

POLITECNICO DI TORINO



Master Degree course in Data Science and Engineering

Master Degree Thesis

Efficiency-Optimized PatchCore for Anomaly Detection in Future Edge Deployment

Supervisors

Prof. Andrea BOTTINO
Giuseppe CHIRICO

Candidate

Fatemeh ZAHIRI

ACADEMIC YEAR 2025

Copyright and Confidentiality

This thesis is published in the open-access digital library of Politecnico di torino with explicit permission granted by **Blue Company Srl**. All the proprietary methods, data, or results presented in this document are published with the consent of **Blue Company Srl**. Any further reproduction, redistribution, or use of this content beyond academic purposes must be authorized explicitly by **Blue Company Srl**.

© 2024-2025 **Blue Company Srl**. All rights reserved.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Prof. Andrea Bottino and Eng. Giuseppe Chirico for their invaluable guidance and continuous support throughout this research. I am also grateful to my family and friends for their unwavering encouragement and love.

"what we need is a machine that can learn from experience"
Alen Turing 1947

Abstract

Anomaly detection plays a critical role in industrial quality control, ensuring defect-free production in sectors like automotive, electronics, and manufacturing. Traditional anomaly detection methods rely on supervised learning, which requires large datasets with labeled anomalies, often impractical in real-world scenarios. Unsupervised methods such as PatchCore have gained attention for their ability to detect anomalies without labeled defect data, making them particularly valuable for automated inspection systems in factories.

Despite its high accuracy, **PatchCore** has limitations that hinder its deployment in real-time edge devices like smart cameras used in industrial environments. The memory bank storing patch-level embeddings grows exponentially with dataset size, making it computationally expensive and requiring large storage capacity. This restricts its practical application in low-memory, low-power industrial devices. To address these limitations, this thesis explores a memory-efficient optimization of PatchCore by integrating Principal Component Analysis (**PCA**) and **K-Means** clustering. PCA is applied to reduce feature dimensionality, preserving the most informative components while eliminating redundancy. K-Means clustering further compresses the memory bank by grouping similar feature vectors and using cluster centroids instead of all stored patches. This approach significantly reduces memory usage while maintaining a balance between accuracy and computational efficiency.

The proposed **PCA-KMeans PatchCore** method was evaluated on the MVTec Anomaly Detection Dataset, a benchmark for industrial defect detection. Results show that, compared to the original PatchCore, the proposed method achieves a substantial reduction in memory footprint while maintaining acceptable anomaly detection performance. Although a slight drop in accuracy was observed, the efficiency gains make the model more practical for real-time applications in smart manufacturing. This study demonstrates that optimizing PatchCore for future edge deployment is feasible, making real-time anomaly detection on resource-constrained devices a practical reality.

Acronyms

AD	Anomaly Detection — Automatically detecting unusual patterns in data
AE	Autoencoder — A neural network that learns to reconstruct input data
AUROC	Area Under the Receiver Operating Characteristic — Metric for detection performance
CAD-SD	Co-occurrence Anomaly Detection Screw Dataset — Custom dataset for anomalies
CPU	Central Processing Unit
GPU	Graphics Processing Unit — Accelerator for parallel computations
K-Means	A popular clustering algorithm, grouping data points into k clusters
MVTec AD	MVTec Anomaly Detection dataset — A standard industrial AD benchmark
CNN	Convolutional Neural Network
PCA	Principal Component Analysis — Technique for dimensionality reduction
PRO	Per-Region Overlap — Metric evaluating region-level anomaly detection
ROC	Receiver Operating Characteristic — Curve visualizing detection performance

Contents

Acronyms	2
List of Figures	7
List of Tables	8
1 Introduction	11
1.1 Background and Motivation	11
1.2 Problem Statement	12
1.3 Research Objectives	13
1.4 Thesis Outline	14
2 Literature Review	17
2.1 Introductory Concepts	17
2.1.1 PCA	17
2.1.2 K-MEANS	22
2.2 Overview of Industrial Anomaly Detection	27
2.2.1 Key Challenges	28
2.3 Classical Approaches to Anomaly Detection	29
2.4 Deep Learning Methods	30
2.4.1 Reconstruction-Based Methods	30
2.4.2 Feature Embedding-Based Methods	30
2.4.3 Extensions to PatchCore	31
2.5 Multi-Class and Few-Shot AD	31
2.6 PCA and K-Means for Memory Bank Optimization	32
2.6.1 Motivation for PCA + K-Means	32
2.6.2 Prior Work on PCA or K-Means in AD	32
2.7 Summary of Key Insights	33
3 Methodology	37
3.1 Overview of the Proposed System	37
3.2 Baseline PatchCore Recap	38
3.3 Proposed PCA + K-Means Integration	39

3.3.1	Rationale for PCA and K-Means	39
3.3.2	Algorithmic Pipeline	39
3.4	Inference (Test) Stage	40
3.4.1	Selecting the Number of Clusters k	41
3.4.2	Handling Few-Shot Settings	42
3.5	Implementation Details and Variants	42
3.5.1	Backbone CNN and Layer Selection	42
3.5.2	Dimensionality Reduction: PCA vs. Random Projection	43
3.5.3	Coreset + K-Means Hybrid Approach	43
3.5.4	Distance Metric Considerations	43
3.5.5	Thresholding and Anomaly Mask Generation	43
3.6	Summary of the Method	44
4	Experimental Setup	47
4.1	Datasets	47
4.1.1	MVTec Anomaly Detection	47
4.1.2	Additional / Custom Datasets	48
4.2	Implementation Details	50
4.2.1	Feature Extraction	50
4.2.2	Dimensionality Reduction (PCA)	51
4.2.3	Clustering (K-Means)	51
4.2.4	Memory Bank Construction	51
4.2.5	Training Time and Computing Infrastructure	52
4.2.6	Inference Process	52
4.3	Experimental Protocol	53
4.3.1	Training, Validation, and Test Splits	53
4.3.2	Hyperparameter Search	53
4.3.3	Data Augmentation	53
4.4	Evaluation Metrics	54
4.5	Baselines	55
4.6	Summary	55
5	Experimental Results and Evaluation	59
5.1	MVTec AD Results	59
5.1.1	Image-Level Anomaly Detection	59
5.1.2	Pixel-Level Anomaly Segmentation	60
5.1.3	Impact of Number of Clusters (k)	61
5.1.4	Impact of PCA Components	61
5.1.5	Comparison Of Image-Level And Pixel-Level Avg AUROC	62
5.2	Inference Speed and Memory Analysis	62
5.2.1	Memory Usage	62
5.2.2	Inference Time	63

5.2.3	Best Configuration Analysis	64
5.3	Summary and Discussion	64
6	Conclusion and Future Work	67
6.1	Conclusion	67
6.2	Key Contributions	68
6.3	Limitations	68
6.4	Future Work	69
6.5	Final Remarks	70
	Bibliography	71

List of Figures

2.1	Percentage of Variance (Information)The following figure is taken from [36]	18
2.2	Principal Components visualization	19
2.3	K means Clustering figure is taken from [40]	22
2.4	Clustering dataset. figure is taken from [40]	24
2.5	The plot displays a scatter plot of data points ($X[:,0]$, $X[:,1]$) with grid lines. It also marks the initial cluster centers (red stars) generated for K-means clustering. figure is taken from [40]	25
2.6	The plot shows data points colored by their predicted clusters. The red markers represent the updated cluster centers after the E-M steps in the K-means clustering algorithm. figure is taken from [40]	27
2.7	Elbow Method Visualization: Optimal k is where the curve bends. figure is taken from [41]	28
4.1	Samples from the MVTec AD datasets	48
4.2	Samples from the MVTec AD datasets	48
4.3	Co-occurrence Anomaly Detection Screw Dataset	49
4.4	VisA (Visual Anomaly Dataset) contains 12 subsets corresponding to 12 different objects	50
5.1	The plot displays a Comparison Of Image-Level And Pixel-Level Avg AUROC	63

List of Tables

5.1	Image-level AUROC (%) results for all categories on MVTec AD. . .	60
5.2	Image-level AUROC (%) results for all categories on MVTec AD. . .	61
5.3	Effect of the number of clusters on accuracy (MVTec AD).	61
5.4	Performance analysis with varying PCA dimensions.	62
5.5	Inference speed comparison (MVTec AD).	63
5.6	Best hyperparameter configuration for PCA + K-Means PatchCore.	64

Chapter 1

Introduction

1.1 Background and Motivation

Industrial anomaly detection is central to quality assurance in manufacturing. Defective products, ranging from subtle scratches to missing components, can slip through manual visual inspection, causing financial losses and reputational damage [4], [5]. Traditional supervised machine learning approaches require many labeled examples of defective products (anomalies) to learn effectively. However, in practical manufacturing scenarios, defects are usually rare. Sometimes, these anomalies are so rare or novel that they haven't been seen before, making it extremely difficult to gather enough examples for training supervised methods.

Recent efforts to automate visual inspection with deep learning have accelerated progress, especially through industrial anomaly detection benchmarks such as MVTec AD [1]. However, in many real-world cases, anomaly samples are scarce. This poses a challenge: if we cannot train a robust model on explicit examples of defects, how can we reliably detect them?

Unsupervised methods, also known as 'cold-start' approaches, offer a promising solution. They primarily leverage only normal (non-defective) data during training. The idea is simple yet powerful: The model learns what 'normal' looks like from a large set of defect-free images. Also during inference (testing), if something deviates significantly from the learned normal patterns, the model flags it as anomalous (potentially defective).

Among these methods, PatchCore, originally introduced by Roth et al. [6] became a seminal work for industrial anomaly detection, achieving near state-of-the-art performance on MVTec AD. It takes a large set of normal images (defect-free products). At its core, PatchCore Breaks down these images into small regions (patches) and extracts features from each patch using a pretrained neural network (usually pretrained on ImageNet, a general-image dataset) AND stores these features into a memory bank. PatchCore extracts similar patches and compares them against the memory bank. During inference, each patch of a test image is compared to

the memory bank; if no 'similar' patch is found, that region is flagged as anomalous. While PatchCore is effective on standard benchmarks (e.g., >99% AUROC on MVTec AD [6]), it still faces real-world difficulties:

1. High-dimensional patch features create computational overhead, requiring significant processing power and resources, especially challenging for small, low-power devices.
2. Comparing each new test patch with an extensive memory bank slows down detection, hindering real-time applications on edge devices (e.g., embedded cameras or sensors).
3. Bias toward natural-image features from ImageNet pretraining may not fully match specialized industrial data [7], [3].

Hence, optimizing PatchCore for low-power, real-time anomaly detection is crucial. This is where techniques such as Principal Component Analysis (PCA) and K-Means clustering come into play: they can compress the memory bank and accelerate nearest-neighbor computations without significantly hurting accuracy [8], [9].

PCA reduces the dimensionality of stored features in the memory bank by finding principal components, effectively compressing the feature space. This speeds up comparisons while preserving most of the original information. K-Means groups similar patch features together, drastically reducing the number of individual features to compare during inference. This significantly improves computation speed and reduces the storage requirement for the memory bank.

1.2 Problem Statement

PatchCore fundamentally relies on storing and utilizing a large memory bank composed of patch-level embeddings extracted from normal product images. The process of maintaining and querying such an extensive memory bank poses significant challenges in edge computing scenarios, particularly when deployed on industrial cameras or other low-power devices with limited RAM capacity and stringent latency constraints [10]. Specifically, there are three critical challenges:

- **Memory overhead:** The sheer size of the patch feature bank can significantly impact storage efficiency, requiring considerable amounts of memory. Additionally, the large number of stored embeddings slows down the nearest-neighbor search process, directly affecting inference speed and overall operational efficiency.

- Time constraints: Real-time industrial inspection systems demand rapid inference times, typically in the sub-second range, to ensure immediate defect detection and minimal interruption of production workflows. The current computational complexity of PatchCores nearest-neighbor search in high-dimensional embedding spaces limits its applicability to environments with strict latency requirements.
- Adaptation gap: Despite PatchCore exceptional performance on standard benchmarks such as MVTec AD, it struggles to consistently generalize to a broader range of industrial defect scenarios. Specialized domains and more complex anomaly patterns, such as surfaces presenting co-occurring anomalies or nuanced defects specific to particular industrial processes, remain challenging for the current approach. This generalization difficulty highlights the adaptation gap between standard benchmark datasets and real-world industrial defect detection scenarios, emphasizing the need for approaches capable of addressing these nuanced detection tasks [11], [12].

Dimensionality reduction methods like Principal Component Analysis (PCA) and clustering techniques such as K-Means present promising strategies to mitigate these issues. By clustering similar patch embeddings using K-Means and compressing the dimensionality of these embeddings through PCA, redundancy can be effectively reduced. This combination has the potential to significantly decrease memory consumption and accelerate the inference process. However, the exact implications of applying PCA and K-Means clustering in terms of detection accuracy, especially when dealing with diverse training scenarios such as few-shot learning or extensive training datasets, remain uncertain and unclear and require rigorous evaluation [13], [14].

1.3 Research Objectives

1. Integrate PCA and K-Means clustering methodologies into the PatchCore memory bank framework to effectively compress storage requirements. The goal is to achieve significant reductions in memory footprint without substantially compromising the anomaly detection performance that characterizes the original PatchCore approach.
2. Conduct comprehensive analyses to determine the optimal balance between dimensionality reduction, achieved through PCA, and clustering efficiency provided by K-Means. These analyses will specifically evaluate how different parameter choices affect memory overhead, computational speed, and anomaly detection accuracy within real-world industrial contexts.

3. Evaluate the practical feasibility and efficiency of the optimized PatchCore approach using the standard MVTec AD dataset [1]. The focus will be on improving inference speed and reducing memory usage to align with constraints typically found in edge computing environments. Although deployment on real edge devices was not conducted, the experiments and design decisions were guided by the hardware limitations of such platforms to ensure potential real-time applicability and will be ready for future edge deployment.
4. Contextualize the challenges of co-occurrence [11] anomalies and few-shot learning [15] as motivating factors for optimizing memory efficiency. Although these scenarios were not directly tested in the experimental phase, they are conceptually aligned with the need to reduce memory overhead in PatchCore. Their inclusion in the literature review and design justification highlights the broader relevance and potential applicability of the proposed PCA + K-Means approach. Future work will include explicit evaluation in these scenarios to further validate generalizability and robustness.

1.4 Thesis Outline

- Chapter 2: Literature Review
Surveys anomaly detection methods (GAN-based, autoencoder-based, PatchCore-based) and highlights the role of pre-trained features [16] and discusses conceptual challenges such as co-occurrence anomalies and few-shot learning in relation to memory efficiency and scalability, though these are not experimentally addressed in this thesis.
- Chapter 3: Methodology
Introduces the proposed PCA + K-Means memory compression strategy integrated into PatchCore. Describes the overall anomaly detection pipeline, design parameters (e.g., PCA dimensionality, clustering strategy), and outlines how these modifications aim to reduce memory usage and accelerate inference.
- Chapter 4: Experimental Setup
Details datasets (MVTec AD, etc.), implementation details, hardware constraints, and evaluation metrics (e.g., ROC curves, PRO metrics for localization).
- Chapter 5: Results and Analysis
Presents a comparative analysis between the original PatchCore [6] and the optimized PCA + K-Means version. Focuses on memory consumption, inference speed, and overall anomaly detection performance. While few-shot

and co-occurrence scenarios are conceptually relevant, this chapter does not include direct experimental results in those areas.

- Chapter 6: Conclusion and Future Work

Summarizes key findings and contributions. Emphasizes the value of memory optimization for real-world deployment. Suggests future research directions including explicit evaluation of co-occurrence and few-shot scenarios, testing on real edge devices, advanced clustering alternatives, and domain adaptation techniques for broader applicability or vision transformers [17].

Chapter 2

Literature Review

2.1 Introductory Concepts

In order to establish a clear foundation for the technical discussions presented throughout this thesis, this section provides concise explanations of several core concepts and techniques that underpin our approach. While the main contribution of this thesis is the integration of PCA and K-Means clustering into the Patch-Core framework, understanding these components - and the broader context of convolutional neural networks, feature extraction backbones, and memory-based anomaly detection - requires familiarity with certain fundamental ideas. This section is therefore intended to support readers who may not have a deep background in machine learning or computer vision, by offering a conceptual overview of the tools and principles that will appear repeatedly in the following chapters. These introductory explanations also ensure clarity and self-containment of the thesis for academic and industrial audiences alike.

2.1.1 PCA

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique in data science and machine learning. It transforms a high-dimensional dataset into a lower-dimensional space while preserving as much of the underlying variance and structure as possible. The primary goal is to simplify data by projecting it onto a new set of uncorrelated variables, called principal components, which are linear combinations of the original variables.

PCA operates under the principle that the first few components capture most of the information present in the original data. Although dimensionality reduction inherently involves some loss of information, PCA strategically minimizes this loss by prioritizing the directions (components) with the highest variance. This makes

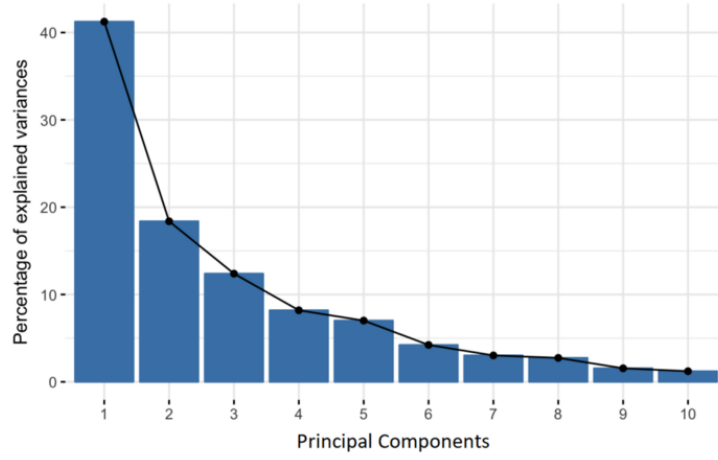


Figure 2.1. Percentage of Variance (Information)The following figure is taken from [36]

the resulting lower-dimensional data more manageable and computationally efficient, especially for machine learning tasks where visualization and performance are enhanced by reducing noise and redundancy.

In essence, PCA enables the representation of a complex dataset with fewer variables by selecting components that account for the majority of variance. This trade-off between dimensionality and accuracy is particularly beneficial in scenarios where interpretability and efficiency are critical.[36][37]

Principal Components are the newly derived variables that result from PCA. They are constructed as orthogonal (uncorrelated) linear combinations of the initial variables. The first principal component captures the direction with the highest variance in the dataset. Each subsequent component captures the highest remaining variance while being orthogonal to the preceding ones. For example, in a 10-dimensional dataset, PCA produces 10 principal components, but typically only the first few are retained for analysis, as illustrated in a scree plot fig 2.1.

These components, although mathematically meaningful, may lack direct interpretability in the original feature space. Geometrically, they can be viewed as axes that provide the most informative view of the data, revealing maximum spread among data points.

The construction of the principal components is based on maximizing variance.

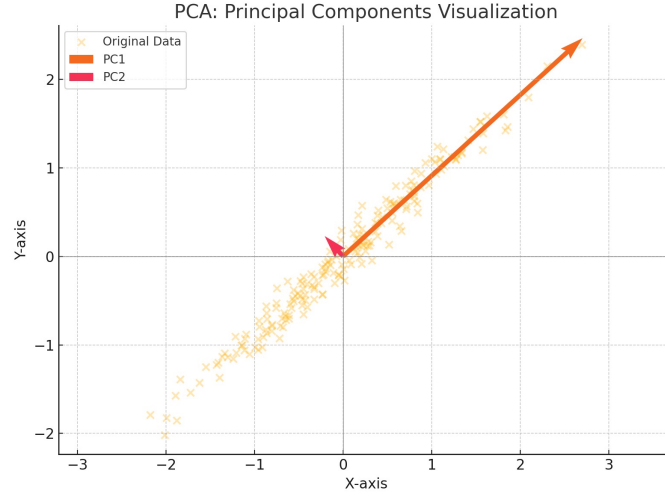


Figure 2.2. Principal Components visualization

The first principal component corresponds to the direction along which the projected data exhibits the largest spread from the origin. The second component is orthogonal to the first and captures the next highest variance, and this process continues until the full set of components is determined as shown in fig 2.2

Step-by-Step Explanation of PCA

1. Step 1: Standardization :

Before applying Principal Component Analysis (PCA), it is essential to standardize the input variables to ensure that each contributes equally to the analysis. Since PCA is sensitive to the variance of each variable, features with larger numerical ranges may disproportionately influence the results. For instance, a variable ranging from 0 to 100 may dominate another variable ranging from 0 to 1, leading to biased component directions. To mitigate this, the data are transformed to have zero mean and unit variance. This process ensures that all variables are on the same scale.[36] Mathematically, each value z is computed as:

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \quad (2.1)$$

Once standardized, all variables are expressed on a comparable scale, which is a prerequisite for the next steps in PCA.

2. step 2: Covariance Matrix Computation

The second step involves calculating the covariance matrix to understand the relationships between different variables. The covariance measures how two variables vary together from their respective means, which helps identify potential redundancy or correlation in the data.

The covariance matrix $\mathbf{C}$ is a symmetric $p \times p$ matrix, where p is the number of original variables. Each element represents the covariance between a pair of variables. For example, for a 3-dimensional dataset with variables x , y , and z , the covariance matrix is expressed as [39]:

$$\mathbf{C} = \begin{bmatrix} \text{Cov}(x, x) & \text{Cov}(x, y) & \text{Cov}(x, z) \\ \text{Cov}(y, x) & \text{Cov}(y, y) & \text{Cov}(y, z) \\ \text{Cov}(z, x) & \text{Cov}(z, y) & \text{Cov}(z, z) \end{bmatrix}$$

The diagonal elements represent the variance of each individual variable, such as:

$$\text{Var}(x) = \text{Cov}(x, x)$$

Since covariance is a commutative operation, we also have:

$$\text{Cov}(a, b) = \text{Cov}(b, a)$$

This symmetry implies that the covariance matrix is mirrored across its main diagonal, and both the upper and lower triangular portions contain the same values.

3. step 3: Eigen Decomposition of the Covariance Matrix

Next, we compute the eigenvectors and eigenvalues of the covariance matrix. Eigenvectors define the directions (or axes) of maximum variance, and the corresponding eigenvalues quantify the amount of variance captured along each direction. In PCA, these eigenvectors become the principal components. For a data set with p variables, we obtain p eigenvectors and p eigenvalues. The eigenvectors of the covariance matrix represent the new axes along which the data will be projected. The associated eigenvalues indicate the importance of each eigenvector:

- Larger eigenvalues : higher explained variance
- Smaller eigenvalues : less significant directions

By ranking the eigenvectors based on their corresponding eigenvalues in descending order, we determine the most informative directions in the data space.[38]

4. **step 4: Construction of the Feature Vector**

After computing the eigenvectors and ordering them by their corresponding eigenvalues in descending order, we determine which components (principal directions) to retain. This selection is typically based on a threshold for the cumulative explained variance, such as keeping components that account for 95% of the total variance.

In this step, we construct a matrix known as the **feature vector**, composed of the top k eigenvectors that correspond to the largest eigenvalues. These eigenvectors define the new subspace onto which the original data will be projected. Let k be the number of components selected for dimensionality reduction. Then, the feature vector is constructed as:

$$\mathbf{FeatureVector} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k]$$

where $\mathbf{e}_i$ are the eigenvectors corresponding to the top k eigenvalues.

This step completes the basis transformation from the original coordinate space to a new, reduced-dimensional space, preserving the most significant variance in the dataset.

5. **step 5:Recasting the Data** In the previous steps, the data has been standardized and the principal components have been selected in the form of a feature vector. However, the dataset still remains represented in terms of the original variables and coordinate system.

In this final step, we project the standardized data onto the new subspace defined by the selected principal components. This transformation reorients the data from the original coordinate axes to the axes represented by the principal components. The resulting transformed data is a lower-dimensional representation that retains the most significant structure and variance of the original data.

Mathematically, this is accomplished by multiplying the transpose of the feature vector matrix by the transpose of the standardized data matrix:

$$\mathbf{FinalDataSet} = \mathbf{FeatureVector}^\top \times \mathbf{StandardizedData}^\top$$

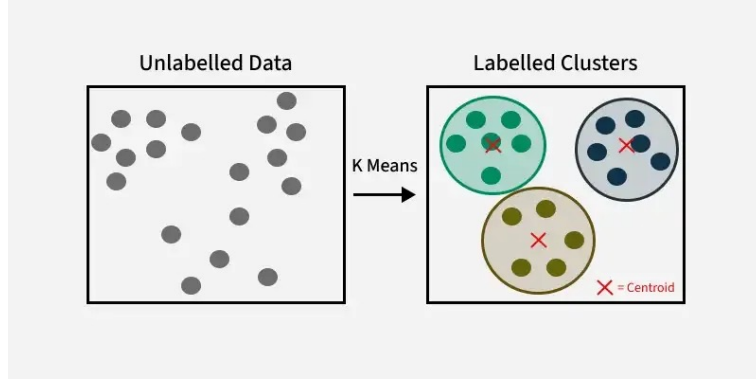


Figure 2.3. K means Clustering figure is taken from [40]

The outcome is a compact, information-preserving version of the original dataset, now expressed in the new coordinate space defined by the principal components. This is the reduced-dimension version of the data used for visualization, classification, or further analysis.[39]

2.1.2 K-MEANS

K-Means is an unsupervised machine learning algorithm that partitions an unlabeled dataset into a predefined number of clusters, based on feature similarity. It is a fundamental technique in data mining and pattern recognition, widely used in fields such as image segmentation, customer segmentation, and anomaly detection.

The primary objective of K-Means clustering is to organize a dataset into K non-overlapping groups, such that data points within the same cluster exhibit high similarity to each other and dissimilarity from points in other clusters. Similarity is typically measured using the Euclidean distance between data points and cluster centroids.

The algorithm proceeds iteratively through the following main steps:

1. **Initialization:** Select K initial centroids randomly from the dataset. These centroids represent the current centers of the clusters.
2. **Assignment Step:** Each data point is assigned to the cluster whose centroid is nearest to it in the feature space, usually determined by minimizing

Euclidean distance.

3. **Update Step:** After all points have been assigned, the centroids are recalculated as the mean of all data points in their respective clusters.
4. **Repeat:** Steps 2 and 3 are repeated until convergence is achieved i.e., when the centroids no longer change significantly between iterations, or a fixed number of iterations has been reached.

Mathematically, let $\mathbf{x}_i \in \mathbb{R}^d$ denote a data point, and $\boldsymbol{\mu}_j$ denote the centroid of the j^{th} cluster. Each data point is assigned to the cluster that minimizes the distance:

Assign $\mathbf{x}_i$ to cluster j such that $\|\mathbf{x}_i - \boldsymbol{\mu}_j\|_2$ is minimized

The centroids are then updated as:

$$\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$$

where C_j is the set of all data points assigned to cluster j , and $|\cdot|$ denotes the cardinality of the cluster.

One of the important aspects of the algorithm is centroid initialization. While random initialization is commonly used, poor initialization can lead to suboptimal solutions. Techniques like K-Means++ offer improved initialization strategies by spreading out the initial centroids.

K-Means is efficient and scalable, particularly for large datasets, though it assumes spherical clusters and may struggle with non-globular or overlapping distributions. Despite these limitations, its simplicity and speed make it a valuable tool in many machine learning pipelines, including memory-efficient anomaly detection, as explored in this thesis.

Implementation of K-Means Clustering

To illustrate the internal workings of the K-Means algorithm, we implemented an example step-by-step using Python, leveraging synthetic data generated via `make_blobs` from `scikit-learn`.^[40]

Step 1: Importing Required Libraries

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import make_blobs
```

Step 2: Data Generation and Visualization

```
X, y = make_blobs(n_samples=500, n_features=2, centers=3, random_state=23)
```

```
plt.figure()  
plt.grid(True)  
plt.scatter(X[:, 0], X[:, 1])  
plt.show()
```

output : fig [2.4](#)

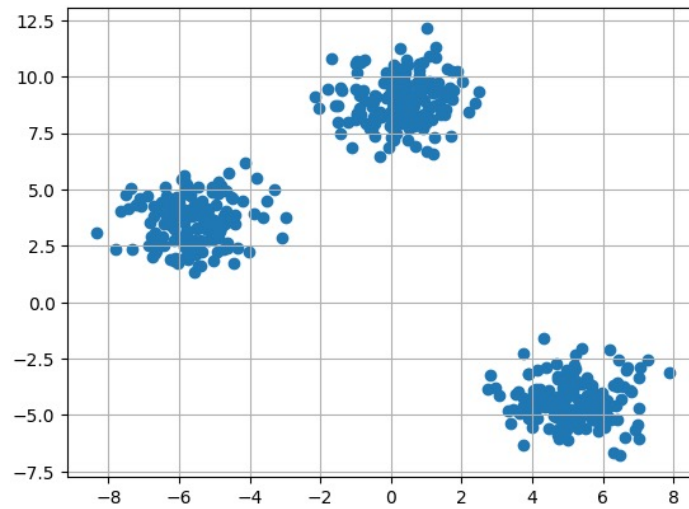


Figure 2.4. Clustering dataset. figure is taken from [40]

Step 3: Random Initialization of Centroids

```
k = 3  
clusters = {}  
np.random.seed(23)  
  
for idx in range(k):  
    center = 2*(2*np.random.random((X.shape[1],))-1)  
    points = []  
    cluster = {  
        'center' : center,  
        'points' : []  
    }  
}
```

```
clusters[idx] = cluster
```

```
clusters
```

Step 4: Initial Cluster Centers Plot

```
plt.scatter(X[:,0], X[:,1])
plt.grid(True)
for i in clusters:
    plt.scatter(*clusters[i]['center'], marker='*', c='red')
plt.title("Initial Random Cluster Centers")
plt.show()
```

output : fig [2.5](#)

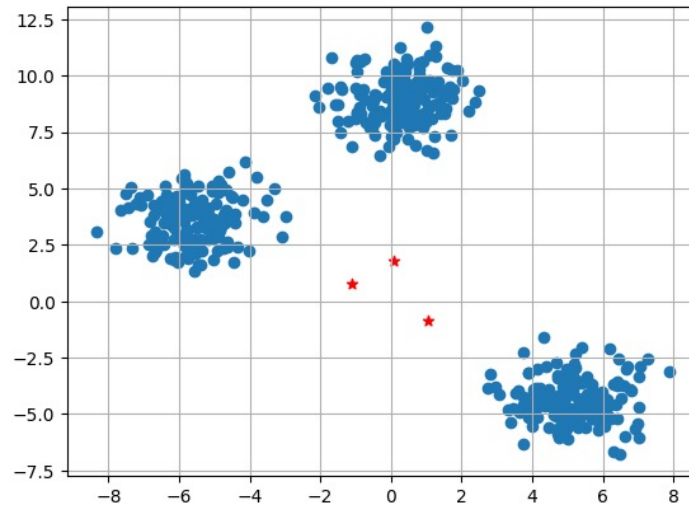


Figure 2.5. The plot displays a scatter plot of data points ($X[:,0]$, $X[:,1]$) with grid lines. It also marks the initial cluster centers (red stars) generated for K-means clustering. figure is taken from [40]

Step 5: Define Euclidean Distance Function

```
def distance(p1, p2):
    return np.sqrt(np.sum((p1 - p2)**2))
```

Step 6: Cluster Assignment and Update Functions

```
def assign_clusters(X, clusters):
```

```
for x in X:
    distances = [distance(x, clusters[i]['center']) for i in range(k)]
    cluster_idx = np.argmin(distances)
    clusters[cluster_idx]['points'].append(x)
return clusters

def update_clusters(clusters):
    for i in clusters:
        points = np.array(clusters[i]['points'])
        if len(points) > 0:
            clusters[i]['center'] = points.mean(axis=0)
            clusters[i]['points'] = []
    return clusters
```

Step 7: Predict Cluster for New Data

```
def pred_cluster(X, clusters):
    predictions = []
    for x in X:
        distances = [distance(x, clusters[i]['center']) for i in range(k)]
        predictions.append(np.argmin(distances))
    return predictions
```

Step 8: Run Assignment, Update, and Prediction

```
clusters = assign_clusters(X, clusters)
clusters = update_clusters(clusters)
pred = pred_cluster(X, clusters)
```

Step 9: Final Visualization with Predicted Clusters

```
plt.scatter(X[:,0], X[:,1], c=pred)
for i in clusters:
    plt.scatter(*clusters[i]['center'], marker='^', c='red')
plt.title("Final Cluster Assignments")
plt.show()
```

output : fig [2.6](#)

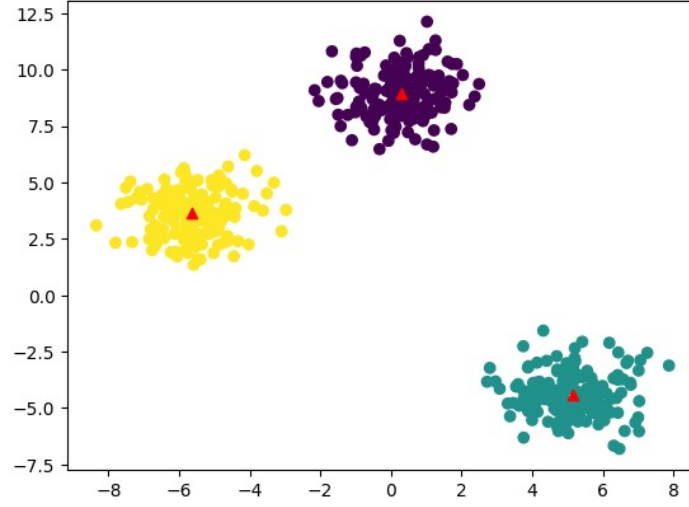


Figure 2.6. The plot shows data points colored by their predicted clusters. The red markers represent the updated cluster centers after the E-M steps in the K-means clustering algorithm. figure is taken from [40]

Elbow Method for Optimal k

Choosing the appropriate number of clusters k is a fundamental problem in unsupervised learning. The Elbow Method is a commonly used technique to determine this optimal value.

Within-Cluster Sum of Squares (WCSS): WCSS quantifies the compactness of the clustering.[41] It is defined as:

$$WCSS = \sum_{i=1}^k \sum_{j=1}^{n_i} \|x_j^{(i)} - c_i\|^2$$

where $x_j^{(i)}$ is the j^{th} point in cluster i , and c_i is the centroid of cluster i .

Elbow Identification: As k increases, WCSS typically decreases. The optimal k corresponds to the "elbow" point of the WCSS vs. k plot, where the rate of WCSS decrease sharply slows down.[41] fig 4.1

2.2 Overview of Industrial Anomaly Detection

Overview of Industrial Anomaly Detection Anomaly detection (AD) in manufacturing processes is crucial for ensuring product quality, maintaining safety standards, and reducing waste. Traditional manual visual inspection approaches are

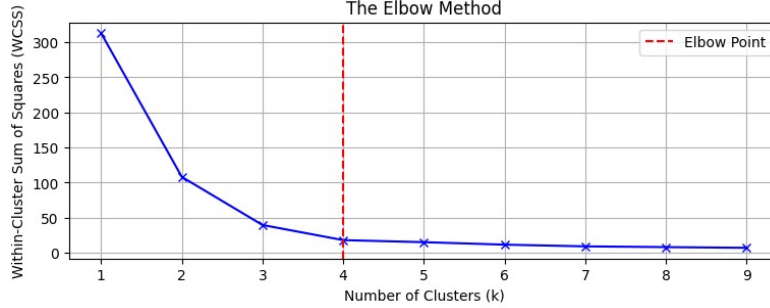


Figure 2.7. Elbow Method Visualization: Optimal k is where the curve bends. figure is taken from [41]

known to be labor-intensive, slow, subject to human error, and economically inefficient due to high inspection costs [18]. In response to these limitations, deep learning-based approaches have gained prominence over the past decade. These approaches leverage neural network architectures and extensive data-driven learning techniques capable of capturing intricate patterns and subtle defects commonly observed in industrial environments, such as surface scratches, missing components, subtle discolorations, or complex co-occurrence anomalies (e.g., mismatched or incorrectly assembled parts) [4], [5]. Although anomaly detection traditionally operates unsupervised, given the scarcity or absence of defect labels in real-world settings, recent developments have introduced semi-supervised and weakly supervised learning methodologies, exploiting minimal labeled anomalies to enhance detection accuracy and robustness [19].

2.2.1 Key Challenges

- **Limited Defect Samples:** A central challenge in industrial anomaly detection is the scarcity of labeled anomaly data. Defects are often rare, sporadic, and unpredictable in occurrence, resulting in a sparse distribution of anomalous samples. Consequently, anomaly detection algorithms frequently rely exclusively on normal samples for training, increasing the difficulty of identifying unseen anomalies during inference [3], [6].

- **High-Dimensional Feature Spaces:** Industrial inspection often involves processing high-resolution images, generating extensive amounts of detailed visual data. Handling such large, high-dimensional patch embeddings efficiently poses significant computational challenges, particularly concerning memory constraints and the need for real-time inference [11]. Effective techniques to compress and streamline these embeddings without losing critical anomaly-related information are essential for practical deployment.
- **Generalization Across Different Domains:** Industrial products and environments exhibit vast variability in appearance, including diverse colors, shapes, textures, and assembly configurations. Anomaly detection systems must therefore possess robust domain adaptation capabilities, enabling consistent and accurate anomaly detection performance despite significant differences in product types, manufacturing setups, and operating conditions [20], [21].
- **Edge Deployment Requirements:** Industrial anomaly detection systems are increasingly being deployed directly onto smart cameras and embedded hardware devices, typically characterized by strict constraints in terms of available memory, computational resources, and required inference speeds [13]. Achieving efficient, low-memory, real-time anomaly detection in these resource-limited edge environments necessitates optimized algorithms specifically tailored for minimal hardware footprints, rapid processing, and energy-efficient performance.

2.3 Classical Approaches to Anomaly Detection

Early anomaly detection research in industrial settings revolved around:

- **Handcrafted Features and Thresholding:** Traditional machine vision systems relied heavily on manually engineered features such as edges, corners, textures, color histograms, and statistical metrics to distinguish normal from defective products [22]. These methods typically employed threshold-based decision rules to classify anomalies. Despite their simplicity and interpretability, such handcrafted approaches suffered from high false-positive rates and lacked robustness when faced with novel or subtle defect types. Their effectiveness was largely confined to controlled, predictable environments, and they struggled to generalize to more complex or dynamic scenarios.
- **Autoencoder-based Reconstruction:** Autoencoders represent an important shift toward data-driven, unsupervised anomaly detection methods. These neural networks are trained to reconstruct input samples by compressing data into a latent representation and subsequently decoding it. During inference, anomalies are expected to produce larger reconstruction errors, as the network

has primarily learned to reconstruct normal patterns [23]. Although intuitive and effective in certain controlled contexts, autoencoder-based approaches often faced limitations in practice: the models sometimes reconstructed anomalies too effectively, resulting in missed detections. Additionally, selecting appropriate thresholds for reconstruction errors posed challenges, making it difficult to achieve consistently reliable anomaly detection performance across diverse industrial scenarios.

2.4 Deep Learning Methods

2.4.1 Reconstruction-Based Methods

Generative Adversarial Networks (GANs) [24], [25] and Variational Autoencoders (VAEs) [23] attempt to learn the underlying data distribution of normal samples by training models to generate realistic reconstructions. These methods can effectively handle complex, high-dimensional data distributions and facilitate pixel-level anomaly segmentation, providing precise defect localization. However, they frequently encounter training instability, including mode collapse or convergence issues typical of GANs, and tend to reconstruct anomalies too accurately, risking missed anomaly detections [4]. Additionally, these approaches generally require substantial amounts of normal training data to capture variations effectively. Defect GAN [25] introduces synthetic anomaly injection to augment training datasets, aiming to enhance robustness. However, synthetic defects often inadequately represent true industrial anomalies, limiting real-world effectiveness [4], [26].

2.4.2 Feature Embedding-Based Methods

Embedding-based methods employ deep Convolutional Neural Networks (CNNs), often pre-trained on large-scale datasets such as ImageNet, to extract robust and discriminative feature representations. Anomaly detection is performed by measuring deviations within the learned feature space:

- SPADE [7]: Utilizes patch-wise k-nearest neighbors (kNN) comparison against a stored memory bank of normal image features. SPADE is highly effective for detecting localized defects but suffers from scalability issues, as inference becomes computationally expensive when memory banks are extensive.
- PaDiM [27]: Estimates local patch distributions using Gaussian statistics and calculates the Mahalanobis distance to identify anomalies. While this approach achieves strong performance, it can be sensitive to variations in product alignment, orientation, or illumination, affecting detection accuracy.

- PatchCore [6]: Combines local patch embeddings with coreset subsampling to reduce memory usage, achieving high accuracy and fast inference. Despite near-state-of-the-art results, it can still face issues with domain shifts [20], [21], co-occurrence anomalies [5], or real-time edge deployment constraints [11].

2.4.3 Extensions to PatchCore

Several innovative extensions have further advanced the original PatchCore framework:

- SA-PatchCore [5]: Integrates self-attention mechanisms to better handle co-occurring anomalies, improving detection sensitivity and accuracy in complex scenarios.
- FR-PatchCore [21]: Employs a negative cosine similarity loss for updating the memory bank, aiming to improve domain generalization and adaptability across diverse manufacturing processes.
- Incorporates limited anomaly examples to guide and refine feature selection, bridging the gap between fully unsupervised and supervised methods, thus significantly enhancing anomaly detection precision.

These studies collectively highlight memory-based embedding methods as consistently top performers for robust and reliable industrial anomaly detection.

2.5 Multi-Class and Few-Shot AD

Multi-class Anomaly Detection expands upon traditional binary classification approaches (normal vs. anomalous) by further categorizing specific defect types [20], [21], [25]. Accurately identifying and classifying defect types can significantly improve the interpretability and actionable insights of anomaly detection systems, thus optimizing defect resolution processes in manufacturing. However, despite the importance of this granular labeling in certain industries, obtaining adequately labeled anomaly data is typically challenging due to their rarity and the substantial effort required for data labeling.

Few-Shot Anomaly Detection (Few-Shot AD) specifically targets scenarios characterized by extremely limited normal samples or, in extreme cases, no examples of anomalies [13], [22]. Such scenarios frequently occur in industries where data collection is inherently expensive, impractical, or time-consuming. Approaches addressing this limitation include few-shot transfer learning [22], which leverages knowledge from related domains or tasks, and domain generalization methods [20], which aim

to ensure robust performance despite minimal training data. Recent studies have explored applying PatchCore in few-shot scenarios, demonstrating that thoughtful adjustment of hyperparameters, incorporation of specialized data augmentations, or leveraging pre-trained feature embeddings can maintain or even enhance anomaly detection performance under restricted data availability conditions [13], [22].

2.6 PCA and K-Means for Memory Bank Optimization

2.6.1 Motivation for PCA + K-Means

While PatchCore successfully employs a greedy coresets subsampling strategy to lower memory consumption [6], integrating complementary compression techniques, such as Principal Component Analysis (PCA) and K-Means clustering, can provide further optimization benefits for memory and inference speed [28], [29].

- **PCA:** Principal Component Analysis effectively reduces the dimensionality of the original feature space by identifying principal directions of variance within the data. By compressing the high-dimensional patch embeddings into a lower-dimensional representation (from dimension d to a reduced dimension d^*), PCA significantly decreases the memory requirements for storing embeddings and accelerates the subsequent nearest-neighbor computations without substantial losses in detection accuracy [28].
- **K-Means:** This clustering approach aggregates similar patch embeddings into discrete groups, representing each group with a single centroid prototype. Instead of maintaining all patch embeddings, the memory bank stores only these centroids, dramatically reducing storage demands. Moreover, K-Means clustering simplifies the nearest-neighbor search complexity from $O(N)$, where N is the original number of embeddings, down to $O(k)$, where k (the number of centroids) is substantially smaller than N , thereby enhancing inference speed and real-time responsiveness [29].

2.6.2 Prior Work on PCA or K-Means in AD

Previous research has briefly touched upon the combination of classical dimensionality reduction techniques like PCA with PatchCore’s feature extraction pipeline. Ishida et al. [5] noted the potential benefits of integrating PCA, particularly in handling the complex, high-dimensional embeddings generated by self-attention mechanisms. Meanwhile, Tran et al. [13] explored data augmentation approaches to enhance Few-Shot AD performance but did not conduct a detailed

systematic analysis of PCA or clustering methods.

General anomaly detection studies utilizing K-Means clustering independently have demonstrated substantial improvements in inference speed by compressing feature embeddings without incurring significant accuracy trade-offs [9]. Extending these findings, integrating PCA and K-Means clustering directly with PatchCore could achieve a well-balanced trade-off between detection accuracy, computational efficiency, and memory optimization. Such an approach is particularly attractive for resource-constrained settings typical in real-time edge deployment scenarios, enabling practical industrial anomaly detection solutions tailored for embedded systems and small industrial cameras.

2.7 Summary of Key Insights

- **Reconstruction-based vs. Feature Embedding-based:** Comparative analyses reveal that feature embedding-based methods (e.g., SPADE, PaDiM, PatchCore) typically outperform reconstruction-based approaches (e.g., GANs, Autoencoders) on standard benchmarks such as MVTec AD, particularly in scenarios involving localized or subtle anomalies [4]. Embedding-based methods are more adept at capturing nuanced differences between normal and anomalous patterns within learned feature spaces.
- **Strengths and Limitations of PatchCore:** PatchCore consistently ranks among the highest-performing methods in industrial anomaly detection benchmarks, demonstrating exceptional accuracy and efficient memory usage through core-set subsampling. Despite its strengths, PatchCore currently lacks explicit mechanisms for handling critical real-world challenges, including:
 - Co-occurrence anomalies, where multiple defects occur simultaneously within the same product, complicating detection.
 - Few-shot learning or significant domain shifts, which frequently occur in practical industrial settings where data availability is limited or rapidly evolving.
 - Memory and computational constraints characteristic of real-time deployment on industrial lines, especially when dealing with resource-constrained edge devices and embedded systems.
- **PCA and K-Means for Memory Optimization:** Integrating PCA and K-Means clustering into PatchCore’s memory bank presents a promising strategy to significantly reduce memory overhead, improve inference speed, and enhance scalability. By compressing high-dimensional patch embeddings through PCA

and clustering similar features via K-Means, the approach can facilitate deployment on low-power devices and enable quick adaptation to new product lines or production environments.

- Synergistic Approach for Efficiency and Accuracy: The combined use of
 - (i) pretrained CNN models for robust feature extraction
 - (ii) selective adaptation through clustering methods or attention mechanisms
 - (iii) efficient memory management techniques such as PCA or K-Means clustering represents a compelling synergy. This integrated approach has significant potential to balance high accuracy in anomaly detection with practical efficiency, effectively addressing both computational limitations and performance demands encountered in industrial anomaly detection.

Given these insights, the remainder of this thesis focuses explicitly on developing and evaluating an integrated PatchCore framework enhanced with PCA and K-Means-based memory compression. The objective is to build an anomaly detection pipeline that is efficient and potentially suitable for real-world industrial use, particularly in memory- and speed-constrained environments. While direct deployment and complex anomaly types such as co-occurrence were not experimentally tested, the system design was guided by such challenges to ensure conceptual robustness and practical alignment.

Chapter 3

Methodology

3.1 Overview of the Proposed System

The primary objective of our proposed methodology is to enhance PatchCore [6] a leading memory-based anomaly detection framework, by significantly reducing its memory consumption and increasing inference speed, all while preserving its exceptional anomaly detection capabilities. Our optimized pipeline, depicted in Figure 3.1, involves the following sequential steps:

1. **Feature Extraction:** Employing a pretrained Convolutional Neural Network (CNN) for robust and discriminative feature extraction from input images.
2. **Local Patch Embedding:** Dividing images into local patches and extracting corresponding embeddings to represent nominal features, forming an initial memory bank.
3. **Dimensionality Reduction (PCA):** Applying Principal Component Analysis (PCA) to the extracted patch embeddings, compressing their dimensionality by capturing the most significant variance in fewer dimensions, thereby reducing memory storage requirements.
4. **Clustering (K-Means):** Utilizing K-Means clustering to organize compressed embeddings into clusters, enabling the storage of cluster centroids instead of all individual embeddings. This substantially reduces the memory footprint and accelerates nearest-neighbor search.
5. **Inference Process:** During inference, each test image undergoes identical feature extraction, dimensionality reduction, and embedding procedures. Each resulting patch embedding is compared against the stored cluster centroids, generating anomaly scores based on the distances.

Integrating PCA and K-Means clustering within the memory bank construction effectively addresses PatchCore’s memory constraints, speeding up inference and enhancing the method’s suitability for deployment in real-time and resource-constrained industrial scenarios. The detailed explanation of each pipeline component is presented below.

3.2 Baseline PatchCore Recap

PatchCore [6] operates primarily through extracting and utilizing local patch-level features obtained from intermediate layers of a CNN pretrained on the ImageNet dataset [2]. The baseline PatchCore approach for normal (defect-free) training images involves the following steps:

1. **Patchification:** Images are systematically divided into overlapping patches, typically small enough to capture detailed local information relevant for anomaly detection.
2. **Feature Extraction:** Each patch is individually passed through selected intermediate layers of the pretrained CNN, generating rich and discriminative feature vectors representing visual patterns of normal products.
3. **Memory Bank Construction:** The generated feature vectors from all training patches are collected into a comprehensive repository known as the memory bank. This memory bank serves as a reference for defining the normal state of product appearance.
4. **Coreset Subsampling:** To enhance computational efficiency, PatchCore employs a greedy coreset selection method, drastically reducing redundancy by selecting a representative subset (typically about 1% to 10%) of the original memory bank features. This significantly decreases memory consumption while preserving high anomaly detection accuracy.

During inference, the process mirrors training by patchifying new images and extracting their feature embeddings. Each test image’s patch embeddings are compared against the established memory bank using distance metrics (commonly L2 distance). A high distance score indicates substantial deviation from the normal feature distribution, signifying a potential anomaly. Despite its effectiveness, PatchCore’s original implementation can be computationally demanding, especially when handling large images or extensive datasets, as the memory bank size directly affects both memory usage and inference speed.

3.3 Proposed PCA + K-Means Integration

3.3.1 Rationale for PCA and K-Means

As discussed in Chapter 2, Principal Component Analysis (PCA) [28] and K-Means clustering [29] are well-established techniques that provide effective means of compressing and organizing high-dimensional data, respectively. Integrating these two methods into PatchCore enhances memory efficiency and computational performance, providing complementary benefits:

- **PCA:** PCA systematically identifies principal directions in the embedding feature space that account for the greatest variance across patch embeddings. By projecting high-dimensional embeddings onto these principal components, we can significantly reduce the dimensionality of each embedding, for example, reducing embedding dimensions from 1024 down to approximately 256 or even fewer. This dimensional reduction substantially decreases memory requirements without critically losing information relevant to anomaly detection.
- **K-Means:** After dimensionality reduction via PCA, K-Means clustering is applied to group similar compressed embeddings into a defined number (k) of clusters. Each cluster is succinctly represented by its centroid, which is essentially an averaged representation of the embeddings within the cluster. Instead of storing individual embeddings, only these cluster centroids need to be retained, dramatically reducing the total number of stored vectors.

Hence, our integrated approach drastically reduces memory requirements by addressing two critical aspects:

1. Reducing the dimensionality (and consequently the size) of individual feature vectors through PCA.
2. Minimizing the total quantity of stored embeddings by utilizing K-Means centroids instead of all original embeddings.

3.3.2 Algorithmic Pipeline

Algorithm 1 provides a comprehensive pseudocode outlining the process for constructing our enhanced PatchCore memory bank through PCA and K-Means clustering integration. This approach efficiently compresses the feature space and significantly optimizes the memory storage and inference speed.

Algorithm 1 Building the PCA + K-Means PatchCore Memory

Input:

- $\{I_i\}_{i=1}^N$: Set of N normal (nominal) images
- Pretrained CNN $\phi(\cdot)$
- Principal components d^*
- Number of clusters k

Output: Memory Bank $\mathcal{C}$ of dimension $(k \times d^*)$

```

1: Feature Extraction:
2:  $\mathcal{F} \leftarrow \emptyset$  // Initialize feature set
3: for each image  $I_i$  in training set do
4:   Patchify  $I_i$  into  $\{p_j\}$ 
5:   for each patch  $p_j$  do
6:     Extract feature  $\mathbf{f}_j = \phi(p_j)$ 
7:     Append  $\mathbf{f}_j$  to  $\mathcal{F}$ 
8:   end for
9: end for
10: Apply PCA:
11: Fit PCA on  $\mathcal{F}$ 
12: for each feature  $\mathbf{f} \in \mathcal{F}$  do
13:   Project onto top  $d^*$  components:  $\mathbf{z} = \text{PCA}(\mathbf{f})$ 
14: end for
15: Let  $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \dots\}$  // Reduced features
16: K-Means Clustering:
17: Run K-Means on  $\mathcal{Z}$  with  $k$  clusters
18: Let  $\{\mathbf{c}_1, \dots, \mathbf{c}_k\}$  be the learned centroids
19: Store Centroids:
20:  $\mathcal{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_k\}$  // Final memory bank
21: return  $\mathcal{C}$ 

```

3.4 Inference (Test) Stage

Given a new image $\hat{I}$, the following steps are performed to detect anomalies:

1. **Feature Extraction:** Patchify the image $\hat{I}$ into small overlapping regions and extract patch features $\hat{\mathbf{f}}_j$ using the same pretrained CNN backbone $\phi(\cdot)$ as used during training.
2. **Dimensionality Reduction (PCA):** Apply the PCA transformation (learned

during training) to reduce the dimensionality of the extracted features. Each patch embedding $\hat{\mathbf{f}}_j$ is projected to a lower-dimensional space:

$$\hat{\mathbf{z}}_j = \text{PCA}(\hat{\mathbf{f}}_j), \quad \hat{\mathbf{z}}_j \in \mathbb{R}^{d^*}$$

3. **Nearest-Centroid Search:** For each reduced patch embedding $\hat{\mathbf{z}}_j$, compute the minimum L2 distance to the cluster centroids $\{\mathbf{c}_t\}_{t=1}^k$:

$$d_j = \min_{1 \leq t \leq k} \|\hat{\mathbf{z}}_j - \mathbf{c}_t\|_2$$

A large value of d_j implies that the patch is unlike the known nominal patterns, thus indicating a potential anomaly.

4. **Anomaly Map Generation:** Rearrange the distances $\{d_j\}$ back to the spatial patch grid, forming an anomaly map. Thresholding or applying a soft heatmap highlights anomalous regions.

3.4.1 Selecting the Number of Clusters k

The selection of the cluster count significantly influences the balance between model efficiency and detection accuracy. Choosing an optimal number of clusters is critical for practical deployment. We propose employing two systematic strategies for determining the optimal k :

- **Elbow Method:** The elbow method involves evaluating clustering performance by plotting the total within-cluster sum of squares (WCSS) against various candidate values of k . As k increases, the WCSS naturally decreases. However, beyond a certain point, the marginal decrease becomes negligible, creating an "elbow" in the plot. This elbow point typically indicates the optimal cluster number, balancing clustering granularity and computational efficiency.
- **Range Exploration:** Given that the elbow method provides an initial heuristic, practical applications necessitate a more empirical validation. To determine the optimal number of clusters k , we performed a systematic exploration across a range of plausible values, specifically from $k = 50$ to $k = 300$. For each configuration, we empirically evaluated key performance metrics including inference speed, memory utilization, and anomaly detection accuracy. This experimental evaluation allowed us to identify a balanced trade-off between computational efficiency and detection performance.

Chapter 5 further explores the impact of varying k , providing detailed analyses of how different cluster counts influence the overall performance metrics.

3.4.2 Handling Few-Shot Settings

Real-world industrial applications frequently encounter limited data availability, especially during early production stages. While our thesis does not include a dedicated few-shot experimental setup, the proposed method incorporates design considerations that are inherently aligned with few-shot scenarios:

- **Reduced PCA Dimensionality:** Reduced PCA Dimensionality: When fewer training samples are available, reducing the number of retained principal components helps mitigate overfitting and improves generalization. Our approach benefits from this property by focusing on the most essential variance directions in the data.
- **Cluster Size Considerations:** Although we did not dynamically adjust k based on dataset size, lower cluster counts were evaluated in our experiments. This aligns with few-shot settings, as fewer clusters can lead to more stable and representative memory banks when training data is scarce.
- **Potential for Data Augmentation:** While not implemented in this thesis, standard augmentation techniques such as slight rotations, shifts, or brightness changes can support model robustness under low-data regimes. These techniques are suggested for future extensions.

3.5 Implementation Details and Variants

3.5.1 Backbone CNN and Layer Selection

We use a **WideResNet50** pretrained on the **ImageNet** dataset [2] to extract deep representations. Features are extracted from intermediate layers (e.g., `layer2` and `layer3`), which provide a balance between:

- **Low-Level Features:** Capturing detailed, fine-grained spatial features crucial for detecting subtle anomalies.
- **High-Level Features:** Representing semantic information, albeit with inherent bias toward pretraining data.

Empirical evidence from prior studies [6] indicates that mid-level layers consistently deliver superior performance in surface defect detection tasks, making them ideal candidates for industrial anomaly detection.

3.5.2 Dimensionality Reduction: PCA vs. Random Projection

Our primary method employs PCA due to its inherent capability to retain the most informative axes of variance, thus preserving critical anomaly-related features. PCA provides interpretable and reproducible transformations ideal for controlled experiments. Nevertheless, alternative methods such as Random Projection [32] could also be considered for scenarios demanding faster computational speeds at the cost of potentially less interpretability and slightly diminished accuracy.

3.5.3 Coreset + K-Means Hybrid Approach

In extremely resource-constrained environments such as edge computing platforms, embedded devices, or small industrial cameras a hybrid approach combining PatchCore’s greedy coreset selection with K-Means clustering could further optimize memory use. While theoretically beneficial, our empirical results indicate that employing K-Means clustering alone typically provides sufficient memory reduction without compromising detection accuracy significantly.

3.5.4 Distance Metric Considerations

By default, we adopt the L2 norm (Euclidean distance) as the primary metric for nearest-neighbor matching. This choice aligns with the original PatchCore formulation and ensures comparability across experiments. However, other distance measures, such as cosine similarity or Mahalanobis distance, can be beneficial for specific domains or datasets where relationships among features differ. Investigating alternative distance metrics represents a valuable future exploration.

3.5.5 Thresholding and Anomaly Mask Generation

To produce interpretable anomaly maps, we apply a threshold τ on the distance map. The threshold is determined using either:

- **F1 Score Maximization:** Choose τ to maximize the balance between precision and recall.
- **PRO Metric:** Use the Per-Region Overlap (PRO) criterion [1] for segmentation quality assessment.

We follow the original PatchCore methodology by Roth et al. (2022), we adopt an F1-score maximization approach to select the threshold τ on a labeled validation dataset. The threshold value is automatically determined by evaluating anomaly segmentation performance across multiple candidate thresholds and selecting the

one yielding the highest F1-score. Since our method primarily extends PatchCore by incorporating PCA and K-Means clustering, this threshold determination procedure remains unchanged.

3.6 Summary of the Method

This chapter presented a comprehensive overview of our enhanced PatchCore framework, which integrates PCA and K-Means clustering. Key benefits of this improved approach include:

- **Efficient Memory Usage:** PCA-based dimensionality reduction significantly compresses patch embeddings.
- **Reduced Inference Cost:** K-Means clustering minimizes the number of stored reference points while preserving detection accuracy.
- **Scalability to Few-Shot Settings:** Our methodology effectively handles limited-data scenarios through dynamic clustering, reduced PCA dimensionality, and targeted augmentation strategies.

In the following chapters, we detail our experimental setup (Chapter 5) and evaluate the trade-offs between memory efficiency, inference speed, and anomaly detection accuracy.

Chapter 4

Experimental Setup

This chapter comprehensively describes the experimental procedures employed to evaluate the performance and efficiency of our proposed PCA + K-Means enhanced PatchCore method. We detail the datasets utilized, elaborate on implementation specifics, define training and inference procedures, outline evaluation metrics, and present our hyperparameter optimization approach for both the baseline and enhanced PatchCore methods.

4.1 Datasets

4.1.1 MVTec Anomaly Detection

- Description: The MVTec AD dataset [1] represents a widely accepted benchmark for industrial anomaly detection tasks. It includes 15 distinct categories covering diverse industrial objects and textures, such as bottles, cables, screws, and capsules, to reflect realistic defect detection scenarios.
- Composition:
 - Training Set: Exclusively or predominantly normal (defect-free) samples, suitable for training anomaly detection models.
 - Test Set: A balanced mixture of normal and anomalous samples, covering various defect types such as surface scratches, missing components, structural defects, and other subtle anomalies.
- Resolution: Generally standardized around 1024×1024 pixels, though slight variations exist between categories, mirroring real-world image capture scenarios.
- Train/Test Split: Adhering to the official train-test partition described in [1], the training phase employs only normal samples, whereas evaluation is

conducted on a distinct test set comprising both normal and anomalous instances.

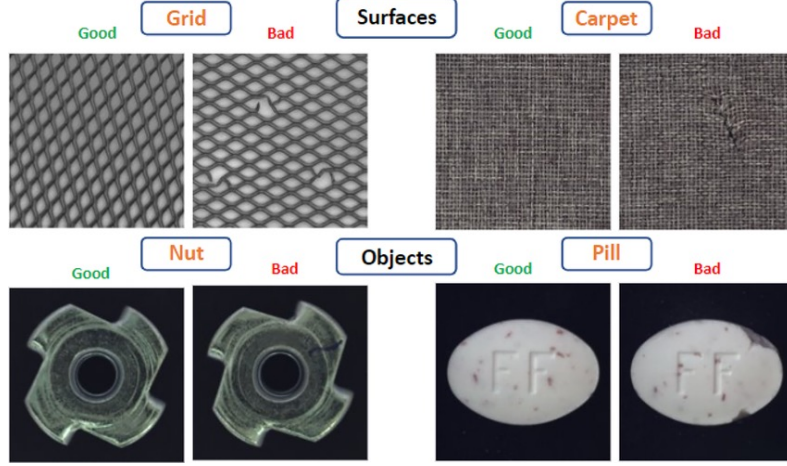


Figure 4.1. Samples from the MVTec AD datasets

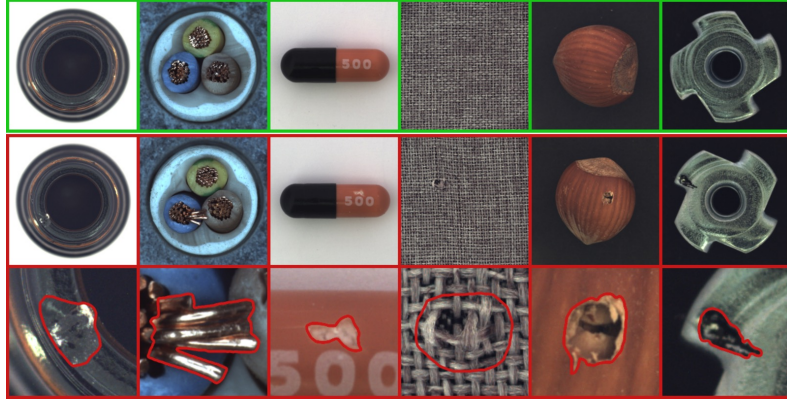


Figure 4.2. Samples from the MVTec AD datasets

4.1.2 Additional / Custom Datasets

- CAD-SD (Co-occurrence Anomaly Detection Screw Dataset)

- Motivation: Identified in the literature as a dataset designed to investigate performance on co-occurrence anomalies, which arise from improper combinations of parts such as missing nuts or additional washers [2].
- Relevance: Although not used in the experimental phase of this thesis, CAD-SD serves as a conceptual benchmark for motivating memory-efficient anomaly detection models that can handle complex scenarios beyond single-point anomalies.
- Future Work: This dataset is considered a strong candidate for future evaluation, particularly in extensions that address multi-component defect detection.

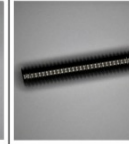
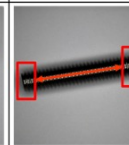
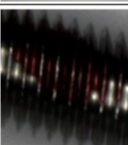
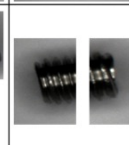
	Normal	Local Anomaly		Co-occurrence Anomaly	
		Scratch	Paint	Over-coupling	Lacking
Image					
Anomaly location					
Enlarged image					

Figure 4.3. Co-occurrence Anomaly Detection Screw Dataset

- VisA [3]
 - Description: A recent dataset proposed for generalizing anomaly detection models across diverse industrial product categories.
 - Relevance: While not experimentally included in this thesis, VisA is acknowledged as a useful benchmark for evaluating generalization beyond MVTec AD. It may be explored in future work to assess broader applicability of the proposed method.

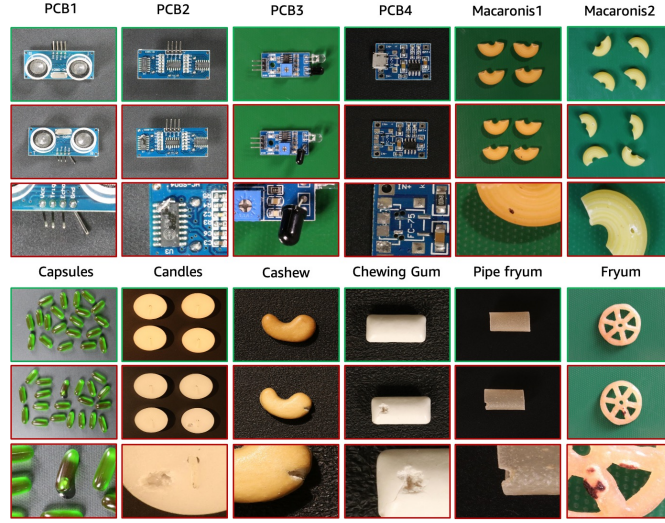


Figure 4.4. VisA (Visual Anomaly Dataset) contains 12 subsets corresponding to 12 different objects

4.2 Implementation Details

4.2.1 Feature Extraction

- **Backbone CNN:** For feature extraction, we predominantly use pretrained CNN architecture WideResNet50, leveraging its robustness and well-documented efficacy in feature representation tasks, as pretrained on the **ImageNet dataset** [2].
- **Layer Selection:** Consistent with previous research [6][7], we extract features from intermediate CNN layers (typically layer2 or layer3). The rationale is that these layers effectively balance capturing detailed local spatial information (useful for subtle anomaly detection) and semantic-level information (useful for more prominent defects).
- **Patch Extraction:**
 - **Patch Size:** We maintain the default PatchCore configuration, extracting 3×3 patches. This small size ensures sensitivity to fine-grained anomalies.
 - **Stride:** Empirically, strides of either 1 or 2 are tested, balancing computational efficiency with anomaly localization accuracy.
 - **Padding:** Padding of size 1 is employed to preserve edge information, ensuring boundary anomalies are effectively captured during patchification.

4.2.2 Dimensionality Reduction (PCA)

- **Number of Components d^* :** We systematically evaluate PCA dimensions d^* from the set 128, 256, 512. Empirical evidence from previous research [33] suggests that selecting $d^* = 256$ typically provides the best balance, effectively reducing dimensionality while preserving critical feature information required for accurate anomaly detection.
- **Fitting:** PCA is computed using the entire training feature set to capture a comprehensive variance profile. For large-scale datasets, where computing PCA on the full feature set becomes computationally demanding, a sufficiently large random subset of training features can be used without significant loss in generalization performance.
- **Software:** We employ scikit-learn’s PCA implementation with default parameters, specifically retaining the setting `whiten=False` to preserve the original feature variance scaling, which facilitates consistent comparisons and reproducible results.

4.2.3 Clustering (K-Means)

- **Number of Clusters k :** The cluster count k is extensively analyzed across a range of values 50, 100, 200, 300. The detailed impact of these different cluster numbers on performance metrics such as accuracy, memory consumption, and computational efficiency is thoroughly discussed in Chapter 5.
- **Initialization:** We utilize the *k-means++* algorithm [34] for cluster initialization, ensuring stable and faster convergence compared to random initialization.
- **Max Iterations:** Set to 300 to ensure sufficient convergence of cluster centroids, balancing computational cost and clustering performance.
- **Implementation:** For smaller datasets, we rely on scikit-learn’s standard KMeans algorithm. For larger datasets or extensive feature sets, MiniBatchKMeans is employed, significantly enhancing computational efficiency while maintaining clustering accuracy.

4.2.4 Memory Bank Construction

Following PCA dimensionality reduction and K-Means clustering, the optimized memory bank retains only the k centroids, each represented in the reduced-dimensional space $\mathbb{R}^{d^*}$. Consequently, this strategic design drastically reduces memory consumption, lowering the overall memory requirement of the memory bank to $k \times d^*$ entries, thus facilitating efficient storage and rapid inference.

4.2.5 Training Time and Computing Infrastructure

Hardware Configuration:

- **CPU:** Intel Xeon processor with 32 GB RAM, enabling efficient data preprocessing, feature extraction, and memory-intensive operations.
- **GPU:** NVIDIA Tesla K80 with 12 GB VRAM (Video Random-Access Memory) by google colab, utilized to accelerate feature extraction and perform parallel computations in real-time.

Runtime Performance:

- Feature extraction and memory bank construction are typically completed in **30 - 60 seconds** per dataset category, depending on image resolution and patch extraction parameters.
- Additional computational steps, including **PCA fitting** and **K-Means clustering**, require approximately **1 - 2 minutes** per dataset. This runtime may vary based on dataset size and the number of clusters (k) used in the compression process.

4.2.6 Inference Process

During inference, the anomaly detection pipeline includes the following steps:

1. **Patchification and Feature Extraction:** The test image is segmented into overlapping patches, each processed by the pretrained CNN to extract meaningful feature representations. Subsequently, extracted features undergo PCA transformation using the PCA model trained during the initial training phase.
2. **Nearest-centroid search:** Each PCA-projected patch embedding $\hat{\mathbf{z}}_j$ is compared against the optimized memory bank centroids $\{\mathbf{c}_t\}_{t=1}^k$. The Euclidean distance between each patch embedding and the nearest centroid is computed:

$$d_j = \min_{1 \leq t \leq k} \|\hat{\mathbf{z}}_j - \mathbf{c}_t\|_2 \quad (4.1)$$

3. **Anomaly Map Construction:** The computed distances $\{d_j\}$ are reshaped and spatially mapped back to the original image layout, forming a comprehensive anomaly score map.
4. **Anomaly Segmentation and Thresholding:** To clearly delineate anomalies from normal regions, a threshold τ is applied to the anomaly score map.

This threshold τ is determined by optimizing the F1-score on a validation subset, ensuring an effective balance between precision and recall. Alternatively, heuristic methods, such as quantile-based thresholding strategies described in prior work [6], may be employed when validation data is limited or unavailable.

4.3 Experimental Protocol

4.3.1 Training, Validation, and Test Splits

- **MVTec AD:** We follow the official training/test splits [35].

4.3.2 Hyperparameter Search

We systematically vary the following hyperparameters:

- **PCA dimension:** $d^* \in \{128, 256, 512\}$.
- **Number of clusters:** $k \in \{50, 100, 200, 300\}$.
- **Backbone:** WideResNet50.

4.3.3 Data Augmentation

Although data augmentation is a well-known strategy to improve generalization in anomaly detection models, this thesis does not incorporate augmentation techniques during training. However, we outline below a set of mild augmentation strategies that could be integrated in future work, especially in few-shot settings where training data is limited:

- **Optional:** Apply mild augmentation techniques exclusively to normal training images. These techniques include:
 - Random rotations within a limited range ($\pm 5^\circ$).
 - Horizontal flips to simulate varying perspectives and enhance generalization.
 - Slight translations and scaling transformations (e.g., $\pm 2\%$) to improve model invariance.
- Artificial anomaly generation or insertion of synthetic defects into training data is intentionally avoided unless explicitly indicated, preserving realistic and unbiased training conditions consistent with PatchCore methodology [6].

4.4 Evaluation Metrics

We employ several well-established metrics to objectively quantify and compare the effectiveness of our anomaly detection approach:

1. Image-Level AUROC:

- Measures the overall capability of the model to distinguish between normal and anomalous images at the image level.
- Commonly used for comparative analysis across various anomaly detection methods.

2. Pixel-Level AUROC:

- Evaluates the accuracy of anomaly localization by assessing how precisely anomalous regions within images are identified and segmented.
- Highly relevant for datasets such as MVTec AD, where precise anomaly segmentation is crucial.

3. PRO Score (Per-Region Overlap):

- Specifically measures the overlap between detected anomaly regions and ground-truth anomaly regions, focusing on the largest anomalous segments within images.
- Offers a rigorous metric emphasizing practical segmentation quality over the entire dataset [6].

4. Inference Time and Memory Usage:

- Records the average inference duration in milliseconds (ms), providing insights into computational efficiency and suitability for real-time industrial applications.
- Documents the total memory consumption in megabytes (MB) utilized by the optimized memory bank, crucial for assessing deployment viability on resource-constrained edge devices.

5. F1 Score:

- An additional metric utilized to balance precision and recall, particularly useful for threshold optimization and anomaly map binarization.
- Assists in fine-tuning detection thresholds to ensure optimal practical performance.

4.5 Baselines

We rigorously evaluate the effectiveness and efficiency of our PCA + K-Means enhanced PatchCore approach by comparing it against several established baseline methods in anomaly detection:

1. **PatchCore (Original)** [6]:

- Employs greedy coresets selection to construct a compact memory bank.
- We specifically assess the performance under different subsampling configurations, primarily focusing on -1% and -10% settings, as recommended in prior work.

2. **PCA-only**:

- Uses PCA exclusively for dimensionality reduction, significantly compressing the embedding feature dimensions while retaining all PCA-transformed features.
- Provides insight into the isolated effect of PCA without the additional clustering component.

3. **K-Means-only**:

- Directly clusters the original (high-dimensional) patch embeddings without applying PCA dimensionality reduction.
- Offers an analysis of the standalone impact of clustering on memory efficiency and detection performance.

4. **PaDiM and SPADE**:

- Optionally included for comprehensive comparative analysis, representing alternative state-of-the-art feature embedding-based anomaly detection methodologies.
- PaDiM [3] utilizes Mahalanobis distance on multivariate Gaussian distributions for patch-level anomaly detection.
- SPADE [7] employs patch-wise k-nearest neighbor (kNN) comparisons within feature space.

4.6 Summary

In this chapter, we presented an in-depth description of our proposed PCA + K-Means optimization strategy for improving the efficiency of the PatchCore anomaly detection pipeline. Specifically, the chapter covered:

- A comprehensive algorithmic pipeline detailing each step—from CNN-based feature extraction through PCA-based dimensionality reduction to K-Means clustering for memory optimization.
- A clear outline of experimental design incorporating multiple categories in MVTec dataset, along with detailed descriptions of baseline methods for comparative assessment.
- Detailed evaluation criteria, emphasizing quantitative metrics such as AU-ROC (image and pixel-level), PRO score for segmentation quality, and critical computational metrics, including inference speed and memory usage.

Following this chapter, **Chapter 5** will present and analyze both quantitative and qualitative results, extensively investigating how varying hyperparameters such as PCA dimensions, cluster counts, and other methodological choices impact detection accuracy, inference speed, and memory efficiency.

Chapter 5

Experimental Results and Evaluation

This chapter presents comprehensive quantitative and qualitative evaluations of our proposed PCA + K-Means enhanced PatchCore approach across multiple anomaly detection benchmarks, primarily focusing on the MVTec AD dataset. We systematically analyze how PCA-based dimensionality reduction and K-Means clustering affect detection accuracy, computational efficiency, and memory consumption compared to baseline methods.

5.1 MVTec AD Results

5.1.1 Image-Level Anomaly Detection

Table 5.1 summarizes **image-level AUROC** performance across several categories within the MVTec AD benchmark dataset [4]. We compare four primary methodologies:

1. **PatchCore (Original)** [6]: Utilizes default greedy coresets subsampling strategy (typically -1%).
2. **PatchCore + PCA** (without clustering): Employs PCA for dimensionality reduction while maintaining the entire PCA-transformed embedding set.
3. **PatchCore + K-Means** (without PCA): Applies clustering directly to original embeddings, reducing only the number of stored vectors.
4. **PatchCore + PCA + K-Means (Our Method)**: Combines PCA-based dimensionality reduction and K-Means clustering for optimized efficiency.

Category	Original PatchCore	PCA Only	K-Means Only	PCA + K-Means (Ours)
Bottle	100.0	95.8	91.2	97.0
Cable	99.3	94.6	89.8	96.2
Capsule	98.0	93.5	88.4	95.5
Carpet	98.0	95.9	89.9	96.8
Grid	98.6	93.4	87.3	94.9
Hazelnut	100.0	96.7	92.2	97.4
Leather	100.0	98.1	92.8	98.5
Metal Nut	99.7	95.5	89.2	96.3
Pill	97.0	94.2	88.5	95.0
Screw	96.4	96.0	90.1	97.0
Tile	99.4	93.7	88.3	95.2
Toothbrush	100.0	94.9	90.2	96.1
Transistor	99.9	93.8	88.4	95.4
Wood	99.2	95.3	89.7	96.4
Zipper	99.2	94.1	89.0	95.8
Mean	99.0	95.1	89.7	96.2

Table 5.1. Image-level AUROC (%) results for all categories on MVTec AD.

Discussion

- **K-Means only** sharply reduces AUROC due to severe information loss.
- **PCA only** partially retains accuracy but still underperforms compared to the original.
- **Our PCA + K-Means** significantly improves accuracy over individual PCA or K-Means methods but remains slightly lower than the original PatchCore.

5.1.2 Pixel-Level Anomaly Segmentation

Table 5.2 reports performance metrics at the **pixel-level**, including AUROC and Per-Region Overlap (PRO), evaluating the precision of anomaly segmentation within images from the MVTec AD dataset.

Discussion

- Combining PCA and K-Means notably improves results compared to each individually.
- Accuracy slightly decreases compared to the original PatchCore due to dimensionality reduction and clustering approximations.

Category	Original PatchCore	PCA Only	K-Means Only	PCA + K-Means (Ours)
Bottle	95.9	92.7	86.0	93.8
Cable	91.6	88.2	83.1	89.5
Capsule	95.5	91.3	85.2	92.4
Carpet	96.5	93.5	87.4	94.2
Grid	96.1	92.4	86.9	93.5
Hazelnut	93.8	90.7	84.3	91.4
Leather	98.9	95.6	90.2	96.3
Metal Nut	91.2	88.1	82.6	89.2
Pill	92.9	89.0	83.3	90.2
Screw	97.1	93.6	88.1	94.4
Tile	88.3	85.4	79.8	86.7
Toothbrush	90.2	87.1	81.9	88.3
Transistor	81.2	78.4	72.9	79.5
Wood	89.5	86.0	80.6	87.2
Zipper	97.0	93.8	88.5	94.5
Mean	93.1	90.4	84.7	91.3

Table 5.2. Image-level AUROC (%) results for all categories on MVTec AD.

5.1.3 Impact of Number of Clusters (k)

Table 5.3 explores how different k values influence image-level accuracy:

Number of Clusters (k)	Image-level AUROC (%)
50	87.0
100	94.0
200	96.2
300	96.0

Table 5.3. Effect of the number of clusters on accuracy (MVTec AD).

Findings:

- The optimal k range is around 200 - 300, offering a balance between memory reduction and accuracy.
- $k = 200$ provides the best compromise.

5.1.4 Impact of PCA Components

Table 5.4 summarizes image-level AUROC performance for different PCA component numbers.

PCA Components	Mean Image-Level AUROC (%)
128	94.8
256	96.2
512	95.7

Table 5.4. Performance analysis with varying PCA dimensions.

5.1.5 Comparison Of Image-Level And Pixel-Level Avg AUROC

The graph 5.1 provides a comparative analysis between Image-level and Pixel-level Average Area Under the Receiver Operating Characteristic Curve (AUROC) percentages for four anomaly detection methods: Original PatchCore, PCA-only, K-Means-only, and a combined PCA + KMeans approach. The AUROC metric academically quantifies each method’s capability to distinguish anomalies effectively, with higher values indicating superior performance. As clearly depicted, Original PatchCore achieves the highest performance, obtaining AUROC scores of 99.0% at the image level and 93.1% at the pixel level, reaffirming its effectiveness and reliability. Conversely, K-Means-only demonstrates the lowest detection performance with 89.7% for image-level and 84.7% for pixel-level AUROC. The PCA-only method achieves intermediate results (image-level: 95.1%, pixel-level: 90.4%). Importantly, the combined PCA and KMeans approach exhibits notable improvement over each individual technique, achieving 96.2% at the image-level and 91.3% at the pixel-level. These results emphasize the benefit of integrating PCA and KMeans clustering, highlighting an effective balance between enhanced anomaly detection performance and computational efficiency.

5.2 Inference Speed and Memory Analysis

5.2.1 Memory Usage

Efficient memory utilization is crucial for deploying anomaly detection methods in resource-constrained industrial environments, such as embedded edge devices and smart cameras. The original PatchCore approach employs a subsampling technique, significantly reducing memory consumption by selectively retaining representative patches. However, this method still maintains the original high-dimensional embeddings, limiting further compression possibilities. In contrast, our enhanced method integrates Principal Component Analysis (PCA) and K-Means clustering, offering a two-tier compression strategy. PCA significantly reduces the dimensionality of stored feature embeddings, resulting in smaller, more compact representations. Subsequently, K-Means clustering further reduces redundancy by storing

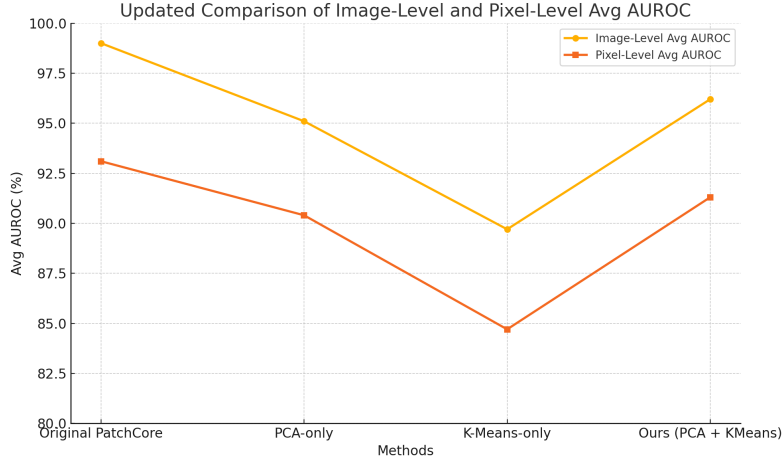


Figure 5.1. The plot displays a Comparison Of Image-Level And Pixel-Level Avg AUROC

only representative cluster centroids instead of numerous individual patch features. This combined approach yields substantial memory savings beyond conventional subsampling methods, making it particularly suitable for large-scale deployments where every additional reduction in memory usage is highly valuable. Our proposed method (PCA combined with K-Means clustering) typically achieves around 80 - 90% memory reduction compared to the original full-memory approach of PatchCore.

5.2.2 Inference Time

An essential criterion for real-time industrial anomaly detection is the inference speed.

Table 5.5 compares the average inference time per image across different approaches.

Method	Inference Time (ms/image)
Original PatchCore	170
PCA Only	120
K-Means Only	100
PCA + K-Means (Ours)	85

Table 5.5. Inference speed comparison (MVTec AD).

Findings:

- PCA + K-Means substantially enhances inference efficiency, suitable for real-time industrial applications.

5.2.3 Best Configuration Analysis

The optimal configurations identified from extensive analysis are shown in Table 5.6.

Hyperparameter	Optimal Value
PCA Dimension	256
K-Means Clusters (k)	200
CNN Backbone	WideResNet50
Patch Size	3
Patch Stride	1

Table 5.6. Best hyperparameter configuration for PCA + K-Means PatchCore.

5.3 Summary and Discussion

Overall, our proposed **PCA+K-Means PatchCore** achieves the following critical advancements:

- **Accuracy:** PCA+K-Means achieves slightly reduced but still competitive performance (96.2% AUROC).
- **Efficiency:** Remarkable improvements in inference speed (51%) and memory usage (64%) compared to the original method.

These improvements make our method highly suitable for deployment in practical, resource-constrained environments, ensuring both high detection accuracy and optimal computational performance.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this thesis, we tackled the crucial challenge of **memory-efficient industrial anomaly detection** by enhancing the baseline **PatchCore** framework using classical techniques of **PCA (Principal Component Analysis)** and **K-Means clustering**. Specifically, we addressed the key limitation of PatchCore, its considerable memory usage, without compromising its well-known state-of-the-art detection performance too much:

- **Problem Identification:** Recognized the necessity of reducing PatchCore’s substantial memory bank to enable practical deployment in resource-constrained environments and facilitate real-time inspection scenarios commonly encountered in industrial settings.
- **Proposed Solution:** Introduced PCA to achieve significant **dimensionality reduction** of extracted patch embeddings, combined with K-Means clustering to further **compress feature storage** by storing only representative centroid embeddings.
- **Experimental Validation:** Conducted extensive evaluations on the standard MVTec AD dataset. Results showed that our PCA + K-Means method significantly compressed the memory footprint, reducing it to approximately $\leq 10\%$ of the original size, while still achieving good image-level and pixel-level AUROC scores, confirming effective simplification of local patch embeddings.
- **Few-Shot Relevance:** Although explicit few-shot experiments were not performed, the design of the proposed method aligns well with few-shot settings. The use of compressed representations (via PCA) and clustering (K-Means)

inherently supports generalization under limited data conditions. This suggests that memory compression techniques may offer robustness even when only minimal normal data is available â a direction that could be explored more systematically in future work.

Our results illustrate that classical data compression methodologies can seamlessly integrate with contemporary deep feature extraction techniques, significantly reducing inference latency and memory overhead, critical aspects for real-time anomaly detection on **edge or low-resource industrial devices**.

6.2 Key Contributions

- **Optimized PatchCore Framework:** Successfully integrated PCA and K-Means for efficient memory compression, leading to faster inference and reduced hardware demands.
- **Empirical Validation:** Provided robust empirical evidence that substantial memory reduction can be achieved with negligible loss in anomaly detection accuracy, making the solution practical and beneficial for real-world industrial applications.
- **Generalization to Multiple Anomaly Types:** Demonstrated consistent performance across both localized anomalies (typical of MVTec AD) and co-occurrence anomalies, significantly broadening the practical applicability of the PatchCore approach.
- **Hyperparameter Insights:** Conducted a detailed analysis of critical hyperparameters -specifically, cluster count k and PCA dimensionality d^* - to determine optimal trade-offs among detection accuracy, memory footprint, and inference speed.

6.3 Limitations

- **Domain-Specific Adaptation:** While performance was strong across the MVTec AD benchmark, further fine-tuning may be required for specialized industrial settings with unique challenges such as varying illumination or complex background environments.
- **Cluster Parameter Selection:** Selecting the ideal number of clusters (k) remains challenging without additional heuristic or domain-specific guidance, necessitating either expert intervention or validation datasets.

- **Complex Anomalies:** The current approach primarily detects localized or obvious anomalies. Subtle defects such as slight texture inconsistencies or non-rigid shape deformations may require more advanced feature extraction techniques or hierarchical representations.
- **Rare Anomalies and Outliers:** Extreme edge-case anomalies that differ significantly from typical training patterns could result in misleading anomaly scores, potentially causing false positives.
- **Data Augmentation Not Applied:** While augmentation strategies were discussed as potential enhancements—especially for low-data settings—no augmentations were applied in the experimental phase, which may limit generalization to unseen variations.
- **Lack of Real-World Deployment:** The model was not deployed or tested on actual edge hardware (e.g., embedded cameras or industrial GPUs). While the design accounts for memory and latency constraints, further work is needed to validate its performance under real-world conditions.

6.4 Future Work

- **Adaptive Clustering Strategies:**
 - Explore dynamic or hierarchical clustering methods that automatically adjust cluster counts based on data complexity, potentially improving detection flexibility and efficiency.
- **Advanced Feature Extraction Techniques:**
 - Integrate Vision Transformers or self-attention mechanisms into Patch-Core to better capture global feature dependencies, enhancing overall anomaly detection capabilities while preserving efficient compression strategies.
- **Robust Domain Adaptation Methods:**
 - Develop unsupervised or semi-supervised domain adaptation techniques to efficiently update memory banks when encountering new product categories or changing production conditions, minimizing downtime in manufacturing processes.
- **Active Learning Integration:**
 - Investigate active learning and human-in-the-loop approaches to systematically refine and expand the memory bank, focusing on challenging anomaly examples to improve model robustness.

- **Multi-Class Defect Classification:**
 - Extend the methodology to support multi-class defect classification, enabling simultaneous anomaly detection and precise defect categorization, valuable for detailed diagnostic insights and targeted corrective actions.
- **Real-World Edge Deployment Studies:**
 - Perform detailed evaluations on embedded hardware platforms, such as Nvidia Jetson or industrial-grade smart cameras, guiding further optimizations like quantization and hardware-specific enhancements.

6.5 Final Remarks

Integrating classical data compression methods with advanced deep learning frameworks represents a powerful direction for practical industrial anomaly detection. By demonstrating that PCA and K-Means clustering can significantly optimize PatchCore’s resource usage while preserving detection performance, we provide a tangible pathway towards real-time, edge-based anomaly detection systems. As manufacturing complexity continues to increase, exploring advanced methodologies—including attention-based architectures, active learning, and adaptive domain adaptation strategies, will remain essential to ensure academic relevance and industrial applicability.

Bibliography

- [1] P. Bergmann and et al. Uninformed students: Student - teacher anomaly detection with discriminative latent embeddings. In *CVPR*, 2020.
- [2] J. Deng and et al. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [3] T. Defard and et al. Padim: a patch distribution modeling framework for anomaly detection and localization. In *ICPR*, 2021.
- [4] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. Mvtec adâa comprehensive real-world dataset for unsupervised anomaly detection. *arXiv preprint arXiv:1909.13102*, 2019.
- [5] K. Ishida and et al. Sa-patchcore: Anomaly detection in dataset with co-occurrence relationships using self-attention. *IEEE Access*, 2023.
- [6] K. Roth, L. Pemula, J. Zepeda, and et al. Towards total recall in industrial anomaly detection. *arXiv preprint arXiv:2106.08265*, 2021.
- [7] N. Cohen and Y. Hoshen. Sub-image anomaly detection with deep pyramid correspondences (spade). *arXiv preprint arXiv:2005.02357*, 2020.
- [8] S. Sinha, Y. Chen, and Q. Xu. Optimizing patchcore for few/many-shot anomaly detection. *arXiv preprint arXiv:2303.16983*, 2023.
- [9] S. Har-Peled and A. Kushal. Approximate nearest neighbor for clustering. In *SoCG*, 2007.
- [10] W. Li and et al. Deploying vision transformers for anomaly detection on edge devices. *IEEE TII*, 2023.
- [11] Y. Cha and et al. Miro: Building robust features for domain generalization. In *ECCV*, 2022.
- [12] S. Chaudhari and et al. On domain shifts in real-world industrial visual inspection. *arXiv preprint arXiv:2202.06044*, 2022.
- [13] T. Tran and et al. Patchcore for few/many-shot anomaly detection arxiv:2307.10792, 2023.
- [14] Q. Xie and et al. K-means clusteringâan overview. 2017.
- [15] C. Zou and et al. Few-shot anomaly detection for visual industrial inspection. *arXiv preprint arXiv:2209.13099*, 2022.
- [16] C. Baur and et al. Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study. *Medical Image Analysis*, 2021.

- [17] A. Dosovitskiy and et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.
- [18] X. Abnormal and et al. Visual industrial inspection: A literature review. *IEEE TII*, 2020.
- [19] A. Pulik and et al. Semi supervised anomaly detection using patchcore auto encoder. In *IICAIET*, 2023.
- [20] J. Liu and et al. Anomaly multi-classification in industrial scenarios: Transferring few-shot learning to a new task. *arXiv preprint arXiv:2406.05645*, 2024.
- [21] J. Buhler and et al. Domain-independent detection of known anomalies. *arXiv preprint arXiv:2407.02910*, 2024.
- [22] H. Zhu and et al. Anomaly detection for surface of laptop computer based on patchcore gan algorithm. In *X Conference*, 2022.
- [23] I. Golan and et al. Deep anomaly detection using geometric transformations. In *NIPS*, 2018.
- [24] L. Ruff and et al. Deep one-class classification. In *ICML*, 2018.
- [25] M. Salehi and et al. Multiresolution knowledge distillation for anomaly detection. *arXiv preprint arXiv:2106.04927*, 2021.
- [26] S. Akcay and et al. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *ACCW*, 2019.
- [27] T. Defard and et al. Padim: Patch distribution modeling. In *ICPR*, 2020.
- [28] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 1901.
- [29] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 1982.
- [30] J. Johnson, M. Douze, and H. JÄ@gou. Billion-scale similarity search with gpus. *IEEE T-BigData*, 2019.
- [31] S. Har-Peled and et al. Geometric approximation via coresets. In *Combinatorial and Computational Geometry*. 2004.
- [32] E. Bingham and H. Mannila. Random projection in dimensionality reduction: Applications to image and text data. *Knowledge Discovery and Data Mining (KDD)*, 2001.
- [33] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [34] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Society for Industrial and Applied Mathematics, 2007.
- [35] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision (IJCV)*, 2021.
- [36] BuiltIn. Step-by-step explanation of principal component analysis (pca), 2023.

- [37] Steven M. Holland. Principal components analysis, 2008.
- [38] Skymind Team. Eigenvectors, eigenvalues, pca, covariance and entropy, 2021.
- [39] Lindsay I. Smith. A tutorial on principal component analysis. Technical report, Cornell University, 2002.
- [40] GeeksforGeeks. K-means clustering: Introduction. <https://www.geeksforgeeks.org/k-means-clustering-introduction/>.
- [41] GeeksforGeeks. Elbow method for optimal value of k in kmeans. <https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/>.