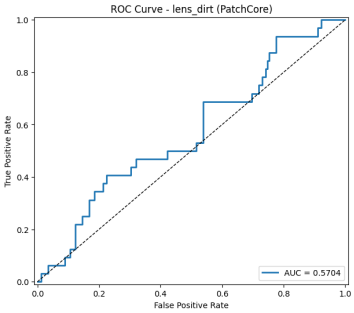
(a) Total (90.79 I-AUROC)



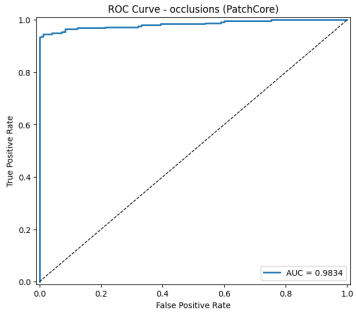
(b) Defocus Blur (93.56 I-AUROC)



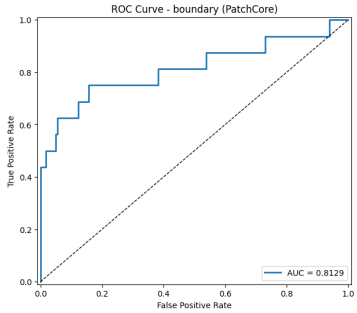
(c) Illumination variations (85.57 I-AUROC)



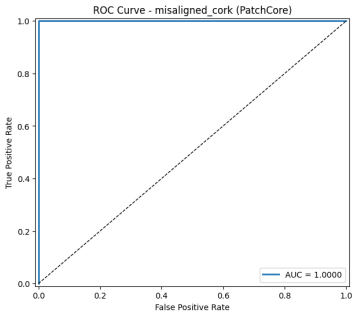
(d) Lens dirt (57.04 I-AUROC)



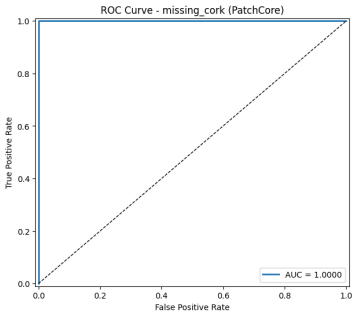
(e) Occlusions (98.34 I-AUROC)



(f) Boundary (81.29 I-AUROC)



(g) Misaligned cork (100.00 I-AUROC)



(h) Missing cork (100.00 I-AUROC)

Figure C.3: Image-level ROC curves obtained with PatchCore on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category.

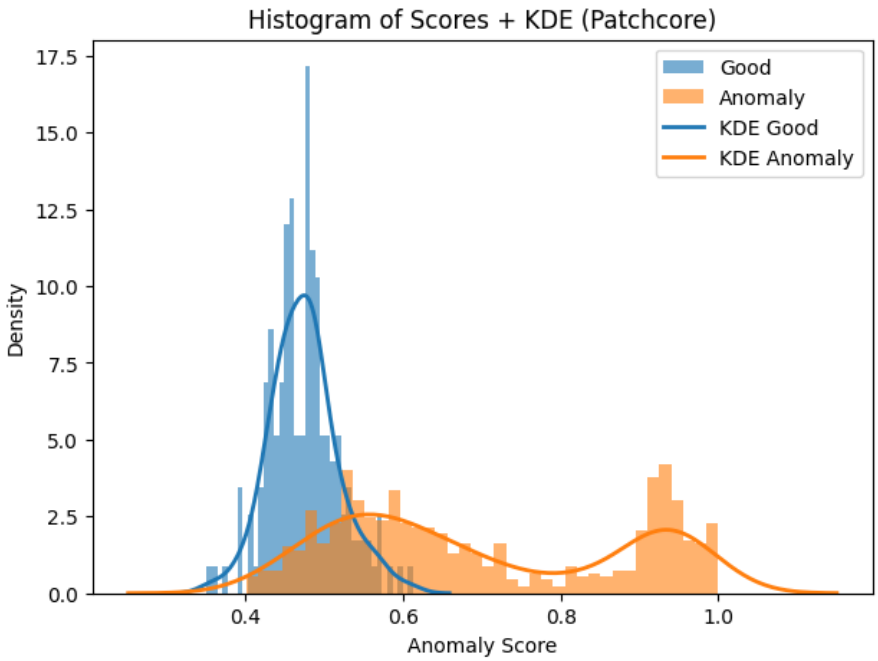
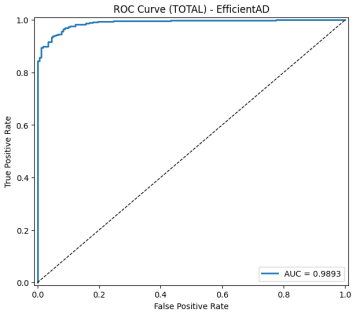
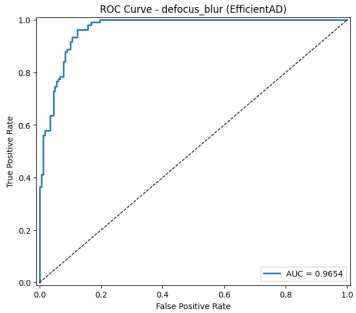


Figure C.4: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with PatchCore on the Acquisition-Related Anomalies Dataset.

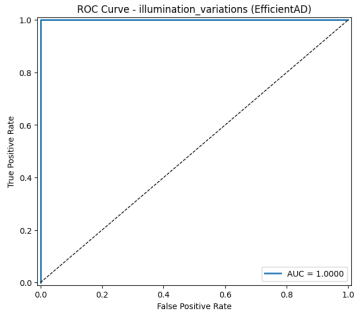
C.3 EfficientAD-S



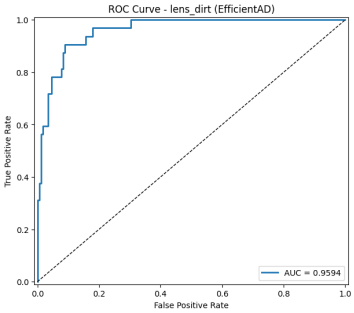
(a) Total (98.93 I-AUROC)



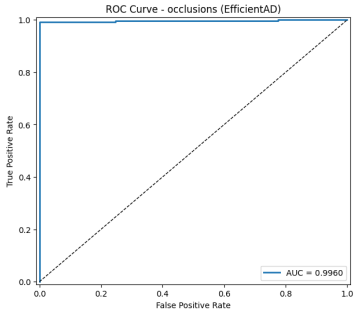
(b) Defocus Blur (96.54 I-AUROC)



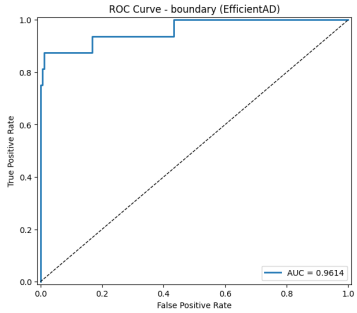
(c) Illumination variations (100.00 I-AUROC)



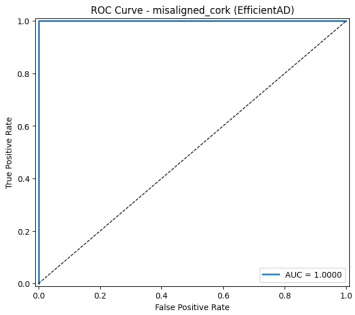
(d) Lens dirt (95.94 I-AUROC)



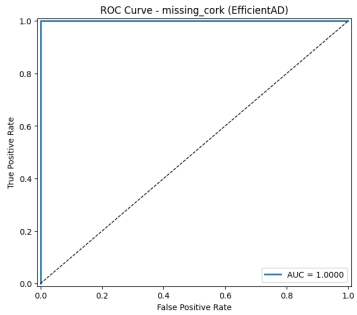
(e) Occlusions (99.60 I-AUROC)



(f) Boundary (96.14 I-AUROC)



(g) Misaligned cork (100.00 I-AUROC)



(h) Missing cork (100.00 I-AUROC)

Figure C.5: Image-level ROC curves obtained with EfficientAD-S on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category.

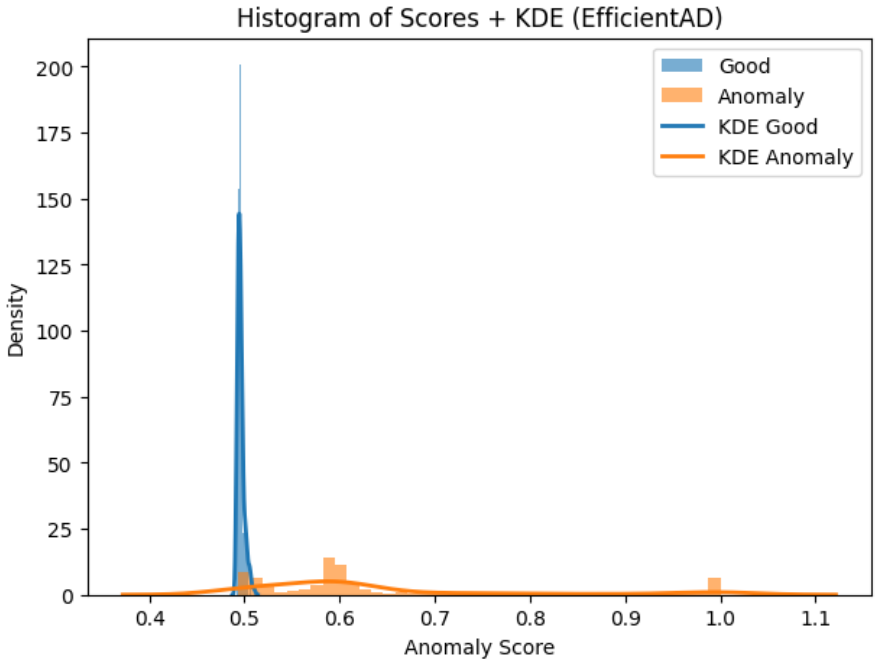
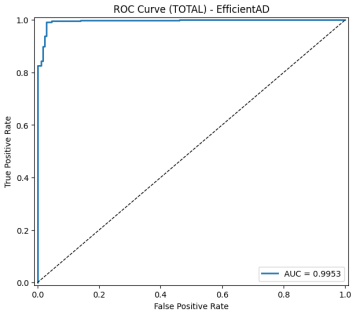
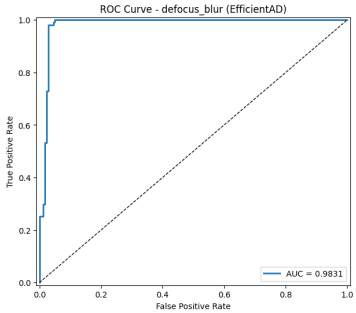


Figure C.6: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with EfficientAD-S on the Acquisition-Related Anomalies Dataset.

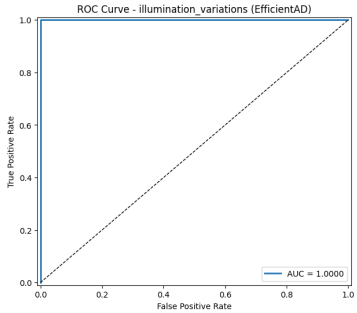
C.4 EfficientAD-M



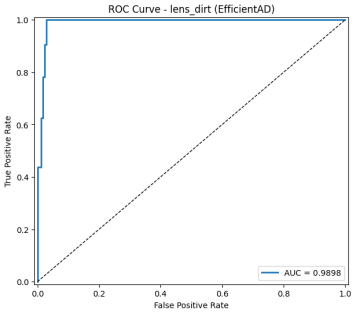
(a) Total (99.53 I-AUROC)



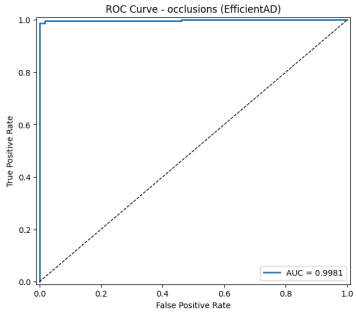
(b) Defocus Blur (98.31 I-AUROC)



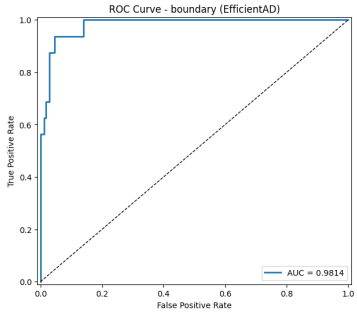
(c) Illumination variations (100.00 I-AUROC)



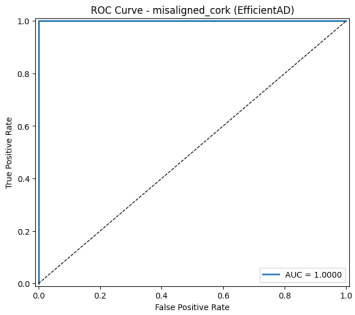
(d) Lens dirt (98.98 I-AUROC)



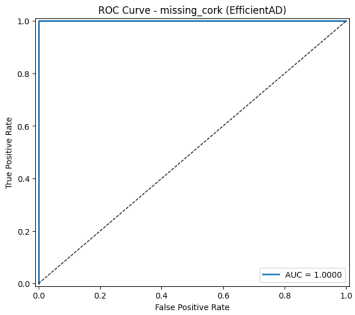
(e) Occlusions (99.81 I-AUROC)



(f) Boundary (98.14 I-AUROC)



(g) Misaligned cork (100.00 I-AUROC)



(h) Missing cork (100.00 I-AUROC)

Figure C.7: Image-level ROC curves obtained with EfficientAD-M on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category.

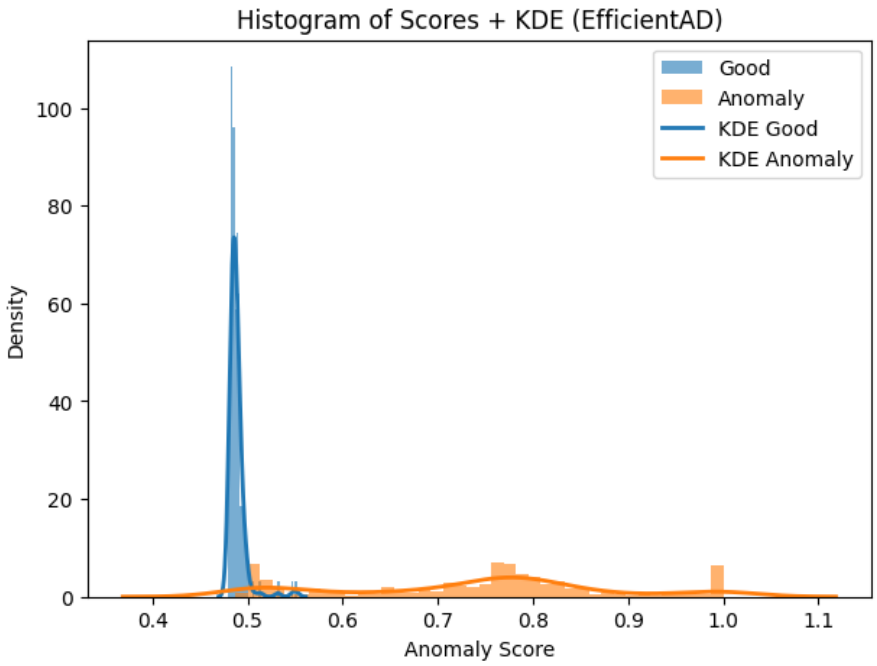


Figure C.8: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with EfficientAD-M on the Acquisition-Related Anomalies Dataset.

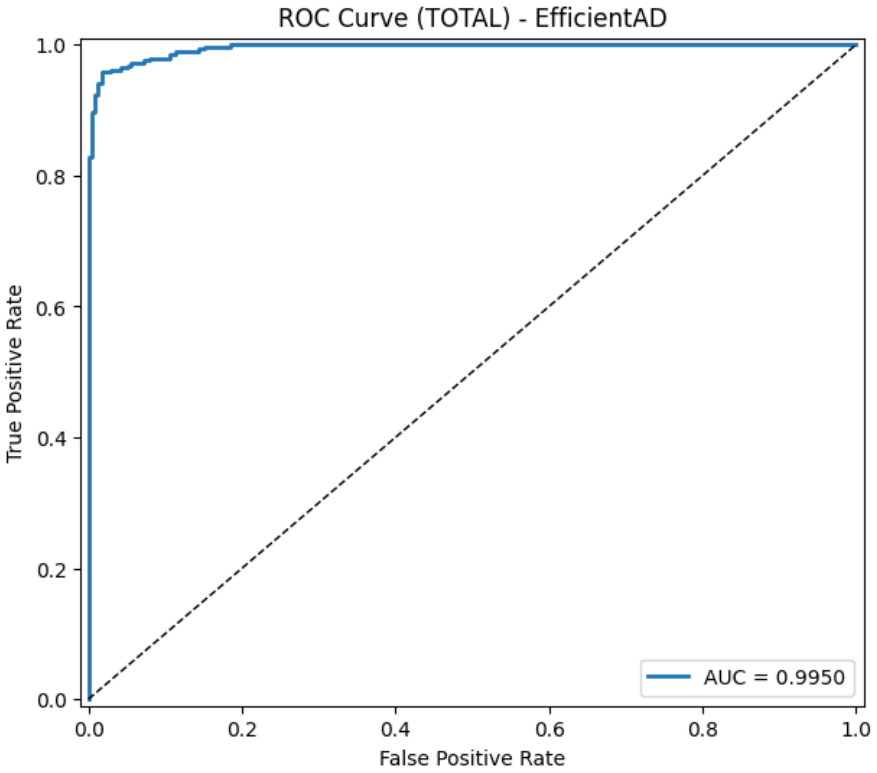


Figure C.9: Image-level ROC curve obtained with EfficientAD-M on the Foreign Object Intrusions Dataset.

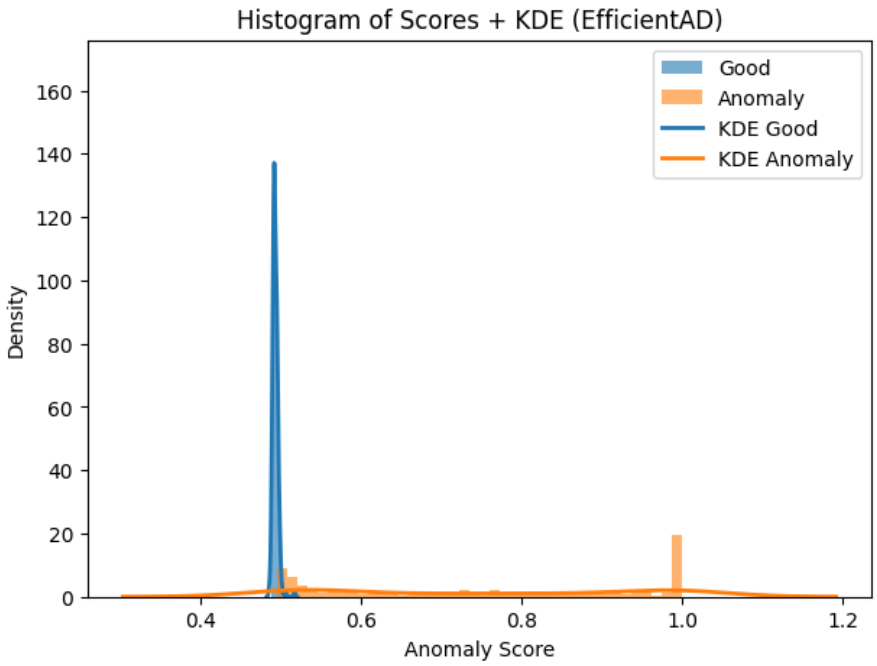


Figure C.10: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with EfficientAD-M on the Foreign Object Intrusions Dataset.

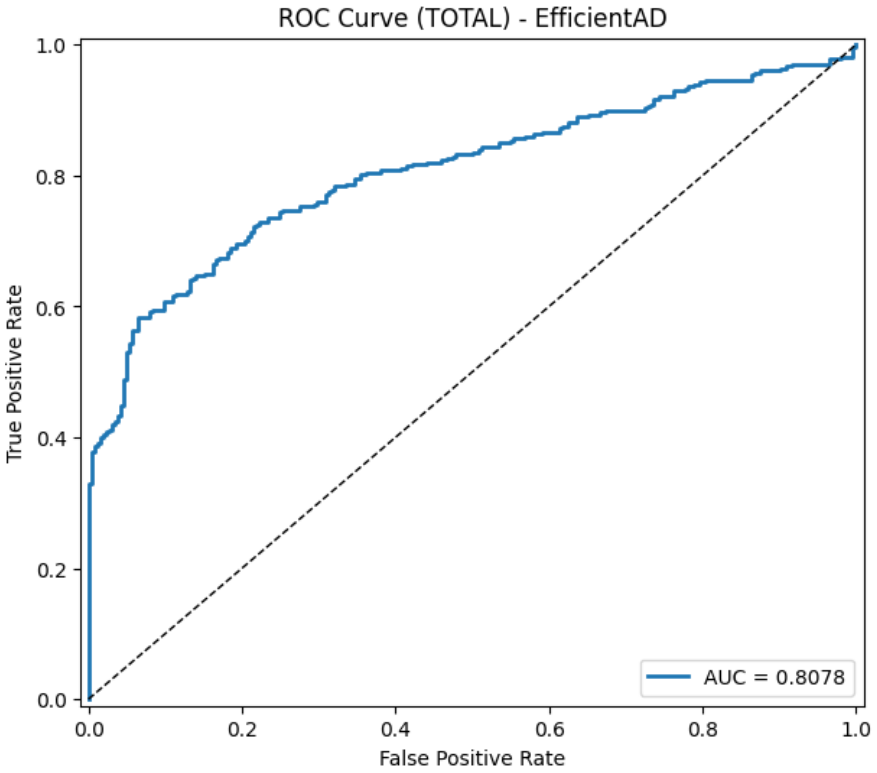


Figure C.11: Image-level ROC curve obtained with EfficientAD-M on the Rolling Inspection Artifacts Dataset.

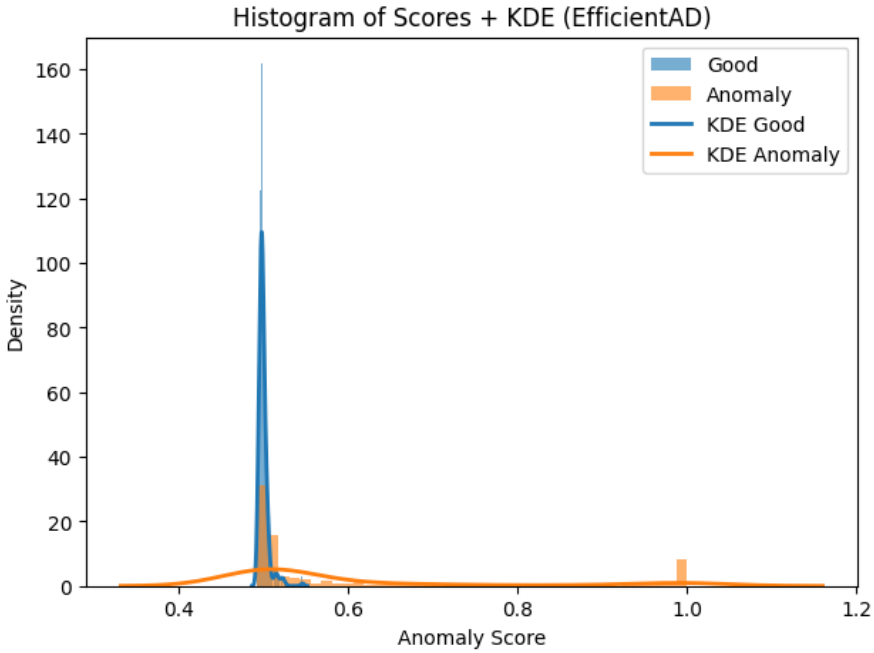
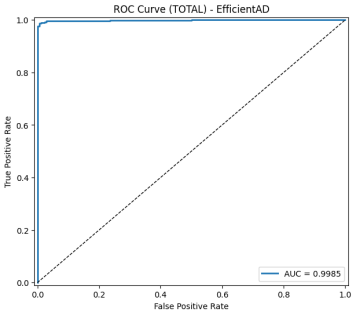
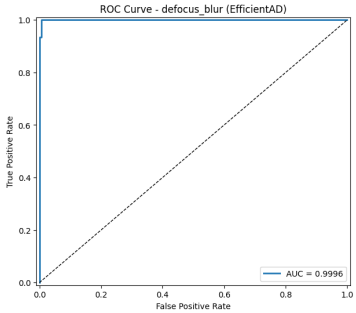


Figure C.12: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with EfficientAD-M on the Rolling Inspection Artifacts Dataset.

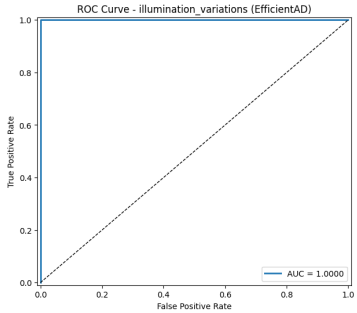
C.5 EfficientAD-M without Autoencoder



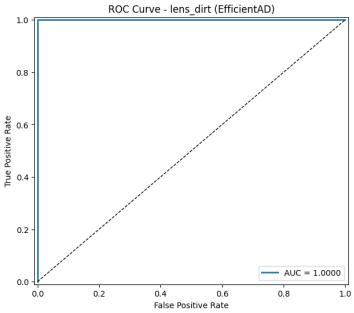
(a) Total (99.85 I-AUROC)



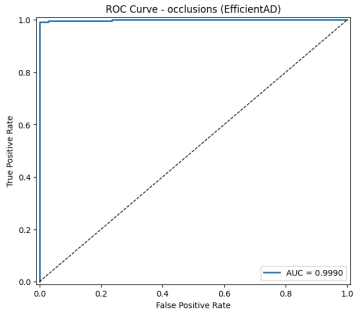
(b) Defocus Blur (99.96 I-AUROC)



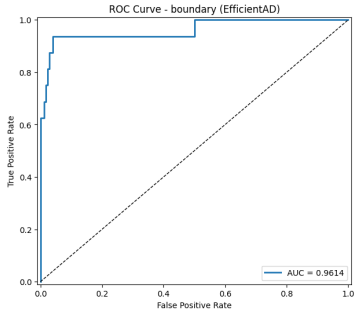
(c) Illumination variations (100.00 I-AUROC)



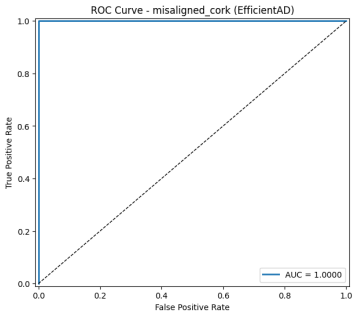
(d) Lens dirt (100.00 I-AUROC)



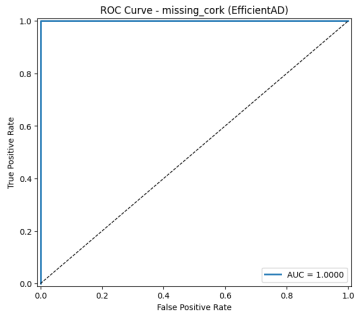
(e) Occlusions (99.90 I-AUROC)



(f) Boundary (96.14 I-AUROC)



(g) Misaligned cork (100.00 I-AUROC)



(h) Missing cork (100.00 I-AUROC)

Figure C.13: Image-level ROC curves obtained with EfficientAD-M without Autoencoder on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category.

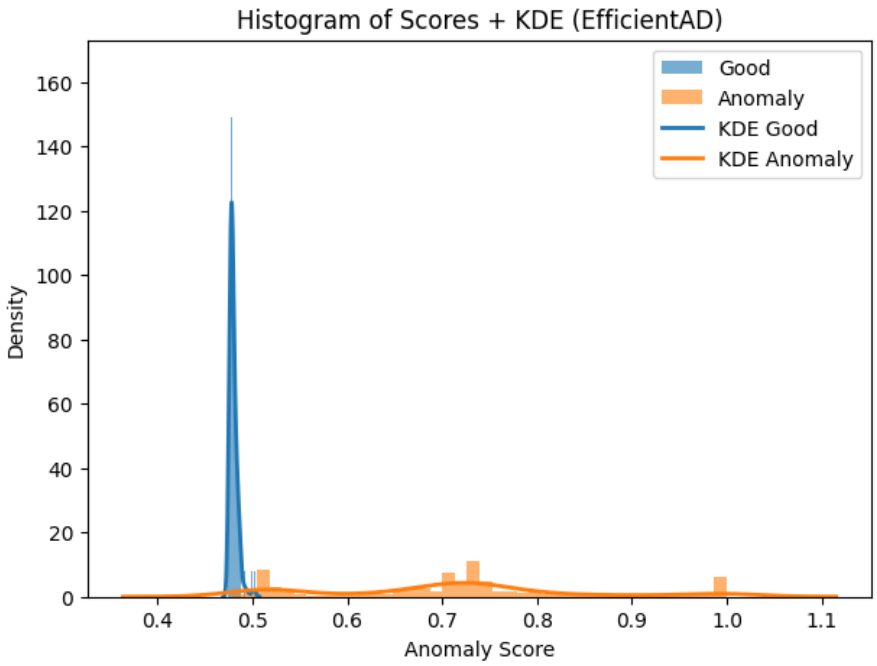


Figure C.14: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with EfficientAD-M without Autoencoder on the Acquisition-Related Anomalies Dataset.

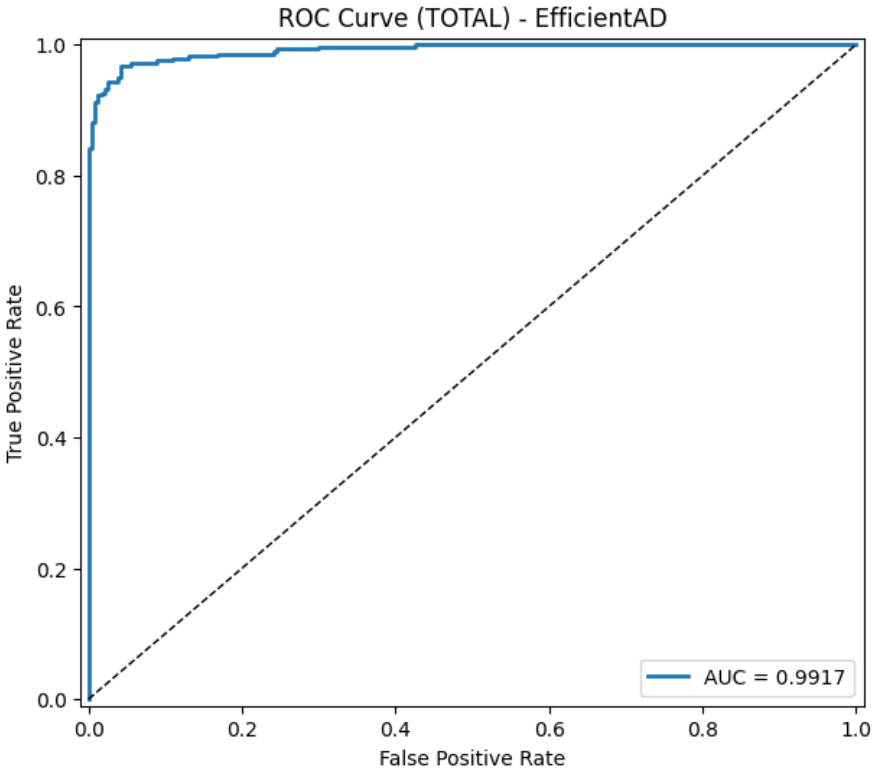


Figure C.15: Image-level ROC curve obtained with EfficientAD-M without Autoencoder on the Foreign Object Intrusions Dataset.

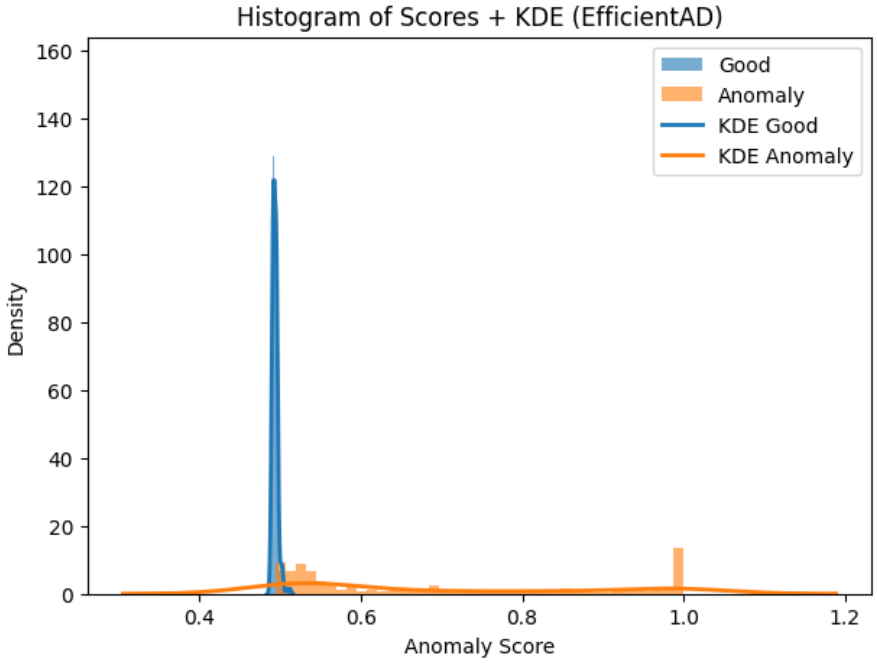


Figure C.16: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with EfficientAD-M without Autoencoder on the Foreign Object Intrusions Dataset.

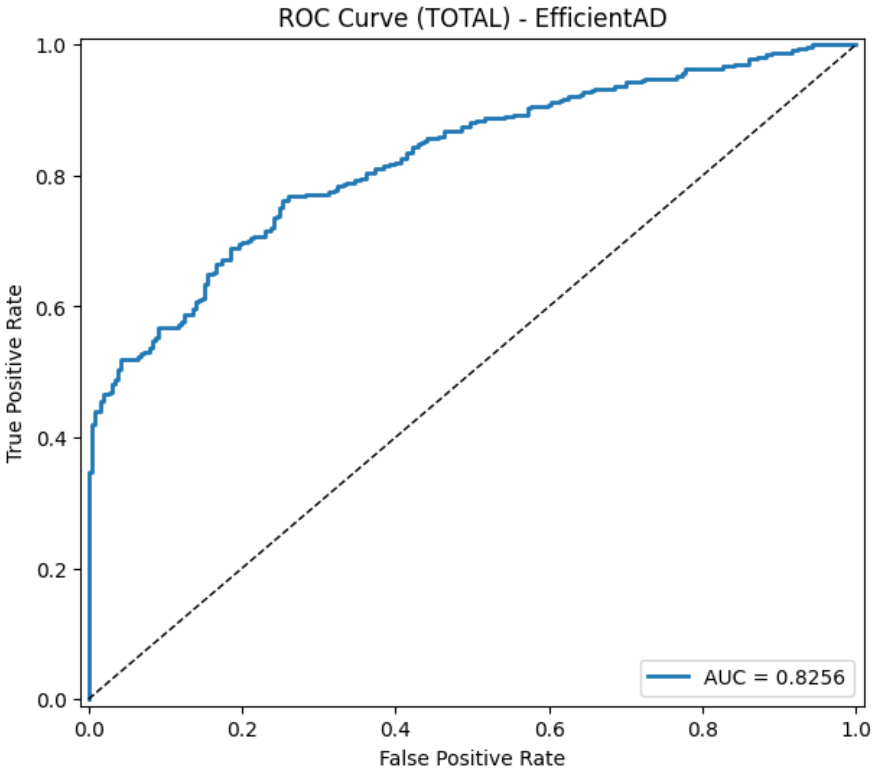


Figure C.17: Image-level ROC curve obtained with EfficientAD-M without Autoencoder on the Rolling Inspection Artifacts Dataset.

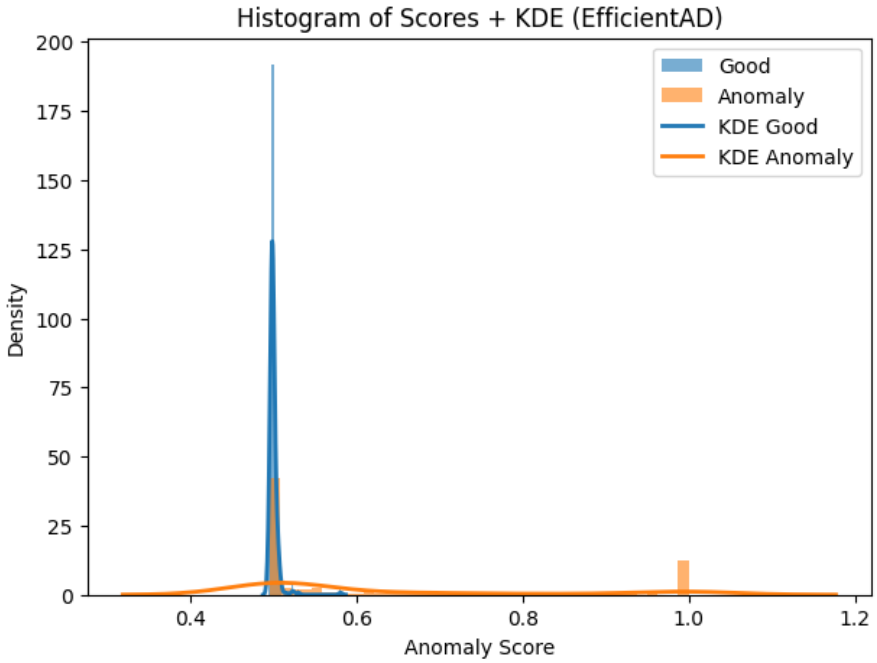
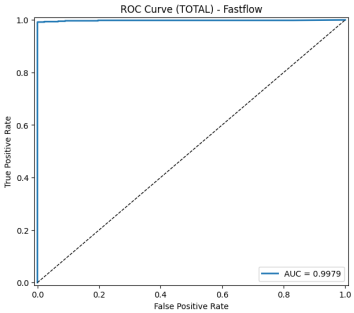
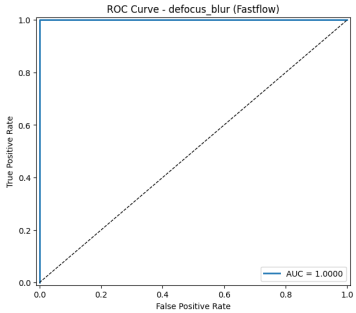


Figure C.18: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with EfficientAD-M without Autoencoder on the Rolling Inspection Artifacts Dataset.

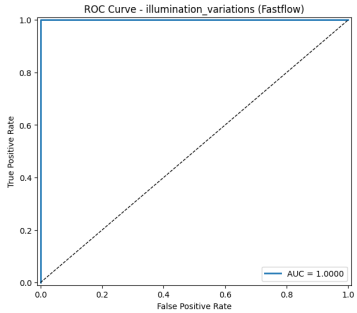
C.6 FastFlow



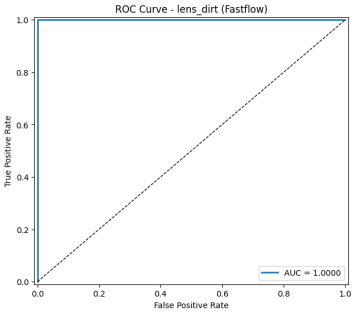
(a) Total (99.83 I-AUROC)



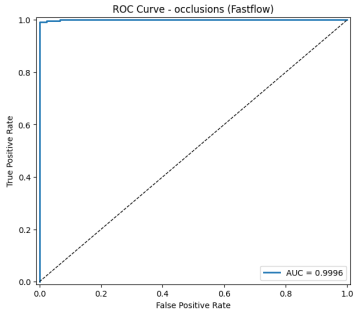
(b) Defocus Blur (100.00 I-AUROC)



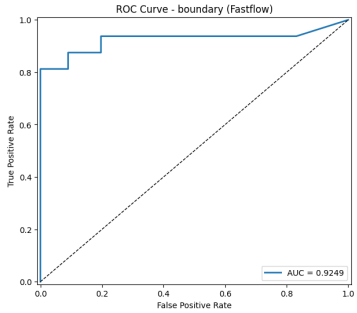
(c) Illumination variations (100.00 I-AUROC)



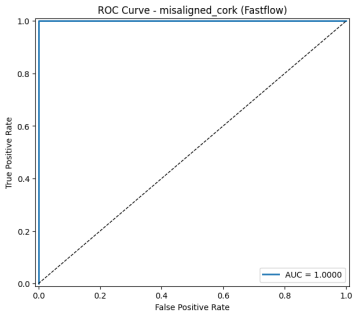
(d) Lens dirt (100.00 I-AUROC)



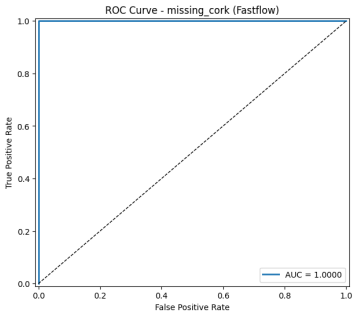
(e) Occlusions (99.77 I-AUROC)



(f) Boundary (97.30 I-AUROC)



(g) Misaligned cork (100.00 I-AUROC)



(h) Missing cork (100.00 I-AUROC)

Figure C.19: Image-level ROC curves obtained with FastFlow on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category.

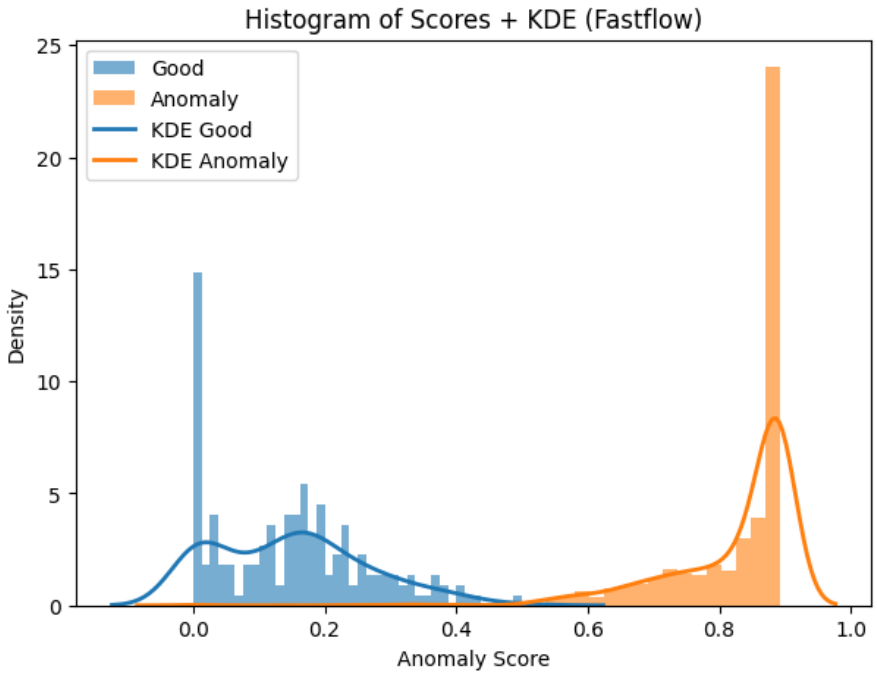


Figure C.20: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with FastFlow on the Acquisition-Related Anomalies Dataset.

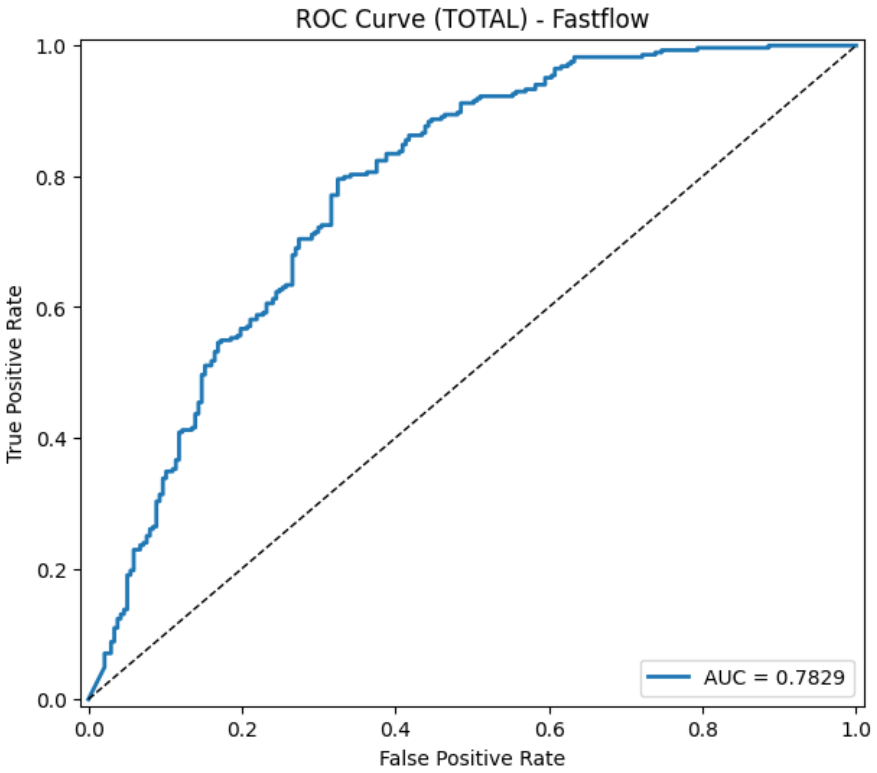


Figure C.21: Image-level ROC curve obtained with FastFlow on the Foreign Object Intrusions Dataset.

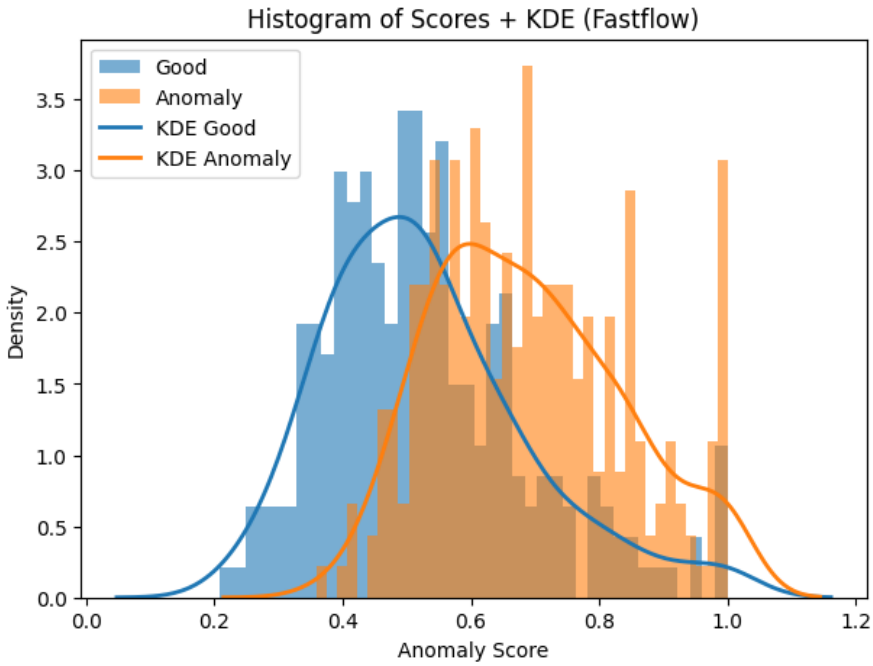


Figure C.22: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with FastFlow on the Foreign Object Intrusions Dataset.

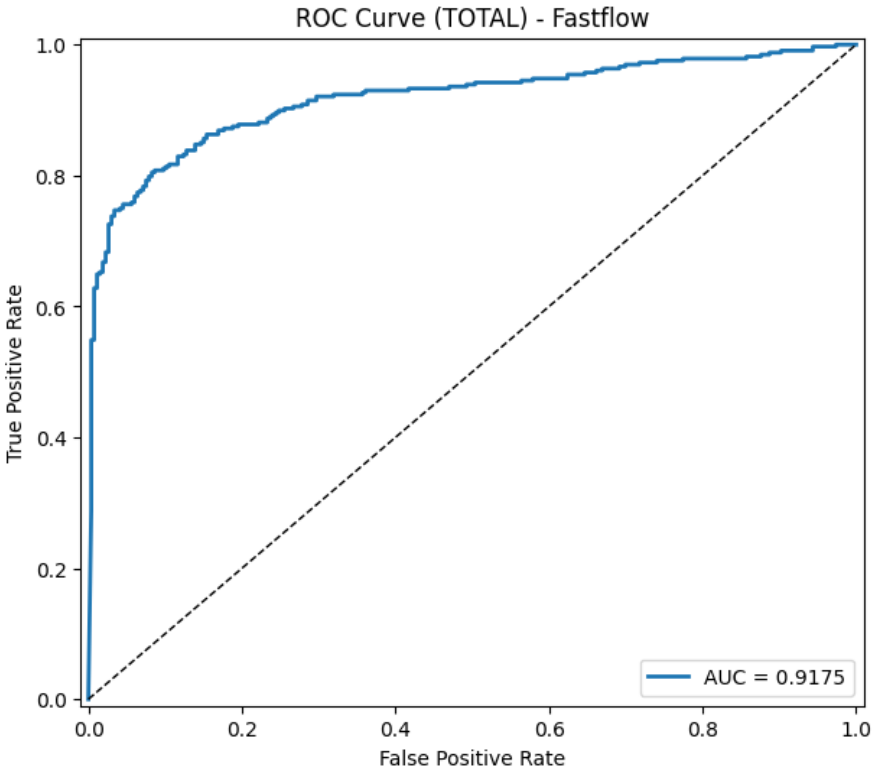


Figure C.23: Image-level ROC curve obtained with FastFlow on the Rolling Inspection Artifacts Dataset.

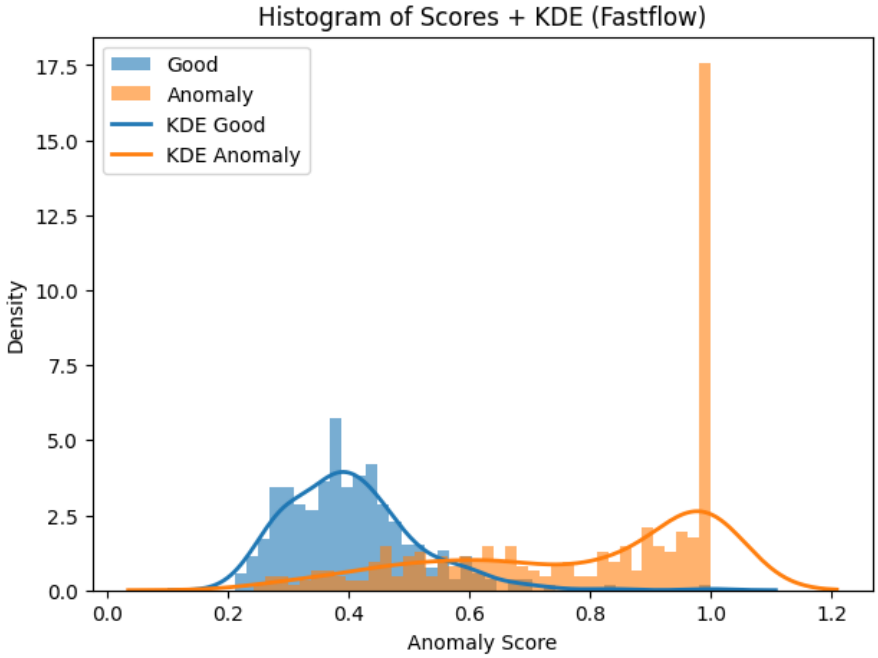
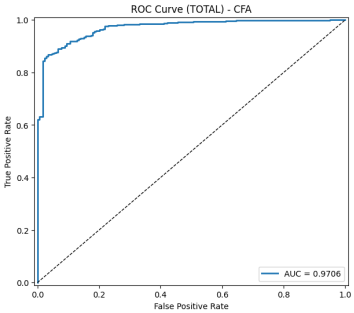
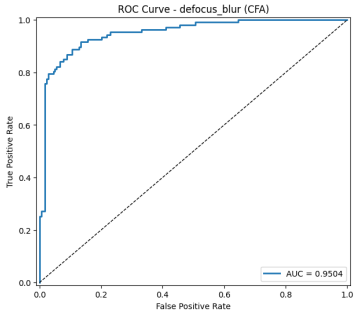


Figure C.24: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with FastFlow on the Rolling Inspection Artifacts Dataset.

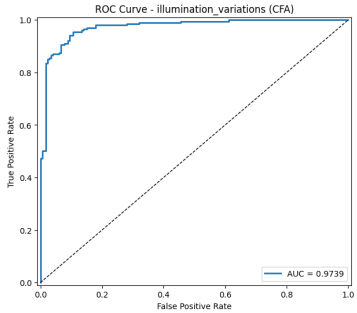
C.7 CFA



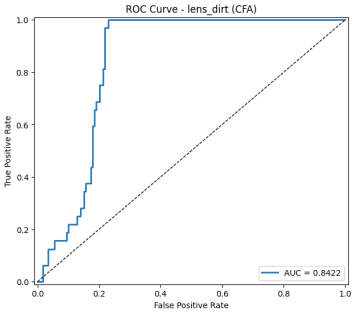
(a) Total (97.06 I-AUROC)



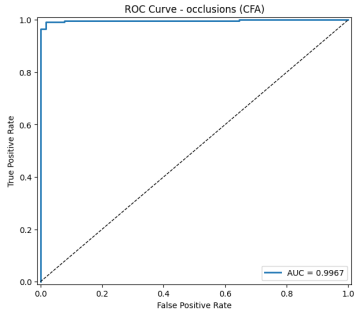
(b) Defocus Blur (95.04 I-AUROC)



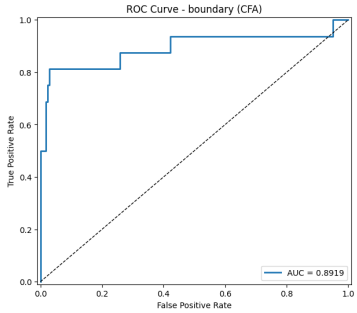
(c) Illumination variations (97.39 I-AUROC)



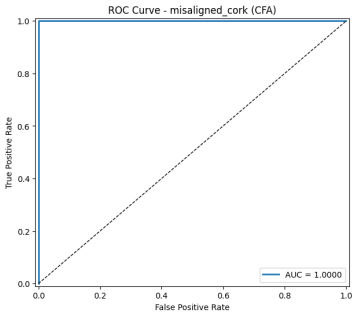
(d) Lens dirt (84.22 I-AUROC)



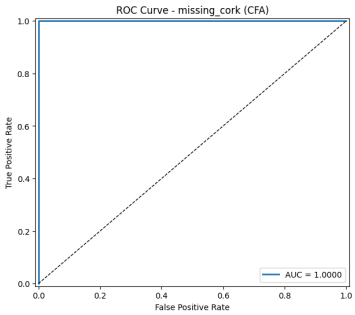
(e) Occlusions (99.67 I-AUROC)



(f) Boundary (89.19 I-AUROC)



(g) Misaligned cork (100.00 I-AUROC)



(h) Missing cork (100.00 I-AUROC)

Figure C.25: Image-level ROC curves obtained with CFA on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category.

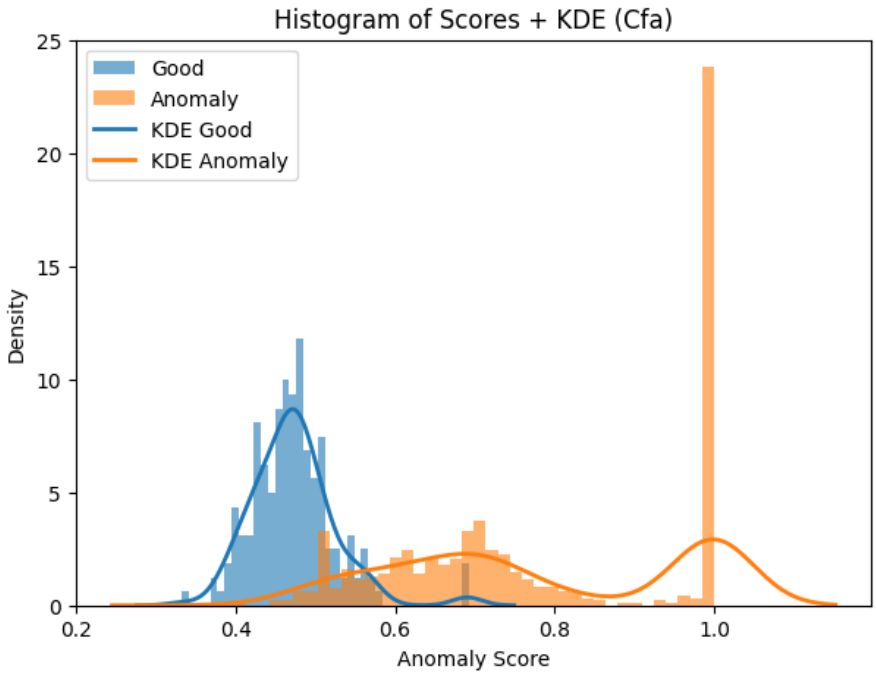
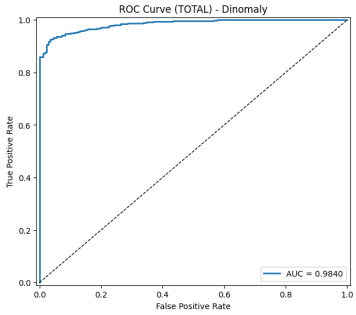
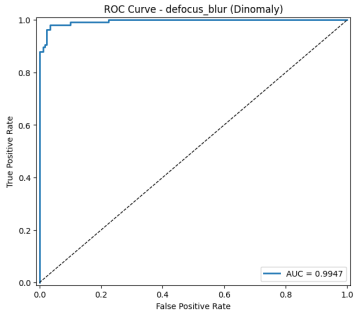


Figure C.26: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with CFA on the Acquisition-Related Anomalies Dataset.

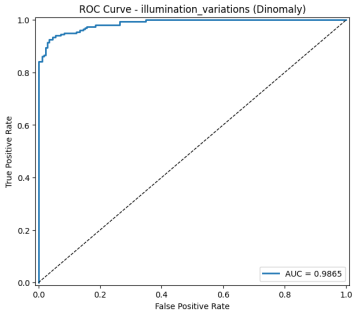
C.8 Dinomaly



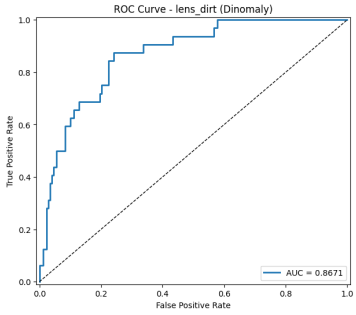
(a) Total (98.40 I-AUROC)



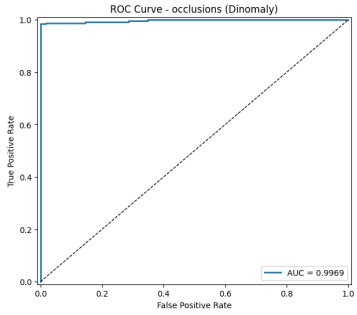
(b) Defocus Blur (99.47 I-AUROC)



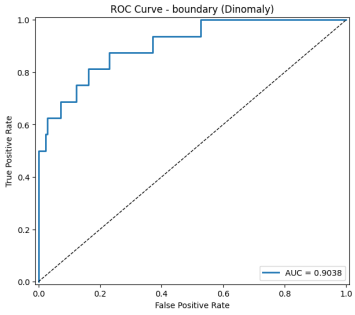
(c) Illumination variations (98.65 I-AUROC)



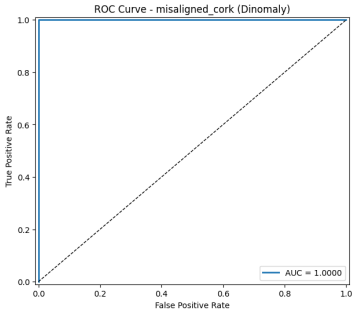
(d) Lens dirt (86.71 I-AUROC)



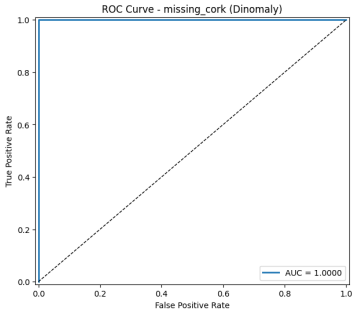
(e) Occlusions (99.69 I-AUROC)



(f) Boundary (90.38 I-AUROC)



(g) Misaligned cork (100.00 I-AUROC)



(h) Missing cork (100.00 I-AUROC)

Figure C.27: Image-level ROC curves obtained with Dinomaly on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category.

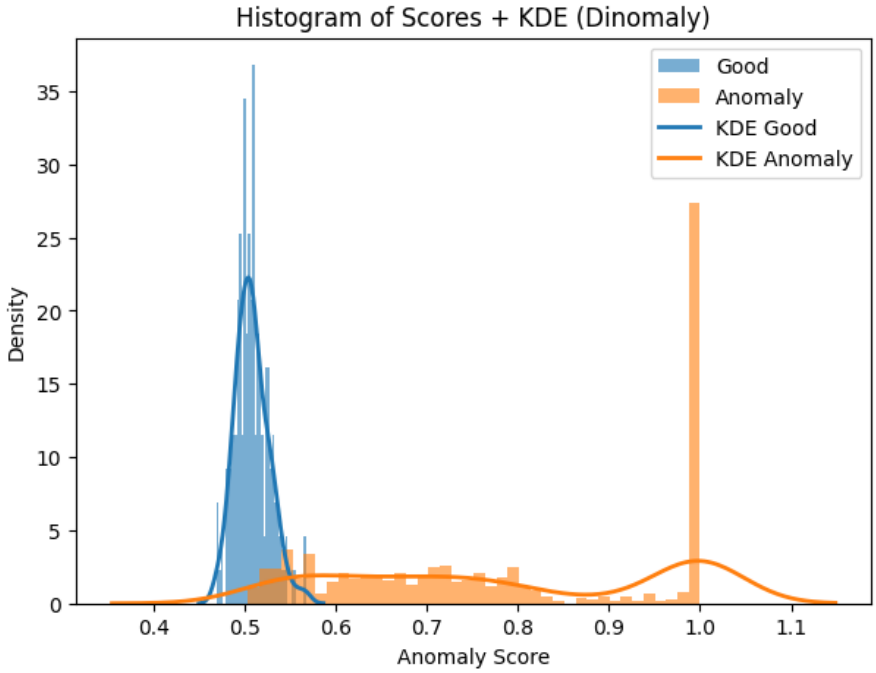
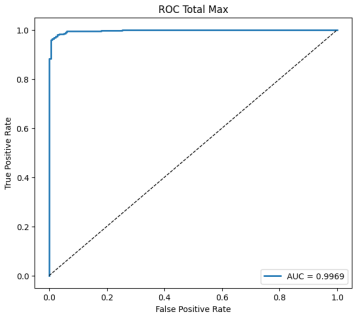
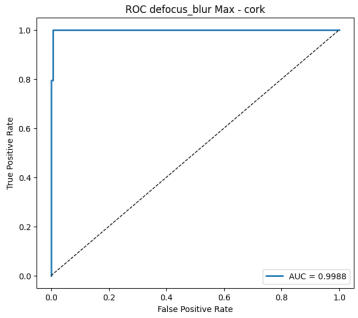


Figure C.28: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with Dinomaly on the Acquisition-Related Anomalies Dataset.

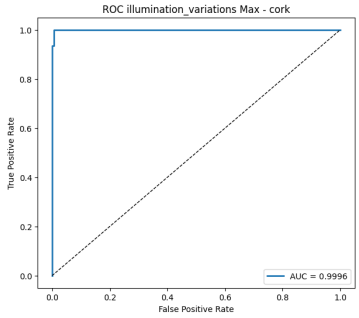
C.9 AST



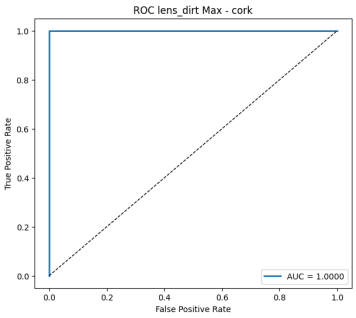
(a) Total (99.69 I-AUROC)



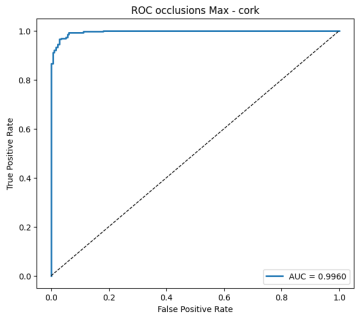
(b) Defocus Blur (99.88 I-AUROC)



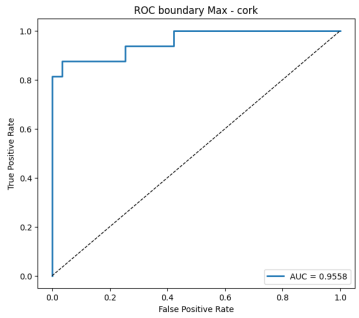
(c) Illumination variations (99.96 I-AUROC)



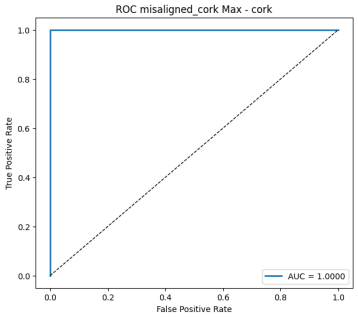
(d) Lens dirt (100.00 I-AUROC)



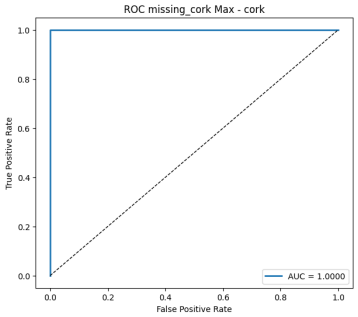
(e) Occlusions (99.60 I-AUROC)



(f) Boundary (95.58 I-AUROC)



(g) Misaligned cork (100.00 I-AUROC)



(h) Missing cork (100.00 I-AUROC)

Figure C.29: Image-level ROC curves obtained with AST on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category.

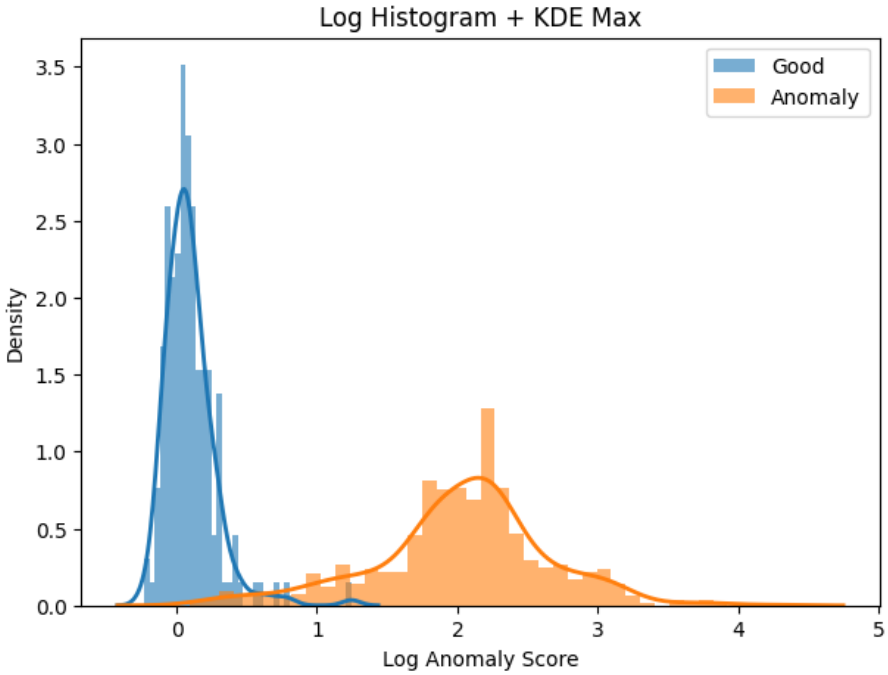


Figure C.30: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with AST on the Acquisition-Related Anomalies Dataset. A logarithmic scale is used for visualization purposes.

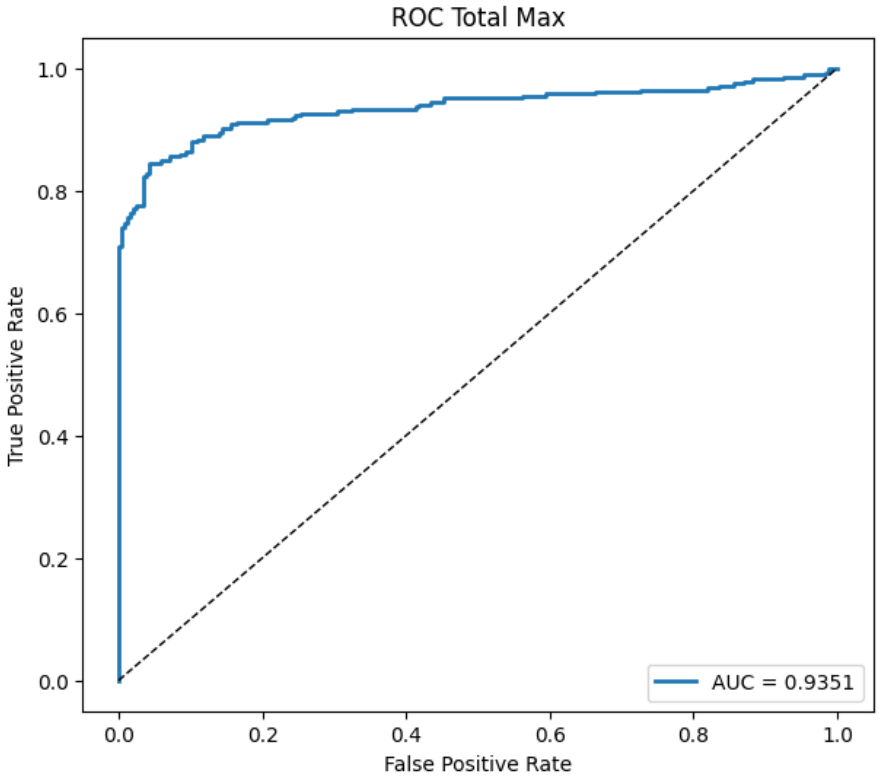


Figure C.31: Image-level ROC curve obtained with AST on the Foreign Object Intrusions Dataset.

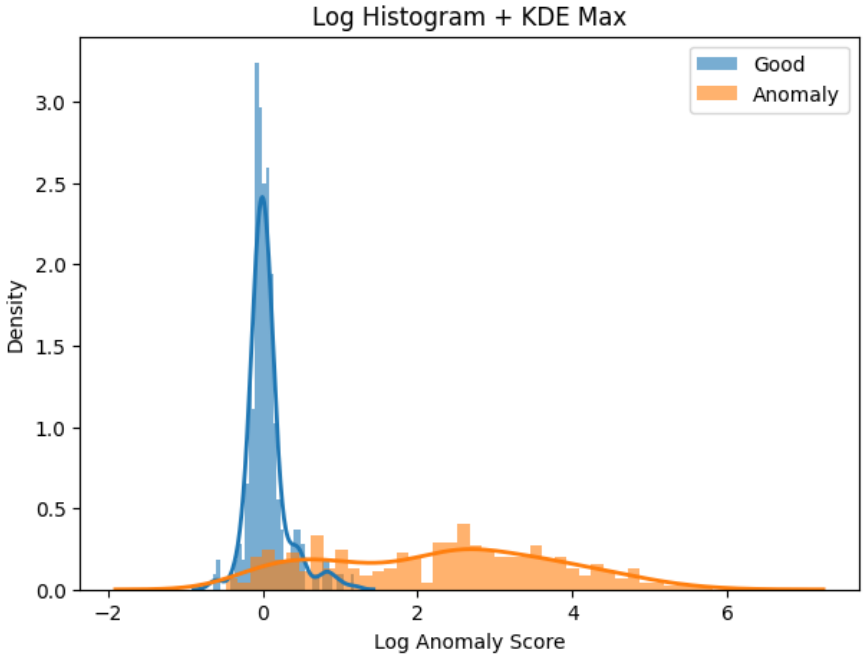


Figure C.32: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with AST on the Foreign Object Intrusions Dataset. A logarithmic scale is used for visualization purposes.

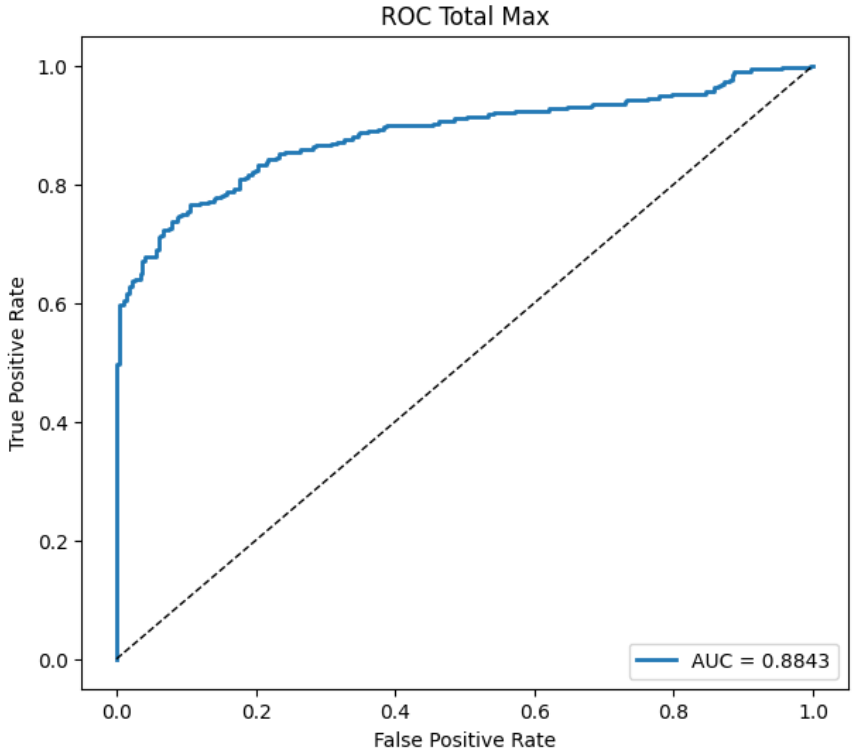


Figure C.33: Image-level ROC curve obtained with AST on the Rolling Inspection Artifacts Dataset.

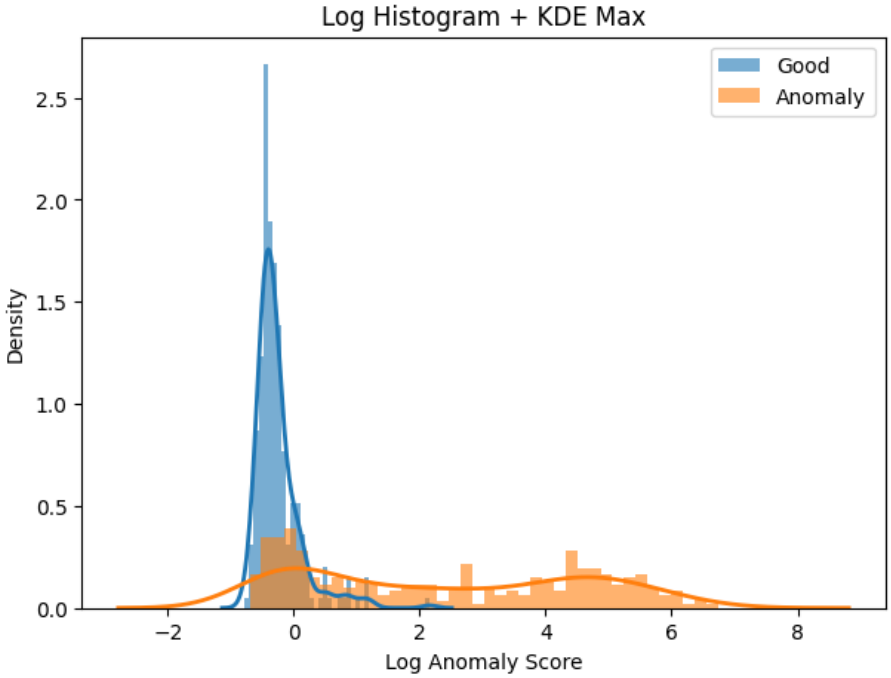
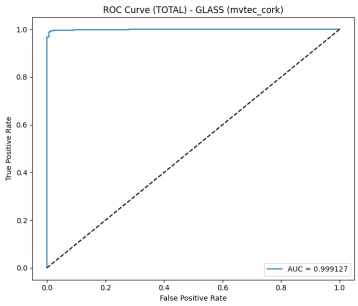
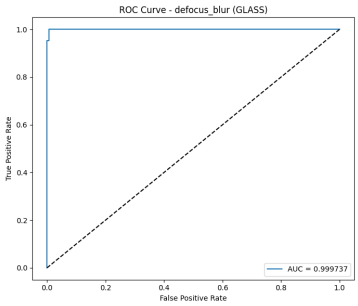


Figure C.34: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with AST on the Rolling Inspection Artifacts Dataset. A logarithmic scale is used for visualization purposes.

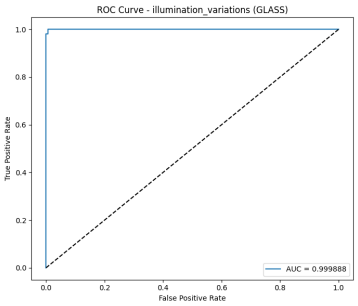
C.10 GLASS



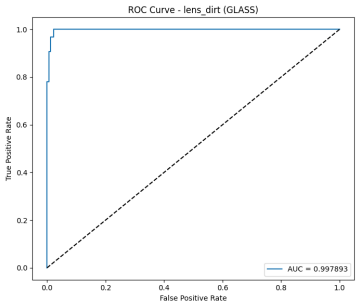
(a) Total (99.91 I-AUROC)



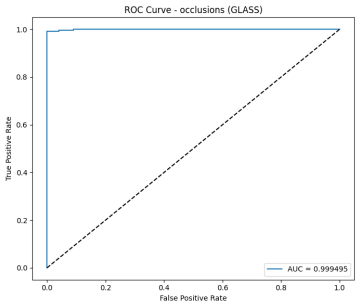
(b) Defocus Blur (99.97 I-AUROC)



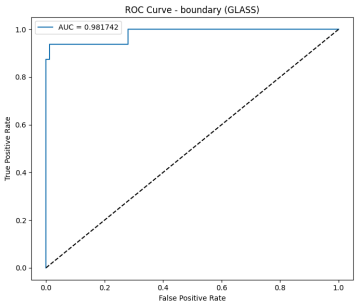
(c) Illumination variations (99.99 I-AUROC)



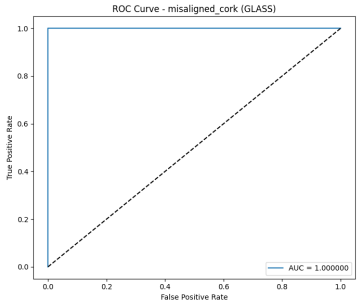
(d) Lens dirt (99.79 I-AUROC)



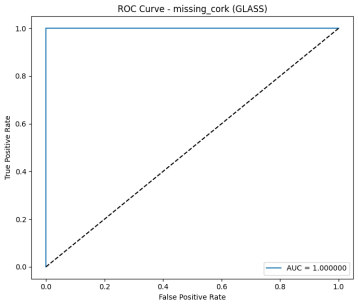
(e) Occlusions (99.95 I-AUROC)



(f) Boundary (98.17 I-AUROC)



(g) Misaligned cork (100.00 I-AUROC)



(h) Missing cork (100.00 I-AUROC)

Figure C.35: Image-level ROC curves obtained with GLASS on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category.

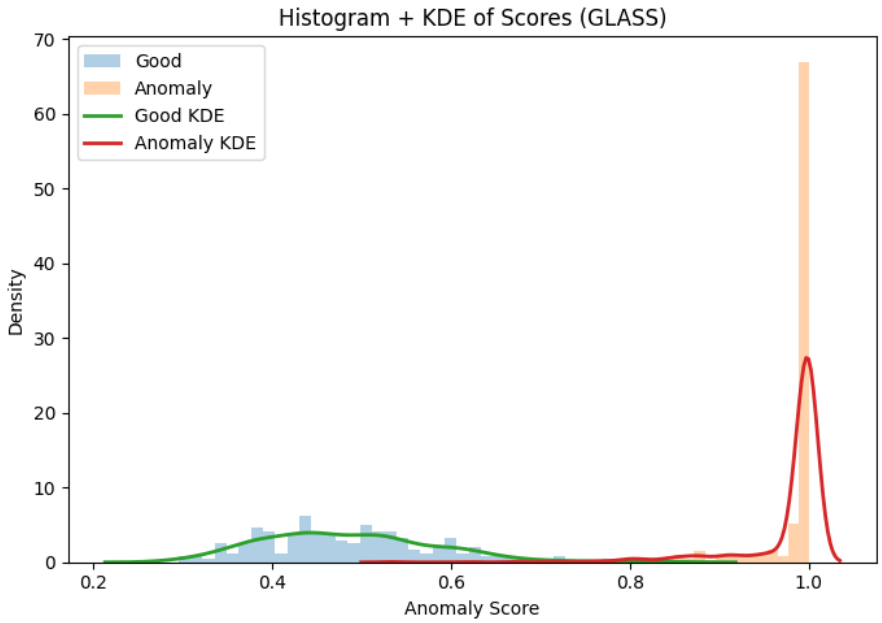


Figure C.36: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with GLASS on the Acquisition-Related Anomalies Dataset.

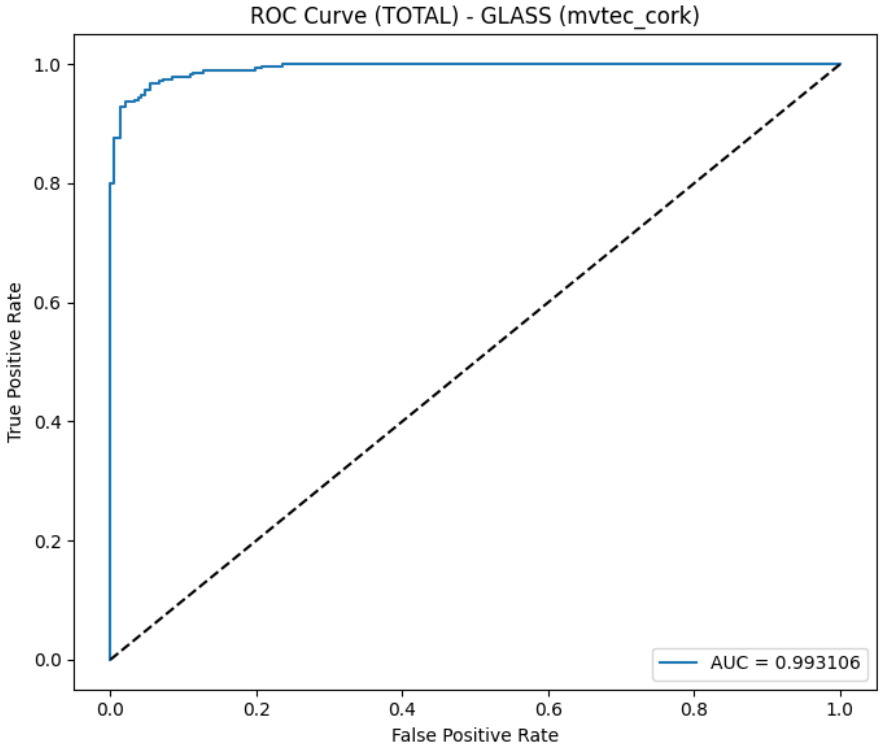


Figure C.37: Image-level ROC curve obtained with GLASS on the Foreign Object Intrusions Dataset.

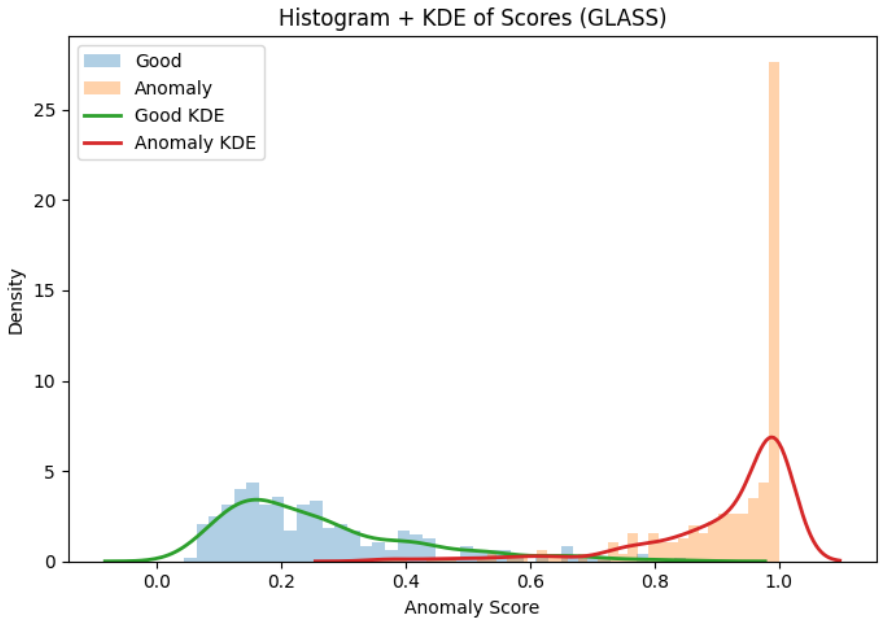


Figure C.38: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with GLASS on the Foreign Object Intrusions Dataset.

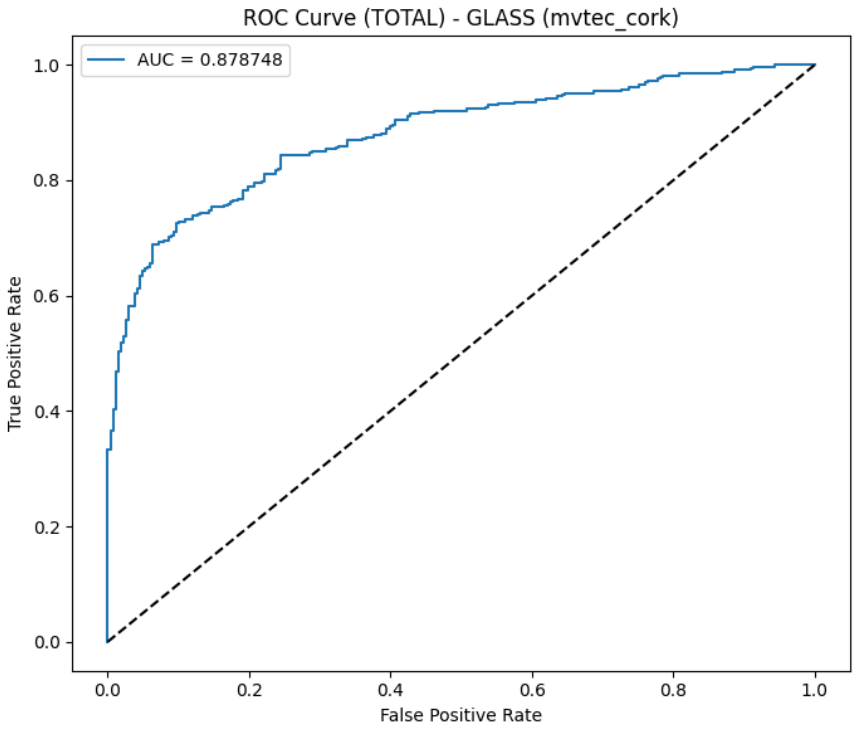


Figure C.39: Image-level ROC curve obtained with GLASS on the Rolling Inspection Artifacts Dataset.

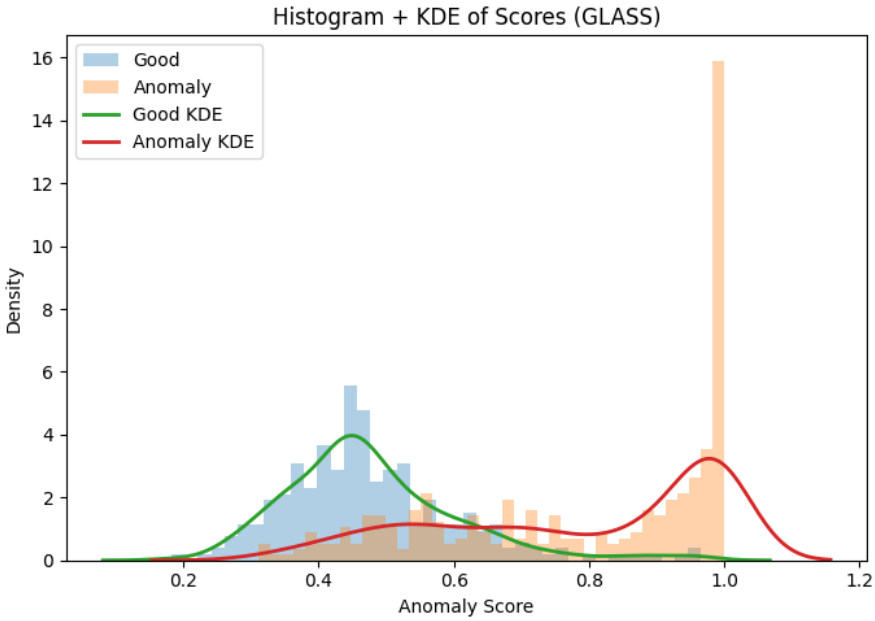
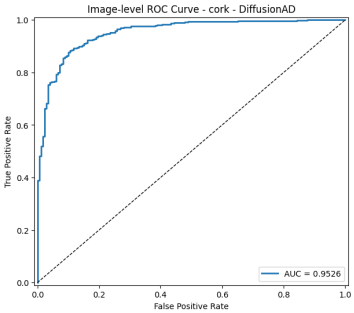
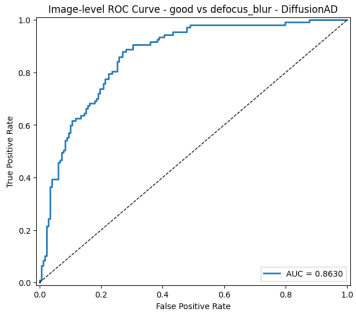


Figure C.40: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with GLASS on the Rolling Inspection Artifacts Dataset.

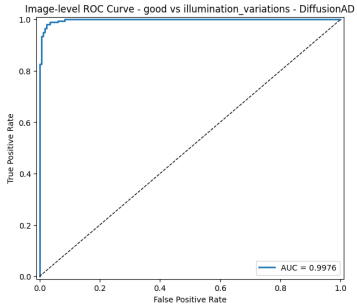
C.11 DiffusionAD



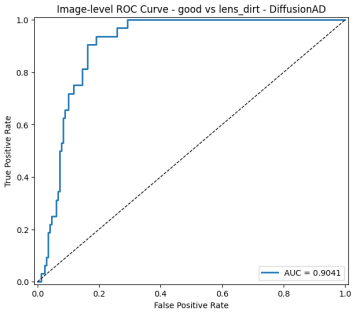
(a) Total (95.26 I-AUROC)



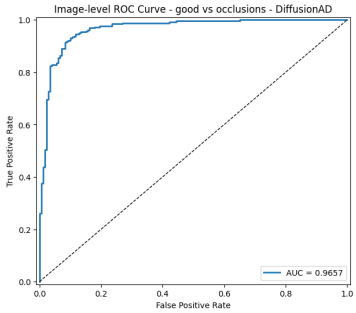
(b) Defocus Blur (86.30 I-AUROC)



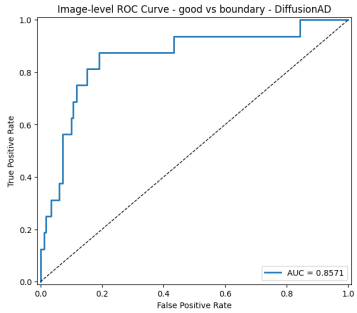
(c) Illumination variations (99.76 I-AUROC)



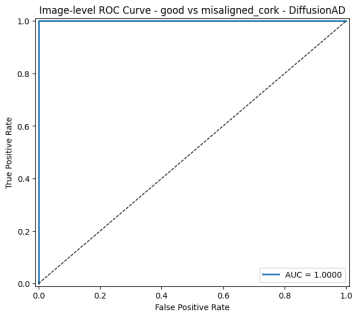
(d) Lens dirt (90.41 I-AUROC)



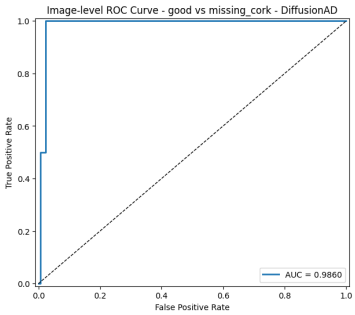
(e) Occlusions (96.57 I-AUROC)



(f) Boundary (85.71 I-AUROC)



(g) Misaligned cork (100.00 I-AUROC)



(h) Missing cork (98.60 I-AUROC)

Figure C.41: Image-level ROC curves obtained with DiffusionAD on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category.

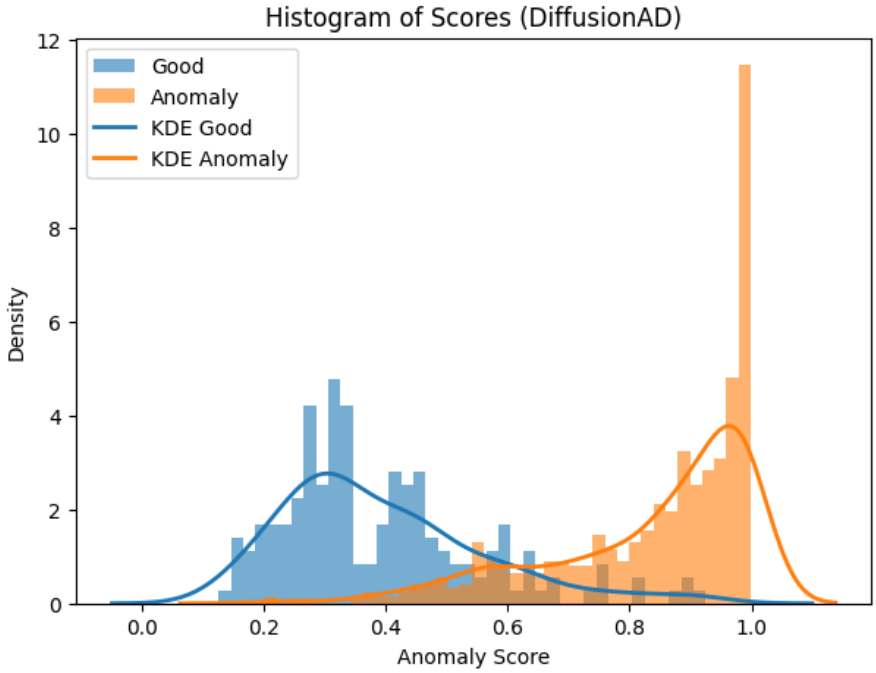
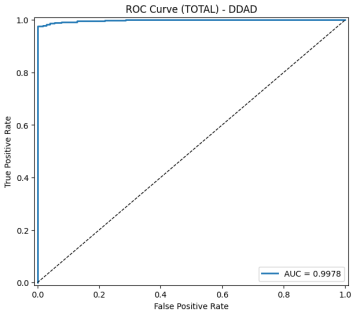
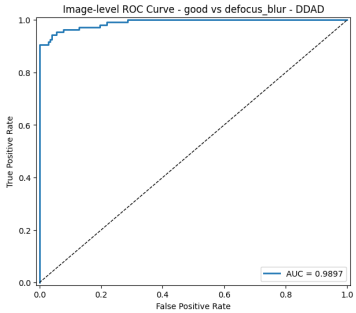


Figure C.42: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with DiffusionAD on the Acquisition-Related Anomalies Dataset.

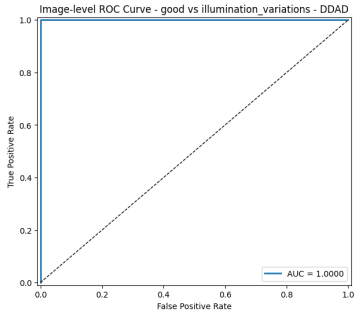
C.12 DDAD



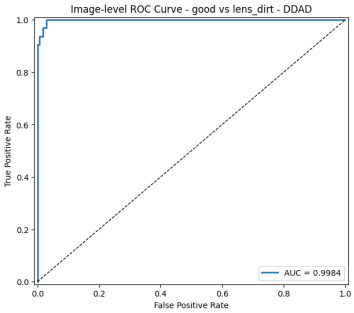
(a) Total (99.78 I-AUROC)



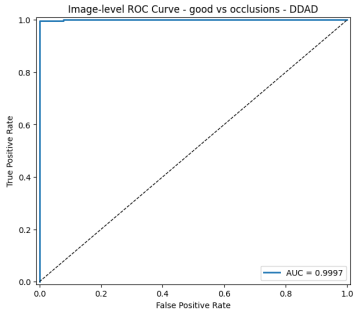
(b) Defocus Blur (98.97 I-AUROC)



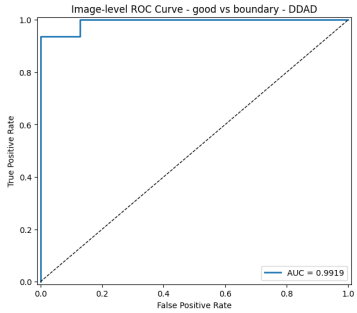
(c) Illumination variations (100.00 I-AUROC)



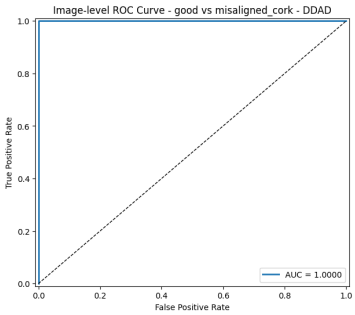
(d) Lens dirt (99.84 I-AUROC)



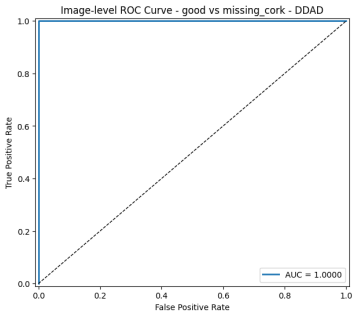
(e) Occlusions (99.97 I-AUROC)



(f) Boundary (99.19 I-AUROC)



(g) Misaligned cork (100.00 I-AUROC)



(h) Missing cork (100.00 I-AUROC)

Figure C.43: Image-level ROC curves obtained with DDAD on the Acquisition-Related Anomalies Dataset, including the overall performance and the performance for each acquisition-related anomaly category.

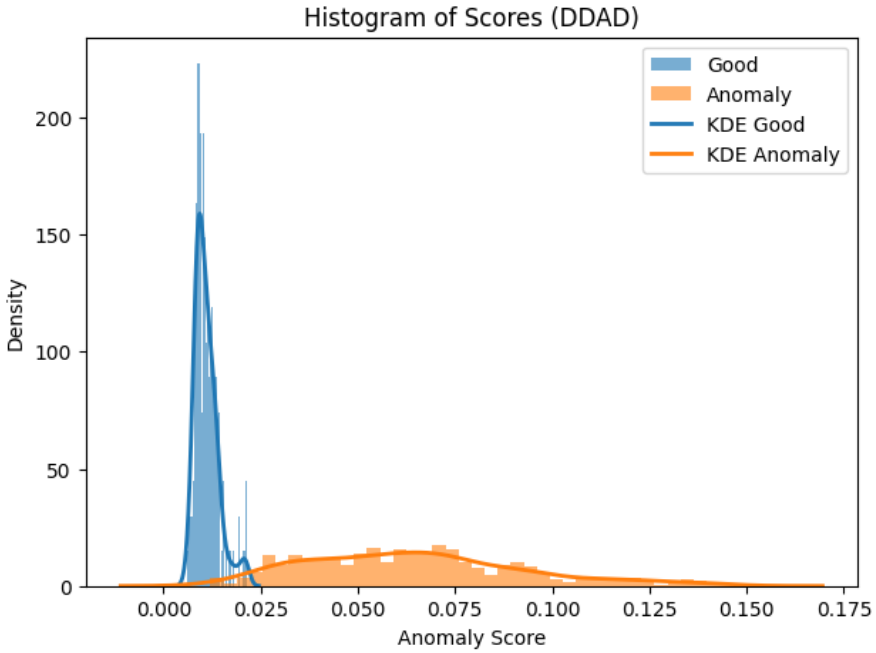


Figure C.44: Histogram of image-level anomaly scores together with the corresponding KDE, obtained with DDAD on the Acquisition-Related Anomalies Dataset.