

Anomaly detection in industrial vision with convolutional autoencoders

Anne-Sophie COLLIN

Thesis submitted in partial fulfillment
of the requirements for the degree of
Docteur en sciences de l'ingénieur

Dissertation committee:

Prof. Christophe DE VLEESCHOUWER (UCLouvain, advisor)

Prof. John LEE (UCLouvain, secretary)

Prof. Pierre GEURTS (ULiège)

Dr. Christophe CHAUDY (Euresys)

Prof. Laurent FRANCIS (UCLouvain, chairman)

October 2024

C'est un petit pas pour l'Homme, et un petit pas pour l'Humanité.

Abstract

INDUSTRIAL VISION greatly benefits from the capacity to identify abnormal samples based on images. Due to the scarcity and variability of anomalies, it is assumed that only anomaly-free images are available during training. This limitation renders supervised methods unsuitable. Consequently, we adopt an unsupervised problem formulation in which a network is expected to reconstruct a normal version out of any input image. In this initial contribution, we consider an autoencoder with skip connections trained to output the normal version of an image corrupted with a homemade synthetic defect. The combination of skip connections with our synthetic defect model enhances the quality of the reconstructed images.

As traditionally done in the field, the reconstruction residual can be used as an anomaly detection metric. However, this metric struggles to detect abnormal structures that exhibit a low contrast with surrounding clean structures. To address this, we explore alternative detection metrics. First, we introduce a metric relying on the difference between the activation tensors of the input image and its clean reconstruction. Anomalies are detected at the voxel level using numerous binary tests, which assess the probability of observing abnormal activations at designated voxels. This metric was developed based on a visual analysis of the propagation of defective structures through the network layers. Second, we propose to train quantization modules, inspired by the VQ-VAE architecture, on top of all skip connections of our baseline network. These quantization modules allow to obtain a multi-level representation of the image, wherein values correspond to indices in a lookup table rather than the initial pixel values. Subsequently, a segmentation network is trained on those lookup tables to produce binary maps that localize defective structures.

This thesis presents a comprehensive analysis of previously cited metrics, both qualitatively and quantitatively. The study is conducted over the MVTec AD dataset, which reflects a practical industrial use case. The focus is on developing a detailed understanding of the properties of these methods.

Acknowledgments

LE TRAVAIL de chercheur est souvent perçu comme un parcours solitaire, où le doctorant s'enferme dans son bureau des heures durant, espérant faire jaillir l'idée miraculeuse. Heureusement, mon expérience a été tout autre! Durant ces années passées à l'UCLouvain, j'ai eu la chance de bénéficier du soutien, des conseils et des encouragements de nombreuses personnes auxquelles je tiens à exprimer toute ma gratitude.

Tout d'abord, je tiens à adresser mes remerciements les plus sincères à mon promoteur, Christophe De Vleeschouwer. Merci pour la confiance que tu m'as accordée tout au long de ce parcours. Ta bienveillance et ton écoute m'ont été précieuses, tout comme tes commentaires avisés et les nouvelles perspectives que tu as su m'apporter. Ta disponibilité et ton optimisme ont fait de ton bureau un espace où trouver une oreille attentive.

Je souhaite également remercier chaleureusement les professeurs John Lee, Pierre Geurts, Laurent Francis, ainsi que Christophe Chaudy, pour avoir accepté de faire partie de ce jury. J'ai eu le plaisir particulier de bénéficier de l'encadrement académique de John, depuis mon mémoire jusqu'à la finalisation de cette thèse, et j'en suis sincèrement reconnaissante.

Tout au long de mon parcours, j'ai eu la chance de côtoyer des collègues formidables qui ont rendu chaque journée stimulante et agréable. Chronologiquement, je commencerai par remercier les *dauphins* (Amir, Antoine V., Béni, François, Gabriel, Maxime, Niels, Pierre-Yves, Simon, Stéphanie, Victor, Vincent, ...). Plus particulièrement, merci à Antoine d'avoir partagé le même bureau et le même appartement pendant plusieurs années. Deux lieux intrinsèquement liés pendant les périodes de confine-

ment ;-)) Merci également à tous les autres membres de la CDV team qui se sont plus sentis *sauterelles* que *dauphins* (Abolfazl, Alexandre, Antoine L., Baptiste, Benoît, Florine, Gilles, Mathieu, Vladimir, ...). Je tiens à remercier Alexandre en particulier, pour son humour débordant qui a rythmé notre quotidien avec ses blagues imprévisibles.

Je ne pourrais oublier mes collègues d'autres équipes, qui m'ont donné envie de venir travailler avec le sourire. Merci à tous les membres du *bureau* et *ses sympathisants* (Brieuc, Clémentine, Eléonore, Karim, Margo, Nico, Noémie, Otilie, Quentin, Roselien, Tom, ... et Loïc). Entre les jeudis pastis, les visites en Wallonie, les repas improvisés et les pauses café, ces moments partagés ont largement embelli le quotidien.

Une part importante de la vie d'assistant se passe en cours. Merci aux collègues avec qui j'ai co-encadré de nombreuses séances (Charles, Damien, Edouard, Jérôme, Simon, ... les autres ayant déjà été mentionnés), ainsi qu'aux professeurs qui m'ont permis d'amener un aspect personnel dans les séances d'exercices (André, Christophe C., Laurent, Luc, Michel, Olivier, Thomas, ...). Un grand merci également à Claude et Jérôme qui, en tant que responsable de la commission ELEC, apportent un soin particulier à la qualité de notre enseignement.

Je voudrais également exprimer ma gratitude envers les membres du PAT, qui nous épaulent au quotidien dans les tâches à réaliser (Anne, Brigitte, Christel, Emilie, François, Isabelle, Marie-Hélène, Nathalie, Patricia, Rémi, ...). Merci également à tous les autres, dont le travail était éloigné mais dont les tasses à café étaient souvent proches (Beni, Christian, Miloud, Nicolas, Patrick, Souley, Valeriya,...).

Enfin, sur une note plus personnelle, je tiens à remercier mes amis et ma famille pour les moments passés en dehors du travail. Un grand merci à mes amis du badminton club de Limal pour les matchs intenses et les troisièmes mi-temps riches en émotions. Une pensée spéciale va aux musiciens de la fraternité grass band et de la royale harmonie la pelouse (ou peut-être est-ce l'inverse ?) pour les concerts, carnivals, festivals et voyages à venir. Sans vous, j'aurais eu beaucoup moins de très bonnes raisons de ne pas travailler.

Contents

List of figures	ix
List of tables	xi
Acronyms and notations	xiii
1 Introduction	1
1.1 What is anomaly detection?	4
1.1.1 <i>Unsupervised problem formulation</i>	5
1.1.2 <i>Overview of unsupervised approaches</i>	6
1.2 Why is this work different?	11
1.3 Contributions and thesis outline	14
2 Background	19
2.1 What are the characteristics of an ideal dataset?	20
2.2 Dataset for unsupervised anomaly detection	21
2.2.1 <i>Autonomous driving</i>	21
2.2.2 <i>Medical imaging</i>	22
2.2.3 <i>Industrial imaging</i>	22
2.3 Performance evaluation	25
3 Detection in the original image space	29
3.1 Related work	31
3.2 Skip connections and synthetic data augmentation	34
3.2.1 <i>Network architecture</i>	34
3.2.2 <i>Stain: a synthetic corruption model</i>	35
3.3 Pipeline description	37
3.4 Quantitative and qualitative results	38

3.5	Comparative study of corruption models	41
3.6	Exploring operational boundaries and working hypotheses . . .	45
3.6.1	<i>Property 1 - Low contrast defects</i>	45
3.6.2	<i>Property 2 - Concept of normality</i>	46
3.7	Summary and fundamental properties	50
4	Detection with internal representations	53
4.1	Related work	54
4.2	Activation tensors visualization	55
4.2.1	<i>Notations</i>	55
4.2.2	<i>Defect propagation across layers</i>	57
4.3	Activation-level anomaly detection and localization	59
4.3.1	<i>Metric definition</i>	59
4.3.2	<i>Pipeline</i>	61
4.3.3	<i>Quantitative and qualitative results</i>	63
4.4	Exploring operational boundaries and working hypotheses . . .	67
4.4.1	<i>Property 1 - False positives in the background</i>	68
4.4.2	<i>Property 2 - Missing object</i>	71
4.5	Summary and fundamental properties	72
5	Detection in a multi-level quantized space	75
5.1	Background on discrete latent representations	77
5.2	Related work	79
5.3	Image projection in a multi-level quantized space	82
5.3.1	<i>Network architecture with quantization modules</i>	83
5.3.2	<i>Space properties</i>	84
5.4	Anomaly detection in the multi-level quantized space	90
5.4.1	<i>Segmentation network</i>	90
5.4.2	<i>Pipeline description</i>	92
5.4.3	<i>Quantitative and qualitative results</i>	96
5.5	Exploring operational boundaries and working hypotheses . . .	99
5.5.1	<i>Property 1 - Empty segmentation masks</i>	99
5.5.2	<i>Property 2 - Impact of the codebook size on the segmentation</i>	102
5.6	Summary and fundamental properties	108
6	Conclusions and perspectives	111
6.1	Underlying assumptions	113
6.2	Key observations	114
6.3	Improvement and perspectives	115

6.4 Final words 117

Appendices 119

A MVTec AD dataset 119

B Impact of skip connections on internal representation 121

C Variational inference 125

D Normal and defective prototypes distributions 127

Author’s publication list 131

Bibliography 133

List of figures

1.1	Application of deep learning in computer vision for anomaly detection on a production line	4
1.2	Unsupervised anomaly detection methods	6
1.3	I-AUROC with state-of-the-art methods	12
2.1	Overall picture of the MVTec AD dataset	24
2.2	Performance assessment with anomaly maps and anomaly scores	26
2.3	Statistics of defective and normal images/pixels in the MVTec AD dataset	27
3.1	AESc architecture and Stain synthetic noise model	36
3.2	Pipeline: AESc	37
3.3	Results with AESc	39
3.4	Impact of the synthetic noise model on AESc	44
3.5	Property 1 - Low contrast defects	46
3.6	Property 2 - Concept of normality	48
3.7	Property 2 - Distribution of screw background	48
3.8	Property 2 - Background removal for the screws	49
4.1	Double inference mechanism	56
4.2	Difference of activation tensors on a bottle	58
4.3	Building block of an Out-of-Distribution Map	60
4.4	Pipeline: OoD	62
4.5	Results with OoD	64
4.6	Relation between AESc and OoD scores	67
4.7	Property 1 - Relation between AESc and OoD scores	69
4.8	Property 1 - False positives detections in the background	70

4.9	Property 2 - Detection of a missing object	71
5.1	Quantization module	79
5.2	q-AESc architecture	83
5.3	Prototype distribution in the quantized space	86
5.4	Reconstruction of AESc vs q-AESc: pill	88
5.5	Reconstruction of AESc vs q-AESc: hazelnut	89
5.6	$(\text{Seg})_q$ architecture	91
5.7	Pipeline: AESc + $(\text{Seg})_q$	95
5.8	Results with AESc+ $(\text{Seg})_q$	97
5.9	Misclassified defective regions by $(\text{Seg})_q$	101
5.10	Impact of the codebook size K on the segmentation produced by $(\text{Seg})_q$ on the wood category	104
5.11	Impact of the codebook size K on the segmentation produced by $(\text{Seg})_q$ on the wood category	105
B.1	Impact of skip connections on internal representations: cable . .	123
B.2	Impact of skip connections on internal representations: carpet .	124
C.1	VAE architecture	125
D.1	Normal/defective prototypes distributions: cable	128
D.2	Normal/defective prototypes distributions: metal nut and tile .	129

List of tables

1.1	Advantages and disadvantages of the four categories of unsupervised anomaly detection methods	10
3.1	Results: AESc	40
3.2	Impact of synthetic noise model on AESc results	43
4.1	Results: OoD	65
5.1	Results: AESc + (Seg) _q	98
5.2	Impact of codebook size on AESc+(Seg) _q results	107
A.1	Details of the MVTec dataset	120

Acronyms and notations

Performance metrics

AUROC	Area Under the Receiver Operating Curve	(p. 25)
AUPRC	Area Under the Precision-Recall Curve	(p. 26)
I-AUROC	Image-wise Area Under the Receiver Operating Curve	(p. 27)
P-AUPRC	Pixel-wise Area Under the Precision-Recall Curve	(p. 27)

Networks and loss functions

AE	Autoencoder network	(p. 34)
AESc	Autoencoder with Skip connections network	(p. 34)
OoD	Out-of-Distribution	(p. 59)
VAE	Variational AutoEncoder	(p. 77)
VQ-VAE	Vector Quantized-Variational AutoEncoder	(p. 78)
q-AESc	AESc with quantization module	(p. 83)
q-Seg	Segmentation network in the quantized space	(p. 90)
MSE	Mean Squared Error	(p. 31)
BCE	Binary Cross Entropy	(p. 92)
KL	Kullback-Leibler divergence	(p. 77)
IoU	Intersection over Union score	(p. 100)

Operators on networks representations

$\mathbf{x}$	input image	(p. 56)
$\hat{\mathbf{x}}$	normal reconstructed version of $\mathbf{x}$	(p. 56)
$l_i(\cdot)$	Activation tensors	(p. 56)
$\Delta l_i(\mathbf{x}, \hat{\mathbf{x}})$	Difference of activation tensors	(p. 56)
$\mu(\cdot)$	Mean of activation tensors along channel axis	(p. 57)
$[\Delta l_{w_i, h_i, c_i}(\mathbf{x}, \hat{\mathbf{x}})]_{\text{OoD}}$	Out-of-Distribution test at the voxel level	(p. 59)
OoDMAP	Out-of-Distribution MAP	(p. 59)
OoDR	Out-of-Distribution Ratio	(p. 60)

1

Introduction

THE INDUSTRIAL SECTOR has undergone a profound transformation over the past few decades, largely driven by the emergence of the Internet of Things (IoT). This technological advancement has resulted in the development of smart factories, where advanced sensors, embedded software, and robotics are integrated into the production process. These innovations facilitate continuous data collection and analysis, enabling real-time decision-making and process optimization. Consequently, industries have witnessed a significant increase in automation, predictive maintenance, self-optimization of processes, and overall efficiency improvements. This digital transformation is the foundation of the latest and ongoing industrial revolution known as "Industry 4.0" [1].

In order to fully comprehend the implications of Industry 4.0, it is imperative to contextualize it within the broader historical landscape. The first revolution (Industry 1.0) began in the latter half of the eighteenth century with the development of mechanics, the exploitation of coal, and the introduction of the steam engine. This era constituted the foundation for the contemporary industrial landscape. The second industrial revolution (Industry 2.0) emerged in the 1870s, bringing with it the widespread adoption of electricity, the development of transportation infrastructure, and the advent of mass production, which significantly reduced production costs. By the 1970s, the third evolution (Industry 3.0) had taken shape, characterized

by the introduction of new information technologies, including electronics and telecommunications, which further advanced the automation of production processes. The term Industry 4.0, first introduced in 2011 [2], promotes the concept of intelligent, integrated, and unmanned factories.

Industry 4.0 represents a paradigm shift in how resources are scheduled and organized, relying on dynamic and integrated systems that manage the entire lifecycle of a product. Nevertheless, the implementation of these state-of-the-art technologies gives rise to a number of challenges. A key issue is the processing and analysis of the vast volumes of data generated by connected sensors. This thesis addresses one of the critical aspects of Industry 4.0, namely the **automatic computer vision-based quality control for real-time anomaly detection**. This refers to techniques that are used to analyze the quality of a product based on a visual inspection of the sample. The objective is to detect deviations from a previously established baseline, which represents the standard behavior that is expected in the production line. By doing so, the system can identify anomalies such as pattern alterations, structural deviations, or other visual domain irregularities.

The demand for automated anomaly detection systems in the industrial sector is significant, as these systems can address many of the limitations associated with manual inspection processes, including low accuracy, poor real-time performance, subjectivity, and high labor intensity [3]. The potential applications of computer vision-based anomaly detection are numerous and diverse, encompassing areas such as metallic surface inspection [4, 3, 5, 6], fabric inspection [7, 8, 9], railway track inspection [10], food inspection [11], and luggage inspection [12]. The nature of the defects can vary considerably, manifesting as scratches, stains, structural deformations, or changes in color and/or texture. In addition, each application domain can present its own specific challenges. For example, luggage inspection often involves X-ray imaging, while metal surface inspection is particularly sensitive to variations in illumination and light reflection. Consequently, the methods employed must be versatile and adaptable to different manufacturing contexts.

In this work, we focus on the problem of anomaly detection, where the decision-making process is based exclusively on the acquisition of a single image per sample. Despite the intrinsic complexity and high dimensionality of images, our approach emphasises the design of a simple yet robust

method that is agnostic to the specific type of product being analyzed. Deep learning, with its capacity to learn directly from data, offers a promising solution to these challenges, eliminating the need for labour-intensive, rule-based systems.

This manuscript introduces **three deep learning methods for anomaly detection that rely on a single image of the product**. The first method is based on the hypothesis that a network trained to reconstruct images of normal products will struggle to accurately reconstruct defective structures that may emerge during the production process. Regions that are poorly reconstructed are identified as defective, with this information being derived from the reconstruction error, *i.e.* the absolute value of the difference between the original image and its reconstructed normal version. One of our key contributions is **the enhancement of reconstructed image quality**, which improves the accuracy of defect detection. Additionally, we propose two alternative approaches to the reconstruction error for identifying defective areas in the image, which operate directly on the **latent representations learned by the network**. We conducted an in-depth analysis of the properties of these latent spaces to comprehensively support the proposed anomaly detection metrics.

While deep learning models have demonstrated remarkable performance across various tasks, this often comes at the expense of transparency. The intrinsic level of complexity and the millions of parameters within these models tend to make them opaque, preventing users from tracing and understanding the steps the algorithm takes to arrive at its final decision. This perception of deep learning models as "black boxes" [13] underscores the critical need for enhanced explainability to facilitate their widespread adoption. To address this need, we employ specialized visualization techniques and conduct comprehensive analyses of the limitations of our proposed approaches on a diverse set of representative real-world examples. This systematic analysis provides intuitive justifications for the underlying decision mechanisms that produce effective solutions for the targeted task. By doing so, we aim to bridge the gap between academic methodologies and industrial requirements, thereby promoting the broader adoption of deep learning methods. It is our conviction that an intuitive understanding of deep learning models is of paramount importance for making these approaches accessible to a non-expert audience.

1.1 What is anomaly detection?

Anomaly detection can be defined as the task of identifying all samples that deviate from the normal distribution of data, which is also referred to as the *clean* data set. In the context of visual inspection of production lines, the objective is to detect all defective samples that arise due to unexpected behaviors in the manufacturing process. This applicative case of deep learning is illustrated in Figure 1.1. The key components of the anomaly detection process include:

Real-time data collection. Production lines are equipped with cameras that continuously capture data to monitor the ongoing production process. In this case, focus is on images of the products being manufactured.

Model Training (Offline). Prior to deploying the system, a deep learning model is trained using images collected before production starts, hereby enabling the model to learn the distribution of normal images.

Anomaly Detection (Online). During the actual production process, images of the products are continuously captured and processed by the model. The model compares each image to the learned normal distribution, and any significant deviations (*e.g.*, unusual patterns or defects) are identified as anomalies.

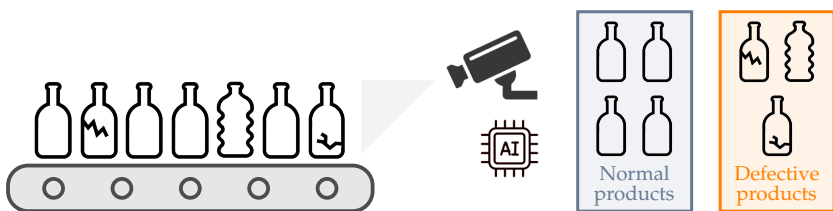


Fig. 1.1 Application of deep learning in computer vision for anomaly detection on a production line.

1.1.1 Unsupervised problem formulation

When considering the task of training a deep neural network to distinguish between normal and defective products, it may seem natural to formulate the problem as a conventional supervised binary classification task. In this scenario, the development of the model would require the construction of a representative labeled training dataset, comprising both normal and defective images with their corresponding annotations. For effective training, it is crucial that the dataset accurately reflects the diversity of scenarios that the model will encounter in production, ideally including an equal proportion of normal and defective samples. However, creating this dataset is challenging for two main reasons:

Scarcity of abnormal events. Defective events are rare by definition since they occur due to an unwanted behavior of the manufacturing process. This scarcity of defective samples frequently results in unbalanced datasets [14]. Furthermore, the financial and temporal costs associated with capturing a sufficient number of defective examples can be prohibitively high, particularly in well-controlled manufacturing environments where defects are minimized.

Incompleteness of defect representation. While the normal behavior of the manufacturing process can be well-defined, it is practically impossible to exhaustively list all potential failure modes. Consequently, this partial sampling of the defective sample distribution introduces uncertainty when the model encounters an unseen failure mode.

In order to circumvent the constraints related to data acquisition in manufacturing settings, the supervised classification problem is redefined as an unsupervised task. In this unsupervised paradigm, **a training set comprising solely normal images is collected**. The objective of the deep learning model is not to directly solve a binary classification problem. In essence, the model is trained to learn the distribution of normal data, effectively capturing the underlying characteristics and patterns that define normality. Once the model has established this baseline, it can then be applied to new, unseen data, identifying anomalies as those samples that significantly deviate from the learned normal distribution. This approach allows for the detection of defects without the need for an exhaustive catalog of potential failure modes, making it more adaptable to real-world manufacturing environments where defects are rare and diverse.

1.1.2 Overview of unsupervised approaches

Multiple problem formulations exist in order to address unsupervised anomaly detection in industrial images, as listed in several surveys [14, 15, 16, 17, 18, 17, 19]. Before going into the details of our proposed framework, we propose a structured and high-level comparison of significant deep learning contributions in the domain. We classify these contributions into four categories: distance-based, reconstruction-based, density-based and self-supervised learning-based methods. For each category, we briefly describe the primary innovation directions that have led to related contributions, building on the initial groundbreaking ideas introduced in the aforementioned papers. Figure 1.2 illustrates and summarizes the main hypotheses underlying each category.

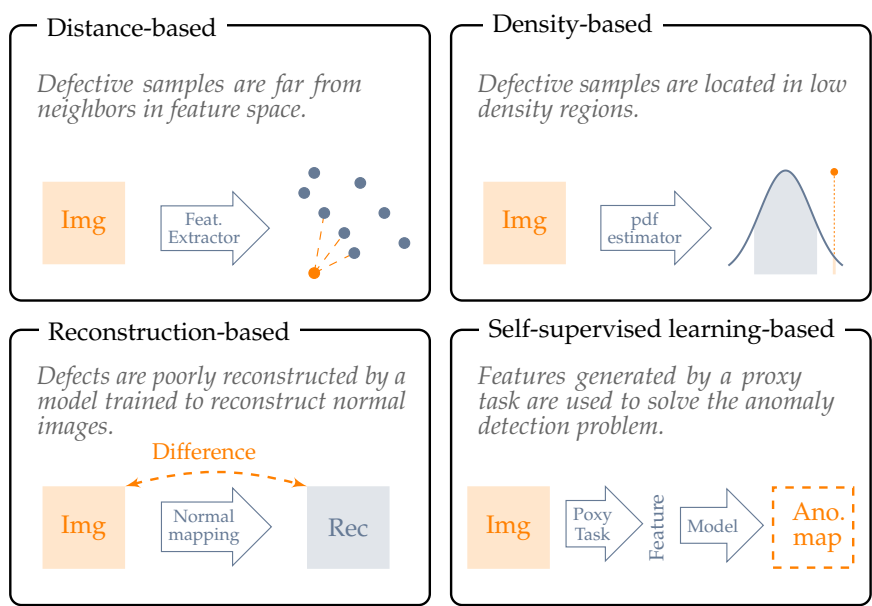


Fig. 1.2 Illustration of four categories of unsupervised anomaly detection methods, each depicted with a schematic pipeline and the corresponding underlying hypotheses.

Distance-based methods

In this framework, images are mapped into an appropriate feature space in which anomalies are detected based on a distance metric. No assumption on the data distribution is made. One disadvantage of these approaches is that it is challenging to obtain a pixel-wise localization of defective structures.

Napoletano et al. [20] employed a pre-trained ResNet-18 [21] to create a dictionary from the extracted feature tensors. An image is classified as defective if the average Euclidean distance between its feature tensor and its m nearest neighbors in the dictionary exceeds a predefined threshold. Later, **SPADE** [22] enhanced this approach by using a multi-resolution feature pyramid, improving localization precision. **Deep SVDD** [23] maps training images into a hypersphere with a constrained volume during training. Embeddings that fall outside this hypersphere are considered as defective.

Reconstruction-based methods

These approaches employ a network trained to reconstruct the normal version of a test image, which may be either clean or defective. The underlying assumption is that a network trained exclusively on normal images will be unable to accurately reconstruct anomalous structures, as these have not been encountered during training. Consequently, by comparing the input image with its reconstructed version, defective structures can be identified and localized.

Outstanding methods following the reconstruction-based paradigm are **AE-L2** [24] and **AE-SSIM** [25], which use vanilla autoencoders trained to minimize the mean squared error and the structural similarity, respectively. **AnoGAN** [26] and **Skip-GANomaly** [27] proposed to use generative adversarial network. However, GAN suffers from mode collapse which can lead to numerous false positive detection in this context. Despite the regularization induced by the bottleneck, such approaches still reconstruct perfectly defective structures instead of replacing them with clean content. To limit generalization ability, **MemAE** [28] added a memory bank to the bottleneck, whose atoms are trained on normal images. During inference, only a limited amount of those learned atoms are used to represent the latent vector to limit reconstruction ability of the network.

Self-supervised learning-based methods

The objective of self-supervised learning is to generate representative visual features from the training data by employing a model that has been trained on an auxiliary unsupervised task. These proxy tasks provide a means of learning meaningful representations without the necessity of labeled data. The aforementioned generated features are then employed in order to address the target tasks of anomaly localization and detection. The proxy tasks employed to train the model may vary considerably, encompassing techniques such as image inpainting, wherein the model is trained to fill in missing portions of an image; position prediction, in which the model infers the spatial configuration of image patches; and contrastive learning, which involves distinguishing between similar and dissimilar pairs of images.

DRÆM [29] utilizes a U-Net-like architecture [30] to reconstruct clean version of input images corrupted with synthetic noise. The original and reconstructed images are fed into a segmentation network, predicting a binary mask to identify defective structures. This inclusion of a segmentation network distinguishes DRÆM from traditional reconstruction-based approaches, as its detection mechanism does not rely on classical reconstruction residuals but rather on segmentation. Building on the concept introduced by the Deep SVDD [23] approach, **Patch SVDD** [31] maps patches extracted from the input image into multi-modal hyperspheres. Additionally, a classifier network is trained within this embedding space to predict the relative positions of the extracted patches in the image. This enhancement facilitates the localization of defective regions, thereby enabling anomaly localization. Another category of methods employs knowledge distillation. **Uninformed Students** [32] involves training multiple student networks to replicate the output of a teacher network on normal data, where the teacher's feature extractor is pre-trained on a large dataset of natural images. Anomalies are localized when the outputs of the student networks deviate from the teacher network and exhibit high variability among the student outputs. However, this approach is time-intensive due to its multi-patch nature. Contrastive learning has been used by **Yoa et al.** [33] by generating pair of clean and locally augmented images. A score of anomaly is computed based regression errors for each layer (between the mean of all training tensors and a test image) as well as the uncertainty in each layer, evaluated with MCDropout [34].

Density-based methods

The fundamental assumption of density-based methods is that normal data are situated in a dense region of a given embedding space. Conversely, abnormal data occupy low-density areas. By estimating the probability density function of the normal data, one can estimate the probability of a test sample belonging to the clean distribution.

There are various methods depending on the distribution assumptions. **PaDiM** [35] used pretrained backbones on ImageNet [36] to extract patch embedding. Then the probabilistic distribution of the the normal class is approximated by a multivariate Gaussian distribution. Normalizing flows [37] are generative invertible networks used to learn a bijective mapping between two distributions, here between a latent distribution and determined prior distribution. **DifferNet** [38] used a pretrained AlexNet[39] on ImageNET to build the latent space. The obtained latent distribution is mapped into a normal distribution which allows to estimate the likelihood of a sample to belong to the normal distribution. To obtain a localization of defective structures, they backpropagate the likelihood score towards the input image space, as done with grad-CAM [40]. **Abati et al.** [41] projected images using a classical encoder into a latent space whose prior distribution is estimated with an autoregressive network. To compute the abnormality of a sample, they combine information from the likelihood of the latent vector and the reconstruction error obtained after resampling the latent variable in the normal distribution. In a similar spirit, **Wang et al.** [42] projected images in a quantized latent space by the use of a VQ-VAE [43]. The discrete latent representation is then approximated with a PixelSNAIL [44] autoregressive network. **VT-ADL** [45] proposed the use of a vision transformer encoder to extract a feature vector out of patches of an image. The obtained embedding is fed to a decoder to obtain a clean reconstruction of the input image and also to a Gaussian mixture density network which is used to localize the anomaly.

General remark

There is no clear subdivision of anomaly detection methods in the literature. We have proposed here four categories that reflect what we believe to be the main trends in the field.

Summary

Each of the four categories of unsupervised anomaly detection methods discussed earlier has distinct advantages and disadvantages, which are summarized in Table 1.1. As expected, these methods often involve a trade-off between the complexity of the formulation and the overall performance. The table provides a concise comparison, highlighting the key strengths and limitations inherent to each approach, thereby offering insights into their applicability across different contexts.

Table 1.1 Advantages and disadvantages of the four categories of unsupervised anomaly detection methods.

Advantages	Disadvantages
Distance-based approaches	
<i>Defective samples are far from neighbors in feature space.</i>	
✓ Feature space can be obtained with a pretrained backbone.	✗ Lacks a pixel-wise localization of defects.
✓ Simple to implement.	✗ Performance is sensitive to the choice of the backbone architecture.
	✗ Entire image is reduced to a single vector, potentially missing subtle defects.
Reconstruction-based approaches	
<i>Defects are poorly reconstructed by a model trained to reconstruct normal images.</i>	
✓ Simple and intuitive problem formulation.	✗ Overgeneralization may result in the reconstruction of defective regions.
✓ Training and network architecture are straightforward.	✗ Difficulties arise in detecting defects with similar intensity to their surroundings.
Density-based approaches	
<i>Defective samples are located in low density regions.</i>	
✓ Captures rich semantic information.	✗ Lacks a pixel-wise localization of defects.
✓ Accurate estimation of data likelihoods for normal samples.	✗ Requires expensive training.
	✗ Model design may necessitate significant refinement.
Self-supervised learning-based approaches	
<i>Features generated by a proxy task are used to solve the anomaly detection problem.</i>	
✓ Captures rich semantic information.	✗ Indirect and complex problem formulation.
✓ Diverse problem formulations offer robustness, potentially addressing shortcomings of other methods.	✗ Requires expensive training.
	✗ Model design may necessitate significant refinement.

1.2 Why is this work different?

In the deep learning literature, the predominant research direction focuses on optimizing performance with the ambition of achieving the holy grail of 100%. A notable example of this performance race is the benchmarking history over the MNIST dataset [46]. Originally introduced in 1998, MNIST is somehow the "hello world" of computer vision datasets. Current state-of-the-art on MNIST ranges accuracy around 99.84% [47, 48], and even 99.87% [49]. Given that human-level performance on this task is estimated to be around 99.80% [50] (which corresponds to 20 errors out of 10,000 test images), deep learning algorithms have reached the top level.

The MVTec AD dataset [51], which is used in this work to evaluate an anomaly detection model, is no exception to this performance race. As presented in Figure 1.3, image-wise anomaly detection performance¹ is approaching that of a perfect classifier denoted by the $y = 1$ horizontal line. To support this claim, four recent contributions achieving the highest anomaly detection performance are introduced. Recently, **EfficientAD** (2024) [52] proposed a student–teacher approach. In this method, a student network trained exclusively on normal images is expected to predict the features computed by a teacher network. The distance between student and teacher outputs is used as anomaly detection metric. **SimpleNet** (2023) [53] relied on a pretrained feature extractor and a trained feature adaptor network to project images in a custom latent space. A discriminator network is then trained to distinguish between normal and noisy embeddings. **PatchCore** (2022) [54] used a memory bank trained on normal patches to evaluate the abnormality of a given patch as a distance measure between the patch embedding and the closest neighbors in the dictionary. Finally, **DDAD** (2023) [55] used a diffusion model to obtain a clean reconstruction of an input image. Anomaly score is computed by combining reconstruction error and distances for some activation layers generated from the input and the train images.

¹The methods under consideration produce a score that reflects the probability of an image being classified as defective. The performance of these methods is evaluated using the Area Under the Receiver Operating Characteristic Curve (AUROC), which assesses the model's ability to distinguish between normal and abnormal samples by plotting the true positive rate against the false positive rate across various threshold settings. The AUROC values range from 0 to 1, where a value of 1 indicates perfect classification, and a value of 0.5 suggests performance equivalent to random guessing. Further details are provided in Section 2.3.

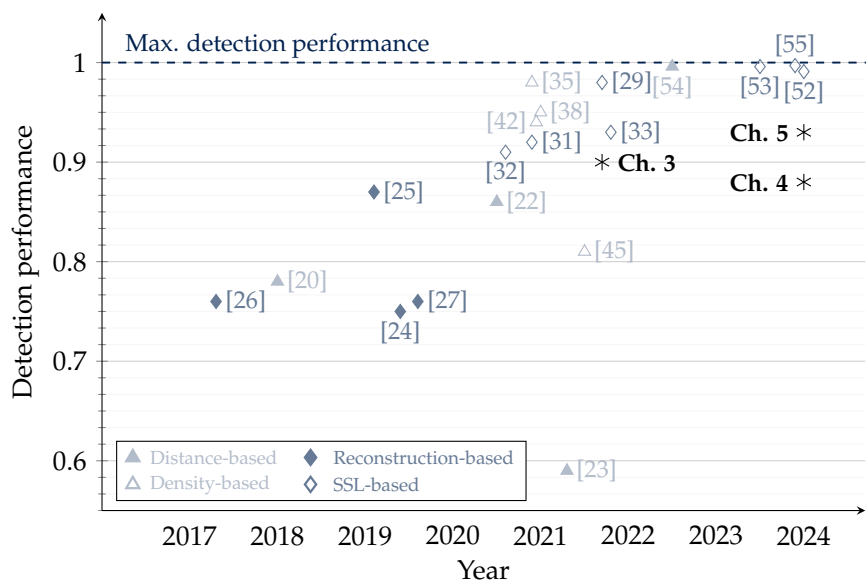


Fig. 1.3 Mean anomaly detection performance over all image categories of the MVTec AD dataset for both methods presented in Section 1.1.2, and more recent state-of-the-art methods associated to top accuracies (blue markers). Performances of the three methods presented in following chapters are shown as comparison (black stars).

Given the impressive performance levels recently achieved in anomaly detection, it may be tempting to assume that the problem has been solved. The values shown in Figure 1.3 suggest that further improvements will be marginal and probably difficult to achieve. First and foremost, it is important to emphasize that the MVTec AD dataset only addresses a limited set of 15 specific use cases within the broader context of industrial anomaly detection. While it has become a standard benchmark for evaluating anomaly detection models, the true scope of potential applications for anomaly detection is much broader and more difficult to delineate. MVTec AD focuses on scenarios that are particularly well defined and structured. Specifically, the dataset encompasses situations where:

- (i) images are consistently well aligned;
- (ii) objects are centered in the image with uniform dimensions;

- (iii) any source of data drift (e.g., changes in illumination, contrast, or color) is discarded during acquisition;
- (iv) the number of normal samples available for training is relatively large, ranging from 60 to 391 images.

In real world settings, such controlled conditions are rarely met. Environmental variability, shifts in data distribution, and diverse object presentations often make it challenging to directly apply state-of-the-art methods, which tend to rely on complex architectures and require extensive hyperparameter tuning to replicate the results achieved in controlled settings. In industrial environments, where efficiency and practicality are paramount, the extensive work required to adapt these methods to a particular use case is prohibitive.

Motivated by these challenges, our research is directed toward developing anomaly detection methods that provide intuitive, easily reproducible workflows grounded in well-defined hypotheses. We aim to develop approaches where each contribution is rigorously validated through ablation studies and networks that are robust in terms of training stability. We focus on simplicity and transparency to ensure that the methods we develop are easy to implement and adapt in practical contexts. We have therefore focused on classical autoencoder architectures, which are known for their training stability, a key component when fine-tuning is required to adapt a method to a new situation. If the most explored research direction emphasizes performance gain, our choice to opt for stability comes at the expense of a slight performance drop as indicated by Figure 1.3. Moreover, we have deliberately kept the frameworks as simple as possible. We have intentionally avoided the use of post-processing techniques, such as morphological operations, and have refrained from incorporating task-specific prior information (e.g., dataset-dependent data augmentation strategies). This approach ensures that any observed differences in results are due to the core contributions of the methods themselves, and not to more complicated hyperparameter tuning processes. The care taken to ensure a reliable comparison between the different approaches is, in our view, an essential element of this study.

In order to convince the reader of the relevance of our contributions, it is crucial to maintain a high degree of transparency in our results analysis, ensuring that the effectiveness of the methods is clearly demonstrated. By

developing and analyzing multiple visualizations of the internal representations of our networks, we aim to propose reliable and intuitive methods for anomaly detection and localization tasks. Additionally, for each method presented in this manuscript, we formulate specific properties that underscore the underlying working hypotheses of the proposed approach. These properties are illustrated using various images from the dataset, taking advantage of the diversity of samples and defects to provide a detailed examination of both the strengths and limitations of our methods. This comprehensive analysis is critical to fostering a deep understanding of the algorithms, enabling even non-experts to select the most appropriate method to address particular problems with known specificities.

By emphasizing an intuitive understanding, reliable and fair comparisons, we aspire to align with industry requirements and convince a non-expert audience of the efficiency of our approaches. Specifically, for each proposed method, we will explore the following questions:

Question 1. What are the foundational assumptions underlying the proposed methodologies?

Question 2. From a comprehensive analysis of success and failure instances, can we predict the scenarios in which our methods are most effective or ineffective?

Question 3. What strategies can we recommend for enhancing our methods, particularly in cases where they do not perform as expected?

1.3

Contributions and thesis outline

This thesis proposes three new frameworks to address the problem of unsupervised anomaly detection for industrial vision, each of which is developed in a separate chapter. The model introduced in chapter 3 serves as a basis for the further developments in chapters 4 and 5. All three frameworks differ in the space in which the detection is performed, namely, in order of appearance: the image space, the space built on internal network representations, and a multi-level quantized space. The organization of these contributions in the present manuscript is explained below.

Chapter 2 - Background

This introductory chapter provides a formal **overview of the dataset** used throughout this thesis and outlines the **performance metrics** used to evaluate the effectiveness of our models in anomaly detection. We make a clear distinction between anomaly detection, which focuses on distinguishing between normal and defective images, and anomaly localization, which aims to identify the specific regions within an image that contain anomalies. Each of these tasks is evaluated using different performance metrics tailored to their respective goals.

Chapter 3 - Detection in the original image space

This chapter introduces a classical reconstruction-based framework that relies on the reconstruction residual between an input image and its normal reconstruction to localize defective structures. The effectiveness of this approach is directly related to the quality of the reconstructed images. To **improve the reconstruction quality**, we propose adding skip connections to an autoencoder and training with synthetically corrupted images. While these modifications significantly improve the reconstruction performance, the approach still struggles to detect **defects that contrast poorly with their surroundings**. For this type of defect, even if the model achieves a perfect replacement of defective structures with normal ones, the residual is still low due to the global intensity similarity between normal and defective structures. This limitation is inherent to reconstruction-based methods and necessitates alternative detection mechanisms.

We argue the choice of a reconstruction-based approach for the intuitiveness of the problem formulation, which is a curcial aspect to help non-experts from the industrial world to have insight into the reasoning of the network. As can be seen in Figure 1.3, reconstruction-based approaches are not the best performing, indicating that there is room for improvement. However, we believe in the robustness and intuitive formulation of the framework to achieve widespread use of the method.

Chapter 4 - Detection with internal representations

This chapter addresses the challenge of detecting defects that contrast poorly with their surroundings, an aforementioned limitation commonly observed in reconstruction-based approaches. This limitation is primarily due to

the detection mechanism operating in the original image space. However, analysis of internal network representations suggests that this problem is mitigated within the network. Motivated by this observation, we present a distance-based approach that exploits the representations of the network trained in the classical reconstruction-based framework. By performing **binary tests at the voxel level**, we quantify the anomaly probability of an image based **on the activation tensors**. This approach not only preserves the ability of the baseline framework to detect defects, but also increases sensitivity to subtle deviations.

Chapter 5 - Detection in a multi-level quantized space

This chapter further extends the reconstruction-based framework by utilizing our initial pretrained network to generate **quantized latent representations**, which are then processed by a **segmentation** network. The resulting segmentation masks are combined with the original reconstruction-based anomaly maps, enhancing the initial framework through a self-supervised problem formulation. This integration offers a more robust approach to anomaly detection and localization, leveraging the strengths of both reconstruction-based and segmentation methodologies.

How do we emphasize a comprehensive understanding of the methods?

In Section 1.2, we highlighted that the primary goal of our methods is not to develop a new framework that outperforms state-of-the-art models, given that the potential performance gains are marginal and increasingly difficult to achieve. Instead, our focus lies in enhancing a traditional reconstruction-based framework by integrating detection mechanisms in alternative spaces beyond the image domain.

Why pursue this approach? Reconstruction-based frameworks offer an intuitively straightforward formulation of the anomaly detection problem: the model reconstructs a normal image from any given input, thereby providing an explanation of the decision-making process by demonstrating how the network interprets normal structures. However, the inherent limitations of detecting anomalies directly in image space underscore the need for alternative approaches, such as leveraging other spaces to identify defective structures. While these alternative spaces may add complexity to the method, especially for non-expert users, they provide a critical advantage

in addressing the limitations of image-space detection. Before presenting the results in Chapters 4 and 5, we first describe the characteristics of the spaces being introduced to make their use less opaque. In each technical chapter, we outline specific properties that illustrate the special use cases of our methods and clarify the limitations of the underlying assumptions.

Our work is guided by the development of methods that emphasize clear, well-understood mechanisms and require minimal hyperparameter tuning to achieve high performance. This approach is particularly important for industrial applications, where access to deep learning experts for fine-tuning is often limited. In such contexts, stable and robust methods that deliver reliable performance with minimal intervention are generally preferred over more complex approaches that may offer higher accuracy but require expert involvement for optimal tuning.

2

Background

THE LITERATURE on unsupervised anomaly detection encompasses a diverse range of applications, such as autonomous driving, medical imaging and industrial imaging, each characterized by its own unique requirements and challenges. The objective of this chapter is to provide a comprehensive analysis of the reasoning behind two critical aspects of our research:

- (i) the selection of the dataset;
- (ii) the choice of the performance metrics.

In the domain of industrial anomaly detection, the MVTec AD dataset [51] has established itself as the primary benchmark, largely due to the absence of a more consistent alternative. This dataset is highly regarded for its high-quality and diverse collection of industrial defect images, making it an essential tool for evaluating the effectiveness of anomaly detection methods within this context.

In regard to the performance metrics, our choices are directly influenced by the key characteristics of the MVTec AD dataset, which is representative of a typical use case of industrial anomaly detection. The first metric assesses image-wise detection, determining whether an entire image is classified as normal or defective. The second metric focuses on pixel-wise detection, evaluating the classification of individual pixels as either normal or defective.

2.1 What are the characteristics of an ideal dataset?

Prior to undertaking a comparison of unsupervised anomaly detection datasets, it is of interest to first consider the key characteristics that contribute to the effectiveness of an ideal dataset. According to the universal approximation theorem [56], artificial neural networks have the capability to approximate any continuous multivariate function. To achieve this approximation, neural networks are trained using example data, enabling the extraction of representative underlying patterns. The quality of the training data is crucial for proper optimization and essential to perform specific tasks effectively. A dataset may be deemed suboptimal if it exhibits any of the following characteristics:

Low quantity. Neural networks generally require large amounts of training data to recognize patterns accurately and make reliable predictions. Furthermore, the use of a larger test set is crucial for reducing variance during performance evaluation, ensuring that the model's performance is assessed in a more robust and representative manner.

Non-representative data. Beyond quantity, training data must be representative of the problem being addressed to ensure generalization in real use cases.

Inaccurate data. If the training data contains numerous errors, the algorithms will learn from these errors, leading to inaccurate predictions. Ensuring data accuracy is therefore essential for reliable model performance.

As discussed previously in Section 1.1.1, the process of collecting datasets for the evaluation of industrial anomaly detection systems presents significant challenges. The probability of defective samples is low, making it challenging to obtain a sufficiently large and diverse set of examples. Additionally, it is nearly impossible to guarantee that all potential failure modes that need to be identified as defects are captured within the dataset. In light of these challenges, it is evident that a meticulous evaluation of dataset quantity, representativeness, and accuracy is essential to ensure that the resulting models are both effective and reliable in real-world industrial applications.

2.2 Dataset for unsupervised anomaly detection

In this section, we provide an overview of datasets commonly used for unsupervised anomaly detection, with a primary emphasis on industrial applications. Additionally, we briefly consider the application of anomaly detection in the contexts of autonomous driving and medical imaging. Unlike industrial settings, where data is typically collected under controlled conditions, road images for autonomous driving are captured in uncontrolled environments, leading to significant data variability and often a lack of normal training data. In the domain of medical imaging, labeled datasets are particularly scarce due to the specialized expertise required for accurate annotation [57]. Furthermore, privacy concerns frequently restrict the availability of publicly accessible medical datasets. By highlighting the specific challenges associated with applying anomaly detection to these diverse domains, we underscore the importance of the quality and suitability of the dataset utilized in this manuscript.

2.2.1 Autonomous driving

A prominent application of unsupervised anomaly detection in autonomous driving is the identification of unexpected events on the road, such as the appearance of animals, lost tires, or construction equipment. The **Road Anomaly Dataset** [58], introduced in 2019, provides a benchmark for this task, consisting of 60 images with pixel-wise annotations of abnormal events. In this dataset, normal training images are synthesized using the pix2pixHD model [59], where unexpected instances are replaced with normal ones in the semantic label maps.

Another significant contribution to this field is the **RoadAnomaly21** dataset [60], released in 2021. This dataset contains 100 images, each with pixel-wise annotations of anomalies. The normal events in RoadAnomaly21 are derived from scenes captured in the Cityscapes dataset [61], providing a basis for comparison and further research in anomaly detection for autonomous driving scenarios.

2.2.2 Medical imaging

Unsupervised anomaly detection methods have been explored for identifying pathological structures in medical images. Potential applications include the detection of pathologies such as breast cancer, strokes, lung lesions, and skin diseases using appropriate imaging modalities [62]. However, no common public benchmark datasets are available for comparing the performance of different methods. As noted by Tschuchnig *et al.* [63], there is a negative bias in the field due to the comparison of new models against older ones, evaluated exclusively on the same dataset. The absence of a publicly available dataset hinders fair comparison among the proposed methodologies.

2.2.3 Industrial imaging

For an extended period, traditional classification datasets such as MNIST [46], CIFAR-10 [64] and ImageNet [39] have been employed due to the lack of a representative dataset for evaluating anomaly detection in industrial contexts. These classical datasets, each comprising over 50000 images across multiple classes, were employed by arbitrarily labeling one or multiple classes as outliers and maintaining the others as inliers [65, 23]. The objective during the testing phase was to differentiate between samples that were previously unseen and those that were known to be inliers. However, these datasets do not accurately reflect real-world anomaly detection scenarios, thereby rendering it unclear how well algorithms would perform in actual industrial applications.

One of the earliest datasets for unsupervised anomaly detection was introduced in 2007 at the **DAGM** (German Association for Pattern Recognition) symposium. This synthetic dataset [66] comprises surfaces and defects classified into ten categories of varying difficulty. Each category contains approximately 1000 normal and 150 defective textures. However, the synthetic defect generation process is relatively coarse, which limits the dataset's relevance. The **MT Defect** dataset [67] proposed in 2020 encompasses five distinct categories of metal defects: porosity, crack, wear, fracture, and non-uniformity, which is a consequence of the grinding process. The dataset comprises 1344 images of magnetic brick surface. A significant limitation of the MT Defect dataset is the restricted range of samples and defects included. In 2021, the **BeanTech Anomaly Detection (BATD)** dataset [45] was

introduced, featuring 2830 real-world images of three normal and defective industrial products, along with corresponding pixel-level defect annotation masks. Training set contain 400, 1000, and 399 training images, for each of the three considered products. Major drawback of the BTAD dataset is the limited variety of samples and defects considered in the dataset.

Several datasets focus on detecting defective structures in textures. For example, the **KolektorSDD** [68] comprises 399 images of electrical commutators, 52 of which have been annotated for microscopic fractions or cracks on the surface of the plastic embedding. In this dataset, a single defect type is the focus of analysis. The **KolektorSDD2** [69] comprises 2979 images of normal samples and 356 images of defective samples. Although the defects exhibit a range of appearances, the dataset encompasses a singular metal surface texture, which is nearly uniform across all samples. Another illustrative example is the **Fabric** dataset, which consists of approximately 2000 fabric samples divided into seven categories. For each category, 140 normal images and 20 defective samples are provided, representing a relatively modest number of defective samples.

The **MVTec AD** dataset was introduced in 2019 as the first large-scale and high-resolution dataset for the task of unsupervised anomaly detection. The dataset includes 5354 images across 15 categories, divided into 5 texture categories and 10 object categories, as represented in Figure 2.1. The training set consists solely of normal images, while the test set contains both normal and defective images, with ground truth binary masks indicating the defective regions provided for the latter. For further details about the number of images contained in each category, please refer to the Appendix A. Despite having a smaller number of images¹ compared to MNIST, CIFAR-10, and ImageNet, the MVTEC AD dataset offers greater representativity of real-world industrial anomaly detection scenarios. This enhanced representativeness mitigates the limitations posed by its smaller dataset size, aligning with real use cases where data collection is often challenging. This motivates our decision to consider this dataset in this manuscript.

¹The biggest image category of MVTEC AD is "hazelnut" with 501 images (391 for training, 110 for testing). The smallest image category of MVTEC AD is "toothbrush" with 102 images (60 for training, 42 for testing).

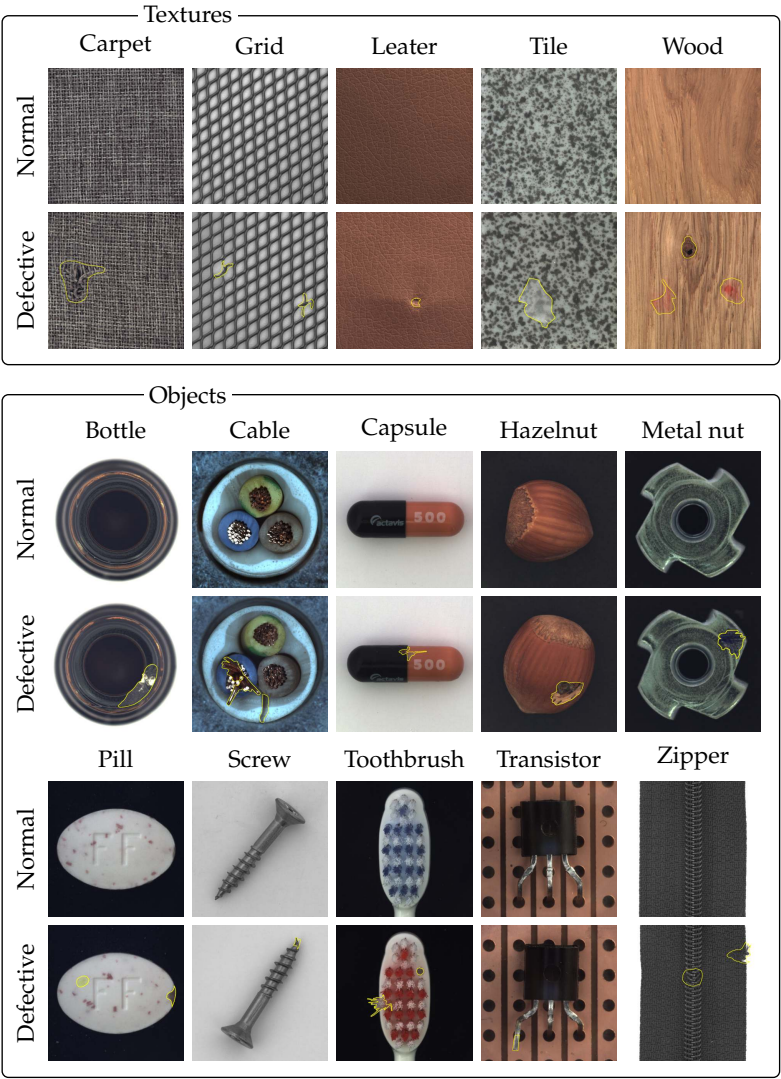


Fig. 2.1 Example of images of the MVTec AD dataset for each of the five texture categories and the ten object categories. For each, the top row shows a normal image while the bottom row shows a defective image. Defective structures are highlighted with the yellow contour obtained with the provided segmentation mask.

2.3 Performance evaluation

This manuscript presents three approaches to address the anomaly detection task. The objective of these methodologies is to construct an **anomaly map** that quantifies the likelihood of each pixel belonging to a defective region. The aforementioned anomaly map can be employed for the purposes of both anomaly detection and anomaly localization. **Anomaly localization** operates at the pixel level, with the objective of identifying all pixels within an image that are indicative of a defect. A pixel is deemed defective if its anomaly likelihood surpasses a predefined threshold. Conversely, **anomaly detection** operates at the image level, addressing the binary classification task of determining whether each test image is normal or defective. To this end, the anomaly map (2D image) is transformed into an **anomaly score** (scalar value) that represents the likelihood of an image being defective.

In order to perform the associated binary classification task, a threshold must be determined for both image-wise and pixel-wise detection. Since the optimal threshold value can vary depending on the application, we utilize the Receiver Operating Characteristic curve (ROC) and the Precision-Recall curve (PR) to present results in a threshold-independent manner. The ROC curve is the plot of the true positive rate (TPR) against the false positive rate (FPR) at each threshold setting. On the contrary, the PR curve is the plot of the precision against the recall at each threshold setting. They are represented in Figure 2.2.

In the context of image-wise detection, we quantify the performance of an anomaly detection model using the **Area Under the Receiver Operating Characteristic curve (AUROC)**. Intuitive interpretation of the obtained AUROC value is the expectation that a uniformly drawn random positive instance is ranked before a uniformly drawn random negative instance. The baseline value of AUROC is set to 0.5 and correspond to the performance of a binary random classifier. At first glance, AUROC can also be used to quantify pixel-detection. However, as illustrated in Figure 2.3, the class imbalance at the pixel level is particularly severe². Usually, defects covers a small region within the image, resulting in a substantial proportion of nor-

²The "screw" category has the highest imbalance with 99.7% of positive and 0.3% of negative pixels.

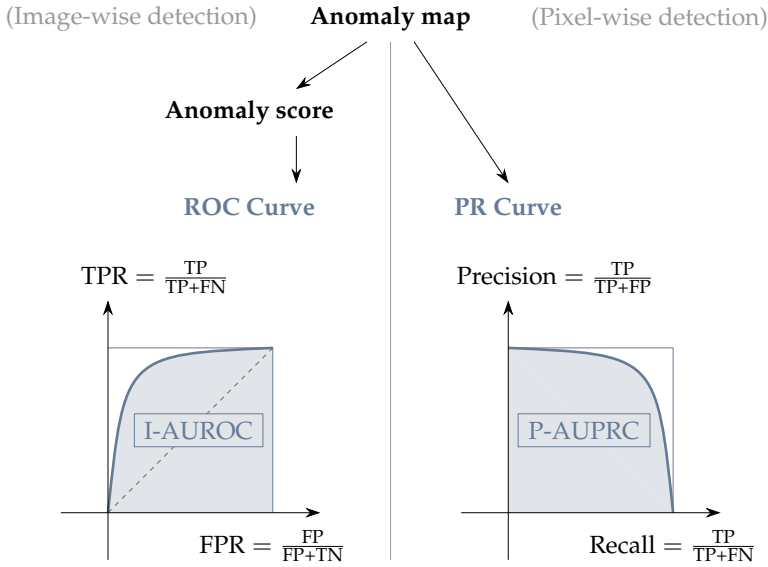


Fig. 2.2 Overview of the performance assessment methodology for both image-wise and pixel-wise detection. For pixel-wise detection, we consider each pixel independently and compute the AUPRC to get the performance of the model on the anomaly localization task. For image-wise detection, we convert the anomaly map into a scalar value, named anomaly score, and compute the AUROC to get the performance of the model on the anomaly detection task. All metrics rely on the true positive (TP), false positive (FP), false negative (FN), and true negative (TN) computed over an entire image category.

mal pixels. In such situations, methods can easily achieve a high TPR while maintaining a low FPR due to the large number of negative pixels, which promotes the prediction of large defective areas. To counterbalance this limitation, we rely on the **Area Under the Precision-Recall Curve (AUPRC)** to quantify pixel-level detection, as precision is significantly impacted when the number of FP is high. If the baseline of AUROC is simply 0.5 in all cases ($y = x$ line), the baseline of AUPRC is equal to the fraction of positives samples ($y = P / (P + N)$ line) [70], which varies with the P:N ratio. For the MVTec AD dataset, the baseline AUPRC for pixel detection ranges between 0.003 (screw) and 0.117 (metal nut).

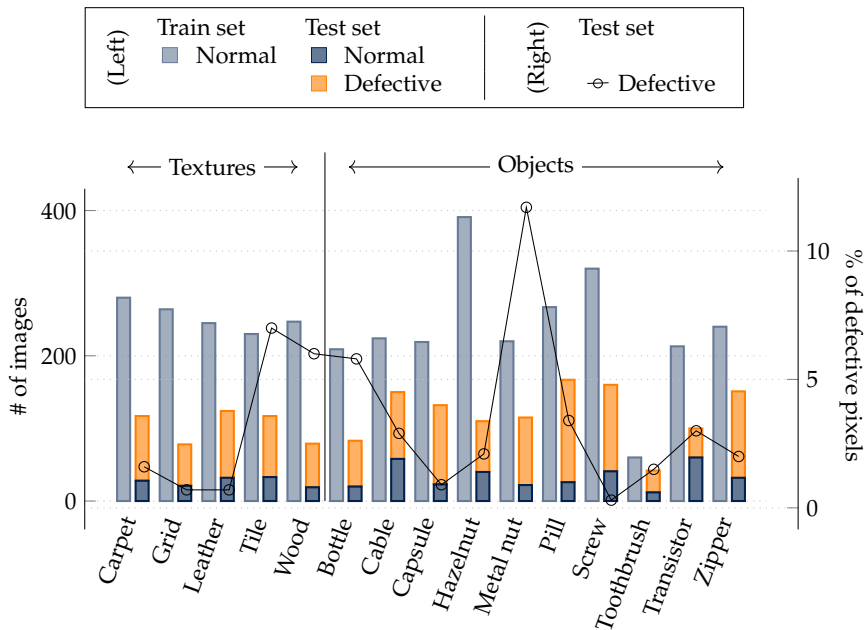


Fig. 2.3 (Bar plot - left axis) For each category, the bar plot indicates the number of train normal images (light blue) and the number of test normal (dark blue) and defective (orange) images. (Line plot - right axis) For each category, the percentage of defective pixels in the test set is indicated by the black line.

At this point, it is useful to clarify how to calculate the area under these two curves. To obtain area under the ROC curve, a straightforward linear interpolation between all measurements provides an accurate estimate. In this case, a trapezoidal rule approximation is suitable for computing the AUROC. However, the computation of AUPRC is more complex because precision does not change linearly with recall [71]. The trapezoidal rule can produce overestimated values and a computation based on average precision is preferred. All AUPRC values presented in this manuscript are estimated with the `average_precision_score` function of scikit-learn [72].

As notation, anomaly detection performance estimated with the AUROC is written as **I-AUROC** and anomaly location performance estimated with the AUPRC is written as **P-AUPRC**.

3

Detection in the original image space

RECONSTRUCTION-BASED approaches are among the most intuitive methods for addressing anomaly detection and localization. The initial objective is to train a model to generate a clean reconstruction of any input image, regardless of whether it contains defects. To achieve this, a network is trained exclusively on normal images to achieve an identity mapping. During inference, it is hypothesized that the network will accurately reconstruct the normal regions of the input while replacing all defective structures with clean content. As a result, the difference between the input and the reconstructed normal image provides information regarding whether or not there is an anomaly in the image. Structures exhibiting high reconstruction residuals are likely to be defective, whereas structures with low residuals are likely to be normal.

The autoencoder architecture offers a number of interesting properties that can be exploited to obtain a clean version of an image. By training the network exclusively on normal, defect-free images to achieve identity mapping, the bottleneck is compelled to learn a compressed latent representation of the training data. This compressed representation encapsulates the essential characteristics of normality, as observed within the normal images that comprise the training set. This latent representation serves as a regu-

larization mechanism, ensuring that the reconstruction process produces output images that align with the learned distribution of normal data. Consequently, any structures within the input that significantly deviate from the typical patterns encountered during training are reconstructed with lower fidelity. This results in higher reconstruction residuals for anomalous regions, thereby enabling their identification as defective.

However, vanilla autoencoders frequently exhibit an inability to accurately reconstruct high-frequency content within images. Given that one of the core working hypotheses of these methods is the perfect preservation of normal structures, any failure to accurately reconstruct such details can result in the misclassification of clean, normal structures as defective. Furthermore, the replacement of defective structures with clean content is not always achieved consistently. The bottleneck alone may not provide sufficient regularization to constrain the output space to the distribution of clean images, and defective structures can be preserved in the reconstruction. This limits the effectiveness of the anomaly detection process. This chapter aims to address the aforementioned limitations by introducing two modifications to the traditional reconstruction-based framework using a vanilla autoencoder:

- (i) Integrating skip connections into the autoencoder architecture;
- (ii) Introducing synthetic noise to corrupt the input images.

A comprehensive quantitative and qualitative analysis demonstrates the benefits of these modifications when applied jointly. The skip connections allow the information to bypass the bottleneck, thereby preserving normal structures. However, this also promotes the convergence of the model towards a pure identity mapping, which is useless for anomaly detection and localization. To counterbalance this effect, the synthetic corruption of training images encourages the networks to replace inappropriate structures observed in the input image.

The methodology presented in this chapter establishes the baseline model for subsequent chapters in the manuscript. This simple yet efficient modification of the original reconstruction-based framework allows a significant improvement in performance while maintaining the intuitive aspect of the approach, which is of particular interest in order to facilitate the use of the method by non-experts in the field.

3.1 Related work

To guarantee accurate detection, it is essential that clean structures are reconstructed in an identical manner while only defective structures are modified. When employing an autoencoder architecture, the bottleneck is typically assumed to regularize the reconstruction towards the normal class. In practice, the reconstruction of clean structures is often suboptimal, and defective structures are sometimes retained in the reconstructed image due to the generalization ability of the network. Various studies have been conducted to constrain the output towards the normal class and enhance the quality of the reconstructed images. This section provides an overview of several extensions to classical reconstruction-based anomaly detection framework, addressing both anomaly detection and localization tasks.

Other loss functions

One category of approaches involves exploring alternative loss functions for training autoencoder networks to achieve identity mapping of the training images. Typically, these methods focus on image-to-image identity mapping using normal data, with the objective of minimizing the **Mean Squared Error (MSE)**. However, the MSE loss is not without its own inherent limitations. It has been demonstrated that MSE tends to produce reconstructions that are blurry in nature [73], which in turn leads to higher residuals in clean structures. This limitation arises because MSE inherently averages out pixel differences, thus failing to preserve fine details.

To improve the sharpness of the reconstruction, **Bergmann et al.** proposed to train an autoencoder to minimize a loss derived from the Structural Similarity Index (SSIM) [25]. The use of the SSIM loss has been motivated by its ability to produce well looking images from a human perceptual perspective [74]. However, the use of SSIM necessitates considering grayscale images and relies on multiple hyperparameters, which can hinder the reproducibility of the results. Due to their capacity to generate realistic images, **Generative Adversarial Networks (GANs)** [75] have been considered to sample the clean distribution of images [24, 26, 27, 76, 77]. However, training GANs poses significant challenges, notably their tendency to converge towards mode collapse. This phenomenon limits the diversity of generated samples. Additionally, the difficulty to exclude abnormal samples from the

generative distribution penalizes performances [27]. Furthermore, some GAN-based methods require an optimization process during inference to find the latent space vector producing the most similar image between an arbitrary query and an output image belonging to the generative distribution [26].

Iterative projections

The conventional methodology for anomaly detection based on reconstruction typically entails a single pass of an image through a network, frequently an autoencoder. During this process, the network is expected to regularize the reconstruction towards the normal class. However, this regularization may be incomplete, and multiple forward passes can facilitate convergence towards a more refined version of the initial reconstruction.

To enhance the quality of the reconstructed images and ensure the replacement of defective structures, **Dehaene et al.** proposed iteratively project the arbitrary input towards the clean distribution of images. The projection is constrained to be similar, in the sense of the $\mathcal{L}^1$ norm, to the initial input [78]. Instead of performing this optimization in the latent space as made with AnoGAN [26], they propose to find an optimal clean input image. If this practice enhances the sharpness of the reconstruction, the optimization step is resource consuming.

Combine coarse and fine reconstructions

The employment of autoencoders in anomaly detection entails a trade-off between the preservation of normal structures and the modification of those that are abnormal.

The **Unsupervised Two-stage Anomaly Detection (UTAD)** [79] framework employs two networks: the Impression Extractor (IE-Net) and the Expert-Net. IE-Net generates a coarse clean image, while the Expert-Net produces a detailed reconstruction based on the input image and the embedding provided by IE-Net. This mechanism serves to prevent the propagation of defective structures to the reconstructed image.

Avoid the comparison with the original input image

The fundamental principle of the reconstruction-based approach is the comparison between the input image and its normal reconstructed version in the original space. This allows for the detection of structures that are poorly reconstructed by the network, which will subsequently be associated with a high likelihood of being defective.

In order to avoid false positive detection due to imperfect reconstruction, **Chung et al.** [80] present an outlier-exposed style distillation network (OE-SDN) which preserve the structure of input images while transferring them to the style of an autoencoder. Anomaly map is obtained by comparing the output of an autoencoder with the output of the OE-SDN network, instead of using the autoencoder residual. This modification has the advantage of reducing the false detection of normal structures.

Image inpainting

To prevent the network from reconstructing defective structures, the reconstruction phase can be formulated as an image completion problem [81, 82]. In this approach, portions of the image are intentionally occluded, and the network is trained to fill in the missing regions with normal structures. This formulation encourages the model to generate plausible, defect-free content, thereby mitigating the risk of reconstructing defective structures.

To make the inference and training phases consistent, it is assumed that the defects are entirely contained in the mask during inference, which limits the practical usage of the method. To circumvent this issue, **Zavrtanik et al.** [83] introduced random inpainting masks constituted by grids of black rectangular regions of size $k \times k$ pixels. Due to their limited receptive field, Convolutional neural networks can struggle at incorporating information from potentially distant regions. This limitation can be addressed by introducing contextual attention in the model [84]. **Inpainting Transformer (InTra)** is trained from scratch transformer used to solve the inpainting problem in the context of anomaly detection and localization. **Mei et al.** [8] also proposed to use a multiscaled denoising autoencoder to reconstruct training images corrupted with salt-and-pepper noise. However they did not discuss the gain brought by this modification, and only considered it for an autoencoder without skip connections.

3.2 Skip connections and synthetic data augmentation

With the objective of addressing the task in the simplest manner, we focus on the formulation involving the training of a single autoencoder for image-to-image mapping. We propose a straightforward approach to enhance the quality of reconstruction obtained with a vanilla autoencoder, addressing the following two criteria simultaneously:

- (i) Enhance the sharpness of the reconstructed normal structures;
- (ii) Replace defective structures with normal ones.

To improve the sharpness of the clean structures reconstruction, we consider an autoencoder architecture augmented with skip connections. They allow the preservation of high-frequency information by bypassing the bottleneck. However, we show that this practice promotes the convergence of the network towards a pure identity mapping when trained exclusively on uncorrupted clean images. Naturally, this penalizes strongly the anomaly detection performances. To mitigate this issue, we introduce an original synthetic noise model, which improves significantly the anomaly detection accuracy for the skipped architecture. This improvement enables the modified architecture to outperform the conventional autoencoder in numerous real-world scenarios.

3.2.1 Network architecture

The reconstruction is based on a convolutional neural network. Our architecture, referred to as **Autoencoder with skip connections (AESc)** and shown in Figure 3.1a, is a variant of U-Net [30]. AESc takes input images with a spatial dimension of 256×256 and projects them onto a latent space of $16 \times 16 \times 256$ dimension. The projection towards the lower dimensional space is performed by 4 consecutive convolutional layers strided by a factor 2. The back projection is performed by 4 layers of convolution followed by an upsampling operator of factor 2. All convolutions have a 5×5 kernel. Unlike the original U-Net version, our skip connections perform an addition, not a concatenation, of feature maps of the encoder to the decoder. For the sake of comparison, we also consider the **Autoencoder (AE)** network which follows exactly the same architecture but from which we removed the skip connections.

All models have been trained during 250 epochs to minimize the MSE loss given by:

$$L_{\text{MSE}}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}_i)^2 \quad (3.1)$$

with n being the number of images. The batch size used is of 16 images. We used the AdamW optimizer [85] with an initial learning rate of 0.01. The learning rate is dynamically adjusted according to the OneCycleLR scheduler [86].

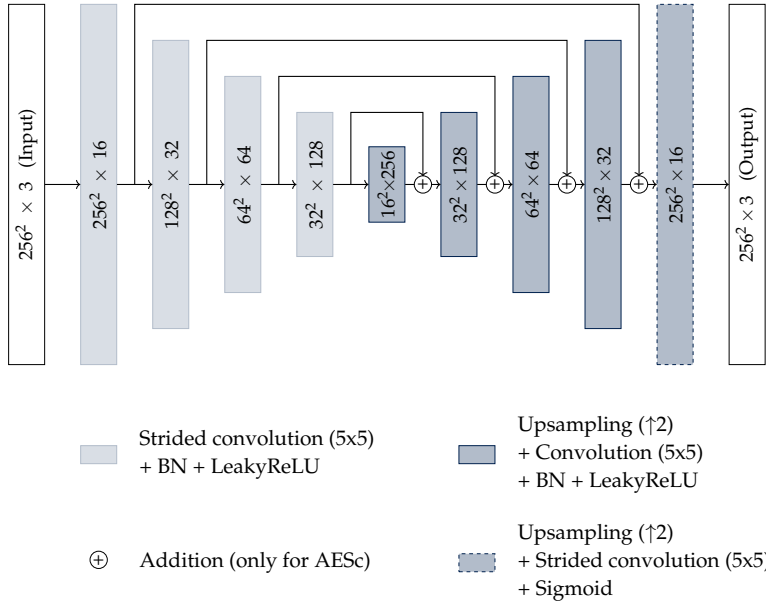
3.2.2 Stain: a synthetic corruption model

Ideally, the autoencoder should preserve clean structures while modifying those that are not. Due to the impossibility of collecting pairs of clean and defective versions of the same sample, we propose to introduce synthetic corruption during training to explicitly constrain the autoencoder to remove this additive noise. Our **Stain** model, illustrated in Figure 3.1b, corrupts images by adding structures whose color is a random perturbation of the mean color value of the entire image and whose shape is an ellipse with irregular edges. A total of 5 stain structures are added to the normal training image to generate a synthetic corrupted version of it.

The intuition behind the definition of this noise model is that occluding parts of the training images is a data augmentation procedure that helps to improve the network training [87, 88]. Due to the skip connections in our network architecture, this form of data augmentation is essential to avoid the convergence of the model towards the identity operator. However, we noticed that the use of regular shapes, like a conventional ellipse, leads to overfitting to this additive noise structure, as also pointed out in a context of inpainting with rectangular holes [89]. Comparative study about the kind of noise that is used is presented in Section 3.5.

We exclusively utilized stain noise for data augmentation. This choice is motivated by the objective of operating without relying on prior knowledge of the defect appearance. For instance, conventional augmentations such as vertical or horizontal flips could inadvertently introduce artifacts that mimic defects in certain image categories. To ensure the integrity of the training process, no additional data augmentation techniques were employed.

3 | Detection in the original image space



(a) AESc architecture performing the projection of an arbitrary 256×256 image towards the distribution of clean images with the same dimension. Note that the AE architecture shares the same specifications with the exception of the skip connections that have been removed.



(b) The Stain noise model is a cubic interpolation between 20 points (blue dots), arranged in ascending order of polar coordinates, located around the border of an ellipse of variable size (blue line). The axes of the ellipse are comprised between 1 and 12% of the smallest image dimension and its eccentricity is randomly initialized.

Fig. 3.1 The two key components of the approach: the autoencoder architecture and the synthetic corruption model.

3.3 Pipeline description

Figure 3.2 illustrates the entire pipeline used to evaluate our approach for the anomaly detection and localization tasks. Our best approach involves corrupting the input image with the Stain noise model. For the sake of comparison, the synthetic corruption step can be omitted, feeding the original clean image directly to the network. Additionally, for further comparison, the skip connections can be removed from the network architecture.

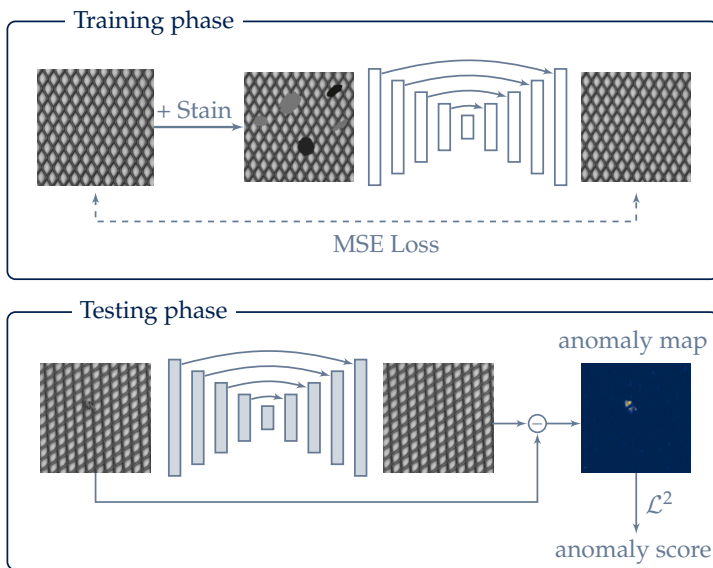


Fig. 3.2 (Training phase) The network (described in Section 3.2.1) is trained to minimize the MSE between the original clean image and its reconstruction. A corrupted version of the original clean image with the Stain noise model (described in Section 3.2.2) is fed to the network. (Testing phase) The network weights are frozen (indicated by the blue filling) and fed with an image, which can either be clean or defective. During testing, no synthetic corruption is added to the input image. An anomaly map is obtained by computing the absolute value of the difference between an image and its reconstructed clean version. An anomaly score is computed with the $\mathcal{L}^2$ norm of the anomaly map.

3.4 Quantitative and qualitative results

The main objective of this study is to highlight the advantages of considering an autoencoder network architecture trained to reconstruct clean images that have been corrupted with our synthetic Stain model. The impact of these two contributions are studied individually and jointly. This investigation encompasses both quantitative and qualitative analyses, emphasizing the gain in performance of the proposed approach. To evaluate the contributions of the skip connections and the synthetic stain model, we conduct comparative analyses across four distinct use cases:

- **AE + None.** An autoencoder without skip connections trained to perform the identity mapping of clean images.
- **AESc + None.** An autoencoder with skip connections trained to perform the identity mapping of clean images.
- **AE + Stain.** An autoencoder without skip connections trained to perform the mapping of corrupted towards clean images.
- **AESc + Stain.** An autoencoder with skip connections trained to perform the mapping of corrupted towards clean images.

Figure 3.3 provides a visual comparison of the four use cases, revealing the general trend of each configuration. When comparing the AE with the AESc network, it can be observed that the AE network results in higher reconstruction residual for clean structures. This phenomenon is observed across both texture images (wood and tile) and object images (cable and hazelnut). On the contrary, the AESc model shows finer reconstruction of the image details, as indicated by nearly zero residuals over the clean areas of the images. However, when AESc is trained without the Stain corruption, the model converges towards an identity operator, as revealed by the close-to-zero residuals of defective structures. Corrupting the training images with the Stain model mitigates this undesirable behavior, leading to high reconstruction residuals in defective areas while maintaining low reconstruction residuals in clean structures.

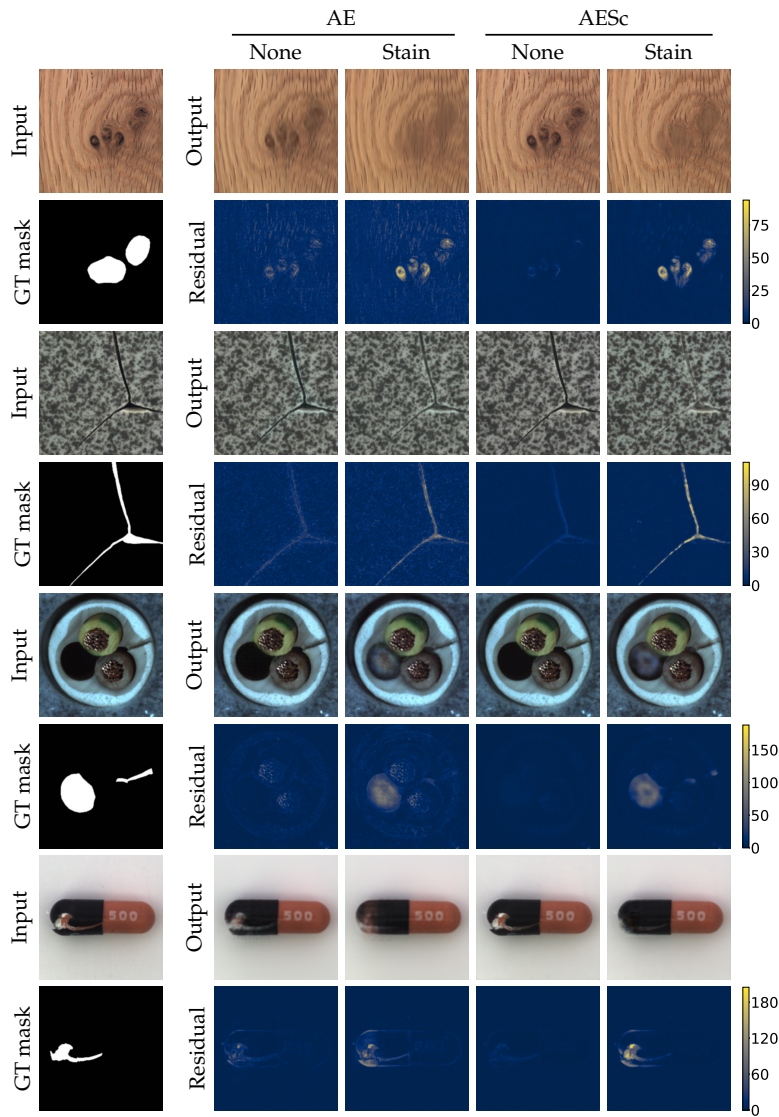


Fig. 3.3 Predictions obtained with either the AESc or AE networks trained with or without our Stain noise. Two defective textures are considered, namely a wood (first sample) and a tile (second sample), as well as two defective objects, namely a cable (third sample) and a capsule (fourth sample). First column shows the image fed in the networks and the defect mask. Odd rows show the reconstructed images and even rows show the anomaly maps obtained with the residual-based strategy.

Table 3.1 Performance obtained with AE + None, AE + Stain, AESc + None and AE + Stain on anomaly detection and localization tasks^a.

		I-AUROC				P-AUPRC			
Network		AE		AESc		AE		AESc	
Corruption		None	Stain	None	Stain	None	Stain	None	Stain
Textures	Carpet	0.36	<u>0.42</u>	0.41	0.83	0.03	<u>0.10</u>	0.02	0.25
	Grid	0.83	<u>0.94</u>	0.79	0.98	0.12	<u>0.40</u>	0.01	0.51
	Leather	0.36	0.39	<u>0.61</u>	0.90	0.08	<u>0.13</u>	<u>0.13</u>	0.38
	Tile	0.86	<u>0.97</u>	0.88	1.0	0.19	<u>0.31</u>	0.19	0.51
	Wood	0.93	<u>0.96</u>	0.90	1.0	0.22	<u>0.26</u>	0.19	0.37
	Mean ^b	0.67	<u>0.74</u>	0.72	0.94	0.13	<u>0.24</u>	0.11	0.40
Objects	Bottle	0.83	<u>0.87</u>	0.58	0.98	0.19	<u>0.23</u>	0.10	0.52
	Cable	0.44	<u>0.63</u>	0.48	0.87	0.04	<u>0.21</u>	0.03	0.43
	Capsule	0.73	<u>0.74</u>	0.51	0.77	0.09	<u>0.13</u>	0.02	0.24
	Hazelnut	0.86	<u>0.89</u>	0.74	0.95	0.42	0.50	0.35	<u>0.51</u>
	Metal Nut	0.23	<u>0.37</u>	0.21	0.82	0.19	<u>0.22</u>	0.19	0.52
	Pill	0.82	0.87	0.78	<u>0.86</u>	0.30	0.30	0.22	<u>0.32</u>
	Screw	<u>0.82</u>	0.73	0.74	0.87	0.03	<u>0.06</u>	0.03	0.15
	Toothbrush	0.84	0.94	0.64	<u>0.91</u>	0.16	<u>0.23</u>	0.08	0.35
	Transistor	0.54	<u>0.70</u>	0.45	0.86	0.08	<u>0.10</u>	0.06	0.16
	Zipper	0.87	0.91	0.59	<u>0.90</u>	0.09	<u>0.12</u>	0.05	0.15
Mean ^c		0.70	<u>0.77</u>	0.57	0.88	0.16	<u>0.21</u>	0.11	0.34
Global mean ^d		0.69	<u>0.76</u>	0.62	0.90	0.15	<u>0.22</u>	0.11	0.36

^a For each row, the best performing approach for anomaly detection and localization is highlighted in boldface and the second best is underlined.

^b Mean I-AUROC and P-AUPRC obtained over the classes of images belonging to the texture categories.

^c Mean I-AUROC and P-AUPRC obtained over the classes of images belonging to the object categories.

^d Mean I-AUROC and P-AUPRC obtained over the entire dataset.

The previous qualitative analysis discussed the principal characteristics of the reconstruction residuals obtained in the four configurations and anticipate their effectiveness as anomaly maps. In this quantitative study, we analyze the impact of the skip connections and the synthetic Stain corruption for both anomaly detection and anomaly localization with the help of dedicated metrics introduced in 2.3. For anomaly detection, which is related to the I-AUROC metric, Table 3.1 highlights the superiority of the AESc + Stain configuration for almost all image categories. When applied to texture images, the AESc + Stain approach improves the I-AUROC score by at least 0.2. When applied to object images, the improvement is up to 0.08, which still represents a significant enhancement. In general, corrupting training images with the Stain noise is beneficial with both AE and AESc architectures. Adding skip connection to the network architecture is only pertinent when combined with the Stain noise corrupted, as reflected by observations made during the qualitative study. Without synthetic corruption, the AESc architecture behaves similarly to an identity mapping, resulting in global performance only slightly better than a random classifier (0.62 vs. 0.50).

For anomaly localization, which is related to the P-AUPRC metric, the AESc + Stain configuration performs the best, with a larger margin of improvement on textures compared to objects. The P-AUPRC improvement is up to 0.16 on textures and 0.15 on objects. Consistent with the observations made for anomaly detection, the Stain noise is beneficial with both AE and AESc architectures. Also, considering the AESc network without Stain corruption results in the worst performances.

3.5 Comparative study of corruption models

The corruption of training images with our synthetic Stain noise promotes the regularization of the reconstruction towards the distribution of normal images. In this comparative study, we emphasize on the impact of the synthetic noise design on this regularization property. For this purpose, we introduce three new noise models illustrated in Figure 3.4:

- **Ellipse.** Corrupt by adding five ellipses of variable color and size. The sampling of color, axes and eccentricity of each ellipse follows the same distribution than the Stain model described in Section 3.2.2.

- **Scratch.** Corrupt by adding one curve connecting two points whose coordinates are randomly chosen in the image and whose color is randomly selected in the gray scale range. The curve can follow a straight line, a sinusoidal wave or the path of a square root function.
- **Gaussian noise.** Corrupt by adding Gaussian noise applied uniformly over the entire image. For normalized intensities between 0 and 1, a corrupted pixel value x' , corresponding to an initial pixel value x , is the realization of a random variable given by a normal distribution with mean x and variance $\sigma^2 \in [0.01, 0.05, 0.1, 0.5]$.

The objective of this analysis is to demonstrate that the purpose of introducing synthetic noise is to train the network on a reconstruction task that facilitates the learning of dedicated latent representations. The focus has not been on designing synthetic defects that mimic real defects, but rather on ensuring that the network can generalize to unknown defect appearances. This approach enables the network to effectively detect and localize anomalies, even when they do not resemble the synthetic noise patterns used during training.

Figure 3.4 presents a comparative analysis of the newly introduced noise models with the Stain one. In both the object (bottle) and the texture (tile) image, an increase in the variance of the Gaussian noise significantly degrades the reconstruction quality of normal structures. This leads to high residual in clean regions, which is particularly detrimental for defect localization and detection. Combined with the fact that defective structures are preserved in the reconstruction, Gaussian noise is a poor synthetic noise model to train the AESc model.

Compared to Gaussian noise, which is non-structural and applied uniformly across the image, we also evaluate several synthetic noises applied in a localized manner. At first glance, it might seem intuitive that scratch noise would be more effective for reconstructing the bottle shown in the second line. The similarity between the appearance of the real defect and the synthetic scratch suggests an optimal training scenario for this specific example. However, the reconstruction achieved with AESc + Scratch does not match the quality of the one obtained with AESc + Stain. This observation highlights the robustness and generalization ability of the Stain noise model to train the AESc network on a reconstruction task. Furthermore, the tile sample corroborates this observation. In the presence of a real defect

resembling a circular yellow drop, the AESc + Stain yields superior reconstruction compared to other noise models. The non-natural aspect of the Stain model does not limit the performance of the approach; rather, this challenging training task enables the network to develop effective internal representations of normal images. The ability of the Stain model to enhance reconstruction quality is validated through these visual examples.

The qualitative impact of the visual effects described previously is presented in Table 3.2. This table summarizes the mean performance metrics for textures, objects, and the entire dataset, providing a comprehensive overview of the comparison. As shown in the qualitative analysis, the AESc + Noise model consistently outperforms others for both textures and objects. The use of regular ellipses instead of the Stain model is detrimental. Training the network with Scratch noise results in the poorest performance among the structural noises evaluated. Corrupting training images with Gaussian noise improves the metrics compared to scenarios without any synthetic corruption. However, the use of non-structural noise still results in less accurate anomaly detection. These observations align with the qualitative findings presented earlier, which supports the superiority of the AESc + Stain model for anomaly detection.

Table 3.2 Performance obtained with the AESc network associated with varying type of synthetic noise model over textures, objects and overall dataset^a.

		None	Stain	Ellipse	Scratch	Gaussian noise (σ)			
						0.01	0.05	0.1	0.5
I-AUROC	Textures	0.72	0.94	<u>0.88</u>	0.87	0.69	0.68	0.73	0.62
	Objects	0.57	0.88	<u>0.86</u>	0.72	0.66	0.74	0.70	0.62
	All	0.62	0.90	<u>0.86</u>	0.77	0.67	0.72	0.71	0.62
P-AUPRC	Textures	0.11	0.40	0.40	<u>0.25</u>	0.15	0.16	0.15	0.10
	Objects	0.11	0.34	<u>0.31</u>	0.15	0.12	0.11	0.18	0.08
	All	0.11	0.36	<u>0.34</u>	0.18	0.13	0.13	0.12	0.09

^a For each row, the best performing approach is highlighted in boldface and the second best is underlined.

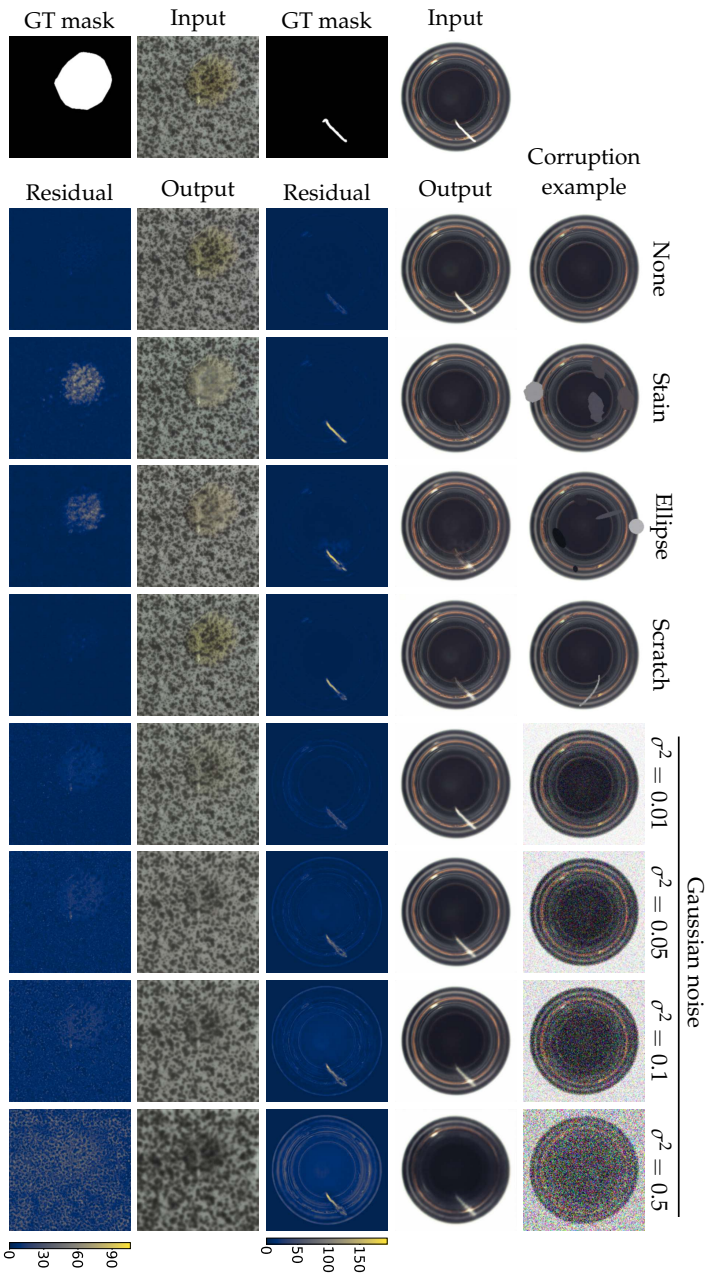


Fig. 3.4 (First line) Training image corrupted with all types of synthetic noise that are compared. (Second and fourth rows) Reconstruction obtained by the AESc network trained to removed the considered synthetic noise on a bottle and a tile. (Third and fifth rows) Anomaly maps corresponding to reconstruction obtain in the above line.

3.6 Exploring operational boundaries and working hypotheses

In the results presented in Section 3.4, we demonstrated the effectiveness of incorporating skip connections into the network architecture, coupled with the introduction of our Stain noise model to corrupt the training images. These modifications substantially enhance the network's ability to accurately reconstruct normal structures while effectively altering those that deviate from the expected patterns. This results in an increase in the residual in defective regions and a decrease in those in normal regions, thereby enhancing the robustness of anomaly localization with respect to the choice of the threshold used to determine whether a structure is normal or defective.

Building on these initial findings, the following sections provide a more detailed examination of the characteristics and underlying mechanisms of the proposed method. Our study is guided by the formulation of specific properties that capture the broader behaviors exhibited by the network. These properties are intended to encapsulate key aspects of the network's functioning, providing insights that extend beyond the particularities of the MVTec AD dataset used in this study. The aim is to derive principles and techniques that can be applied in diverse industrial contexts.

3.6.1 Property 1 - Low contrast defects

The intensity of the anomaly map is linked to the pixel-wise difference between a defective structure and its normal reconstructed version. If the clean reconstruction contrasts poorly with surrounding structures in the image, the intensity of the anomaly map will be low.

The anomaly map is obtained with the absolute value of the difference between the input image and the clean image reconstructed by the network. This approach is based on the assumption that, when trained exclusively on normal data, the network will effectively replace any defective structures in the input image with plausible, clean content that closely resembles the normal distribution. It is crucial to acknowledge that the accuracy of this methodology is dependent of the contrast between the defective structures and the clean content generated by the network. In particular, when an anomaly displays only slight discrepancies in comparison to its recon-

structured counterpart, the resulting anomaly map may exhibit low intensity values in those regions.

Figure 3.5 illustrates this limitation, demonstrating how low-contrast anomalies may not be effectively highlighted in the anomaly map. Understanding and addressing this challenge is crucial for improving the robustness of anomaly detection methods, particularly in applications where detecting subtle defects is of critical importance. This will be addressed in the remainder of the manuscript.

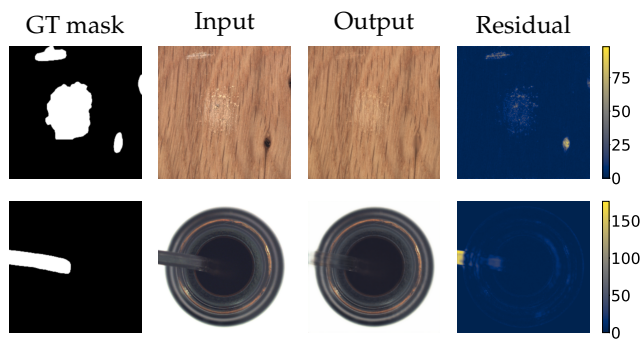


Fig. 3.5 (First row) Anomaly localization of defects with variable contrast relative to their neighborhood in a wood image. In the anomaly map, the pixels corresponding to the black knot are related to higher values. (Second row) Anomaly localization in a bottle image of a defect whose neighborhood exhibits structural variability. The anomaly map shows more effective detection over the white background compared to the object structures, highlighting the impact of contrast on performances.

3.6.2 Property 2 - Concept of normality

The concept of normality depends on the variability that is reflected in the training set, containing normal samples exclusively.

In the considered problem formulation, which addresses anomaly localization and detection in an unsupervised setting, only normal images are available for training. This implies that the network must implicitly learn the normal image appearance. In contrast, a supervised problem setting

would lead to explicit learning of defect appearance. To perform efficiently, the implicit training of normal image appearance relies on the challenging trade-off between:

- (i) Learning a wide normal distribution and considering only large variation as defective;
- (ii) Learning a narrow normal distribution and considering any subtle variation as defective.

The concept of normality is inherently dependent on the specific characteristics of the dataset and the types of defects being considered for detection. As such, the definition of what constitutes "normal" can be quite nuanced and context-specific. This notion is particularly important when dealing with complex or subtle defects, where the distinction between normal and defective structures is not always straightforward. Figure 3.6 illustrates a test capsule for which the absence of the "500" is considered defective, while the absence of the "activis" is not. In order to categorize the absence of "500" as defective but not the absence of "activis," the training set must contain training images that exemplify this scenario. Therefore, ensuring consistency in the training set is essential to facilitate accurate learning of the concept of normality.

This second property may also present challenges in scenarios where domain shift occurs between the training and test sets. In such instances, the variations introduced into the test set may be incorrectly identified as defects. For example, in the screw image category, we observe slight variations in background intensity, likely attributable to changes in illumination conditions. This distribution shift is illustrated on Figure 3.7 by the difference between the distribution of background pixels in the train and the test sets. The implication of this difference on anomaly detection is represented on Figure 3.8 with multiple images from the screw category.

As illustrated by these visual examples, the background leads to a significant increase in the number of false positive detections. As the domain shift has a detrimental impact on all the methods presented in this thesis, we present the results obtained on the screw dataset, which has undergone preprocessing with a background removal technique. This process has been conducted using the approach presented in [90]. This includes the results previously showcased in Table 3.1.

3 | Detection in the original image space

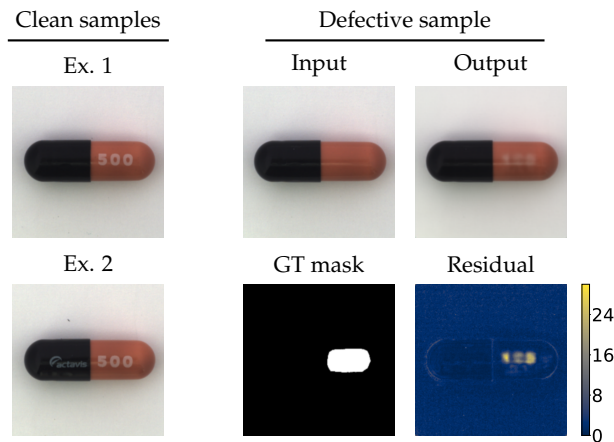


Fig. 3.6 (Left) Two samples, considered as normal, are presented. Example 1 shows a capsule with a "500" mark on the right orange part. Example 2 shows a capsule with a "500" mark on the right orange part and a "activis" logo on the left black part. In this dataset, the absence of the "500" print marking is considered as defective, while the lack of the "activis" logo is considered as normal. (Right) Anomaly localization on a defective capsule lacking both the "500" mark and the "activis" logo. This illustrates that the model captures the absence of the "500" mark as defective and the lack of the "activis" logo as normal.

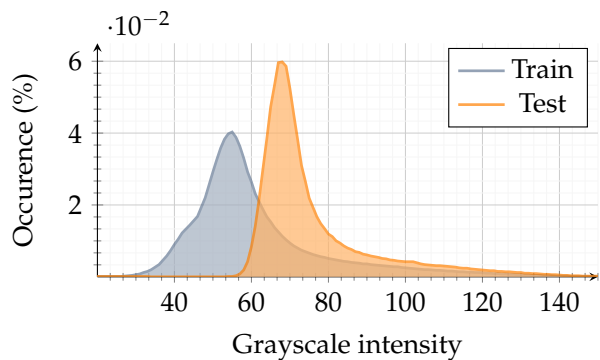


Fig. 3.7 Distribution of grayscale intensities between the background of the train normal samples (blue) and the test normal samples (orange).

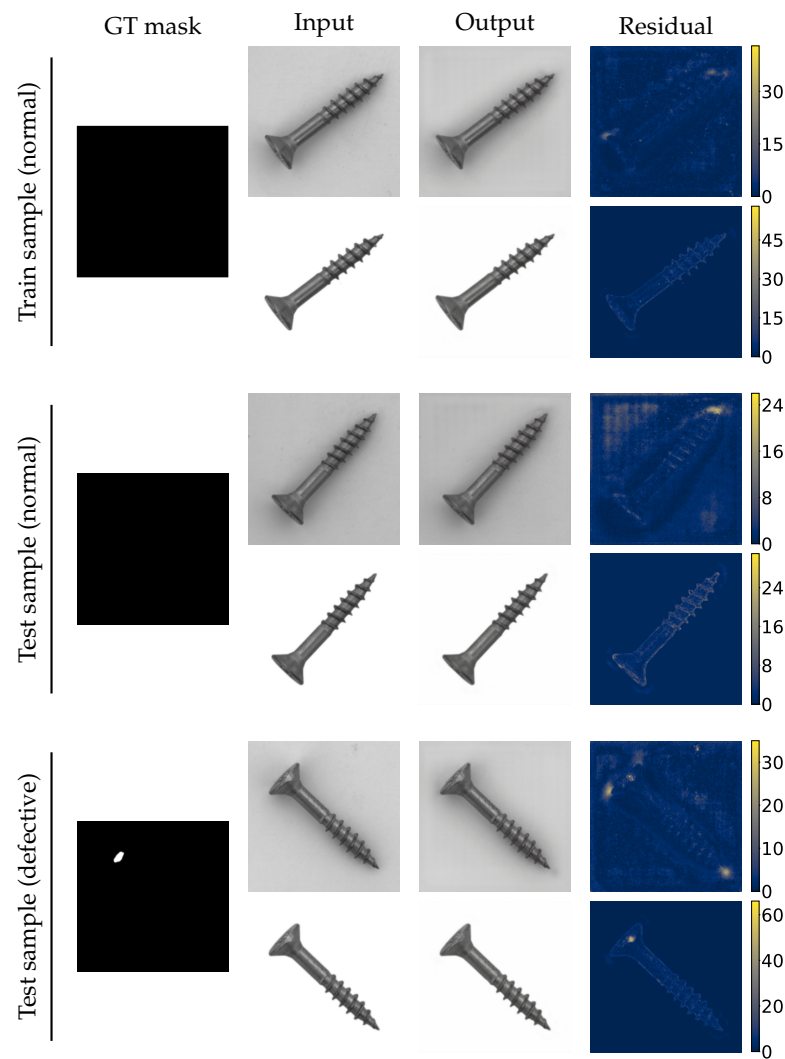


Fig. 3.8 (Odd rows) Localization results obtained on the original screw dataset under change in background illumination. In a normal train image (first row), the background is considered as normal as reflected by the low intensities of the anomaly map. In both normal and defective samples (third and fifth rows), the background contributes to higher anomaly map intensities. (Even rows) Removal of the background from the entire dataset eliminates false positive detections associated with the background.

3.7 Summary and fundamental properties

Reconstruction-based approaches are a widely used methodology in anomaly localization and detection, where the primary objective is to reconstruct a normal version of an input image, regardless of whether it contains defects. Typically, this involves training an autoencoder to perform identity mapping, aiming to recreate the input image as accurately as possible. The detection mechanism is then based on analyzing the reconstruction residuals. For effective anomaly detection, it is essential that the model accurately reconstructs normal structures while replacing any defective elements with normal content. The autoencoder's bottleneck serves as a regularization mechanism, constraining the network's capacity to memorize the input, thus encouraging the model to generalize well. However, this often leads to suboptimal reconstructions, especially in the presence of complex or subtle anomalies.

To address these challenges and enhance both the reconstruction quality and detection accuracy, we implemented two key modifications in our framework, applied jointly:

1. **Integrating skip connections into the autoencoder architecture.** This modification allows the network to bypass the bottleneck for certain feature information, thus preserving high-frequency details that are crucial for accurately reconstructing normal structures.
2. **Introducing synthetic noise to corrupt the input images.** By training the network on corrupted data, we force it to learn more robust features, improving its ability to differentiate between normal and defective structures.

Compared to a classical reconstruction-based framework, which typically involves an autoencoder without skip connections trained on identity mapping of uncorrupted images, these modifications significantly enhance anomaly detection. Specifically, the I-AUROC improved from 0.69 to 0.90. This enhancement is particularly notable for texture images, where achieving a high-quality reconstruction of normal structures requires preserving high-frequency details. In terms of anomaly localization, the P-AUPRC increased from 0.15 to 0.36.

A deeper analysis of the detection mechanism revealed challenges in identifying defective structures that poorly contrast with their surroundings. Even when these structures are perfectly reconstructed, they may produce low residuals, leading to potential misdetections. This limitation underscores the difficulty of detecting subtle anomalies when operating solely in the original image space.

Furthermore, we observed that the method is sensitive to small distribution shifts, such as changes in illumination, as demonstrated in the screw category. The concept of normality is defined by the distribution of normal data, and any deviation, such as a change in background lighting, can be mistakenly classified as defective. However, when dealing with objects, employing background removal techniques could effectively address this challenge and reduce false positives due to environmental changes.

In summary, this study underscores both the strengths and limitations of reconstruction-based methods for anomaly detection. The enhancements introduced—specifically, the use of skip connections and synthetic noise—have proven effective in improving detection accuracy. However, the study also highlights the need for further refinement, particularly in handling subtle defects and adapting to environmental changes. These insights provide a valuable foundation for future research aimed at developing more robust and accurate anomaly detection frameworks in complex, real-world scenarios.

4

Detection with internal representations

RECONSTRUCTION-BASED anomaly detection and localization methods are typically based on the reconstruction error of a network that has been trained exclusively to output clean images. In order to achieve optimal performance, the network must be able to identify defective structures and replace them with normal content. However, a notable limitation of this approach is that defective structures that have been replaced by clean content with similar intensity may not be reflected in the anomaly map. This does not indicate that the network did not identify those structures as anomalous; rather, their replacement with normal content does not allow their detection with the reconstruction residual.

The initial focus of this chapter is to develop an understanding of information propagation through the network by considering each voxel within every activation tensors. Given the high dimensionality of these tensors, visualization presents a challenge, necessitating techniques to enhance the desired signal. Our visualization strategy employs a double inference mechanism. Instead of examining the activation tensors of an image directly, we analyze the absolute value of the difference between the activation tensors of the image and its reconstructed version. This method allows us to observe the activations of defective structures, illustrating how the

network intrinsically discriminates between clean and defective images. Subsequently, we introduce a formal definition of an anomaly map and an anomaly score based on observed voxel-level phenomena. Relying on the difference of activation tensors, we propose applying a binary test to each voxel to distinguish between normal and abnormal values. The aggregation of all binary tests allows the computation of anomaly maps and anomaly scores. This new method addresses the limitations of traditional approaches by providing a more nuanced detection mechanism that accounts for subtle differences in voxel activations.

4.1 Related work

Convolutional neural network visualizations encompass techniques designed to visually represent the features or patterns learned by various layers within a trained network. These visualizations are crucial for gaining insights into how the network processes input images. In the general context of computer vision, visualization approaches based on integrated gradients have become some of the most widely used methods. These techniques generate heat maps that highlight the most significant input locations for predictions by computing the loss with respect to feature maps of a convolutional layer. The back-propagated gradients are combined to obtain the importance weights map [91].

In the domain of anomaly detection, several notable examples of such methods can be cited. As mentioned previously, **DifferNet** [38] is a density-based anomaly detection framework augmented with anomaly localization by the use of a back propagation mechanism. **CutPaste** [92] trains a network to distinguish between normal images and defective images created by cutting and pasting portions of normal images. Anomaly localization in this approach relies on Grad-CAM [40], while anomaly detection is formulated as a density-based task. **Convolutional Adversarial Variational autoencoder with Guided Attention (CAVGA)** [93] employs a discriminative latent variable, trained in an adversarial setting, to ensure that the feature representation of the latent variables encode all the normal regions. This process encourages the attention map generated with Grad-CAM to cover the entire normal image. **Salehi et al.** [94] proposed a student-teacher approach where the anomaly score is computed by summing all distances

between the student and teacher tensors. To extract an anomaly map out of this score, they use the gradients of the loss with respect to the input. The resulting map is then refined using Gaussian blur and morphological opening filters.

In addition to integrated gradients, activation tensor visualization is another valuable approach for internal visualization. We define an activation tensor as the output of a convolutional layer, represented in memory as a multidimensional matrix where each element is termed a voxel. Visualizing these tensors can be helpful for understanding which features of the input image are detected or preserved in the network's internal representations. Numerous tutorials cover the practical programming aspects of tensor visualization [95, 96, 97]. However, to our knowledge, there are no contributions that specifically address their application to anomaly detection and localization tasks.

4.2 Activation tensors visualization

In this section, we develop an intuitive understanding of the propagation of defective structures through the AESc + Stain network, or AESc in short, introduced in Chapter 2. We emphasize on the fact that no network training is considered in this chapter; all network weights correspond to previously established results. In addition to providing intuitive insights on the workings of trained autoencoders, this analysis proposes a valuable basis for extending the anomaly detection mechanism beyond the sole use of the reconstruction residual.

4.2.1 Notations

The high dimensionality of activation tensors makes their visualization challenging. In this study, we aim to provide a visualization strategy that highlights the propagation of defective structures through the network. To minimize the signal related to the propagation of normal structures, we propose to consider the difference between activation tensors obtained by inferring an image and its reconstruction. Inferring the reconstruction requires a second pass through the network. To define this mechanism represented in Figure 4.1, we introduce multiple notations.

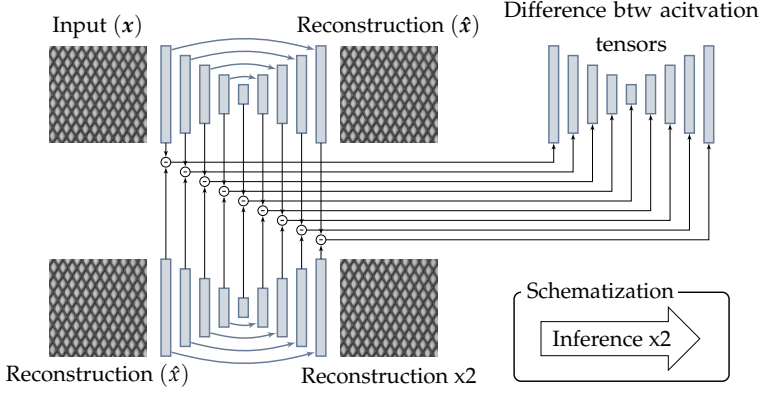


Fig. 4.1 We use a double inference mechanism to obtain activation tensors highlighting defective structure propagation through the network. We compute the absolute value of the difference between the activation tensors generated by the inference of an input image and the activation tensors generated by the inference of its normal reconstructed version.

Definition 4.1 (Input images). We denote an input image by

$$x \in \mathbb{R}^{W \times H \times C},$$

where W , H and C stand for the width, height and number of channels of x . As a reminder, we exclusively consider color images with $W = H = 256$ and $C = 3$. During training, those images have been corrupted with synthetic Stain noise. Here, we consider test images only, without any synthetic corruption. The reconstruction of the image x is referred to as $\hat{x}$.

Definition 4.2 (Activation tensors). We define the operator

$$l_i(\cdot) : \mathbb{R}^{W \times H \times C} \rightarrow \mathbb{R}^{W_i \times H_i \times C_i}$$

as the projection of an input image onto its activation tensor in the i^{th} layer. W_i , H_i , and C_i respectively denote the width and height (spatial dimensions) and the number of channels in layer i . According to our setup, $l_0(x)$ corresponds to the activation tensor obtained after one convolutional layer while $l_4(x)$ is the activation tensor in the autoencoder bottleneck.

Definition 4.3 (Difference of activation tensors). Given two activation tensors at a given layer i , we define the difference of activation tensors between an image $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$ as

$$\Delta l_i(\mathbf{x}, \hat{\mathbf{x}}) = |l_i(\mathbf{x}) - l_i(\hat{\mathbf{x}})|.$$

$\hat{\mathbf{x}}$ being the closest clean reconstruction of $\mathbf{x}$, obtained by inferring $\mathbf{x}$ through the AESc autoencoder. If $\mathbf{x}$ is sampled from the clean distribution, the images $\mathbf{x}$ and $\hat{\mathbf{x}}$ should be almost identical, as their activation tensors. This implies that the resulting Δl_i tensor of $\mathbf{x}$ should contain close-to-zero values. On the contrary, if $\mathbf{x}$ is sampled from the defective distribution, the corresponding Δl_i tensor should deviate from the zero tensor.

Definition 4.4 (Mean of activation tensors along channel axis). We define the operator

$$\mu(\cdot) : \mathbb{R}^{W_i \times H_i \times C} \rightarrow \mathbb{R}^{W_i \times H_i},$$

as the mean of a 3D tensor along channel axis, resulting in an image with spatial dimension $W_i \times H_i$. We will both consider the application of this operator to visualize an activation tensor $l_i(\mathbf{x}) \in \mathbb{R}^{W_i \times H_i \times C_i}$ or a difference of activation tensors $\Delta l_i(\mathbf{x}, \hat{\mathbf{x}}) \in \mathbb{R}^{W_i \times H_i \times C_i}$.

4.2.2 Defect propagation across layers

Figure 4.2 presents a comparative analysis of latent representations for an image of a defective bottle. The second and third rows illustrate the mean of activation tensors along the channel axis of the input and output images, respectively. The objective of the second inference is to subtract the activation associated with clean structures from the activation of the input image. This process emphasizes the activations corresponding to defective structures, as depicted in the last row. In this example, the defective structure covers a large area of the object and exhibits relatively low contrast with its surrounding structures. The AESc model outputs a clean reconstruction of the input image, effectively replacing defective structures with clean content. The orange circle, for instance, has been successfully recovered on the entire object. However, despite the quality of the obtained output image, the anomaly map obtained by computing the reconstruction residual does not lead to the detection of the entire defective structure. For example, even if the orange circle is correctly reconstructed, this defective structure does

not lead to high residual intensities, a limitation previously discussed in Chapter 3.

Defective structures correlate with higher intensities in the mean of the difference between activation tensors. In the shallow layers ($i = 0, 1, 7, 8$), the highest intensities correspond to defective structures with high contrast relative to their surroundings, such as the light stain visible in the input image. In the deeper layers ($i = 3, 4, 5$), high activation is distributed throughout the defective region, with the correlation to neighboring contrast being diminished. This characteristic is particularly relevant for anomaly detection and supports the development of metrics for anomaly localization and detection using latent representations.

For the sake of comparison, Appendix B comments the impact of the skip connections in the network architecture on the internal representations.

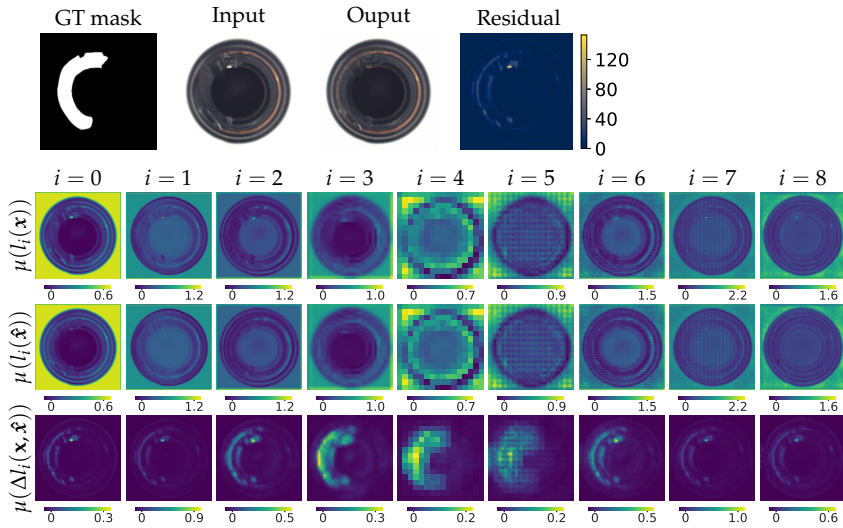


Fig. 4.2 (First row) Anomaly detection method on a bottle as described in Chapter 3. (Second row) Mean activation tensors along channel axis of the input image x . (Third row) Mean activation tensors along channel axis of the output image $\hat{x}$. (Fourth row) Mean of difference of activation tensors along channel axis between the input image x and the output image $\hat{x}$.

4.3 Activation-level anomaly detection and localization

Based on the visualization presented above, it seems promising to take advantage of the way the network propagates an image in order to address both anomaly detection and localization. To avoid distillation of salient information, we rely on binary tests performed on the smallest possible scale, *i.e.* the voxel-level. We then summarize the contribution of all binary tests applied on a tensor to obtain a image in which pixel values represent the likelihood of abnormality of the considered tensor.

4.3.1 Metric definition

It is expected that all values of the difference of activation tensors $\Delta l_i(\mathbf{x}, \hat{\mathbf{x}})$ of voxels localized in normal structures are close-to-zero. However, we observed that, for a set of clean images, those values diverge more or less from zero. In order to take this phenomenon into account, we introduce a binary test of abnormality for each single voxel as *is the current voxel's activation greater than the maximum activation of all training images at this voxel location?* This test is designated as the **Out-of-Distribution (OoD)** test, and we refer to this corresponding detection approach as the OoD detection method.

Definition 4.5 (Out-of-Distribution test (OoD)). For a difference of activation tensors $\Delta l_i(\mathbf{x}, \hat{\mathbf{x}}) \in \mathbb{R}^{W_i \times H_i \times C_i}$, we apply the binary OoD test at each voxel location w_i, h_i, c_i of the network:

$$[\Delta l_{w_i, h_i, c_i}(\mathbf{x}, \hat{\mathbf{x}})]_{\text{OoD}} = \begin{cases} 1, & \text{if } \Delta l_{w_i, h_i, c_i}(\mathbf{x}, \hat{\mathbf{x}}) > \max_{\mathbf{x}_t \in \text{Train}} \Delta l_{w_i, h_i, c_i}(\mathbf{x}_t, \hat{\mathbf{x}}_t) \\ 0, & \text{else.} \end{cases}$$

This process converts the 3-D tensor $\Delta l_i(\mathbf{x}, \hat{\mathbf{x}}) \in \mathbb{R}^{W_i \times H_i \times C_i}$ into a 3-D binary tensor $[\Delta l_i(\mathbf{x}, \hat{\mathbf{x}})]_{\text{OoD}} \in \{0, 1\}^{W_i \times H_i \times C_i}$.

Definition 4.6 (Out-of-Distribution MAP (OoDMAP)). The anomaly map considered in this chapter is the OoDMAP. For each pixel location, we compute the percentage of positive OoD tests as described by:

$$\text{OoDMAP}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{I} \sum_i^I \text{interpolate}(\mu([\Delta l_i(\mathbf{x}, \hat{\mathbf{x}})]_{\text{OoD}})).$$

Since all layers i have equal or smaller spatial dimension than the original image ($\forall i, W_i \times H_i \leq W \times H$), we perform a bilinear interpolation before combining the contribution of all layers. The computation process of the OoD MAP at a given layer i is illustrated in Figure 4.3.

Definition 4.7 (Out-of-Distribution Ratio (OoDR)). The anomaly score considered in this chapter is the OoDR and is the percentage of voxels in the entire network having a positive OoD test:

$$\text{OoDR}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{W \times H} \sum_{w_i, h_i}^{W, H} \text{OoD MAP}_{w_i, h_i}.$$

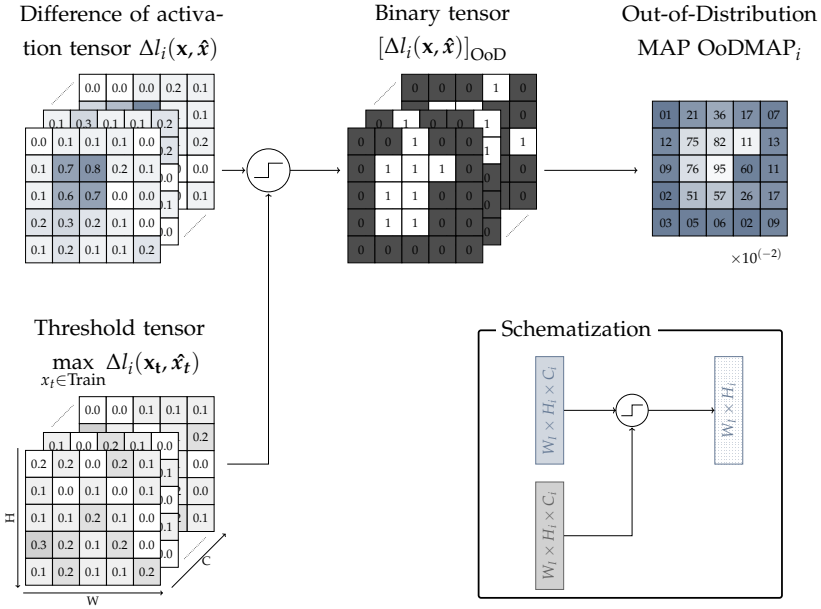


Fig. 4.3 Out of an activation tensor and a corresponding threshold tensor, we apply a voxel-wise binary thresholding process. This results in a binary tensor that represents whether or not the voxel's activation is abnormal. The mean over the channel axis of this binary tensor leads to the out-of-distribution map.

4.3.2 Pipeline

Figure 4.4 illustrates the entire pipeline generating anomaly maps and anomaly scores based on our activation tensors visualization mechanism. Two main components leverage this approach:

- (i) The double inference mechanism (described in Figure 4.1) that relies on the absolute value of the difference between an input image and its normal reconstruction, in the activation tensor space.
- (ii) The binary tests at the voxel level (described in Figure 4.3) that allows to identify all abnormal activations.

The network used to extract the activation tensor is the AESc trained to reconstruct normal images corrupted with the Stain model, and is the AESc + Stain model presented in Chapter 3. Note that all trainable parameters have been frozen, since we want to compare two detection mechanisms and not the training quality.

In order to generate an anomaly map with the same spatial dimensions than the original image, it is necessary to map the multiscale activation tensors onto a single image. A significant challenge in this process is the variation in global intensities across different network layers observed in Figure 4.2, which complicates their integration into a single anomaly map. Initial attempts at normalization yielded suboptimal results. The proposed solution involves applying a binary test on each voxel to transform each intensity value into a binary indicator of normality. To this purpose, we approximate the normal model by inferring the training images. Consequently, the mean of binary tensors along the channel axis represents the percentage of positive tests. This approach provides a robust mechanism for combining contributions from different network layers into a coherent anomaly map.

The binary test methodology demonstrates significant robustness in integrating multiscale activation tensor data, effectively highlighting anomalies with spatial coherence to the original image dimensions. The simplicity and intuitiveness of this approach offer a reliable means of detecting defects by synthesizing multiscale activation data into an interpretable format.

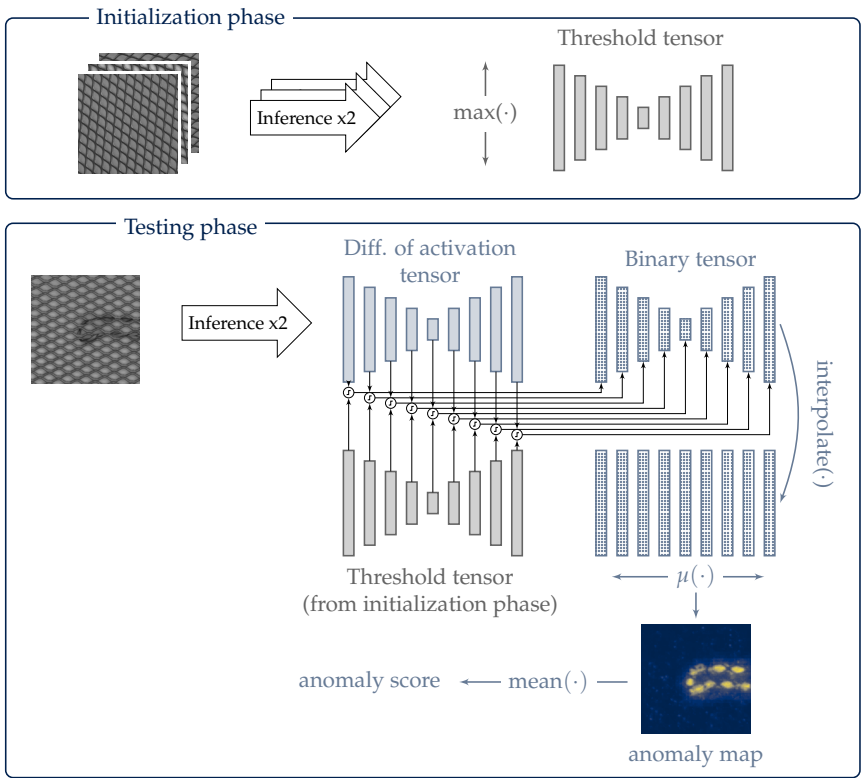


Fig. 4.4 (Initialization phase) Each training image and respective reconstruction are inferred to get the activation tensors. Voxel-wise maximum over all images of the absolute value of the difference between original image and reconstruction activation tensors are used as a normality model. (Testing phase) All images and their respective reconstructions are inferred to obtain difference of activation tensors. Binary tests are applied on each voxel and obtained binary tensors are interpolated into a unique spatial dimension. Anomaly map reflects the percentage of abnormal tests for a given spatial location while anomaly score is the global percentage of abnormal tests.

4.3.3 Quantitative and qualitative results

This analysis focuses on comparing the current activation tensor-level approach (OoD) with the initial approach (AESc) operating in the original image space. As previously mentioned, the networks used for the OoD approach share the same weights as AESc + Stain. This design choice aims to highlight the differences between the two detection mechanisms, independent of the hyperparameters used for training the autoencoders. Both quantitative and qualitative comparisons are conducted to evaluate performance.

In Figure 4.5, we visually compare the anomaly localization results obtained using AESc and the OoD approach. The anomaly map for the former method is the reconstruction residual, while for the latter, it is the OoDMap. To conduct this comparison, we selected one sample from each out of four image categories: capsule, zipper, bottle, and leather. In all four cases, the OoDMap demonstrates superior defect localization, as evidenced by the high intensities in the anomaly map corresponding to the defective structures. In contrast, the reconstruction residual shows less discrepancy between the values in normal and defective regions. As previously discussed, the intensity of the residuals diminishes when a defective structure has low contrast with its surroundings, making the AESc method less effective for the selected images. Conversely, the OoD approach does not depend on this contrast, which is reflected in the OoDMap results.

These examples confirm that achieving good reconstruction alone is insufficient for detecting defective structures based on the reconstruction residual. Despite the nearly perfect reconstruction of the zipper using AESc, defect localization based on the residual is not straightforward. Similarly, although the S-shaped defect in the bottle is partially reconstructed, the residual is only slightly higher in the defective structure. In the capsule example, the scratch is partially reconstructed but remains undetectable based on the residual. However, those partial reconstructions suggest that AESc has the capacity to detect defects, even if the network does not reconstruct those regions perfectly. This intuition implies that the trained internal representations could be effective in detecting these structures. This hypothesis is confirmed by the OoDMaps, which show higher rates of abnormal activations in the defective regions, enabling the detection of these defects by the use of internal representations.

However, a notable drawback of the OoDMAP method is the increase in false positives that we associate with two reasons. First, the defective regions are larger due to the multiscale nature of the detection. Second, the design of the binary tests makes the detection highly sensitive to small variations. While this sensitivity facilitates defect detection in the capsule, zipper, and bottle samples, it also leads to false positive detections when normal structures slightly deviate from the training set, as evidenced by the high OoDMAP intensities near the shadow (upper part of the image) in the leather example.

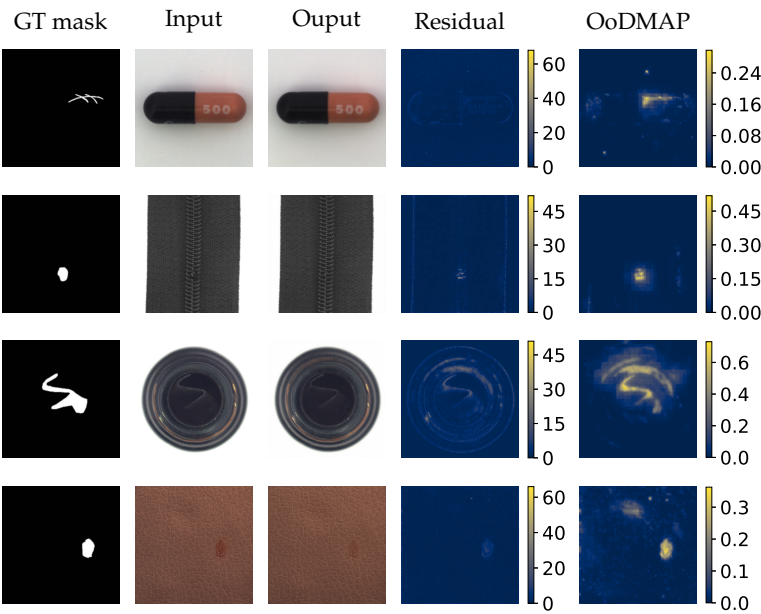


Fig. 4.5 Comparison of anomaly maps obtained with the reconstruction residual (fourth column) or the OoDMAP (fifth column). Three defective objects are considered, namely a capsule (first row), a zipper (second row) and a bottle (third row), as well as one defective texture, a leather (fourth row). First column contains the defective mask, second column shows the image fed in the networks while third column is the reconstructed clean version.

Table 4.1 Performance obtained with AESC and the OOD method on anomaly detection and localization tasks^a.

Method		I-AUROC		P-AUPRC	
		OOD	AESc	OOD	AESc
Textures	Carpet	0.69	0.83	0.57	0.25
	Grid	0.96	0.98	0.55	0.51
	Leather	0.84	0.90	0.40	0.38
	Tile	1.0	1.0	0.77	0.51
	Wood	1.0	1.0	0.58	0.37
	Mean ^b	0.90	0.94	0.58	0.40
Objects	Bottle	0.95	0.98	0.73	0.52
	Cable	0.83	0.87	0.52	0.43
	Capsule	0.83	0.77	0.36	0.24
	Hazelnut	0.88	0.95	0.51	0.51
	Metal Nut	0.75	0.82	0.46	0.52
	Pill	0.90	0.86	0.26	0.32
	Screw	0.67	0.87	0.10	0.15
	Toothbrush	0.97	0.91	0.21	0.35
	Transistor	0.91	0.86	0.45	0.16
	Zipper	0.96	0.90	0.34	0.15
	Mean ^c	0.87	0.88	0.40	0.34
Global mean ^d		0.88	0.90	0.45	0.36

^a For each row, the best performing approach for anomaly detection and localization is highlighted in boldface.

^b Mean I-AUROC and P-AUPRC obtained over the classes of images belonging to the texture categories.

^c Mean I-AUROC and P-AUPRC obtained over the classes of images belonging to the object categories.

^d Mean I-AUROC and P-AUPRC obtained over the entire dataset.

The difference between the AESc and the OoD approaches is significant when comparing their quantitative performance on anomaly detection and localization tasks. The OoD approach for anomaly localization results in a notable increase in the mean P-AUPRC, from 0.36 to 0.45. This substantial improvement is attributed to the ability of the OoD method to detect defects that contrast poorly with neighboring regions. However, for four image categories (meta nut, pill, screw, and toothbrush), the use of the OoD mechanism is detrimental compared to the original residual-based localization method. An detailed analysis of the failure cases occurring with those image categories will be discussed in Section 4.4. In terms of anomaly detection, the trend is less pronounced. The AESc approach performs better in 8 out of 15 categories, while the OoD approach outperforms the AESc approach in 5 categories. Across the entire dataset, transitioning from the OoD to the AESc approach results in a slight decrease in I-AUROC by 0.02.

This final phase of our study investigates the relationship between AESc and OoD methodologies in both anomaly localization and detection. Our objective is to determine whether failure cases are common across both methods, indicating challenging scenarios, or if each method operates effectively under distinct conditions. Specifically, we examine the question: *Does a high residual value imply a high OoD score, and vice versa?* Figure 4.6 represent the anomaly scores/anomaly maps intensities in corresponding images/pixels obtained with the AESc and OoD on the bottle category. Firstly, in the pixel-wise scenario, OoDMap intensities are distributed across the entire domain, whereas residual intensities are concentrated near lower values. The rug plots along the axes illustrate that OoDMap achieves better separation between normal and defective pixels, aligning with the quantitative anomaly localization results. The point cloud is not localized near the diagonal but instead aligns vertically, indicating that many pixels with low residuals are mapped to pixels with high OoDMap values. This shift is advantageous for the defective class, although it results in increased false positive detections for normal pixels. This increase in false positives is evident in the visual example in Figure 4.5 and is reflected in the graph by the presence of blue crosses in the upper left quadrant. Secondly, there is a strong correlation between image-wise anomaly scores based on residuals and OoDR, as demonstrated by the alignment of the point cloud along the diagonal. This similarity in anomaly score computation is consistent with the quantitative results presented earlier.

In summary, the analysis of the relationship between AESc and OoD for anomaly detection and localization reveals the complementary nature of these two approaches, particularly for the defect localization.

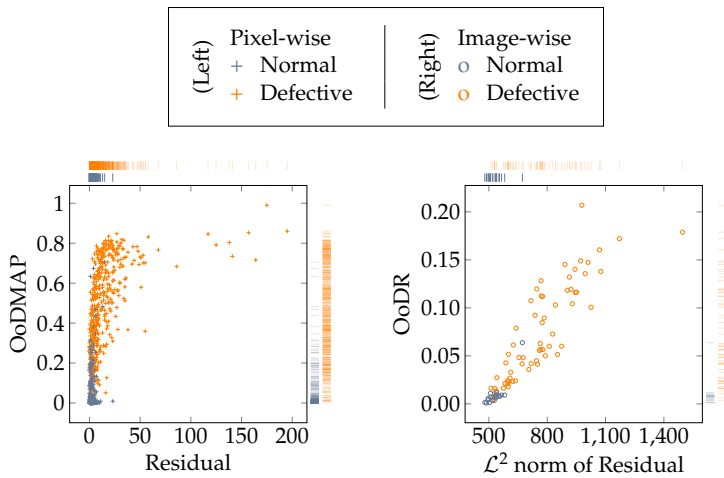


Fig. 4.6 (Left) For 500 normal and defective pixels sampled randomly across the bottle category¹, a 2-D scatter plot represents the Residual intensities versus the OoDMAP intensities. (Right) The $\mathcal{L}^2$ norm of the reconstruction residual and respective OoDR for all images of the bottle category.

4.4 Exploring operational boundaries and working hypotheses

The previous result analysis highlighted disparities between AESc and OoD methods, especially when addressing the anomaly localization task. This section explores two distinct properties of the OoD method that emphasize these differences. Unlike AESc, OoDMAP excels in detecting defective structures that are less contrasted with their surroundings. While previous illustrations and discussions have illustrated this ability, our focus here shifts to exploring more nuanced scenarios where these methods diverge.

¹Random sampling of 500 normal and defective pixels was employed to improve the clarity of the graphical representation. Given the extremely high number of pixels involved, particularly normal pixels, visualizing the distribution of both classes on the same graph would have been challenging otherwise.

4.4.1 Property 1 - False positives in the background

The sensitivity of the OoD method is influenced by threshold tensors, which are determined based on the training images. In the presence of a smooth background, these threshold values become extremely low, increasing the likelihood of OoD predicting false positives in these unremarkable regions of the image.

To illustrate this property, we compare two objects categories with dissimilar backgrounds: pill and capsule. The background of the pill images is dark and uniform, whereas the background of the capsule images is lighter and noisy. Figure 2.1, previously introduced, offers a visual overview of the MVTec AD dataset, enabling a comparison of the backgrounds across all image categories. Regarding the quantitative results, the OoD method outperforms AESc approach in anomaly localization for the capsule category (0.36 vs. 0.24) but not for the pill category (0.26 vs. 0.32). Figure 4.7 focuses on anomaly localization by presenting relationship between the OoDMap and residual pixel intensities, for both pill and capsule categories. Two main observations can be made:

- (i) The distribution of residuals for defective pixels has a high density near small values. This indicates that defective structures are either poorly contrasted with their surroundings or that defects are preserved in the reconstructed normal image. In the capsule category, using OoDMap instead of residuals helps to spread the defective intensities across the entire range of the anomaly map. This effect is also present in the pill category, to a lesser extent.
- (ii) The distribution of residuals for normal pixels should always be concentrated around small intensities. However, we observe that a significant number of pixels with small residuals are mapped to high OoDMap intensities, leading to false positive detections. This effect is more pronounced in the pill than in the capsule images.

A deeper analysis of the anomaly localization results allowed us to identify the origin of these false positive detections in the pill category when using the OoD method. We found several pill images where background pixels are associated with high OoDMap intensities. One example is shown in Figure 4.8. For comparison, we also identified a capsule image where the background is associated with high OoDMap intensities, in contrast with the majority of capsule images. The second row of the upper part (capsule) and the lower part (pill) in Figure 4.8 represents the mean of the difference

in activation tensors along the channel axis. In both capsule and pill images, higher intensities are observed in the background, particularly in deeper layers ($i = 3,4,5,6$) of the network. On those intensities, we apply the binary tests with threshold values represented in the third row and obtain the mean percentage of positive tests along channel axis represented in the fifth row. For the capsule category, high threshold values are associated with the background, mitigating the percentage of positive binary tests in these regions. Conversely, in the pill category, lower threshold values are linked to the background, resulting in a higher percentage of positive binary tests.

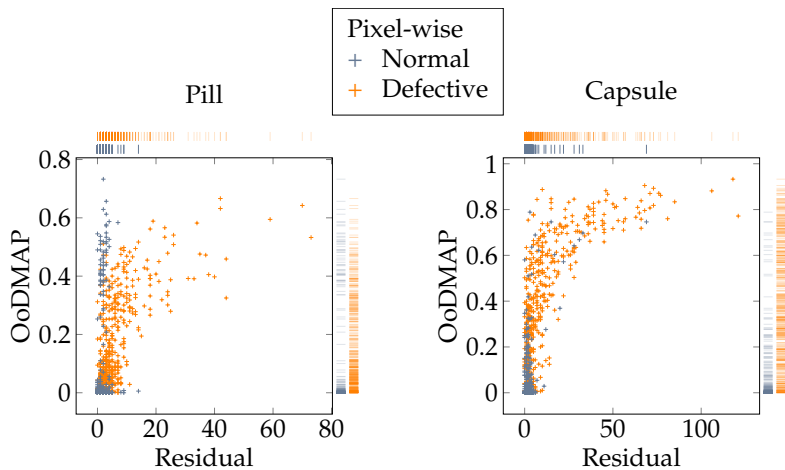


Fig. 4.7 For 500 normal and defective pixels sampled randomly, a 2-D scatter plot represents the Residual intensities versus the OoD MAP intensities. Left graph is samples from the pill category while right graph from the capsule category.

This example underscores the role of binary tests in ignoring minor variations in the reconstruction of normal image regions. However, it also highlights the sensitivity of setting thresholds. To address this, if the applicative task allows, we recommend increasing threshold values in regions unlikely to contain objects, particularly in cases where the background can be disregarded.

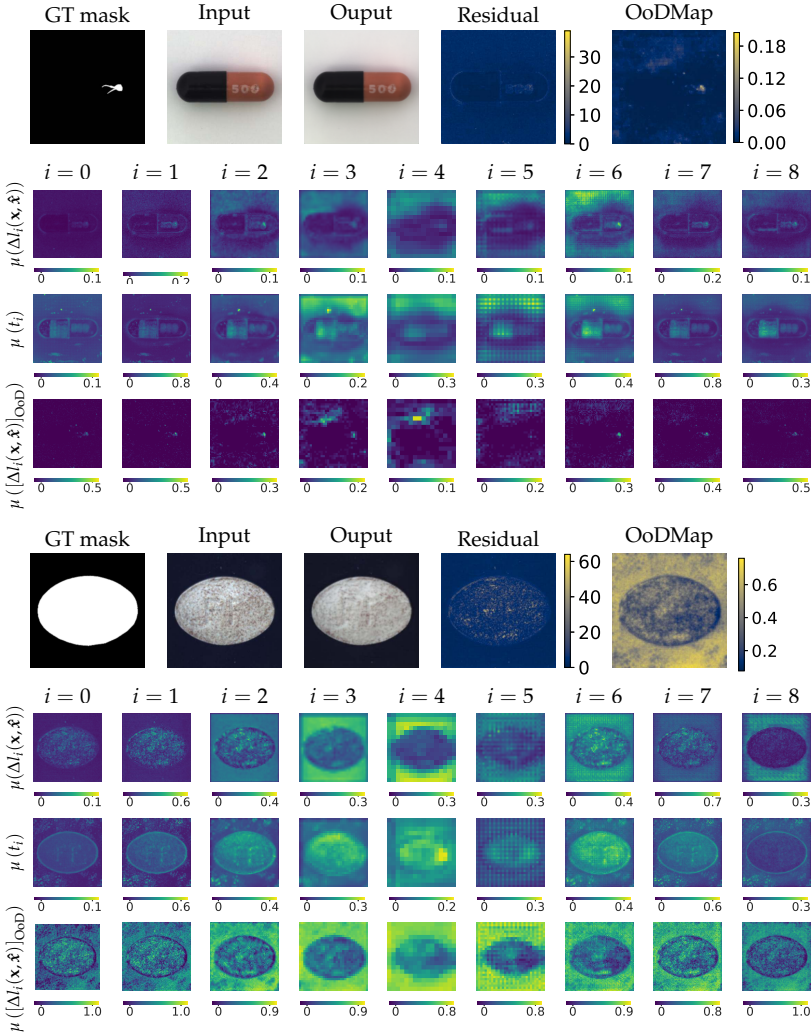


Fig. 4.8 (First row) ASc and OoD anomaly maps on a defective input object, a capsule for the upper part and a pill for the lower part. (Second row) Mean of the difference in activation tensors along the channel axis in every layer. (Third row) Mean of threshold tensors along the channel axis. Here use the notation $t_i = \max_{\mathbf{x}_t \in \text{Train}} \Delta l_{w_i, h_i, c_i}(\mathbf{x}_t, \hat{\mathbf{x}}_t)$. (Fourth row) Map of the percentage of positive binary tests in every layer.

4.4.2 Property 2 - Missing object

If objects are aligned in the training set, the OoD approach can be sensitive to translation, even if AESc if not.

The transistor category, depicted in Figure 2.1, consists of small transistors with three pins aligned vertically on a non-homogeneous orange background. The black casing is typically the most accurately reconstructed part of the image by AESc. In contrast, the background and the pins yield higher residuals, even in the training set. This discrepancy is shown in the third line of Figure 4.9. In the defective image, the entire transistor is missing, and AESc reconstructs the background across the entire image, aligning with translation equivariance property of convolutional networks. We observe that reconstructing the background in the area usually occupied by the black casing leads to an increase in the number of positive binary tests. We interpret it as *the network reconstructs a complex structure in place of a simpler one, where small residuals tend to activate positively the OoD tests*.

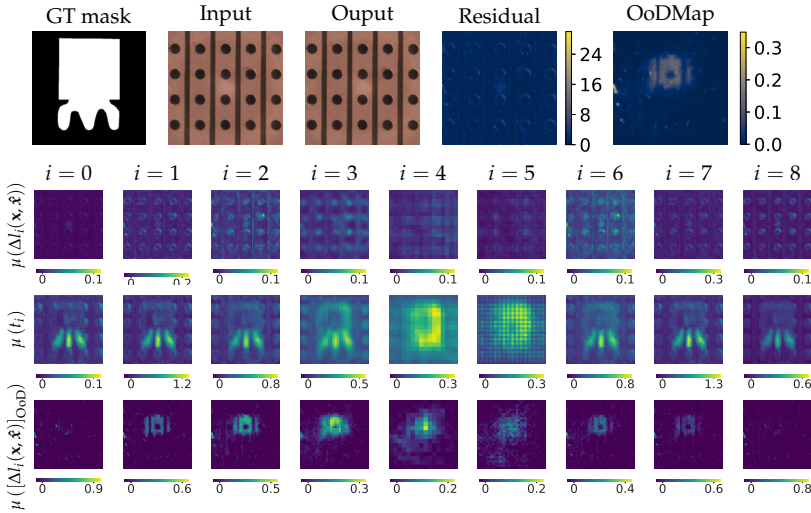


Fig. 4.9 (First row) AESc and OoD anomaly maps on a defective transistor. (Second row) Mean of the difference in activation tensors along the channel axis in every layer. (Third row) Mean of threshold tensors along the channel axis. Here use the notation $t_i = \max_{\mathbf{x}_t \in \text{Train}} \Delta_{w_i, h_i, c_i}(\mathbf{x}_t, \hat{\mathbf{x}}_t)$. (Fourth row) Map of the percentage of positive binary tests in every layer.

4.5 Summary and fundamental properties

The detection of anomalies in the image space is highly sensitive to the contrast between defective structures and their surrounding neighborhood. Even when a network achieves perfect reconstruction, subtle defects may still go undetected if they lack sufficient contrast. However, if the network can produce an accurate reconstruction, it suggests that the internal learned patterns are capable of discriminating between normal and defective structures.

The aim of this chapter is to leverage the rich information embedded within the tensor space to improve anomaly detection. To achieve this, we employ two key mechanisms:

1. **Activation Difference Analysis.** We examine the difference in activation tensors, denoted as $\Delta l_i(x, \hat{x})$, between an input image x and its reconstructed normal version $\hat{x}$. This approach highlights activations corresponding to defective structures while attenuating those linked to normal features. By focusing on the differences, this technique effectively emphasizes the unique patterns associated with anomalies.
2. **Binary Tests on Activations.** We apply binary tests to the activation values to identify deviations from the normal range observed in the training set. This method allows us to detect activation values that fall outside the typical distribution, signaling potential defects.

This approach offers a significant advantage in detecting low-contrast defects that the network captures internally but are challenging to identify in the image space. Nevertheless, a consequence of this approach is an increase in the number of false positive detections, particularly in areas that are well-reconstructed in the training set, such as the background.

Quantitatively, this method has led to a significant improvement in performance for anomaly localization. The overall P-AUPRC score on the MVTec AD dataset increased from 0.36 to 0.45. This approach shows superior performance across all texture categories and five object categories. The observed decrease in performance for the remaining object categories can largely be attributed to the method's sensitivity to slight variations in background elements. When the background in the training images is particularly smooth or consistent across samples, the threshold for activation

values becomes extremely low, making the method prone to false positives in these regions.

While this sensitivity may result in the misdetection of elements in the background, it can also facilitate the identification of more complex structural defects, such as the absence of an expected object. This capability underscores the potential of the proposed approach to discern subtle anomalies that might be undetected by conventional image-based detection methods.

Overall, this chapter demonstrates the effectiveness of leveraging tensor space information for anomaly detection, offering a more nuanced and detailed understanding of defects. The insights gained from this approach pave the way for further research into refining and optimizing anomaly detection techniques, particularly in scenarios where subtle and complex defects are present.

5

Detection in a multi-level quantized space

ANOMALY detection within the original image space presents significant challenges due to high dimensionality and intractable data distributions. Traditional reconstruction-based approaches rely on the residual between the input and a normal reconstruction to detect defective structures. However, reconstruction residuals are notably ineffective for defective structures with poor contrast against their surroundings. Despite this limitation, networks trained with classical reconstruction-based approaches can still extract representative patterns within the data. As done in Chapter 4, modifications to the detection process can leverage the shortcomings of the residual metric to exploit the rich internal representations of the network.

In this chapter, we focus on projecting the input image into a multi-level, low-dimensional space, which is parametrized by the autoencoder with skip connections that was introduced in Chapter 3. This approach is designed to capture the essential features of the input image at various levels of abstraction, facilitating more effective anomaly detection. To accomplish this, we enhance the network architecture by integrating a quantization module into each skip connection of the autoencoder. These quantization modules are inspired by the Vector Quantized-Variational AutoEncoder (VQ-VAE)

architecture [43], a model known for its ability to learn discrete representations. Each module includes a learned codebook composed of multiple prototype vectors, which represent the fundamental building blocks or patterns derived from the training data.

During each forward pass through the quantization module, the 3D input tensor is transformed by replacing each 1D vector within the tensor with the closest matching prototype vector from the learned codebook. This process results in the generation of a 2D map, where each element represents the index of the prototype vector that was selected to replace the corresponding 1D vector from the original input. These 2D quantized index maps, which serve as low-dimensional representations of the input image, are pivotal in capturing the essential features of the data while significantly reducing its dimensionality. The integration of a quantization module into each skip connection of the autoencoder enables the generation of multiple 2D quantized index maps for each input image. These maps are derived from various stages of the network, thus reflecting different levels of abstraction and feature representation within the model. A particularly noteworthy characteristic of these index maps is their inherent agnosticism with respect to the contrast between the input structures. The values within these maps correspond to indices within the codebook, which are inherently unordered. This attribute allows the model to encode images in a manner that is less sensitive to the variations in contrast that might exist between different structures in the input image.

The 2D quantized index maps obtained at multiple levels within the autoencoder are subsequently fed into a segmentation network that has been specifically trained on synthetic defects to localize abnormal structures within the input images. By leveraging these multi-level representations, the segmentation network is able to detect deviations from the normal index distribution, effectively identifying regions that differ from the learned normal patterns. The output of this segmentation process is a binary mask that highlights the areas of potential defects. This binary mask is then combined with the reconstruction residual to obtain the anomaly map, which characterizes this detection mechanism. The incorporation of the information contained in the binary masks improves the performance of the baseline model in both anomaly localization and detection tasks.

5.1 Background on discrete latent representations

A vanilla autoencoder (AE) is composed of two parts: the encoder, which maps the high-dimensional input x to a low-dimensional latent space z , and the decoder, which maps z back to an output $\hat{x}$. The bottleneck forces the network to learn a compressed version of the input, comprising high-semantic information. The latent space is unconstrained, which results in suboptimal generative performance when decoding vectors sampled randomly in this low-dimensional space.

In order to ensure optimal reconstruction performance for vectors sampled randomly in the latent space, data representations in the bottleneck can be constrained by a prior distribution. These networks are referred to as **Variational AutoEncoders (VAEs)** [98] and extend the traditional AE architecture by introducing a probabilistic framework for generating the compressed representation of the input data. Rather than mapping the input x to a fixed vector z , x is mapped into a parametrized distribution p , which is typically a multivariate Gaussian distribution. VAEs comprise three parts:

1. An encoder e that parametrizes the posterior distribution $p(z|x)$;
2. A prior distribution $p(z)$ representing initial assumptions about the latent space;
3. A decoder d that parametrizes the likelihood $p(x|z)$.

Objective of the training is to describe the distribution of the latent variables z given the observation of the variables x . This can be done with variational inference detailed in Appendix C. The loss function of a VAE become:

$$L_{\text{VAE}} = \underbrace{\log p(x|z)}_{\text{Reconstruction loss}} - \underbrace{KL(p(z|x), p(z))}_{\text{KL divergence}}. \quad (5.1)$$

The reconstruction term maximizes the (log-)likelihood of generating real data while the second term minimizes the **Kullback-Leibler (KL) divergence** between the posterior and the prior distributions.

The **Vector Quantized-Variational AutoEncoder (VQ-VAE)** [43], introduced by Van den Oord *et al.* in 2017, represents an extension of the traditional VAE. Unlike VAEs, which learn a continuous random variable in the latent space, VQ-VAE is designed to learn a discrete, constrained latent variable. Another key distinction between VQ-VAE and traditional VAEs is the way in which the prior distribution is handled. In VAEs, the prior is fixed, usually as a standard Gaussian distribution. In contrast, VQ-VAE does not rely on a fixed prior; instead, it learns the prior during the training process, allowing the model to adaptively determine the most suitable latent space representation based on the data.

The VQ-VAE architecture incorporates a quantization module, illustrated in Figure 5.1, within the bottleneck of an autoencoder. The module relies on a learned codebook containing K prototype vectors $e_k \in \mathbb{R}^{1 \times 1 \times C}$. Out of an input latent tensor $z_e \in \mathbb{R}^{W \times H \times C}$, the latent variable z is obtained with a nearest-neighbor lookup in the codebook. The output tensor $z_q \in \mathbb{R}^{W \times H \times C}$ is then obtained by concatenating vectors from the codebook. At a spatial location, the corresponding vector in z_q is the codebook vector whose index is determined by the latent variable z at this location.

Due to the nearest neighbor lookup in the codebook, it is not possible to train this quantization module through backpropagation. Hence, gradients are copied from the output tensor $z_q(x)$ to the input tensor $z_e(x)$. The overall loss function is actually composed of three elements:

$$L_{\text{VQ-VAE}} = \underbrace{\log p(z_q(x))}_{\text{Reconstruction loss}} + \underbrace{\| \text{sg}[z_e(x)] - e \|_2^2}_{\text{VQ loss}} + \underbrace{\beta \| z_e(x) - \text{sg}[e] \|_2^2}_{\text{Commitment loss}}, \quad (5.2)$$

where $\text{sg}[\cdot]$ is the stop gradient operator¹ and β a parameter set by default to 0.25, as suggested by the authors of [43]. The reconstruction loss ensures that the output and input distributions are similar. The VQ loss pushes the codebook prototypes towards the z_e tensor. The commitment loss prevents the embeddings from growing arbitrarily by pushing the z_e outputs towards the embeddings. Note that the training of the layers preceding z_e is ensured by the first and the third terms; the training of the layers following z_q is ensured by the first term only.

¹The stop gradient operator means that no gradient flows to the structure on which it is applied.

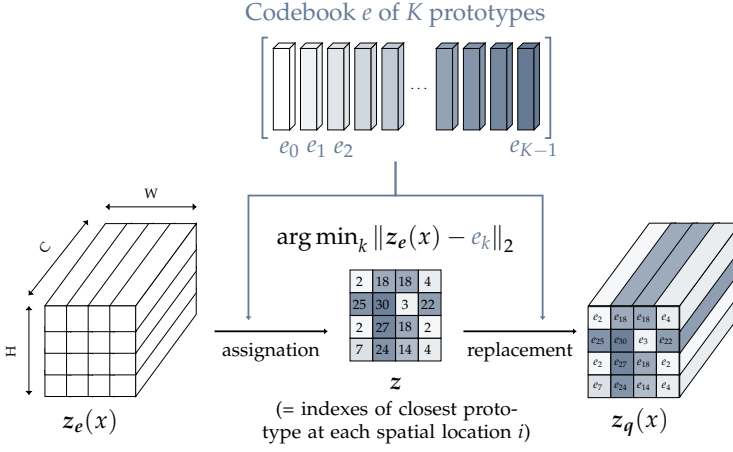


Fig. 5.1 Quantization module of the VQ-VAE (figure inspired from [43]). The latent variable z is a index map containing the index of the element of the codebook which is the closest of each i^{th} element of the input tensor z_e , with $i \in \mathbb{R}^{W \times H}$. Each i^{th} vector of the output tensor is replaced by the k^{th} prototype of the codebook whose index is encoded in z .

5.2 Related work

When anomaly detection is addressed with a reconstruction-based problem formulation, a significant limitation emerges due to the network's capacity to reconstruct previously unseen defective structures. To restrict the network's generalizability, memory modules or memory banks can be utilized. These modules employ codebooks comprising learned prototypes of varying spatial dimensions, spanning from $1 \times 1 \times C$ vectors to $W \times H \times C$ tensors, or even $w \times h \times C$ tensor patches where $1 < w < W$ and $1 < h < H$. Accessing the memory module or bank entails the construction of a weighted linear combination of prototype vectors, resulting in the generation of a regularized tensor that is fed to the subsequent layers of the network. This formulation represents the most common usage of quantized modules. These modules are typically not cited as an extension of the work of Van den Oord *et al.* and are more commonly referred to as memory modules. However, they share some technical aspects with the quantization module of the VQ-VAE.

A memory module in the bottleneck

In several proposed works [28, 99, 100], a memory module is introduced into the bottleneck of an autoencoder trained in a similar way to traditional reconstruction-based frameworks. The objective of this memory module is to implement explicit regularization in the latent space that favors the normal class. It is therefore anticipated that the network’s generalizability will be diminished, and that defective structures are less likely to be preserved in the reconstruction. The fundamental difference with the VQ-VAE module concerns the memory addressing. In contrast to the VQ-VAE module, which only considers the closest prototype for replacement, the memory modules under consideration here employ a linear combination of prototypes. This process is written as follows:

$$\hat{\mathbf{z}}_e = \mathbf{w}\mathbf{e} = \sum_{k=1}^K w_k \mathbf{e}_k, \quad (5.3)$$

where $\hat{\mathbf{z}}_e$ denotes the regularized tensor after the pass in the memory module (similarly to the $\mathbf{z}_q$ tensor described in Section 5.1), $\mathbf{e}$ represents the memory (or codebook), and $\mathbf{w}$ is a row vector with non-negative entries that sum to one. The values of the vector $\mathbf{w}$ are usually computed as follow:

$$w_i = \frac{\exp(d(\mathbf{z}_e, \mathbf{e}_i))}{\sum_{k=1}^K \exp(d(\mathbf{z}_e, \mathbf{e}_k))}, \quad (5.4)$$

where $d(\cdot, \cdot)$ denotes a similarity measurement, such as the cosine similarity. In several previous works, a sparsity constraint on the weight vector, denoted here as $\mathbf{w}$, was considered. This was done to ensure that only a few prototypes are used to approximate the latent representation, and ensure regularization. The rationale behind this sparsity constraint is to explicitly reduce the network’s generalizability.

A first example of this approach is **MemAE** [28], which integrated a memory module into the bottleneck of an autoencoder designed to reconstruct the input image. The prototypes in this codebook share the same dimensions as the embedding tensor. During memory access, a sparsity constraint is enforced, ensuring that at most $k < K$ prototypes are utilized in the weighted linear combination. This restriction curtails the network’s expressivity and ensures that the memory captures the most prevalent patterns of

normal data. **MemKD** [99] adapted this approach within a student-teacher framework. Also, **Park et al.** [100] adopted a similar strategy, employing a memory module containing $1 \times 1 \times C$ prototypes to detect anomalies in videos. Unlike the previously mentioned methods, they introduced a loss term that maximizes the discrepancy between prototypes. Here, the z_e and $\hat{z}_e$ tensors are concatenated and fed into the decoder.

A memory module on each skip connection

Autoencoders are known for producing blurry outputs, an issue that can be mitigated by adding skip connections, as discussed in Chapter 3. While previous studies have focused on introducing memory modules solely in the bottleneck, these modules can also be incorporated into skip connections.

Divide-and-Assemble Anomaly Detection (DAAD) [101] proposed integrating a patch-wise memory module into skip connections. The training process involves a combination of reconstruction loss, adversarial loss, and alignment loss, which aligns the feature representations of the input with those of the reconstructed image. To further regularize the reconstruction, a sparsity constraint ensures that only a small number of memory items are accessed at any given time. **Prototypical Residual Network (PRN)** [102] similarly incorporates memory modules into skip connections. Each prototype matches the spatial dimensions of the input vector, and a sparsity constraint is applied when accessing the codebook. Additionally, the memory addressing computes the element-wise difference between the query tensor and its closest prototype, generating residual tensors. All multi-scale residuals and original query tensors are concatenated and fed into a secondary network, which outputs a segmentation mask for anomalies. Synthetic defects are used to train this segmentation process.

Latent space resampling

A separate category of techniques involves a resampling in the discrete latent space. These approaches feed the latent discrete maps z of a VQ-VAE into a deep autoregressive model, which learns the probability distribution of the discrete latent space. This process allows for the computation of the likelihood of occurrence for each index in the discrete map z , which can then be replaced by the most probable index at each location. This mechanism regularizes explicitly the latent space towards a normal distribution.

Wang *et al.* [42] trained a VQ-VAE to reconstruct normal images, with the resulting index maps fed into PixelSnail [44], a convolutional autoregressive model. Instead of detecting abnormal structures based on the reconstruction residual, the difference between reconstructions with and without resampling is considered. Similarly, **Pinaya *et al.*** [103] used this mechanism but replaced PixelSnail with a transformer [104]. They also retained a binary mask with indices modified by the autoregressive model. An upsampled version of this mask is used to filter the reconstruction residual, thereby reducing false positive detections.

This chapter is devoted to a comprehensive analysis of VQ-VAE modules, employed as a projector into a quantized space for the purpose of anomaly detection. In contrast to memory modules, the objective here is to utilize index maps for image segmentation, rather than to achieve a more regularized version of the reconstructed image. This choice is driven by the aforementioned limitation of detection in the original image space.

5.3 Image projection in a multi-level quantized space

Related work incorporates a codebook module into the architecture of neural networks. While the concepts are similar, the utilization of these modules is driven by two distinct underlying principles:

- (i) For the first category, those module are referred to as memory modules or memory banks. Their objective is to regularize the reconstruction towards the normal class. The underlying rationale is that, since the codebook is trained exclusively on normal data, it captures only prototypical patterns of clean structures. This inherently limits the expressivity of the model, thereby diminishing its capability to reconstruct previously unseen structures.
- (ii) For the second category, the quantized latent space is associated with a deep autoregressive models that resamples index maps in the learned normal distribution. In this context, the quantization module does not explicitly regularize reconstruction. Instead, it projects input images into a low-dimensional discrete space, enabling the use of autoregressive models.

Our research aligns with the second theoretical foundation. Our aim is not to improve the reconstruction via a quantification module but rather to project input images into a low-dimensional quantized space. We will demonstrate that this representation is particularly effective for addressing the challenges of anomaly detection and localization.

5.3.1 Network architecture with quantization modules

Figure 5.2 illustrates the **q-AESc** network, an AESc network for which each skip connection is equipped with a quantization module. These quantization modules adhere to the explanation provided in Figure 5.1 as introduced in VQ-VAE [43]. For our experiments, we used codebooks containing $K = 32$ prototypes vectors with dimensions ranging from $1 \times 1 \times 16$ to $1 \times 1 \times 256$, depending on the layer to which the module is applied.

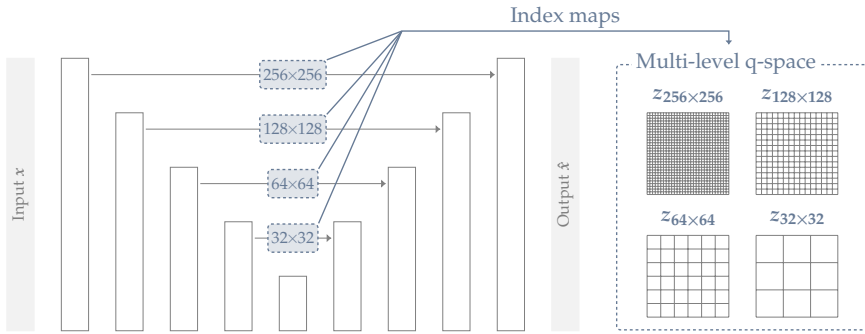


Fig. 5.2 Architecture of the q-AESc network with a representation of the multi-level quantized space. This space comprises four index maps, each obtained after inference through one of the four codebooks. We denote each index map as $z_{W_i \times H_i}$, where W_i and H_i represent the spatial dimensions of the corresponding layer.

To initialize our q-AESc, we copy and freeze all the weights of the convolutional layers shared with the AESc architecture, ensuring that only the quantization modules are subject to training. This strategy enables us to isolate and directly assess the effect of incorporating these modules on the reconstruction process. The codebooks are trained over 250 epochs to minimize the L_{VQ-VAE} loss function defined in Equation 5.2.

VQ-VAEs are known to suffer from *codebook collapse* [105]. This phenomenon occurs when a limited subset of prototypes is repeatedly selected for embedding input vectors, rendering the majority of prototypes inactive and thereby reducing the information capacity of the quantization module. To address this issue, we adopt a random restart strategy for inactive codes, as proposed in Jukebox [106]. If a prototype remains unselected for an entire epoch, we employ a random reinitialization method, resetting its value to that of another randomly selected prototype from the codebook. The likelihood of selecting a particular prototype is proportional to its utilization frequency within the epoch.

5.3.2 Space properties

In this chapter, we explore the process of projecting images into a multi-level quantized space using the q-AESc network, a key component of our proposed methodology. These quantized projections are subsequently utilized as inputs to a segmentation network, the architecture of which will be thoroughly described in Section 5.4.1. Prior to developing the integration of these quantized projections into our overall pipeline for anomaly detection and localization, we undertake a comprehensive study to underscore the properties of this quantized space.

It is important to emphasize that, within the context of our anomaly detection pipeline, the reconstructed images generated by the q-AESc network are not utilized in the final detection process. Instead, the focus is on the quantized representations, which serve as the basis for subsequent processing. The reconstructed images are included in this study to enhance our understanding of the q-AESc architecture and to provide a deeper comprehension of how the network transforms input data into its quantized form.

First insight into the quantized latent space.

Figure 5.3 illustrates results obtained with the original reconstruction-based framework based on the reconstruction residual. Additionally, the figure depicts the quantized latent representations derived from images of a cable and a bottle. In both instances, the reconstruction generated by q-AESc exhibits a lesser tendency towards regularization towards the normal distribution in comparison to the reconstruction produced by AESc. Specifically,

in the cable image, the external cut of the object is partially retained in the reconstruction, whereas in the bottle image, the white scratch is scarcely reconstructed and exhibits a resemblance to the original.

The q-AESc output exhibits a less detailed reconstruction. This is to be expected, since only a limited set of latent codes can be used to infer the input image. In the cable example, the green and blue rings exhibit notable detail loss, resulting in a uniform color reconstruction across the entire structure. This observation is consistent with the index maps z , suggesting that one prototype encodes the green color and another encodes the blue color. It should be noted that the relative grayscale intensity of the index maps z is not meaningful, as the prototype index is a categorical value. However, this representation effectively illustrates the consistency in prototype selection across a given structure.

An intriguing observation is revealed in the bottle image, wherein the inner circle of the object is not encoded by a single prototype. A slightly lighter circular structure is represented by a distinct prototype in all index maps. Despite its minimal visibility in the original image, the quantized map representation provides an embedding space where structures with minor color variations are encoded with different prototypes.

A further observation regarding the latent quantized space of the q-AESc network concerns the frequency of prototype utilization. The graphs in Figure 5.3 depict the frequency of use for each prototype across all four codebooks, with prototypes ordered in descending order of frequency. This distribution follows an exponential decay function, indicating that a small number of prototypes predominantly drive the reconstruction process. To illustrate, a single prototype may encode the background of the bottle image. Given the extensive area this structure covers, the corresponding prototype is frequently utilized. Concurrently, we promote the use of all prototypes with our random restart strategy to ensure that the entire capacity of the codebook is employed in generating latent codes.

To go further, a detailed analysis of the utilization of prototypes for encoding normal and defective structures is presented in Appendix D. This analysis highlights three samples where defective structures are encoded using prototypes that have a low likelihood of being selected in normal regions but a high likelihood of being selected in defective regions.

5 | Detection in a multi-level quantized space

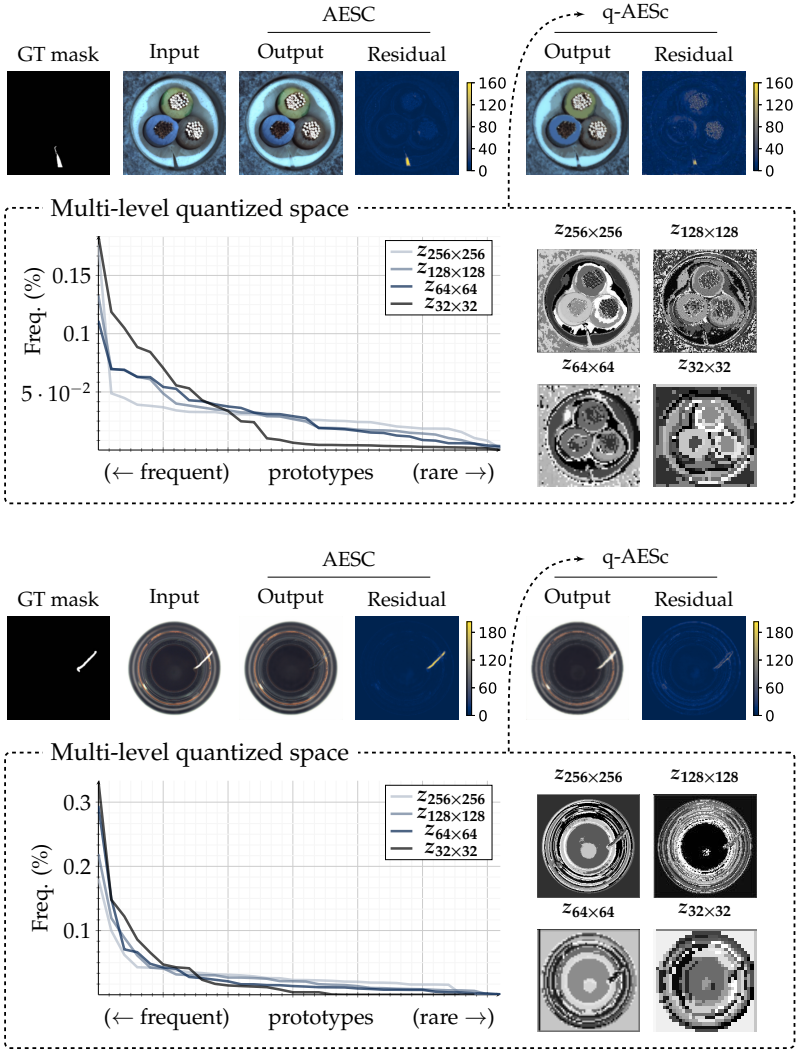


Fig. 5.3 Comparison between AESc and q-AESc for anomaly localization using residual over a cable (upper part) and a bottle (lower part). In the dotted frame, right image display the index maps, denoted as z_i . Index values range arbitrarily from black (0) to white (31). The graph shows the frequency distribution of prototype usage across the entire test set, with prototypes ordered from most to least frequently used.

Impact of the codebook size on latent representations

The second analysis, conducted prior to the introduction of our anomaly detection pipeline, concerns the impact of the codebook size on both latent representations and reconstruction. Intuitively, a smaller codebook will promote regularization towards a normal distribution, whereas larger codebooks will allow for greater generalizability in the reconstructed images. To ensure the veracity of this intuition, we will examine several examples. Specifically, we compare the effects across different codebook sizes: 8, 16, 32, and 64.

Figure 5.4 presents a visual comparison of the reconstruction of a defective pill using AESc and q-AESc with varying codebook sizes. In this example, the defective structures are identified by the presence of five small yellow dots. The reconstruction obtained with AESc partially replaces the yellow dots with normal structures, yet some yellow stains persists, indicating an imperfect reconstruction. The reconstructions obtained with q-AESc exhibit varying properties depending on the codebook size. When using 8 prototypes, the yellow dots are effectively removed and replaced with clean structures. The limited expressivity of this smaller codebook hinders the reconstruction of yellow structures, a color not encountered during training. As the number of prototypes increases, the dots are reconstructed, but their original yellow color is replaced with gray.

Figure 5.5 provides a comparable illustration of a defective hazelnut. As with the observations made regarding the pill example, a smaller codebook size encourages the reconstructed image to be regularized towards a normal distribution. However, even with a codebook size of 8, the replacement is not flawless, as the black hole is only partially removed. This contrasts with the pill example, where no yellow structures were seen during training, suggesting that black structures have been encountered during training.

In both examples, the residuals in normal areas for AESc are lower compared to those obtained with an autoencoder equipped with skip connections and quantization modules. This indicates that even with a larger codebook, the tensors passing through the quantization modules are subject to slight modifications. As a result, the residual obtained with q-AESc is not well-suited for anomaly detection using a reconstruction-based approach. Consequently, we do not utilize the q-AESc reconstructions in our pipeline.

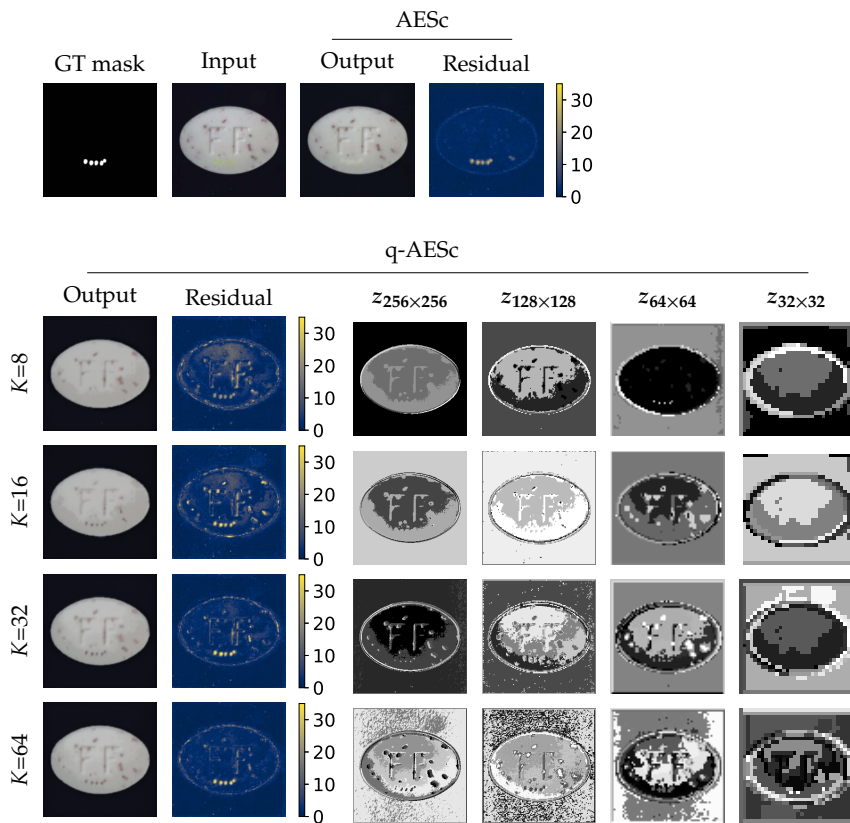


Fig. 5.4 Comparison of reconstructed normal versions of a defective pill using AESc and q-AESc with varying codebook sizes. The first row presents the ground truth mask, input image, reconstructed output, and residual obtained with AESc. The second, third, fourth, and fifth rows display the reconstructed outputs and residuals generated by q-AESc for codebook sizes of 8, 16, 32, and 64, respectively. Additionally, for each codebook size, the corresponding multi-level index maps are provided.

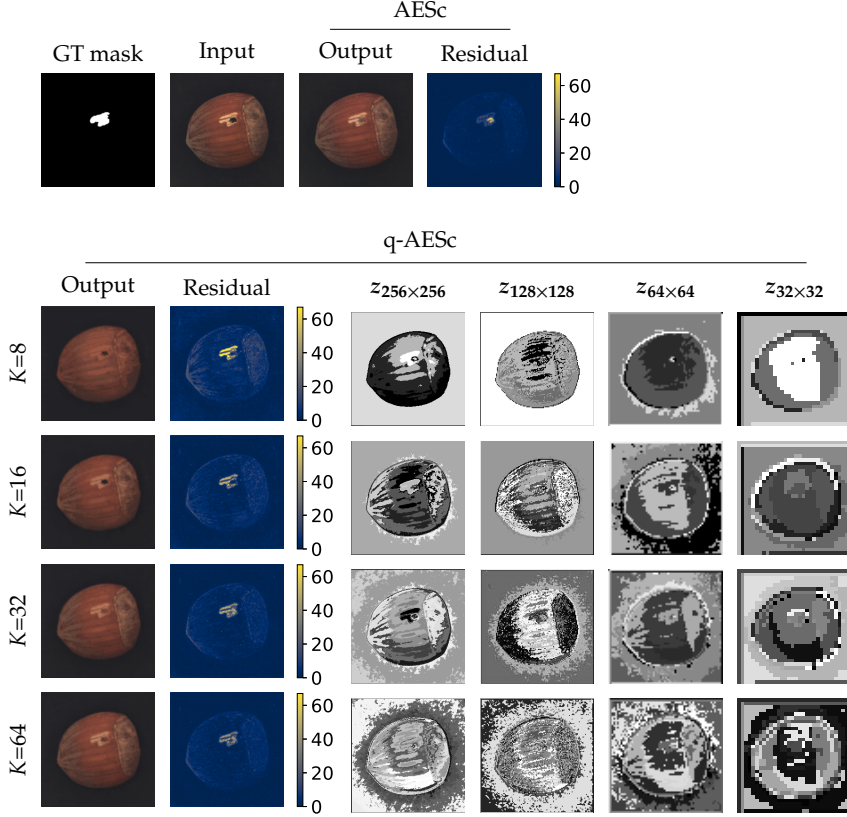


Fig. 5.5 Comparison of reconstructed normal versions of a defective hazelnut using AESc and q-AESc with varying codebook sizes. The first row presents the ground truth mask, input image, reconstructed output, and residual obtained with AESc. The second, third, fourth, and fifth rows display the reconstructed outputs and residuals generated by q-AESc for codebook sizes of 8, 16, 32, and 64, respectively. Additionally, for each codebook size, the corresponding multi-level index maps are provided.

5.4 Anomaly detection in the multi-level quantized space

Building upon the insights gained from the previous analysis, we propose the utilization of this multi-level quantized representation for the task of anomaly detection. In scenarios where structures exhibit subtle variations in color intensity, these differences can be challenging to detect using traditional methods based solely on reconstruction residuals. Such minor variations may not produce sufficiently large residuals to be identified as anomalies. However, within the multi-level quantized space, these subtle differences can be effectively encoded by distinct prototypes, enabling the detection process to be as straightforward as identifying anomalies between structures with completely different color intensities.

The approach presented in this chapter leverages a segmentation model that operates directly on the quantized latent space of an image. The primary objective of this model is to predict a binary mask that accurately identifies defective structures within the input image. This segmentation model is tasked with learning the complex distribution of normal index maps generated by our quantization modules. By doing so, the network becomes capable of distinguishing between normal and defective regions within the quantized space, even when the anomalies are characterized by only slight deviations from the normal distribution. By capitalizing on the distinct encoding capabilities of the quantized space, the segmentation model is able to effectively identify defective structures, thereby enhancing the overall performance of the anomaly detection framework.

5.4.1 Segmentation network

The segmentation network, termed **(Seg)_q** and whose architecture is depicted in Figure 5.6 consists of an encoding and a decoding path. The encoder transforms the input data into a bottleneck representation of dimensions $32 \times 32 \times 128$ through a series of four strided convolutional layers. At each layer, the encoder merges information from the previous layer with the quantized index map at the corresponding resolution, employing a custom fusion module for this combination. The decoder then reconstructs the embedding back to the original spatial dimensions via four successive up-convolution layers, producing a binary output of size 256×256 .

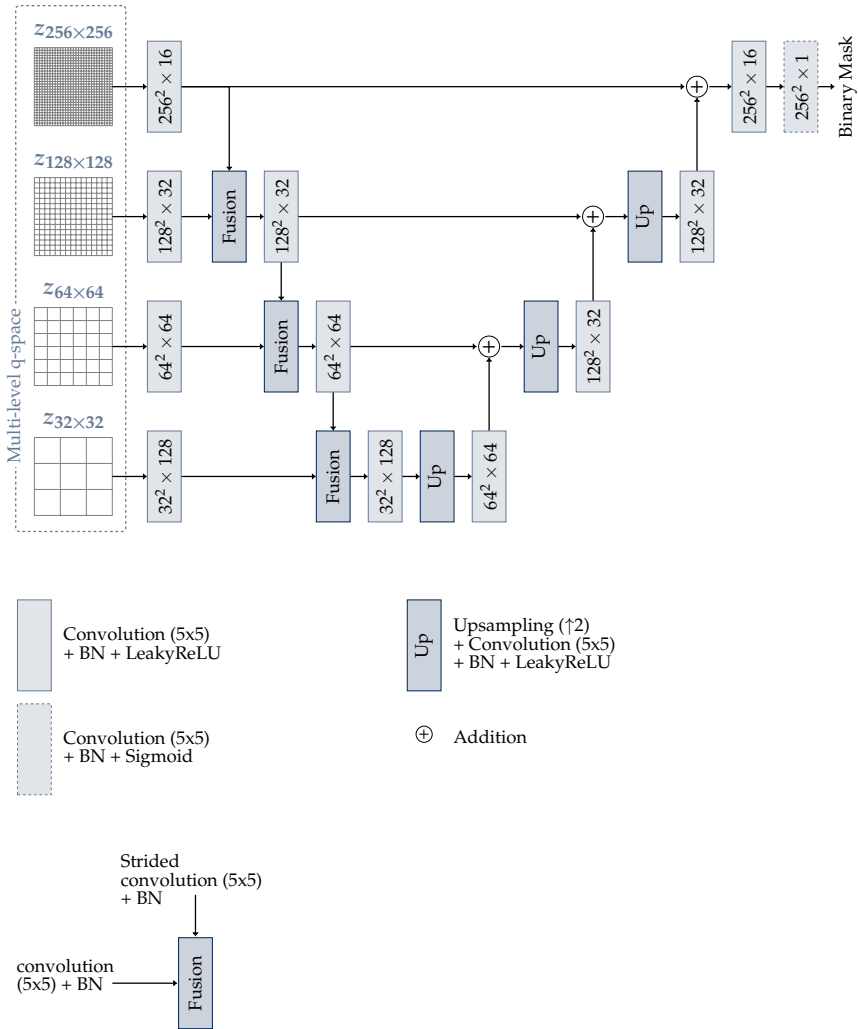


Fig. 5.6 The $(\text{Seg})_q$ architecture which takes the multi-level quantized space as input and outputs a binary segmentation mask with the same spatial dimensions than the original image.

To train the network, we exclusively utilized training images corrupted with the synthetic Stain noise model, as described in Section 3.2.2. This approach enabled the creation of pairs of input image and binary mask that identify the location of synthetic defects. The training process was conducted over 100 epochs using the AdamW optimizer [85], starting with an initial learning rate of 0.01. This was adjusted according to the OneCycleLR scheduler algorithm [86]. The objective function minimized during training was the BCE-Dice Loss, which is an aggregate of two components given by:

$$L_{\text{BCE-DICE}}(\mathbf{x}, \hat{\mathbf{x}}) = L_{\text{BCE}} + L_{\text{DICE}}. \quad (5.5)$$

The first component, the **Binary Cross Entropy (BCE)** Loss, is defined as:

$$L_{\text{BCE}}(\mathbf{x}, \hat{\mathbf{x}}) = -(\mathbf{x} \log(\hat{\mathbf{x}}) + (1 - \mathbf{x}) \log(1 - \hat{\mathbf{x}})), \quad (5.6)$$

and the second, the Dice Loss, is defined as:

$$L_{\text{DICE}}(\mathbf{x}, \hat{\mathbf{x}}) = 1 - \frac{2\mathbf{x}\hat{\mathbf{x}} + 1}{\mathbf{x} + \hat{\mathbf{x}} + 1}. \quad (5.7)$$

This combination leverages the diversity provided by the two loss components while maintaining the stability inherent to BCE.

5.4.2 Pipeline description

Having provided a comprehensive overview of the fundamental elements of our approach, we are now able to integrate these components into a unified framework for anomaly detection and localization. The aforementioned components include the AESc network, the q-AESc network, which projects images into a multi-level quantized latent space, and the $(\text{Seg})_q$ network, which utilizes these projections to produce a binary mask, effectively localizing defective regions in the original image space.

Figure 5.7 illustrates the pipeline designed for anomaly localization and detection, which enhances our classical reconstruction-based framework

with a detection in a multi-level quantized latent space. This workflow integrates both a classical reconstruction-based approach and a segmentation method, providing a robust framework for anomaly detection. The pipeline is composed of two parts:

- A. **The reconstruction pathway** has been designed with the objective of producing a reconstruction residual based on the AESc network, which was introduced in Chapter 3.
- B. **The segmentation pathway** has the objective of producing a binary mask that localizes defective structures based on the $(\text{Seg})_q$ architecture introduced in Section 5.4.1. This network takes as input the projection of the image in the multi-level quantized space described in Section 5.3, which was obtained with the q-AESc network.

The two outputs are combined to create the anomaly map under consideration in this chapter. This merging can be viewed as an enhancement of the original reconstruction-based approach.

Training phase

The following steps are required to train the elements of our pipeline:

- A. **The reconstruction pathway.** We imported the weights of the pre-trained AESc network from Chapter 3.
- B. **The segmentation pathway.** This is a two steps process:
 1. **Training the codebooks of q-AESc.** As outlined in Section 5.3.1, all convolutional layer weights are copied from AESc and maintained as frozen parameters, with only the codebook undergoing training. In the pipeline figure, trainable components are illustrated with dotted structures for enhanced clarity.
 2. **Training $(\text{Seg})_q$.** After establishing the multi-level quantized space, we proceed to train $(\text{Seg})_q$ to generate binary masks using synthetic examples. This step focuses on learning the mapping from quantized latent representations to defect localization. This process was outlined in Section 5.4.1.

It should be noted that the two pathways are trained independently on their respective tasks. The merging of both outputs is only considered during the inference stage.

Testing phase

During the inference stage, the pipeline generates two distinct yet complementary outputs: a reconstruction residual and a binary mask. These outputs are generated through the two aforementioned pathways:

- A. The reconstruction pathway.** The AESc network outputs a normal reconstruction of the input image, with the residual serving as the initial building block.
- B. The segmentation pathway.** The $(\text{Seg})_q$ network outputs a binary segmentation mask produced out of the projection of images in the multi-level quantized space obtained with q-AESC. This projection is the only output of the q-AESC network that is considered.

Let R denote the reconstruction residual obtained via the reconstruction pathway and M the binary mask obtained via the segmentation pathway. The anomaly map of this pipeline, designated as q-AnoMap, is a combination of both outputs defined by:

$$\text{q-AnoMap} = R * (1 - M) + \max(R) * M, \quad (5.8)$$

where $\max(R)$ is a positive value representing the maximum of the reconstruction error in the image residual. In other words, within regions predicted as normal by the binary mask, the intensity of the anomaly map corresponds to the reconstruction residual. Conversely, in regions predicted as defective, the intensity of the anomaly map is set to the maximum value of the reconstruction residual within the entire image. The anomaly score is then calculated as the $\mathcal{L}^2$ -norm of the anomaly map, in accordance with the methodology delineated in Chapter 3.

This definition is based on two underlying principles. From one perspective, the segmentation rarely results in the identification of false positives. The increased likelihood of abnormality in the positive regions of the mask is advantageous. Conversely, the segmentation mask is more prone to false negatives than the residual, which is why the anomaly map is retained as the residual in those regions. This approach uses the segmentation mask as a means of enhancing the residual, resulting in improved anomaly localization. This methodology enables a direct comparison with the baseline model, employing the I-AUROC and P-AUPRC metrics, thus providing a clear benchmark for evaluating performance enhancements.

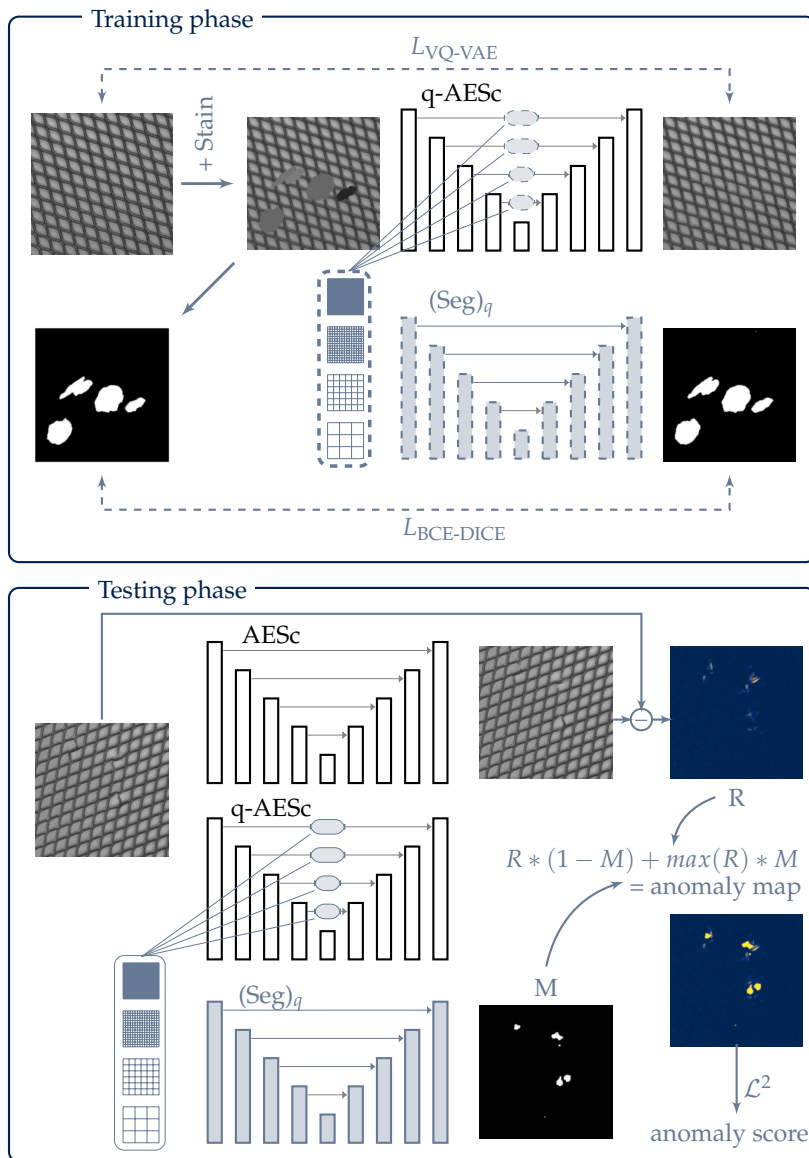


Fig. 5.7 Pipeline of the anomaly detection method based on the projection in a quantized space. For explanations, refer to the text in Section 5.4.2.

5.4.3 Quantitative and qualitative results

The primary objective of this study is to demonstrate the benefits of enhancing our original reconstruction-based approach with a segmentation network that operates within a multi-level quantized space. As discussed in Section 5.3.2, the quantized space is not intended to improve the reconstructed image itself, but rather to project the image into a different space that captures high-quality semantic information useful for anomaly detection. This information is then integrated into the reconstruction residual in the form of a binary mask.

Figure 5.8 provides a visual comparison between the traditional anomaly localization method based solely on reconstruction residuals and the proposed method, which combines the residuals with the binary masks generated by $(\text{Seg})_q$. For the four selected samples, the AESc reconstructions appear visually similar to the input images, with anomalies being subtly contrasted, making it difficult to differentiate between defect-free and defective areas based on reconstruction alone. Except for the hazelnut image, the $(\text{Seg})_q$ network clearly and accurately identifies defective regions, closely aligning with the ground truth masks. The multi-scale quantized representation is particularly beneficial for capturing the full extent of defects and enhancing the detection of anomalies that contrast poorly with their surroundings, such as the red stains in the wood image.

Table 5.1 provides a quantitative comparison of the three methods presented in this manuscript. For anomaly detection, AESc+ $(\text{Seg})_q$ emerges as the best-performing approach overall for both objects and textures. Notably, the decrease in I-AUROC from AESc+ $(\text{Seg})_q$ to AESc is minimal, with a worst-case difference of only 0.02. For anomaly localization, the OOD approach generally performs best. However, AESc+ $(\text{Seg})_q$ offers a significant improvement over AESc in specific categories, such as the tile category, where the P-AUPRC increases from 0.51 to 0.74. A visual illustration of this improvement can be seen in Figure 5.8. AESc achieves a higher P-AUPRC value compared to AESc+ $(\text{Seg})_q$ in only two categories: grid and toothbrush.

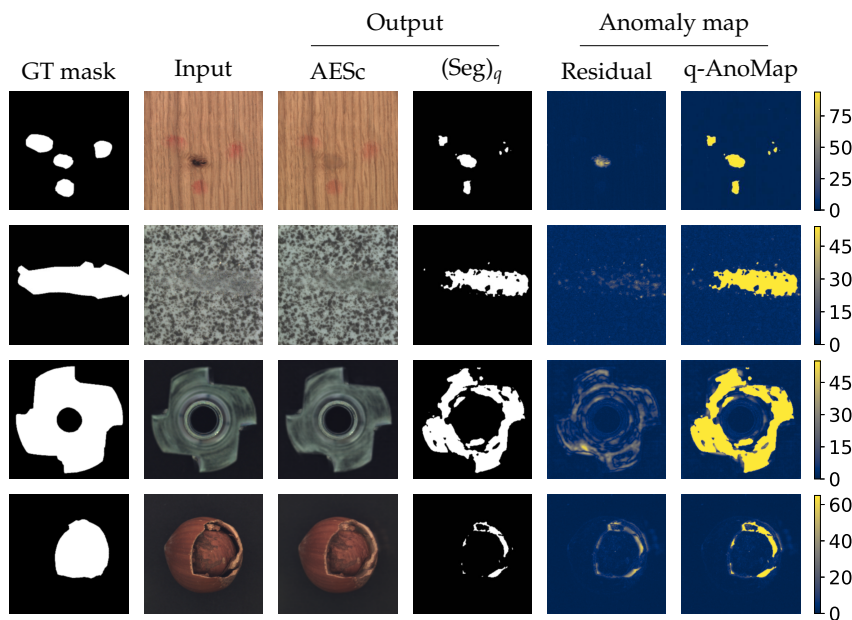


Fig. 5.8 Comparison between AESc and AESc+ $(\text{Seg})_q$ for anomaly localization over four examples. Two defective textures are considered, namely a wood (first sample) and a tile (second sample), as well as two defective objects, namely a metal nut (third sample) and a hazelnut (fourth sample). The figure is organized as follows: the first column displays the ground truth binary segmentation masks (Mask column). The second column shows the input images provided to both AESc and q-AESc (Input column). The third and fourth columns present the outputs, with the third column depicting the reconstruction results from AESc and the fourth column displaying the binary masks obtained using $(\text{Seg})_q$. Finally, the fifth and sixth columns illustrate the anomaly maps, comparing the classical reconstruction-based framework with the current approach that incorporates information from the multi-level quantized space.

Table 5.1 Performance obtained with AESC + Stain (AESC in the Table), the OOD method and AESc+(Seg)_q on anomaly detection and localization tasks^a.

		I-AUROC			P-AUPRC		
Method		OOD	AESC	AESC+ (Seg) _q	OOD	AESC	AESC+ (Seg) _q
Textures	Carpet	0.69	0.83	0.94	0.57	0.25	0.32
	Grid	0.96	0.98	0.99	0.55	0.51	0.45
	Leather	0.84	0.90	0.96	0.40	0.38	0.41
	Tile	1.0	1.0	1.0	0.77	0.51	0.74
	Wood	1.0	1.0	0.99	0.58	0.37	0.44
	Mean ^b	0.90	0.94	0.98	0.58	0.40	0.47
Objects	Bottle	0.95	0.98	0.98	0.73	0.52	0.53
	Cable	0.83	0.87	0.92	0.52	0.43	0.46
	Capsule	0.83	0.77	0.78	0.36	0.24	0.28
	Hazelnut	0.88	0.95	0.96	0.51	0.51	0.52
	Metal Nut	0.75	0.82	0.87	0.46	0.52	0.61
	Pill	0.90	0.86	0.88	0.26	0.32	0.34
	Screw	0.67	0.87	0.87	0.10	0.15	0.18
	Toothbrush	0.97	0.91	0.92	0.21	0.35	0.32
	Transistor	0.91	0.86	0.88	0.45	0.16	0.24
	Zipper	0.96	0.90	0.98	0.34	0.15	0.24
	Mean ^c	0.87	0.88	0.90	0.40	0.34	0.37
Global mean ^d		0.88	0.90	0.93	0.45	0.36	0.41

^a For each row, the best performing approach for anomaly detection and localization is highlighted in boldface.

^b Mean I-AUROC and P-AUPRC obtained over the classes of images belonging to the texture categories.

^c Mean I-AUROC and P-AUPRC obtained over the classes of images belonging to the object categories.

^d Mean I-AUROC and P-AUPRC obtained over the entire dataset.

5.5 Exploring operational boundaries and working hypotheses

In the preceding section, we examined the prevailing trends and typical scenarios observed when projecting images into a multi-level quantized space for the purpose of generating binary segmentation masks. In most images, the q-AnoMap enhances the signal associated with defective structures. The projection into the quantized space alters the relative intensities of structures with comparable color intensities in the original image space, thereby facilitating the detection of structures that exhibit poor contrast with their surroundings. In this section, we provide a more detailed illustration of some key properties of the proposed method.

5.5.1 Property 1 - Empty segmentation masks

Some binary masks produced by $(\text{Seg})_q$ may contain only false negative instances. In cases where $(\text{Seg})_q$ fails to detect defective structures, no additional information is incorporated into the reconstruction residual.

The binary masks generated by the $(\text{Seg})_q$ network demonstrate a low incidence of false positives, which is advantageous in that it prevents the unwarranted increase of anomaly map values within non-defective regions. However, the binary segmentation output of the $(\text{Seg})_q$ network is prone to false negatives. As illustrated in Figure 5.9a, the darker bar plot represents the proportion of predicted empty masks associated with defective images, relative to the total number of defective images in the test set. This phenomenon is particularly evident in the screw and capsule categories, where defects tend to be relatively small in area, thereby increasing the difficulty for the network to accurately detect and segment them.

Figure 5.9b illustrates two typical scenarios of false negative detection. The upper portion of the figure depicts a defective screw exhibiting an irregularity in the thread. This irregularity affects only a minor portion of the screw, as indicated by the ground truth binary mask. In this instance, the alteration in the object's general shape is barely discernible in the $z_{256 \times 256}$ index maps and becomes undetectable in the lower resolution index maps due to spatial downsampling. Consequently, only a limited number of indices in the $z_{256 \times 256}$ index map capture the defective structure, which presents a challenge for detection due to the restricted number of values

that deviate from the normal distribution. The $(\text{Seg})_q$ segmentation network did not account for this minor distribution shift, as evidenced by the empty segmentation mask that was produced.

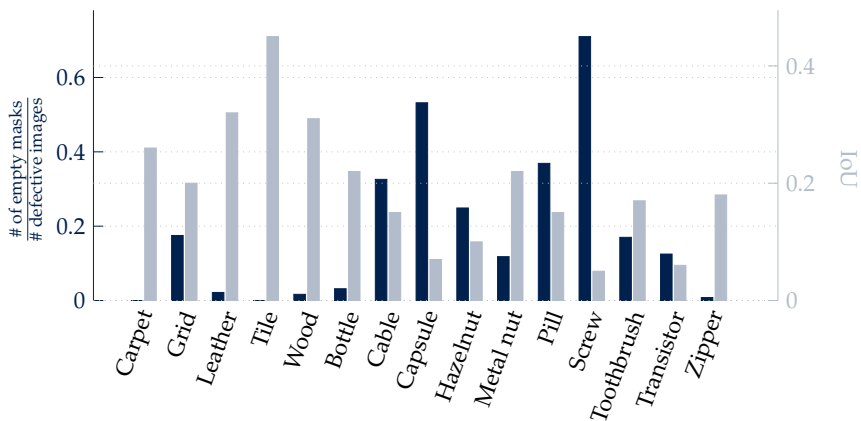
The second kind of false negative detection is observed in the hazelnut and transistor categories. The lower portion of the Figure 5.9b illustrates a typical scenario over a hazelnut sample. To better understand this phenomenon, we utilize the **Intersection over Union (IoU)** metric, defined as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}. \quad (5.9)$$

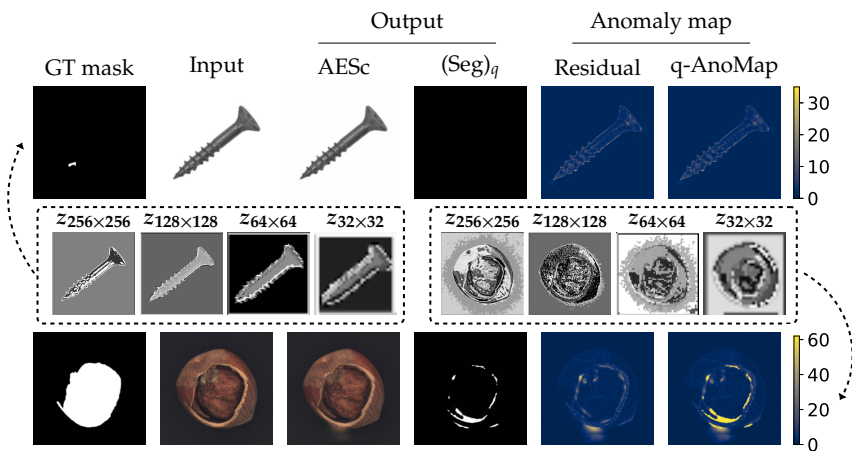
This metric quantifies the overlap between the predicted segmentation mask and the ground truth, where the number of pixels common between the ground truth and inference masks (True Positive (TP)) is divided by the total number of pixels present across both masks (TP + False Negative (FN) + False Positive (FP)).

For the hazelnut and transistor categories, the IoU scores associated with defect segmentation are relatively low, which indicated a poor quality of the segmentation². As depicted in Figure 5.9b, the binary masks often fail to fully capture either portions or the entirety of large defective regions. Unlike the screw sample shown in the figure, the defective area in the hazelnut is encoded differently from normal structures within the multi-level quantized space. This difference in encoding should theoretically facilitate their segmentation by the $(\text{Seg})_q$ network as defective structures. However, this is not consistently observed. A plausible explanation for this misclassification by $(\text{Seg})_q$ is the nature of the synthetic defects used during training, which tend to be more uniform in color in both the image space and the index maps. As a result, $(\text{Seg})_q$ struggles to generalize to these more complex and non-uniform structural defects due to the variability in defect appearance within the multi-level quantized space. This discrepancy highlights the challenges associated with accurately identifying anomalies in real-world scenarios.

²One limitation of IoU is observed when characterizing non-overlapping masks, wherein the number of true positives is equal to zero. It is noteworthy that the relative number of empty masks in the hazelnut and transistor categories does not appear to be anomalously high. Consequently, we correlate a low IoU with a high number of false negatives.



(a) (Left axis) Number of empty predictions of $(\text{Seg})_q$ that correspond to a defective sample. (Right axis) IoU over the test set.



(b) Traditional AESc detection method based on the residual, supplemented by segmentation in the multi-level quantized latent space. The figure considers two defective samples: a screw (top row) and a hazelnut (third row). The second row represents the quantized space used for segmentation.

Fig. 5.9 Analysis of misclassified defective regions by $(\text{Seg})_q$.

5.5.2 Property 2 - Impact of the codebook size on the segmentation

Depending on the image category, the optimal codebook size which achieve the best segmentation performance can vary. The impact of the codebook size is mitigated at the level of the entire dataset for anomaly localization and detection.

Providing an analysis of the impact of the codebook size on the anomaly localization and detection performance is not an easy task. It is imperative to maintain rigorous and meticulous methodology in order to prevent the misinterpretation of findings and the drawing of hasty conclusions. This analysis will be conducted in two steps. Initially, the impact of the codebook size on the segmentation mask produced by the $(\text{Seg})_q$ network will be examined. Subsequently, the impact of this hyperparameter on the overall performance of the method will be investigated.

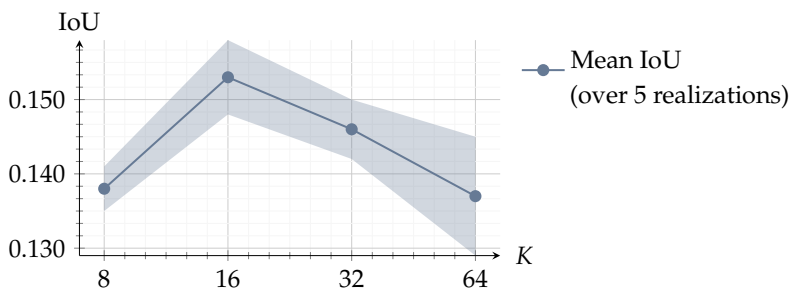
Impact on the segmentation mask produced by $(\text{Seg})_q$

The number of prototypes in the codebook represents a hyperparameter that can be adjusted in order to optimize the performance of the model. This parameter determines the level of detail that is retained when the input image is projected into the multi-level quantized space. In Section 5.3.2, we investigated the impact of this hyperparameter on the latent representations and the reconstructed image generated by the q-AESc network, despite the fact that the reconstructed image itself is not utilized directly within the pipeline. In the present study, we extend these observations to assess the influence of varying the number of prototypes on the segmentation pathway within the proposed pipeline. This approach enables us to quantify the impact of the granularity of the quantized representations on the segmentation performance, thereby providing insights into the optimal configuration for effective anomaly detection and localization.

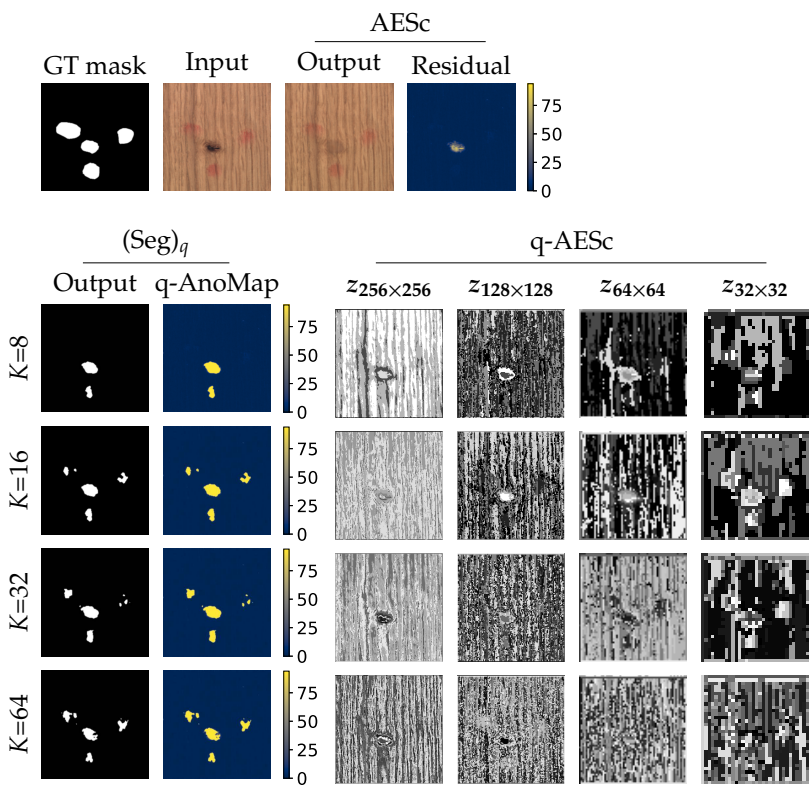
Figures 5.10 and 5.11 illustrate the impact of codebook size on the IoU score of the $(\text{Seg})_q$ network across the wood and capsule categories. In the first section of these figures (Subfigures 5.10a and 5.11a), the mean IoU values corresponding to each codebook size K (8, 16, 32, and 64) are depicted. These mean IoU values are derived from five independent experimental runs, with the associated variance illustrated by the shaded blue region. The subsequent part of the figures (Subfigures 5.10b and 5.11b) provides a comparative evaluation of the performance on defective wood and capsule samples using AESc and AESc+ $(\text{Seg})_q$ across varying codebook sizes.

The first row in these parts presents the ground truth mask, input image, reconstructed output, and residual obtained from the AESc method. The following rows (second through fifth) display the segmentation masks generated by $(\text{Seg})_q$, along with the corresponding anomaly maps, which are derived by incorporating information from the residuals at each codebook size (8, 16, 32, and 64). Furthermore, for each codebook size, the respective multi-level index maps are also provided.

As depicted in Figure 5.10a, an increase in codebook size from 8 to 16 results in an improvement in the mean IoU for the wood category. However, further increases in codebook size beyond 16 tend to decrease the segmentation accuracy. This quantitative trend is intriguing, yet it is challenging to fully interpret the underlying causes based solely on these numerical observations. To gain a more comprehensive understanding, we go deeper into a qualitative analysis by examining a sample from the wood category that contains four distinct defective stains: one prominently contrasted dark stain and three less contrasted red stains. In alignment with our preliminary hypothesis, the dark stain is consistently identified by both the reconstruction residual and all segmentation masks, regardless of the codebook size. This indicates that the dark stain, due to its high contrast with the surrounding area, is easily identifiable by the model and that the segmentation process is relatively straightforward. In comparison, the segmentation of the red stains defects presents a more complex challenge. These stains are almost entirely retained in the reconstructed image produced by the AESc model, indicating that the reconstruction process does not effectively differentiate these anomalies from the normal texture. In this scenario, the $(\text{Seg})_q$ network plays a crucial role in enhancing the detection capability. With a smaller codebook size ($K=8$), the index maps exhibit minimal variation between the normal structures and the red stains, which leads to their misdetection. This indicates that with a restricted codebook size, the model encounters difficulties in adequately capturing and differentiating subtle features, which results in reduced effectiveness in anomaly detection. As the codebook size increases, the index maps begin to reflect more distinct encodings between the red stains and the surrounding normal structures, which enable their segmentation as defective areas. From a visual perspective, the encoding of the red stains with distinct prototypes in comparison to the neighborhood is particularly evident at $K = 16$ and $K = 32$.

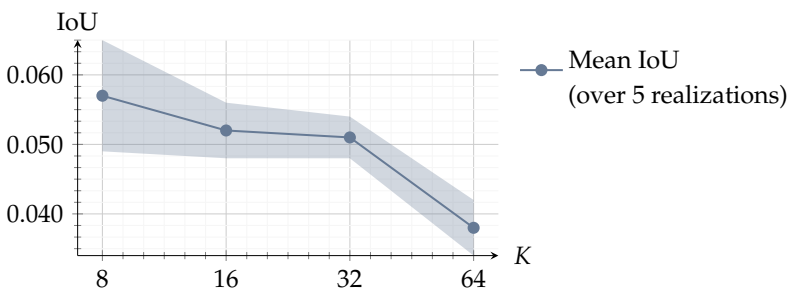


(a) Impact of the codebook size K on the IoU.

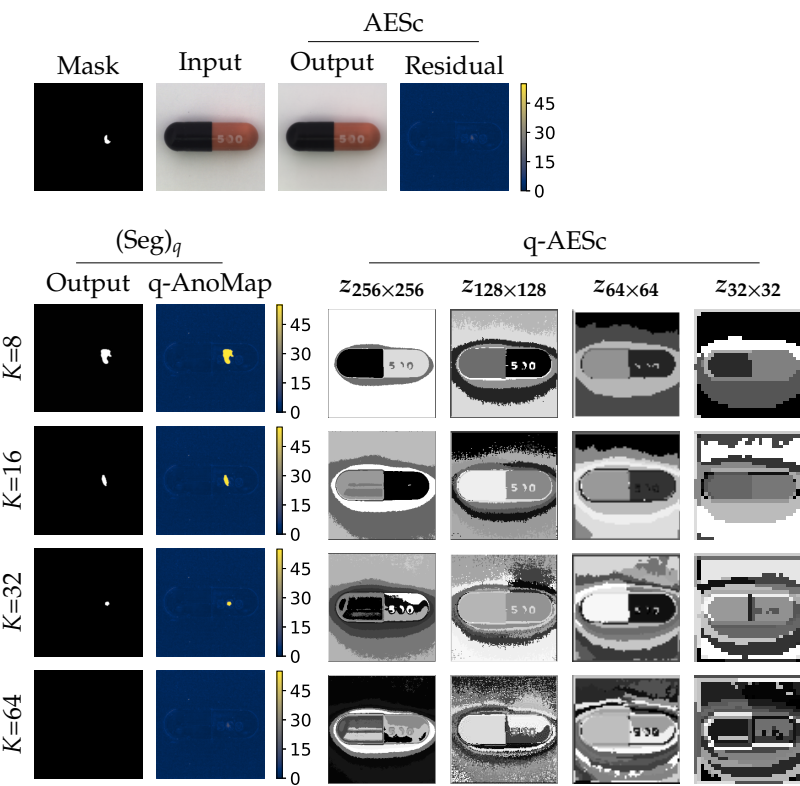


(b) Sample illustrating this impact.

Fig. 5.10 Impact of the codebook size K on the segmentation produced by $(\text{Seg})_q$ on the wood category. For further details, see Section 5.5.2.



(a) Impact of the codebook size K on the IoU.



(b) Sample illustrating this impact.

Fig. 5.11 Impact of the codebook size K on the segmentation produced by $(\text{Seg})_q$ on the capsule category. For further details, see Section 5.5.2.

In comparison, the behavior observed in the capsule category presents a completely different trend. As illustrated in Figure 5.11, smaller codebook sizes are associated with higher IoU scores. Capsule images typically feature more regular and less detailed structures compared to wood samples, which could account for the differing impact of codebook size on segmentation performance. In the selected example, a misprinted "0" digit within the "500" mark is effectively detected by segmentation masks generated from index maps corresponding to codebook sizes of 8, 16, and 32. The $z_{128 \times 128}$ maps for these codebook sizes reveal that the misprinted region is encoded using the same prototype as the underlying orange structure. This consistent encoding suggests that the $(\text{Seg})_q$ network identifies this anomaly by detecting a recurring prototype in an unexpected location. However, as the codebook size increases to $K = 64$, the encoding becomes less consistent and more difficult to interpret visually. The distribution shift becomes more challenging to detect, likely due to the introduction of noise and increased variability in the index maps.

These findings highlight the complex relationship between codebook size and the performance of the $(\text{Seg})_q$ network in image segmentation. Importantly, the impact of codebook size varies across different image categories, suggesting that the optimal codebook size may depend on the specific dataset. Our hypothesis is that smaller codebook sizes may be advantageous for images with regular patterns and minimal details, whereas larger codebook sizes may be more beneficial for images with increased complexity.

Impact on the anomaly localization and detection performance

In our initial analysis, it was observed that the codebook size influences the quality of the segmentation masks produced by the $(\text{Seg})_q$ network. These segmentation masks, when combined with the reconstruction residuals from AESc, contribute to the construction of the final anomaly map. To comprehensively understand the impact of codebook size on the overall anomaly localization and detection performance of our proposed method, we conducted a detailed investigation into this hyperparameter.

Table 5.2 presents a comparative analysis of the performance in anomaly localization and detection across various codebook sizes: 8, 16, 32, and 64. The results highlight notable performance differences among the three

evaluated methods: OoD detection, AESc, and AESc+(Seg)_q. In comparison, the impact of codebook size on the overall performance of AESc+(Seg)_q is relatively modest. Specifically, for anomaly detection, the maximum observed variation in the I-AUROC across different codebook sizes is 0.01. For anomaly localization, the impact is slightly more pronounced, with a maximum difference of 0.04 in localization accuracy. The configuration with $K = 32$ appears to provide the best balance between anomaly detection and localization performance. Nevertheless, this conclusion should be approached with caution, as the magnitude of the observed differences across codebook sizes is relatively small.

Table 5.2 Performance obtained with AESc, OoD and AESc+(Seg)_q involving multiple codebook sizes on anomaly detection and localization tasks over textures, objects and overall dataset^a.

		OOD	AESc	AESc+(Seg) q			
				8	16	32	64
I-AUROC	Textures	0.90	0.94	0.96	<u>0.97</u>	0.98	<u>0.97</u>
	Objects	0.87	0.88	0.91	<u>0.90</u>	<u>0.90</u>	<u>0.90</u>
	All	0.88	0.90	0.93	<u>0.92</u>	0.93	0.93
P-AUPRC	Textures	0.58	0.40	0.44	0.43	<u>0.47</u>	0.45
	Objects	0.40	0.34	0.31	0.35	<u>0.36</u>	0.35
	All	0.45	0.36	0.35	0.37	<u>0.39</u>	0.38

^a For each row, the best performing approach is highlighted in boldface and the second best is underlined.

In the proposed pipeline, the segmentation mask produced by the (Seg)_q network is integrated with the reconstruction residual from AESc to generate the final anomaly map, which is central to both anomaly detection and localization. This indirect utilization of the segmentation mask complicates the assessment of the direct impact of codebook size on the overall performance of the method. However, experimental results suggest that the method exhibits robust performance regardless of the specific choice of codebook size, indicating that the method is not overly sensitive to the number of prototypes selected.

5.6 Summary and fundamental properties

In this chapter, we proposed an enhancement to the classical reconstruction-based anomaly detection method by incorporating a segmentation approach utilizing images projected into a multi-level quantized space. This projection is achieved through quantization modules integrated into the skip connections of our autoencoder. Each input tensor is encoded by juxtaposing prototype vectors, which are exclusively learned from normal data. While such modules are typically used to impose explicit regularization on reconstructed images towards the normal class, our approach leverages them in a novel manner. Specifically, we fed the indexes of the closest selected prototypes into a segmentation network, which is trained to produce a binary mask that localizes unexpected prototype selections. By capturing the underlying data distribution within these quantized spaces, the resulting binary mask effectively enriches the anomaly localization process, complementing the reconstruction residual-based method.

Quantitative evaluations demonstrate that this approach significantly enhances anomaly detection performance, improving the I-AUROC from 0.90 to 0.93. This improvement is particularly notable for texture categories, where the mean I-AUROC increased from 0.94 to 0.98. In terms of anomaly localization, the global P-AUPRC over the entire MVTec AD dataset rose from 0.36 to 0.41, with the proposed method outperforming the baseline in 13 out of the 15 image categories. However, it should be noted that the Out-of-Distribution (OoD) detection method introduced in Chapter 4 still achieves the best overall P-AUPRC metric.

A key advantage of our approach is that the output segmentation mask modifies the residual only in regions where positive pixels are detected. As a result, false negative detections do not adversely affect the detection mechanism, as they leave the original reconstruction-based detection unaltered. However, we observed that relatively small defects or defects with complex appearances covering larger areas could occasionally lead to empty segmentation masks. In these cases, the segmentation pathway does not impact the original detection mechanism, ensuring robustness.

Compared to the other methods described in this manuscript, this ap-

proach introduces a more complex system that involves training both the quantization modules and the segmentation network. Nevertheless, the introduced hyperparameters, such as the codebook size, do not critically impact performance, indicating a degree of robustness in our model's configuration. Overall, this chapter presents a sophisticated and effective enhancement to the classical reconstruction-based anomaly detection framework, offering substantial improvements in both detection accuracy and localization precision.

6

Conclusions and perspectives

PERFORMANCE metrics are not all... In the introduction to this thesis, we underscored the necessity of grasping the fundamental causes underlying our findings, rather than focusing exclusively on the comparison of methods based on their performance. Simplicity was a guiding principle in achieving this objective. To clearly demonstrate the impact of our contributions, we maintained straightforward pipelines throughout the study. We intentionally avoided any post-processing techniques, such as morphological operations on predicted segmentation masks, and did not rely on prior knowledge of defect characteristics, such as appearance, shape, or size. Furthermore, we avoided performing hyper-parameter tuning tailored to each image category. This approach ensured that variations in results were due to the inherent design choices of the methods rather than an extensive search in the hyper-parameter tuning process. Our goal was to help develop the reader's intuition and provide confidence in the robustness of our results and contributions.

The proposed methods are intended to be applicable in real industrial use cases. To bridge the gap between academic methodologies and industrial requirements, we focused on analyzing our results with the aim of understanding the reasons behind our findings. Therefore, we conducted detailed analyses of our findings, focusing on the underlying mechanisms and properties of our approaches. Characterizing the properties of our

methods has been central to this work. More specifically, we addressed three key questions throughout this manuscript:

Question 1. What are the foundational assumptions underlying the proposed methodologies?

Question 2. From a comprehensive analysis of success and failure instances, can we predict the scenarios in which our methods are most effective or ineffective?

Question 3. What strategies can we recommend for enhancing our methods, particularly in cases where they do not perform as expected?

Having addressed these research questions, we now present a summary of the findings from the three methodologies explored in each technical chapter of this thesis:

AESc (AutoEncoder with Skip connections) method. This framework is an extension of classical reconstruction-based methods operating on input data in the original image space. To enhance the quality of the reconstructed images, skip connections were integrated into the network architecture, and synthetic corruption, termed Stain, was applied to the input images.

The details of this pipeline were provided in Chapter 3.

OoD (Out-of-Distribution) method. Building on the AESc method, this approach was designed to detect subtle variations in activation patterns through binary tests conducted at the voxel level.

The implementation and nuances of this method were discussed in Chapter 4.

AESc+(Seg)_q (Quantized-AESc + segmentation). In this framework, quantization modules were incorporated into the skip connections of the AESc network, projecting input images into a multi-level quantized space. The resulting projections were then fed into a segmentation network that produced a binary mask, which was subsequently combined with the reconstruction residual to accurately localize defective structures.

The complete pipeline was outlined in Chapter 5.

6.1 Underlying assumptions

The first research question posed was: *What are the foundational assumptions underlying the proposed methodologies?* In this section, we outline the principal assumptions shared by the three approaches investigated.

Small variations

In this context, anomalies refer to variations that alter part of the image or the general appearance of the object or texture under consideration. The proposed frameworks are not intended for detecting outlier samples from completely different data distributions. Global aspect of the image is assumed to be preserved.

Concept of normality

Given our unsupervised problem formulation, we avoid introducing prior information about the appearance of defects to be detected. As a result, the model must autonomously learn to define the natural variation of normal examples based solely on the provided data. This approach has both advantages and drawbacks:

Advantage. Unlike supervised methods, this indirect detection approach renders our methods agnostic to specific defect appearances, enabling them to detect a wide range of unexpected abnormal structures.

Drawback. Any structural deviation not represented in the training set is likely to be classified as defective. Therefore, it is essential to have a sufficiently large and diverse training set of normal samples to capture the full spectrum of normal variations.

To ensure that the model accurately captures the concept of normality, the quality of the data collection process is crucial. The training set must encompass the full range of what can be considered clean by the model. The presence of defective samples in the training set can significantly impair the model's detection performance.

Although no manual labeling of the training set is required, dataset sizes remain limited. Ensuring a comprehensive and representative dataset is crucial for optimal model performance.

Equivariance to translation

The proposed methods rely on convolutional neural networks, which do not explicitly encode the spatial localization of objects. Consequently, similar results are expected across the entire image. Only the OoD method discussed in Chapter 4 is potentially sensitive to positional shifts, accounting for changes in object placement.

6.2 Key observations

The second research question posed was: *From a comprehensive analysis of success and failure instances, can we predict the scenarios in which our methods are most effective or ineffective?* Throughout the manuscript, a significant number of examples were analyzed, and various properties of the three studied methods were highlighted. Below is a summary of the key observations.

Low contrasted defects

A primary limitation of the AESc method is its difficulty in detecting defective structures that contrast poorly with their surroundings. Even if the network successfully replaces defective structures with normal content, these regions may remain undetected due to the pixel-wise difference mechanism in the image space. To address this challenge, we proposed two extensions of the baseline AESc approach.

First, the OoD method enhances the detection of low-contrast anomalies. The binary tests in this framework allow for the detection of subtle activations, where even a slight deviation above the threshold contributes equally to the anomaly score as a significant deviation.

Second, we introduced the AESc+(Seg) q framework, which operates in a multi-level quantized space. Images are transformed into multiple quantized index maps at different resolutions. Similar pixels are mapped to the same value, but the concept of contrast is randomized in this space. When processed by a segmentation network, these maps enable the identification of low-contrast defects. The combination of this detection mechanism with the residuals from the AESc method results in a substantial improvement in overall performance.

Textures versus objects

The MVTec AD dataset comprises five texture categories and ten object categories. Textures are characterized by sharp details and repetitive structures, whereas objects typically exhibit fewer details. We observed that the proposed methods perform particularly well on texture images. Detecting defects in these types of images with a reconstruction-based approach necessitates the preservation of fine normal details, a challenge effectively addressed by integrating skip connections into the autoencoder architecture. Coupled with our Stain corruption model, the AESc method demonstrated a satisfactory balance between preserving normal structures and replacing defective ones with normal content. This combination proved especially effective in maintaining the integrity of textures while accurately identifying anomalies.

Sensitivity to false positives

The sensitivity of the detection methods involves a trade-off between detecting subtle variations in the images and accounting for the inherent variability among normal samples. To enhance the detection of low-contrast defects, we proposed the OoD approach, which is designed to capture more subtle variations. While this approach improves the detection capabilities compared to the original AESc method, it also results in a higher false positive detection rate, occasionally identifying anomalies in non-relevant areas of the image, such as the background.

6.3 Improvement and perspectives

The third research question posed was: *What strategies can we recommend for enhancing our methods, particularly in cases where they do not perform as expected?* Below, we outline several areas for potential improvement concerning the three methods proposed in this manuscript.

Introduction of prior knowledge

The objective of the proposed analysis was to investigate the impact of performing anomaly detection in spaces other than the image space to enhance a classical reconstruction-based framework. To validate the proof of concept, we emphasized ensuring a fair comparison between our approaches. However, in real industrial scenarios, incorporating prior knowledge about the

target task into the initial pipeline could further improve results. Without requiring the collection of defective images for training, several modifications can be considered:

Post-processing of network output. Morphological operators can be employed to eliminate the detection of structures whose area falls below a predefined threshold, thereby reducing false positives.

Define an area of interest. In object categories, the background often occupies a significant portion of the image. The reconstruction loss penalizes object structures and the background equally, as well as all parts of the image contribute equally to the anomaly score. To address this issue, implementing a background removal technique can help focus the detection on the relevant areas.

Adapt the synthetic noise. In Chapter 3, we introduced the stain noise model to prevent our reconstruction network from converging to identity mapping. The proposed corruption model was designed to be generic, helping the network capture complex patterns in the data. However, if prior knowledge about specific defects is available, customizing the data augmentation process can be advantageous. For instance, if variations in lighting conditions are anticipated during image acquisition, incorporating data augmentation techniques that simulate contrast and brightness changes can help the model recognize these variations as part of the normal distribution.

Taking global information into account

In the reconstruction-based framework, variations between the input image and its reconstruction are detected by computing pixel-wise differences. This method inherently relies on a pixel-by-pixel comparison and does not explicitly account for interactions between pixels at different spatial locations. As a result, detecting shifts, bending or the absence of entire structures is challenging with a reconstruction-based framework. Incorporating spatial relationships and interactions between pixels across different locations could enhance the detection mechanism.

6.4 Final words

Unsupervised anomaly detection is a challenging problem that can be effectively addressed with a reconstruction-based framework. By reconstructing a normal version of an input image, the model offers a visual explanation of its decision, thereby mitigating the black-box nature of deep learning and increasing user trust. However, a significant limitation of reconstruction-based methods is their inability to detect low-contrast defects effectively, due to the pixel-wise difference in the image space. To address this issue, we proposed two extensions that utilize alternative spaces and incorporate the network's internal representations. By analyzing these spaces, we could highlight their unique properties and understand how they improved the effectiveness of the original detection mechanism.

Appendices

A MVTec AD dataset

The MVTec AD dataset comprises 5354 high-resolution color images across various object and texture categories listed in table A.1. In our analysis, we focus exclusively on images that have been resized to a spatial resolution of 256×256 . These images are intended to be captured under controlled lighting conditions to ensure consistency. Each category includes normal images for training, as well as both normal and defective images for testing. The dataset features over 70 distinct types of defects, including scratches, stains, oil drops, and structural alterations. These anomalies vary in severity, ranging from minor, subtle defects such as scratches or spots to more significant deformations or contaminations, effectively simulating the kinds of defects that are likely to occur in real-world industrial settings. Furthermore, the MVTec AD dataset provides pixel-level annotations for these anomalies in the defective images, allowing for detailed evaluation not only of the anomaly detection performance but also of the accuracy in localizing these defects within the images.

Table A.1 Repartition of normal/defective images (in absolute number) and normal/defective pixels (in percentage) of the MVTEC AD dataset over each category.

		Train (#Img)	Test (#Img)		Test (%Pxl)	
		Clean	Clean	Defect	Clean	Defect
Textures	Carpet	280	28	89	98.4	1.6
	Grid	264	21	57	99.3	0.7
	Leather	245	32	92	99.3	0.7
	Tile	230	33	84	93.0	7.0
	Wood	247	19	60	64.0	6.0
Objets	Bottle	209	20	63	94.2	5.8
	Cable	224	58	92	97.1	2.9
	Capsule	219	23	109	99.1	0.9
	Hazelnut	391	40	70	97.9	2.1
	Metal Nut	220	22	93	88.3	11.7
	Pill	267	26	141	96.6	3.4
	Screw	320	41	119	99.7	0.3
	Toothbrush	60	12	30	98.5	1.5
	Transistor	213	60	40	97.0	3.0
	Zipper	240	32	119	98.0	2.0

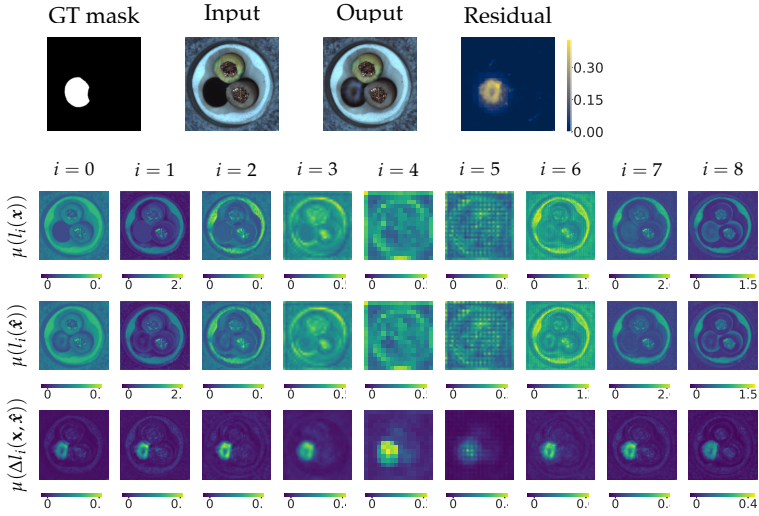
B Impact of skip connections on internal representation

In this analysis, we investigate the impact of skip connections on the internal representations generated by an autoencoder. To explore this effect, we provide a visual comparison of the internal representations of both a defective object (a cable) and a defective texture (a carpet), as shown in Figures B.1 and B.2, respectively. These comparisons are made between the traditional autoencoder (AE) and the autoencoder with skip connections (AESc), both trained on input images corrupted with the stain noise model.

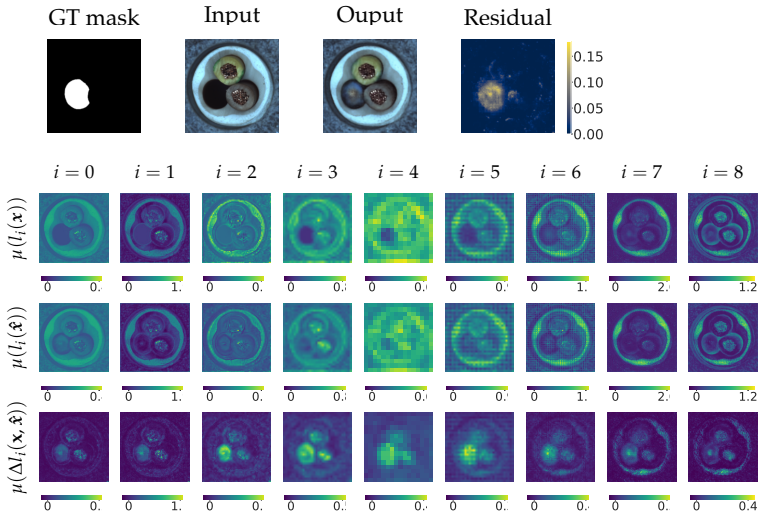
First, we examine the reconstruction of the defective cable using the AESc network (Figure B.1a) and the AE network (Figure B.1b). Both networks successfully replace the missing cable with cleaner content, indicating that they capture high-level semantic information effectively. Notably, the AE network's reconstruction appears to be even more structurally realistic. When analyzing the differences in activation tensors between the AE and AESc networks, two key observations emerge. First, the AESc network's superior ability to reconstruct normal structures results in lower intensity activations in regions corresponding to normal parts of the image. Second, intensities associated with the defective structure are consistently visible across all layers of the AESc network, whereas, in the AE network, the defective region only stands out in the deeper layers. This can be intuitively explained by considering the nature of autoencoders with and without skip connections. An autoencoder without skip connections must project both normal and defective embeddings toward a similar tensor to produce a clean reconstructed image. In contrast, with skip connections, some information about the defective structure bypasses the bottleneck, and the decoder's role is to effectively compensate for this by generating an output that corrects the defective structures to match the clean image.

Next, we consider the reconstruction of defective textures using the AESc network (Figure B.2a) and the AE network (Figure B.2b). Similar patterns emerge as with the cable image. The activation differences in normal structures are more pronounced with the AE network than with AESc, and the defective activations in the AE network tend to merge with those of normal structures as they pass through the decoder.

The purpose of this appendix is to substantiate the use of the AESc network, as discussed in Chapter 4. Achieving better detection performance hinges on obtaining salient representations that effectively distinguish between normal and defective structures. In this regard, the ability of the AESc network to preserve the integrity of normal structures during reconstruction is a critical property that enhances its capacity to differentiate between normal and defective regions in the internal representation space. This characteristic underscores the importance of the AESc network in the overall framework for improved anomaly detection.

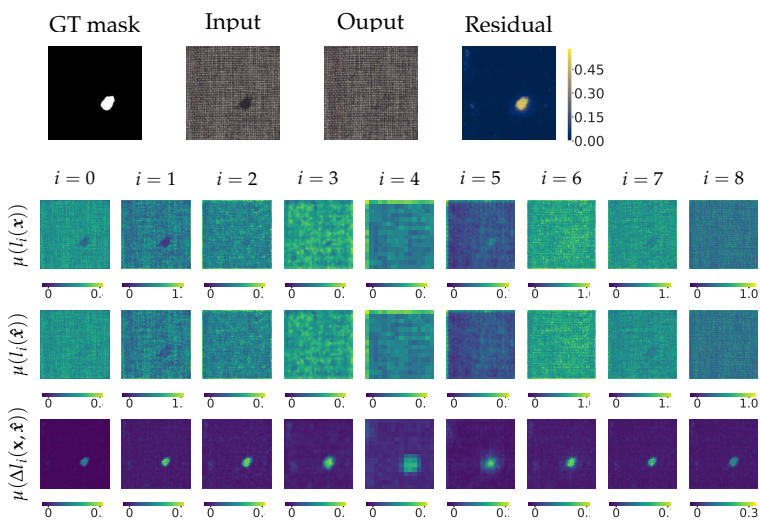


(a) Internal representations with the AESc network.

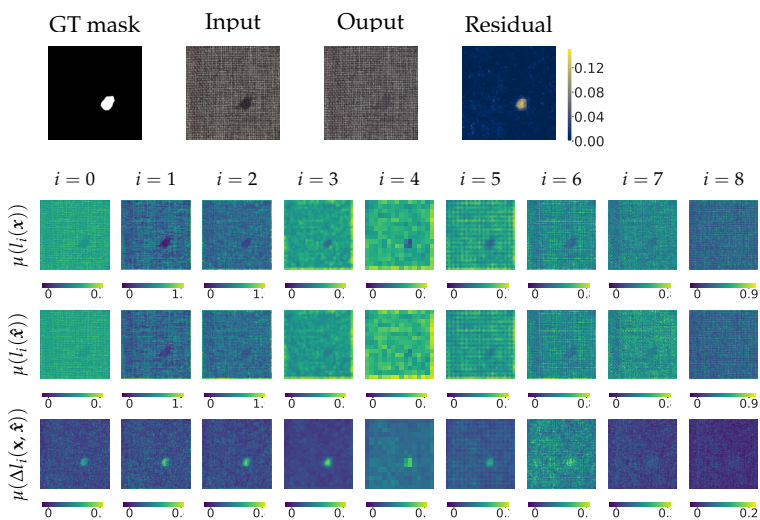


(b) Internal representations with the AE network.

Fig. B.1 Comparison of internal representations obtained on an autoencoder with/without skip connections over a defective cable. For row meaning, refer to the caption of Figure 4.2.



(a) Internal representations with the AESc network.



(b) Internal representations with the AE network.

Fig. B.2 Comparison of internal representations obtained on an autoencoder with/without skip connections over a defective carpet. For row meaning, refer to the caption of Figure 4.2.

C Variational inference

Objective of the training of a VAE is to describe the distribution of the latent variables z given the observation of the variables x . According to the Bayes theorem, we express $p(z|x)$ as:

$$p(z|x) = \frac{p(x|z) p(z)}{p(x)}, \quad (\text{C.1})$$

with $p(x)$ being the evidence, $p(x|z)$ being the likelihood, $p(z)$ being the prior and $p(z|x)$ being the posterior. The prior is a normalization factor required to be computed as:

$$p(x) = \int p(x|z)p(z)dz. \quad (\text{C.2})$$

In practice, the evidence becomes intractable in high dimension, as it is very expensive to check all the possible values of z and sum them up. This term is then approximated, which can be done with variational inference. It consists in finding the best approximation of a distribution among a parametrized family. The parametrized functions are neural networks determined by the network architecture represented in Figure C.1.

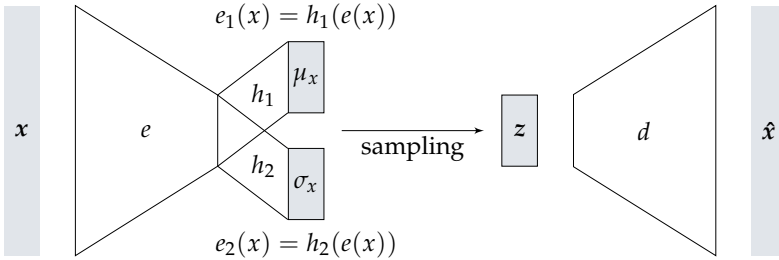


Fig. C.1 VAE architecture composed of an encoder e and a decoder d . The encoder maps an input x into a random variable parametrized with mean μ_x and variance σ_x . A vector x is sampled from this latent distribution and fed to the decoder d to produce an output $\hat{x}$.

Here, we approximate $p(z|x)$ by a Gaussian distribution $q_x(z)$ whose mean and covariance are defined by two functions, e_1 and e_2 :

$$q_x(z) \equiv \mathcal{N}(e_1(x), e_2(x)), \quad e_1 \in E_1 \text{ and } e_2 \in E_2 \quad (\text{C.3})$$

Among the families E_1 , E_2 and D , we search for the best candidate e_1^* , e_2^* and d^* which minimize the Kullback-Leibler (KL) divergence between the approximation and the target $p(z|x)$. The optimization problem, adapted from [107], is written as follow:

$$(e_1^*, e_2^*, d) = \arg \min_{(e_1, e_2, d) \in E_1 \times E_2 \times D} KL(q_x(z), p(z|x)) \quad (\text{C.4})$$

Bayes;

$$= \arg \min_{(e_1, e_2, d) \in E_1 \times E_2 \times D} KL\left(q_x(z), \frac{p(x|z)p(z)}{p(x)}\right) \quad (\text{C.5})$$

$$KL(P|Q) = \mathbb{E}_{z \sim P(z)} \log \frac{P(z)}{Q(z)};$$

$$= \arg \min_{(e_1, e_2, d) \in E_1 \times E_2 \times D} \left(\mathbb{E}_{z \sim q_x} (\log q_x(z)) - \mathbb{E}_{z \sim q_x} \left(\log \frac{p(x|z)p(z)}{p(x)} \right) \right) \quad (\text{C.6})$$

Basic rule of logarithm;

$$= \arg \min_{(e_1, e_2, d) \in E_1 \times E_2 \times D} \left(\mathbb{E}_{z \sim q_x} (\log q_x(z)) - \mathbb{E}_{z \sim q_x} (\log p(z)) \right. \\ \left. - \mathbb{E}_{z \sim q_x} (\log p(x|z)) + \mathbb{E}_{z \sim q_x} (\log p(x)) \right) \quad (\text{C.7})$$

Basic rule of logarithm + suppress last term;

$$= \arg \min_{(e_1, e_2, d) \in E_1 \times E_2 \times D} \left(\mathbb{E}_{z \sim q_x} \left(\log \frac{q_x(z)}{p(z)} \right) - \mathbb{E}_{z \sim q_x} (\log p(x|z)) \right) \quad (\text{C.8})$$

$$\text{Change to maximization} + KL(P|Q) = \mathbb{E}_{z \sim P(z)} \log \frac{P(z)}{Q(z)};$$

$$= \arg \max_{(e_1, e_2, d) \in E_1 \times E_2 \times D} \left(\underbrace{\mathbb{E}_{z \sim q_x} (\log p(x|z))}_{\text{Reconstruction loss}} - \underbrace{KL(q_x(z), p(z))}_{\text{KL divergence}} \right) \quad (\text{C.9})$$

The optimization reflects a trade off between the confidence we have in the data and the confidence we have in the prior, balanced with two terms:

1. The reconstruction term ensuring that the latent variable z explains the data;
2. The negative KL divergence between the approximation and the prior that tends to adjust the parameters in order to make the approximation close to the prior distribution.

D Normal and defective prototypes distributions

In this analysis, we examine samples where the selected prototypes encoding defective structures are among the least frequently chosen across the test set. Figure D.1 provides an illustration of the quantized latent space for a capsule, with a specific focus on the frequency distribution of prototypes associated with normal structures (blue line in the graphs) versus those associated with defective structures (orange line in the graphs). The primary observation from these graphs is the discrepancy between the two lines, indicating that certain prototypes may frequently encode defective structures but are rarely associated with normal ones, and vice versa. The last row of the figure illustrates the spatial distribution of the frequencies corresponding to the selected prototypes. In the first row of the dotted block, it is evident that the same prototypes are consistently used to encode defective structures, which manifest as black holes resulting from missing inner cables. The third row demonstrates that these prototypes have a high likelihood of being selected in defective structures. Graph above suggest these prototypes are also linked to a low likelihood of being in a normal structures. This observation can be attributed to the relatively low occurrence of black pixels in the normal cables.

Figure D.2 depicts a similar phenomenon for samples of a metal nut and a tile. Compared to the previous cable sample, the defective structures in these samples are nearly undetectable based on the reconstruction residual. However, the prototypes used to encode the defective structures are linked to a high probability of belonging to defective structures. Notably, for $z_{32 \times 32}$, these prototypes are associated with an extremely low likelihood of appearing in normal structures.

This analysis suggests that a segmentation network trained to output a binary segmentation mask for anomalies, utilizing this quantized space as input, could leverage the relative frequency of prototype occurrences to make predictions. Naturally, this approach does not account for larger structural contexts and focuses on each pixel individually. Consequently, such a network would also depend on detecting unseen patterns in the quantized maps to make accurate predictions.

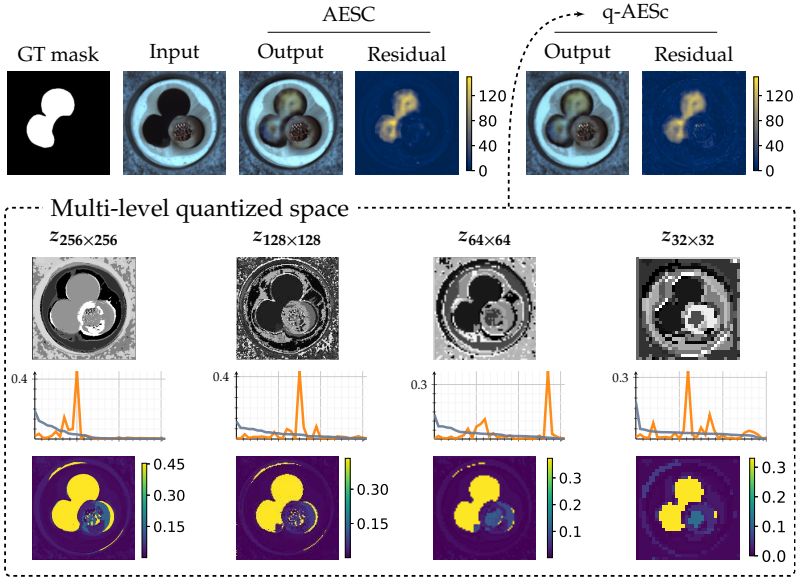


Fig. D.1 Comparison between AESc and q-AESc for anomaly localization using residual over a defective cable. In the dotted frame, the first row displays the index maps, denoted as z_i , with index values arbitrarily ranging from black (0) to white (31). The second row shows the frequency distribution of prototype usage across the entire test set, with prototypes ordered from most to least frequently used. The blue line represents prototypes located in normal regions, while the orange line indicates prototypes in defective regions, with the order defined by the normal distribution. The third row illustrates the frequency of the selected prototypes according to the defective distribution (orange curve above).

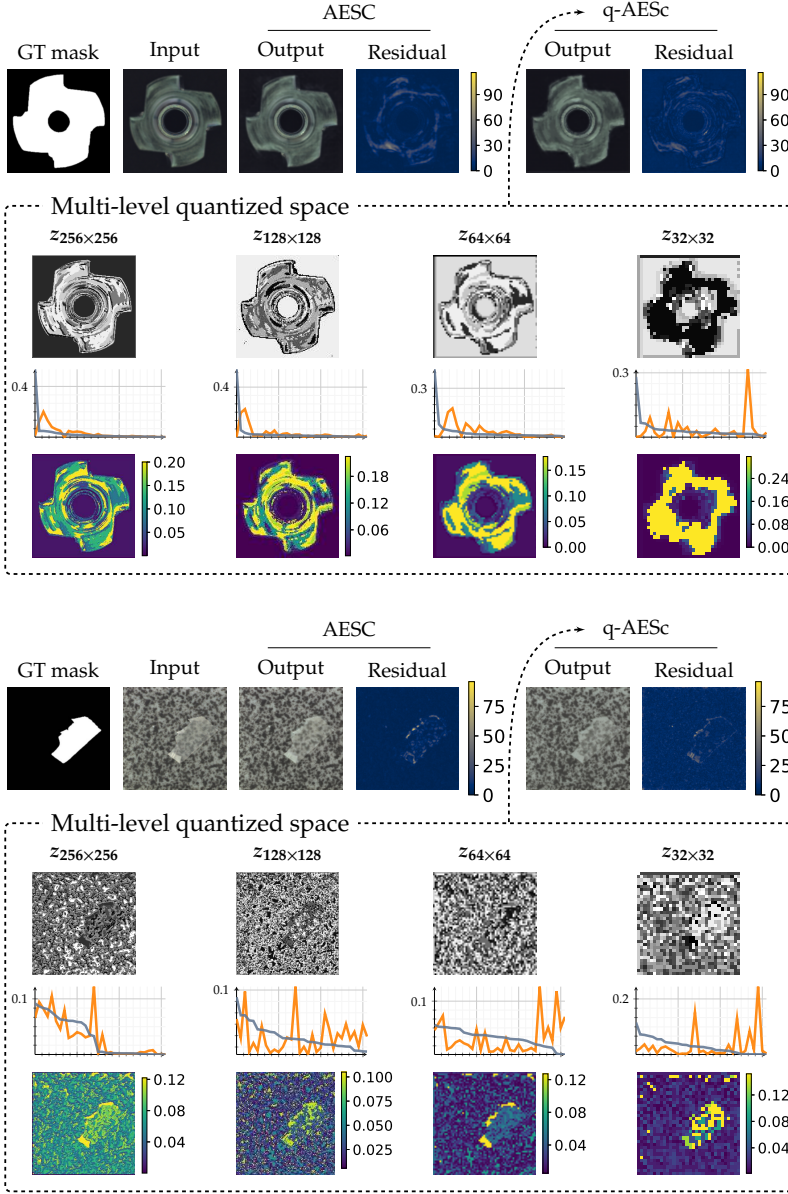


Fig. D.2 Comparison between AESc and q-AESc for anomaly localization using residual over a metal nut (upper part) and a tile (lower part). The explanation for the content within the dotted frame follows the same description as provided in Figure D.1.

Author's publication list

Related contributions

Parts of the work presented in this manuscript have been published during the doctoral training, either in conferences (CP) or in journals (JP). The content of Chapter 3 is related to the conference paper CP1. In addition to this initial conference, this article was also presented at a local french conference (CP1-a) and at a workshop aimed at highlighting the work of women in the field of computer vision (CP1-b). The content of Chapter 4 is related to the journal article JP1 as well as the conference article CP2. The content of Chapter 5 has not yet been published but will be submitted to a conference in the near future.

[CP1] [A.-S. Collin](#) and C. De Vleeschouwer, "Improved anomaly detection by training an autoencoder with skip connections on images corrupted with stain-shaped noise", in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2021.

[CP1-a] [A.-S. Collin](#) and C. De Vleeschouwer, "Synthetic corruption of images for anomaly detection using convolutional autoencoders", in *Proceedings of the Congrès Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP)*, 2020.

[CP1-b] [A.-S. Collin](#) and C. De Vleeschouwer, "Synthetic corruption of images for anomaly detection using convolutional autoencoders", in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Women in Computer Vision (WiCV) Workshop*, 2021.

- [JP1] A.-S. Collin and C. De Vleeschouwer, "Joint Use of Skip Connections and Synthetic Corruption for Anomaly Detection with Autoencoders", in *Control Charts and Machine Learning for Anomaly Detection in Manufacturing*, pp. 139-144, 2022.
- [CP2] A.-S. Collin, C. de Bodt, D. Mulders and C. De Vleeschouwer, "Don't skip the skips: autoencoder skip connections improve latent representation discrepancy for anomaly detection", in *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2023.

Unrelated contributions

Collaborations extending beyond the work presented in this manuscript are ongoing. The first collaboration, within Prof. De Vleeschouwer's research team, involves benchmarking vision-language pretraining on bone X-ray and french medical reports for multiple downstream task of interest. The article (JP2) has been published on ArXiv and is currently under review in a journal. A second collaboration with Prof. D. Bol's team aims at applying machine learning model to predict the surface of semiconducting material based on the measurement of non time-intensive quantities. The subject is crucial for the for the Life-Cycle-Assessment (LCA) community, yet there is currently a need to introduce a deeper understanding of machine learning to propose an effective predictive model whose performance is evaluated with an appropriate methodology. The preliminary findings of this work were presented during a talk at a LCA conference (CP3).

- [JP2] A. Englebert, A.-S. Collin, O. Cornu and C. De Vleeschouwer, "Self-supervised vision-langage alignment of deep learning representations for bone X-rays analysis", in *arXiv preprint arXiv:2405.08932*, 2024.
- [CP3] A. Wattiez, R. Dethienne, A.-S. Collin and D. Bol, "Parametric modeling for improved die area estimation in the life cycle assessment of electronic systems", Talk in *Electronics Goes Green 2024+: Berlin*, 2024.

Bibliography

- [1] E. Oztemel and S. Gursev, "Literature review of Industry 4.0 and related technologies," *Journal of Intelligent Manufacturing*, vol. 31, pp. 127–182, 2020.
- [2] H. Kagermann, W.-D. Lukas, and W. Wahlster, "4.0-Internet-Dinge," *Industriellen Revolution. VDI nachrichten*, vol. 13, pp. 1090–1100, 2011.
- [3] X. Tao, D. Zhang, W. Ma, X. Liu, and D. Xu, "Automatic Metallic Surface Defect Detection and Recognition with Convolutional Neural Networks," *Applied Sciences*, vol. 8, p. 1575, 2018.
- [4] E. A. Kholief, S. H. Darwish, and M. N. Fors, "Detection of steel surface defect based on machine learning using deep auto-encoder network," *Industrial engineering and operations management*, pp. 218–229, 2017.
- [5] X. Fang, Q. Luo, B. Zhou, C. Li, and L. Tian, "Research progress of automated visual surface defect detection for industrial metal planar materials," *Sensors*, vol. 20, pp. 1–35, 2020.
- [6] B. Tang, L. Chen, W. Sun, and Z. kang Lin, "Review of surface defect detection of steel products based on machine vision," *IET Image Processing*, vol. 17, pp. 303–322, 2023.
- [7] X. Xie, "A Review of Recent Advances in Surface Defect Detection using Texture analysis Techniques," *Electronic Letters on Computer Vision and Image Analysis (ELCVIA)*, vol. 7, pp. 1–22, 2008.
- [8] S. Mei, Y. Wang, and G. Wen, "Automatic fabric defect detection with a multi-scale convolutional denoising autoencoder network model," *Sensors*, vol. 18, p. 1064, 2018.

- [9] C. Kampouris, S. Zafeiriou, A. Ghosh, and S. Malassiotis, "Fine-grained material classification using micro-geometry and reflectance," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 778–792, 2016.
- [10] J. Gan, Q. Li, J. Wang, and H. Yu, "A Hierarchical Extractor-Based Visual Rail Surface Inspection System," *IEEE Sensors Journal*, vol. 17, pp. 7935–7944, 2017.
- [11] S. A. D. Araújo, J. H. Pessota, and H. Y. Kim, "Beans quality inspection using correlation-based granulometry," *Engineering Applications of Artificial Intelligence*, vol. 40, pp. 84–94, 2015.
- [12] L. D. Griffin, M. Caldwell, J. T. A. Andrews, and H. Bohler, ""unexpected item in the bagging area": Anomaly detection in x-ray security images," *IEEE Transactions on Information Forensics and Security*, vol. 14, pp. 1539–1553, 2018.
- [13] V. Terziyan and O. Vitko, "Explainable AI for Industry 4.0: Semantic Representation of Deep Learning Models," *Procedia Computer Science*, vol. 200, pp. 216–226, 2022.
- [14] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014.
- [15] X. Tao, X. Gong, X. Zhang, S. Yan, and C. Adak, "Deep Learning for Unsupervised Anomaly Localization in Industrial Images: A Survey," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, 2022.
- [16] Y. Cui, Z. Liu, and S. Lian, "A Survey on Unsupervised Anomaly Detection Algorithms for Industrial Images," *IEEE Access*, vol. 11, pp. 55 297–55 315, 2023.
- [17] J. Liu, G. Xie, J. Wang, S. Li, C. Wang, F. Zheng, and Y. Jin, "Deep Industrial Image Anomaly Detection: A Survey," *Machine Intelligence Research*, vol. 21, pp. 104–135, 2024.
- [18] R. Chalapathy and S. Chawla, "Deep Learning for Anomaly Detection: A Survey," *arXiv preprint arXiv:1901.03407*, 2019.
- [19] M. Markou and S. Singh, "Novelty detection: A review - Part 2: Neural network based approaches," *Signal Processing*, vol. 83, pp. 2499–2521, 2003.

- [20] P. Napoletano, F. Piccoli, and R. Schettini, "Anomaly detection in nanofibrous materials by CNN-based self-similarity," *Sensors*, vol. 18, p. 209, 2018.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- [22] N. Cohen and Y. Hoshen, "Sub-Image Anomaly Detection with Deep Pyramid Correspondences," *arXiv preprint arXiv:2005.02357*, 2020.
- [23] P. Liznerski, L. Ruff, R. A. Vandermeulen, B. J. Franks, M. Kloft, and K. R. Muller, "Explainable Deep One-Class Classification," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [24] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain MR images," *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) Brainlesion Workshop*, pp. 161–169, 2018.
- [25] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving Unsupervised Defect Segmentation by Applying Structural Similarity to Autoencoders," *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, pp. 372–380, 2019.
- [26] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," *Proceedings of the International conference on Information Processing in Medical Imaging (IPMI)*, pp. 146–157, 2017.
- [27] S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon, "Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection," *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2019.
- [28] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. van den Hengel, "Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1705–1714, 2019.

- [29] V. Zavrtanik, M. Kristan, and D. Skočaj, “DRÆM - A discriminatively trained reconstruction embedding for surface anomaly detection,” *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8310–8319, 2021.
- [30] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.
- [31] J. Yi and S. Yoon, “Patch SVDD: Patch-level SVDD for Anomaly Detection and Segmentation,” *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020.
- [32] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, “Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4183–4192, 2020.
- [33] S. Yoa, S. Lee, C. Kim, and H. J. Kim, “Self-supervised learning for anomaly detection with dynamic local augmentation,” *IEEE Access*, vol. 9, pp. 147 201–147 211, 2021.
- [34] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1650–1659, 2016.
- [35] T. Defard, A. Setkov, A. Loesch, and R. Audigier, “PaDiM: a Patch Distribution Modeling Framework for Anomaly Detection and Localization,” *arXiv preprint arXiv:2011.08785*, 2020.
- [36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.
- [37] D. J. Rezende and S. Mohamed, “Variational Inference with Normalizing Flows,” *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1530–1538, 2015.

- [38] M. Rudolph, B. Wandt, and B. Rosenhahn, "Same same but DifferNet: Semi-supervised defect detection with normalizing flows," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1907–1916, 2021.
- [39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Proceedings of the Advances in Neural Information Processing Systems conference (NeurIPS)*, 2012.
- [40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.
- [41] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 481–490, 2019.
- [42] L. Wang, D. Zhang, J. Guo, and Y. Han, "Image anomaly detection using normal data only by latent space resampling," *Applied Sciences*, vol. 10, pp. 8660–8679, 2020.
- [43] A. V. den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," *Proceedings of the Advances in Neural Information Processing Systems conference (NeurIPS)*, pp. 6306–6315, 2017.
- [44] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, "PixelsNail: An improved autoregressive generative model," *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 864–872, 2018.
- [45] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti, "VT-ADL: A Vision Transformer Network for Image Anomaly Detection and Localization," *Proceedings of the International Symposium on Industrial Electronics (ISIE)*, pp. 01–06, 4 2021.
- [46] Y. Lecun, E. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–1324, 1998.
- [47] V. Mazzia, F. Salvetti, and M. Chiaberge, "Efficient-CapsNet: capsule network with self-attention routing," *Scientific Reports*, vol. 11, 2021.

- [48] D. Hirata and N. Takahashi, "Ensemble learning in CNN augmented with fully connected subnetworks," *IEICE TRANSACTIONS on Information and Systems*, vol. 106, pp. 1258–1261, 2023.
- [49] A. Byerly, T. Kalganova, and I. Dear, "No Routing Needed Between Capsules," *Neurocomputing*, vol. 463, pp. 545–553, 2021.
- [50] Y. Lecun, L. J. Lawrence D Eon Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. A. Muller, E. Sackinger, P. Simard, and V. Vapnik, "Learning algorithms for classification: A comparison on handwritten digit recognition," *Neural networks: the statistical mechanics perspective*, vol. 261, 1995.
- [51] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD - A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9592–9600, 2019.
- [52] K. Batzner, L. Heckler, and R. König, "EfficientAD: Accurate Visual Anomaly Detection at Millisecond-Level Latencies," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 128–138, 2024.
- [53] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "SimpleNet: A Simple Network for Image Anomaly Detection and Localization," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20 402–20 411, 2023.
- [54] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14 318–14 328, 2022.
- [55] A. Mousakhan, T. Brox, and J. Tayyub, "Anomaly Detection with Conditioned Denoising Diffusion Models," *arXiv preprint arXiv:2305.15956*, 2023.
- [56] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Chapter 6, 2016.
- [57] J. Weese and C. Lorenz, "Four challenges in medical image analysis from an industrial perspective," *Medical Image Analysis*, vol. 33, pp. 44–49, 2016.

- [58] K. Lis, K. Nakka, P. Fua, and M. Salzmann, "Detecting the Unexpected via Image Resynthesis," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2152–2161, 2019.
- [59] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs," *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8798–8807, 2018.
- [60] R. Chan, K. Lis, S. Uhlemeyer, H. Blum, S. Honari, R. Siegwart, P. Fua, M. Salzmann, and M. Rottmann, "SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation," *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [61] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016.
- [62] C. I. Bercea, B. Wiestler, D. Rueckert, and J. A. Schnabel, "Towards Universal Unsupervised Anomaly Detection in Medical Imaging," *arXiv preprint arXiv:2401.10637*, 2024.
- [63] M. E. Tschuchnig and M. Gadermayr, "Anomaly Detection in Medical Imaging – A Mini Review," *arXiv preprint arXiv:2108.11986*, 2021.
- [64] A. Krizhevsky and G. Hinton, "Learning Multiple Layers of Features from Tiny Images," *Toronto, ON, Canada*, 2009.
- [65] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture on IE*, vol. 2, pp. 1–18, 2015.
- [66] W. Matthias, H. Tobias, and H. Fred, "Weakly supervised learning for industrial optical inspection." [Online]. Available: <https://hci.iwr.uni-heidelberg.de/content/weakly-supervised-learning-industrial-optical-inspection>
- [67] Y. Huang, C. Qiu, and K. Yuan, "Surface defect saliency of magnetic tile," *The Visual Computer*, vol. 36, pp. 85–96, 2020.

- [68] D. Tabernik, S. Šela, J. Skvarč, and D. Skočaj, "Segmentation-based deep-learning approach for surface-defect detection," *Journal of Intelligent Manufacturing*, vol. 31, pp. 759–776, 2020.
- [69] J. Božič, D. Tabernik, and D. Skočaj, "Mixed supervision for surface-defect detection: From weakly to fully supervised learning," *Computers in Industry*, vol. 129, 2021.
- [70] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, 2015.
- [71] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 233–240, 2006.
- [72] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [73] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss Functions for Image Restoration With Neural Networks," *IEEE Transactions on Computational Imaging*, vol. 3, pp. 47–57, 2016.
- [74] J. Snell, K. Ridgeway, R. Liao, B. D. Roads, M. C. Mozer, and R. S. Zemel, "Learning to generate images with perceptual similarity metrics," *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 4277–4281, 2017.
- [75] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, pp. 139–144, 2020.
- [76] 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.
- [77] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially Learned One-Class Classifier for Novelty Detection," *Proceedings of the IEEE/CVF Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3379–3388, 2018.

- [78] D. Dehaene, O. Frigo, S. Combrexelle, and P. Eline, "Iterative energy-based projection on a normal data manifold for anomaly localization," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.
- [79] Y. Liu, C. Zhuang, and F. Lu, "Unsupervised Two-Stage Anomaly Detection," *arXiv preprint arXiv:2103.11671*, 2021.
- [80] H. Chung, J. Park, J. Keum, H. Ki, and S. Kang, "Unsupervised Anomaly Detection Using Style Distillation," *IEEE Access*, vol. 8, pp. 221 494–221 502, 2020.
- [81] M. Haselmann, D. P. Gruber, and P. Tabatabai, "Anomaly Detection Using Deep Learning Based Image Completion," *Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1237–1242, 2018.
- [82] A. Munawar and C. Creusot, "Structural inpainting of road patches for anomaly detection," *Proceedings of the IAPR International Conference on Machine Vision Applications (MVA)*, pp. 41–44, 2015.
- [83] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognition*, vol. 112, p. 107706, 2021.
- [84] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative Image Inpainting with Contextual Attention," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5505–5514, 2018.
- [85] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *arXiv preprint arXiv:1711.05101*, 2017.
- [86] L. N. Smith and N. Topin, "Super-convergence: very fast training of neural networks using large learning rates," *Artificial intelligence and machine learning for multi-domain operations applications*, pp. 369–386, 2019.
- [87] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13 001–13 008, 2020.

- [88] R. Fong and A. Vedaldi, "Occlusions for effective data augmentation in image classification," *Proceedings of the International Conference on Computer Vision Workshop (ICCVW)*, pp. 4158–4166, 2019.
- [89] G. Liu, F. A. Reda, K. J. Shih, T. C. Wang, A. Tao, and B. Catanzaro, "Image Inpainting for Irregular Holes Using Partial Convolutions," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 85–100, 2018.
- [90] X. Qin, H. Dai, X. Hu, D.-P. Fan, L. Shao, and L. V. Gool, "Highly Accurate Dichotomous Image Segmentation," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 38–56, 2022.
- [91] R. Kozma, C. Alippi, Y. Choe, and F. C. Morabito, *Artificial intelligence in the age of neural networks and brain computing*. Elsevier, 2023.
- [92] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "CutPaste: Self-Supervised Learning for Anomaly Detection and Localization," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9664–9674, 2021.
- [93] S. Venkataramanan, K.-C. Peng, R. V. Singh, and A. Mahalanobis, "Attention Guided Anomaly Localization in Images," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 485–503, 11 2019.
- [94] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution Knowledge Distillation for Anomaly Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14 902–14 912, 2021.
- [95] J. Brownlee, "How to Visualize Filters and Feature Maps in Convolutional Neural Networks," 2019. [Online]. Available: <https://machinelearningmastery.com/how-to-visualize-filters-and-feature-maps-in-convolutional-neural-networks/>
- [96] D. Deepak, "CNN Visualization Techniques: Feature Maps, Gradient Ascent," 2023. [Online]. Available: <https://medium.com/@deepeshdeepakdd2/cnn-visualization-techniques-feature-maps-gradient-ascend-aec4f4aaf5bd>
- [97] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, vol. 2, 2017.

- [98] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.
- [99] Z. Gu, L. Liu, X. Chen, R. Yi, J. Zhang, Y. Wang, C. Wang, A. Shu, G. Jiang, and L. Ma, "Remembering Normality: Memory-guided Knowledge Distillation for Unsupervised Anomaly Detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16 401–16 409, 2023.
- [100] H. Park, J. Noh, and B. Ham, "Learning Memory-guided Normality for Anomaly Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14 372–14 381, 2020.
- [101] J. Hou, Y. Zhang, Q. Zhong, D. Xie, S. Pu, and H. Zhou, "Divide-and-Assemble: Learning Block-wise Memory for Unsupervised Anomaly Detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8791–8800, 2021.
- [102] H. Zhang, Z. Wu, Z. Wang, Z. Chen, and Y.-G. Jiang, "Prototypical Residual Networks for Anomaly Detection and Localization," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16 281–16 291, 2023.
- [103] W. H. Pinaya, P. D. Tudosiu, R. Gray, G. Rees, P. Nachev, S. Ourselin, and M. J. Cardoso, "Unsupervised brain imaging 3D anomaly detection and segmentation with transformers," *Medical Image Analysis*, vol. 79, 2022.
- [104] K. Choromanski, V. Likhoshesterov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. Colwell, and A. Weller, "Rethinking Attention with Performers," *arXiv preprint arXiv:2009.14794*, 2020.
- [105] F. Mentzer, D. Minnen, E. Agustsson, and M. Tschannen, "Finite Scalar Quantization: VQ-VAE Made Simple," *arXiv preprint arXiv:2309.15505*, 2023.
- [106] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A Generative Model for Music," *arXiv preprint arXiv:2005.00341*, 2020.

- [107] J. Rocca, "Understanding Variational Autoencoders (VAEs)," 2019. [Online]. Available: <https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73>