Robust Patch Distribution Modeling for Sensory Anomaly Detection

Kamila Kare, Marc Swynghedauw

Robust Patch Distribution Modeling for Sensory Anomaly Detection

Kamila KARE
EXPLEO FRANCE
kamila.kare@expleogroup.com

Marc SWYNGHEDAUW
EXPLEO FRANCE
marc.swynghedauw@expleogroup.com

Abstract

This paper addresses the challenging issue of unsupervised sensory anomaly detection (AD) in noisy image datasets, which is highly relevant in real-world applications where training images are often noisy and exhibit variability. Current approaches for detecting anomalies in such scenarios are limited and do not account for the noisy nature of the data. To address this limitation, we propose a new method called RPaDiM (Robust Patch Distribution Modeling), which leverages a pretrained CNN and then filter out contaminated patches and computes robust parameters of multivariate Gaussian distributions to summarize the entire training set. Our experiments demonstrate that RPaDiM outperforms existing techniques such as PaDiM and Patchcore in noisy-free settings, and achieves comparable performance with the state-of-the-art SoftPatch method in noisy frameworks. Our results also show that RPaDiM is able to effectively localize anomalies of different sizes, shapes, and positions in the images, making it a promising solution for industrial applications.

1 Introduction

Given a dataset, an anomaly can be thought of as a point or a group of points, whose characteristics are so different from the other points that they must have come from a different data generating process. Therefore AD consists of finding interesting patterns that deviate from the expected behavior in a dataset.

To address AD in images, several frameworks have been considered, including Unsupervised Learning [8, 19, 11, 23] , Self-Supervised [12, 22], Supervised [4, 14, 25], Weakly Supervised [24, 6]. Among these, Unsupervised AD is the most commonly used technique as it identifies anomalies or outliers without any prior knowledge of what constitutes an anomaly. Since anomalous data is typically hard to come by or insufficient, and may exhibit unpredictable patterns, AD is commonly formulated as a one-class learning problem, that relies solely on normal data for training. In visual inspection, AD attempts to determine whether an image corresponds to a regular or an atypical instance, usually generating an anomaly score to guide the decision.

To improve the reliability of anomaly detection results, it can be beneficial to locate the defective areas, referred to as defects, within images at the pixel level, thus providing more precise and explainable results. This undertaking is commonly referred to as anomaly localization, or anomaly segmentation. However, achieving high-precision anomaly detection and localization without relying on annotated training data remains a formidable challenge.

In one-class learning settings [2, 20], where only normal images are available for training, the most common approach is to analyze the statistical properties of the extracted feature maps representing the normality over the entire training set and use them to look for patterns that do not fit with these properties. However, in industrial applications, the training images may be subject to various sources

Preprint. Under review.

of noise and heterogeneity. For example, images of manufactured products may contain surface defects, variations in lighting, or noise from the image capture process. In surveillance applications, images may be captured under challenging lighting conditions or in cluttered environments, which can affect the quality of the resulting images. These factors can result in noisy and non-homogeneous image datasets, which can pose significant challenges for AD algorithms. Despite the importance of this noisy framework, it has received very little attention, especially in industrial applications.

To address this problem, this paper presents a novel method called RPaDiM for unsupervised sensory anomaly detection in images. RPaDiM operates at the patch level and learns the properties of the extracted embeddings thanks to a pretrained convolutionnal neural networks and then filters the potentially contaminated noisy patches in order to compute a robust parameter of multivariate Gaussian distributions to obtain a summary of the entire training set.

Compared to existing techniques such as PaDiM [8] and PatchCore [19] in a non-noisy setting, RPaDiM outperforms them in terms of detection accuracy. Moreover, in a noisy framework, RPaDiM achieves comparable performance with the state-of-the-art SoftPatch [11].

The proposed RPaDiM method has several advantages over existing techniques. First, it is robust to noisy training data, making it more suitable for real-world scenarios. Second, it is computationally efficient and scalable, making it feasible for large-scale industrial applications.

The proposed RPaDiM method has the potential to have a significant impact on various industrial applications, including quality control, surveillance, and medical imaging.

Our major contributions are summarized below:

- We propose RPaDiM, a novel unsupervised sensory anomaly detection consisting of PaDiM, a denoising strategy and a robust parameter estimation for robust anomaly detection.
- We conduct experiments on two benchmark datasets, including MVTec Anomaly Detection (MVTecAD) [2], BTAD [15], to highlight the robustness of RPaDiM as compared to the current state of the art.

2 Related method

AD in images has been a widely studied problem in the field of machine learning, and several methods have been proposed to address this problem. In one-class learning settings, since the most common goal is to model the distribution of the normal data, a popular technique to do so is based on patch-level statistics.

2.1 Unsupervised anomaly detection in noisy-free images

2.1.1 Embeddings methods

Patch-level statistics-based methods such as SPADE [7], PaDiM [8], PatchCore [19] have been proposed to address the problem of unsupervised sensory anomaly detection in images using a pretrained CNN. An underlying concept of these systems is that an image can be considered abnormal if even a single patch within it is classified as abnormal. SPADE and PatchCore have much in common in that they both use a representative memory bank of nominal features extracted by a pretrained CNN. However PatchCore does not store the extracted feature maps but a transformation of them thanks to a local aggregation technique. In addition, the memory bank is subsampled during inference to ensure low inference cost at higher performance. The anomaly score is taken as the maximum distance between the test patch in the test patch collection and its respective nearest neighbor.

PaDiM models the distribution of the normal patches by computing, at the patch level, the parameters of a Gaussian distribution of the extracted patch embeddings. This makes PaDiM independent of the size of the training dataset at the inference stage, unlike PatchCore or SPADE. PaDiM then computes the anomaly score of an image thanks to the Mahalanobis distance of the patch embeddings with respect to the Gaussian distribution at the patch level.

While these methods have shown promising results in detecting anomalies in non-noisy settings, their performance degrades when the training data is noisy and not clean.

2.1.2 Autoencoder-based methods

Autoencoders and Variational Autoencoders (VAEs) are widely used in sensory anomaly detection. These techniques involve projecting images into a lower-dimensional latent space and reconstructing them directly by using the vectors of the latent space (simple Autoencoders) or by sampling according to the distribution of the latent space (VAEs). Studies using the simple Autoencoders approach include [1, 3, 9], while those using VAEs include [13, 21]. During training, the network learns a representation of normality in which a normal image has a reconstruction score close to zero. During inference, a poorly reconstructed image is labeled as abnormal, and this can be done at the patch level to localize anomalies. Despite their intuitive nature, the results of these methods have not always been satisfactory [8]. One of the issues with these approaches is that although they are trained only on normal images, they can still reconstruct abnormal images [17].

2.2 Learning with noisy data

Only a few studies have aimed at developping unsupervised anomaly detection techniques that can effectively handle noisy settings. Among them, [26, 16] focus on semantic anomaly detection, which aims to detect deviations in context, and are thus not directly related to our topic. The study by [11], on the other hand, is relevant to our topic because it proposes a more robust technique for unsupervised anomaly detection in noisy settings.

Softpatch[11] uses a soft thresholding approach based on a denoising strategy to adaptively select the most informative patches for modeling the normal data distribution. The training noise discriminator, which detects training contamination is based on a density-based method: LOF [5]. However, Softpatch is essentially based on the PatchCore model, which uses a bank of memory, and therefore it is not independent of the size of the training dataset at prediction stage and may suffers from high computational complexity and scalability for large-scale datasets.

In contrast to existing methods, the proposed RPaDiM method takes a novel approach by first filtering the contaminated noisy patches after embedding and then computing a robust parameter of multivariate Gaussian distributions to obtain a summary of the entire training set. This approach has several advantages over existing techniques, including improved robustness to noisy data, computational efficiency, and scalability, making it suitable for various industrial applications.

3 Background

3.1 Problem setting

We consider a one-class learning setting, where both normal and potentially noisy normal images are available for training. The goal is to learn a model of normality and use it to detect deviations from that model that correspond to anomalous events. Specifically, we assume that we have a training set of $N = N_1 + N_2$ images $\{\boldsymbol{x}_i\}_{i=1}^{N}$ where N_1 are drawn i.i.d from an unknown distribution $p(\boldsymbol{x})$ and N_2 from $p(\boldsymbol{x}) + \epsilon(\boldsymbol{x})$, where $\boldsymbol{x}$ is an image in $\mathbb{R}^{C\times H\times W}$, H and W are the height and width of the image, and C is the number of channels and ϵ is a noise distribution.

In the test set, we denote the subset of anomalous images by $A = \{\boldsymbol{y}_i\}_{i=1}^{m}$, where m is the number of anomalous images. We assume that the anomalous images are drawn from a different and unknown distribution $q(.)$, which may be significantly different from the normal distribution $p(.)$. Our goal is to detect and localize the anomalous images in the test set without any prior knowledge or supervision about their specific nature or location.

To better understand our method, let us describe some concepts.

3.2 Geometric median

The geometric median is a robust estimator of the center of a dataset that is less sensitive to outliers and contamination than the arithmetic mean [18]. In the context of machine learning, the geometric median is often used as a robust estimator of the center of a distribution, especially when the data is contaminated with outliers.
It generalizes the classical median to multidimensional spaces. Formally, given N points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$ with weights $w_1, w_2, \ldots, w_n \geq 0$, the geometric median is defined as

$$\boldsymbol{x}_{\text{GM}} = \underset{\boldsymbol{z} \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} w_i \|\boldsymbol{z} - \boldsymbol{x}_i\|_2.$$

Figure 1 illustrates the robustness of geometric median to the mean. This median can be used in a

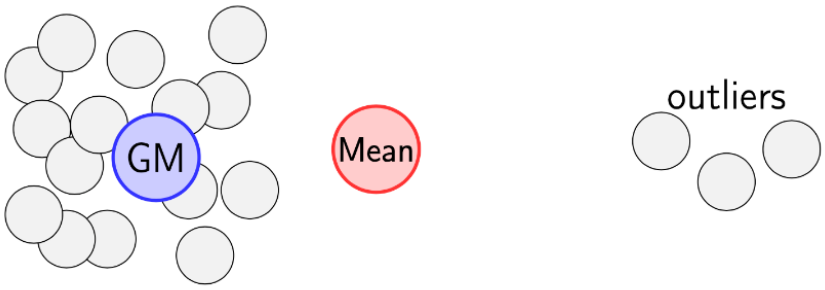

Figure 1: Geometrical median illustration.

Mahalanobis distance.

3.3 Mahanalobis distance

Given a probability distribution $\boldsymbol{P}$ on $\mathbb{R}^n$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the Mahanalobis distance between $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{P}$ is a reweighted Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$ defined as

$$d_M(\boldsymbol{x}, \boldsymbol{P}) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}. \tag{1}$$

This was used in the original PaDiM to measure an abnormality score over patches. In a noisy framework, computing a distance to the mean that is not robust can be misleading. In our work, the geometric median $\boldsymbol{x}_{\text{GM}}$ is the basis of the Mahalanobis distance and (1) yields to

$$d_M(\boldsymbol{x}, \boldsymbol{P}) = \sqrt{(\boldsymbol{x} - \boldsymbol{x}_{\text{GM}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{x}_{\text{GM}})}. \tag{2}$$

3.4 Local outlier factor

It is an outlier detection method that can be applied when some regions in a dataset have different densities. In such a dataset, distance and density-based outlier detection may miss outliers. The key idea of LOF is to compare the local density of a point with the average of the local density of its neighbors, and points with relatively low density will be spotted as outliers.

The LOF of a patch (h, w) is defined as:

$$\text{LOF}(h,w) = \frac{1}{\left|\mathcal{N}_k\big(\phi(h,w)\big)\right|} \sum_{(h_p,w_p)\in\mathcal{N}_k\left(\phi(h,w)\right)} \frac{\text{lrd}(h_p,w_p)}{\text{lrd}(h,w)} \quad \text{where}$$

$$\text{lrd}(h,w) = \left(\frac{1}{\left|\mathcal{N}_k\big(\phi(h,w)\big)\right|} \sum_{(h_p,w_p)\in\mathcal{N}_k\left(\phi(h,w)\right)} \text{ReachDist}\big(\phi(h,w) \leftarrow \phi(h_p,w_p)\big) \right)^{-1}$$

$$\text{ReachDist}\big(\phi(h,w) \leftarrow \phi(h_p,w_p)\big) = \max\left(d_k(\phi(h_p,w_p)), d(\phi(h,w),\phi(h_p,w_p))\right)$$

where $d(a,b)$ is the classical Euclidean distance between a and b and $d_k(\phi(h_p,w_p))$ is the core size of kth-neighbor.

The LOF of a point can be interpreted as the ratio of the total distance of its k neighbors to the average of the total distances of these k neighbors from their k neighbors in turn. In simple words, the LOF tells you about how much influence or dense neighborhood a data point has, as compared to other data points in the same dataset.

3.5 Locally aware patch features

Locally aware patch features are an image representation technique that can capture local information and have the potential to improve the performance of anomaly detection methods [19]. Let ϕ be a CNN, e.g. Resnet18 [10] and ϕ_j be a mid-level feature map. Given a location (h, w), we define its neighborhood as

$$\mathcal{N}_p^{(h,w)} = \Big\{(a,b)\big|a \in \big[h - \lfloor p/2 \rfloor, \ldots, h + \lfloor p/2 \rfloor\big], b \in \big[w - \lfloor p/2 \rfloor, \ldots, w + \lfloor p/2 \rfloor\big]\Big\},$$

and its local aware features at the same location as

$$\phi_{i,j} := \phi_j\big(\boldsymbol{x}_i,\ \mathcal{N}_p^{(h,w)}\big) = f_{\text{agg}}\Big(\phi_j\big(\boldsymbol{x}_i; a, b\big)\big|(a,b) \in \mathcal{N}_p^{(h,w)}\Big) \tag{3}$$

where p is the neighborhood size. As in PatchCore, we use f_{agg} the adaptive average pooling as the aggregation function.

We are now ready to present our method.

4 Proposed method

4.1 Overview

Inspired by PaDiM and SoftPatch, this work aims at learning robust representative features from normal images.

Our model can been described in three main steps.

- **Embedding extraction**: when a pretrained network is applied to an image, it uses its learned weights to extract high-level features and patterns from the image. The extracted features can be used as input for downstream tasks such as anomaly detection.
More formally, given an image $\boldsymbol{x}_i$ from the training set and a CNN ϕ, applying ϕ to $\boldsymbol{x}_i$ gives $\phi_j\big(\boldsymbol{x}_i;\ h, w\big) \in \mathbb{R}^{c_j}$ for $(h, w) \in [1, H] \times [1, W]$ which by virtue of (3) will result in $\phi_{i,j} \in \mathbb{R}^{c_j, H_j, W_j}\ j = 2, 3$. Lastly, $\phi_{i,2}$ and $\phi_{i,3}$ will be concatenated to give $\phi_i \in \mathbb{R}^{c^*, H^*, W^*}$ to form the collection $\{\phi_i\}_{i=1}^N$ over the training set. It is also important to remember that unlike PaDiM where all 3 intermediate layers of the backbone followed by a dimension reduction, we only consider the last 2 intermediate layers in order to leverage the training context and avoid relying on features that are either too generic or too heavily biased towards ImageNet classification [19].

- **Denoising strategy**: In this step, the noisy patches are determined by computing the LOF score for each patch and removing (or setting to zero) the patches with the top τ percent scores as in [11]. Thus, the embedding patches are $\{\tilde{\phi}_i\}_{i=1}^N$

- **Parameters distribution Estimation**: The geometric median $\boldsymbol{x}_{\text{GM}}$ and covariance matrix $\boldsymbol{\Sigma}$ are computed after flattening the last two dimensions of $\{\tilde{\phi}_i\}_{i=1}^N$ into one, i.e $H^* \times W^*$. Therefore, $\boldsymbol{x}_{\text{GM}} \in \mathbb{R}^{c^*, H^* \times W^*}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{c^*, c^*, H^* \times W^*}$. Let us notice that a regularization term $\xi\, I$ where I is the identity matrix is added to $\boldsymbol{\Sigma}$ to ensures its invertibility. Figure 2 depicts an overview of the proposed framework.

4.2 Anomaly detection based on RPaDiM

Given a test image $\boldsymbol{x}$, the embedding extracted denoted $\phi \in \mathbb{R}^{c^*, H^* \times W^*}$, so that the formula (2) can be used to compute the anomaly score at the patch level. So we have a matrix of anomaly score $\boldsymbol{M} \in \mathbb{R}^{H^*, W^*}$. Anomalous areas in the image are identified by high scores in the anomaly map $\boldsymbol{M}$. The final anomaly score of the entire image is determined by selecting the maximum score from the anomaly map.

5 Results and discussions

5.1 Experimental details

- **Datasets** : Our experiments focused primarily on two benchmark datasets: MVTecAD [2] and BTAD [15]. The MVTecAD dataset consists of 15 categories with a total of 3629 training images

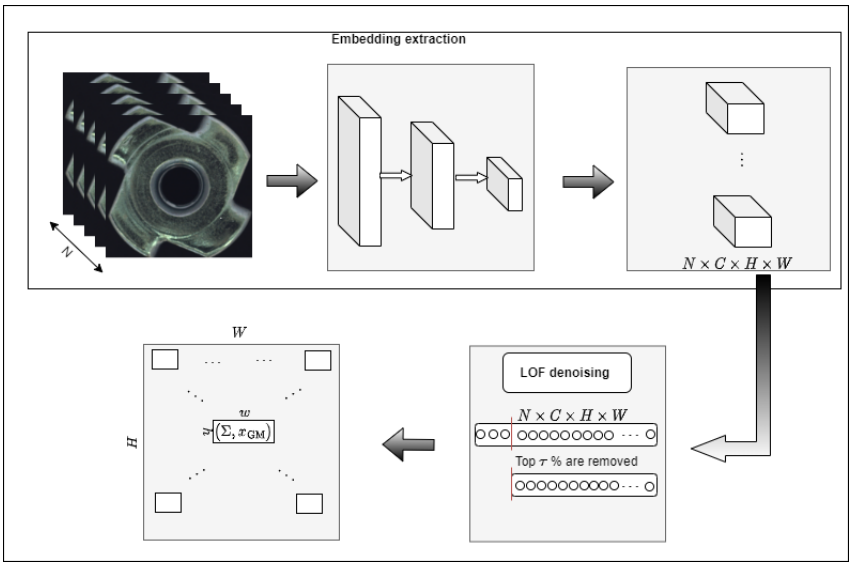

Figure 2: Our framework will learn the geometric median and covariance at each patch (h, w) after a denoising strategy that filters out contaminated patches.

and 1725 test images, while the BTAD dataset has three categories with 1799 images. This provides a comprehensive challenge due to the different classes of industrial production, including objects, textures, and different types of rotations. To test the robustness of our method in the noisy case, we follow the same approach as [11] since MVTecAD essentially contains clean, homogeneous and noiseless images. This approach consists of creating a noisy training set by randomly sampling anomalous images from the test set and mixing them with the existing training images, while maintaining the original number of normal samples in the training set. To avoid training and testing on the same dataset, the injected anomalous samples are not evaluated. This approach is more representative of real-world applications where we very often find heterogeneous images. We did not introduce any artificial noise into the BTAD dataset, as [11] observed that the training samples already contain some noise (typically small scratches) upon visual inspection, which is consistent with the characteristics of our problem setting and further emphasizes the relevance of our proposed approach.

- **Evaluation metrics**: Our evaluation of the proposed approach uses both image-level and pixel-level AUROC as the metrics of choice for each category in both the MVTecAD and BTAD datasets, which we then average to derive the average image/pixel level AUROC.

- **Implementation details**: We compare the performance of RPaDiM with three state-of-the-art (SOTA) methods for unsupervised sensory anomaly detection: PaDiM, PatchCore, and SoftPatch. On the one hand, all four methods use ResNet18 as the backbone for embedding extraction to ensure a perfect comparison, as performance can vary between backbones. On the other hand, since the BTAD dataset is more challenging specially the category BTAD-02, we used WideResnet50 as backbone for all four methods. We evaluate the methods on both noise-free and noisy scenarios. For the noisy setting, we set a seed to ensure that the same anomalous images are injected into the training set for all methods. In our experiments, we have kept the same hyperparameters used in the literature for preprocessing: thus we use 256×256 resolution for MVTecAD images and center-crop them to 224×224, followed by normalization. For BTAD, we use a resolution of 512×512. Note that a separate model is train for each class in MVTecAD and BTAD. Regarding the LOF hyper parameters namely K and τ, please refer to the sub-section 5.4 for more details.

All experiments are performed on a machine with an NVIDIA A100 20GB GPU and 35GB RAM.

5.2 Analysis results

Noisy-free MVTecAD: The Table 1 shows us a clear improvement of PaDiM and that we can reach or even exceed the performances of greedy methods with memory bank (PatchCore, SoftPatch) with

Table 1: Comparison of RPaDiM with the SOTA methods for both pixel-level and image-level anomaly detection performance on MVTecAD dataset. The results are reported with AUROC.

	Noisy-free				Noisy			
Category	PaDiM	PatchCore	SoftPatch	RPaDiM	PaDiM	PatchCore	SoftPatch	RPaDiM
bottle	0.989	0.989	**0.991**	0.989	0.989	0.990	**0.992**	0.989
cable	0.902	**0.972**	0.956	0.962	0.852	0.854	**0.947**	0.932
capsule	0.926	**0.968**	0.957	0.952	0.896	**0.967**	0.958	0.946
carpet	0.986	0.986	0.985	**0.993**	0.982	0.986	0.985	**0.992**
grid	0.889	0.917	0.938	**0.959**	0.849	0.920	0.948	**0.960**
hazelnut	0.940	**0.989**	0.987	0.973	0.864	**0.990**	0.987	0.970
leather	0.946	**0.996**	0.994	0.993	0.975	**0.997**	**0.997**	0.995
metalnut	0.971	**0.983**	0.979	0.980	0.945	0.896	**0.979**	0.961
pill	0.869	0.905	0.918	**0.946**	0.858	0.881	0.901	**0.953**
screw	0.858	**0.990**	0.968	0.951	0.807	**0.992**	0.970	0.952
tile	0.936	0.933	0.924	**0.953**	0.845	0.907	0.926	**0.946**
toothbrush	0.967	**0.993**	0.961	0.991	0.974	**0.994**	0.969	0.993
transistor	0.947	0.960	0.925	**0.990**	0.896	0.944	0.920	**0.984**
wood	0.960	0.956	**0.962**	0.957	0.903	0.955	**0.960**	0.953
zipper	0.859	**0.977**	0.959	0.959	0.794	**0.973**	0.962	0.952
Average	0.930	0.967	0.960	**0.970**	0.895	0.950	0.960	**0.965**

Table 2: Comparison of RPaDiM with the SOTA methods for both pixel-level and image-level anomaly detection performance on BTAD dataset. The results are reported with AUROC.

Category	PaDiM	PatchCore	SoftPatch	RPaDiM
01	**1.000**	**1.000**	0.999	0.968
02	0.871	0.871	**0.934**	0.911
03	0.971	**0.999**	0.997	0.996
Average	0.947	0.957	**0.977**	0.958

a simple method like RPaDiM which keeps the computational efficiency of PaDiM. Indeed, the fact of integrating to PaDiM, this notion of patch aggregation in a neighborhood allows it to better take into account the nomimal context and thus increase its performances. In addition, focusing on the last two intermediate layers of the Resnet backbone allows us to effectively leverage the training context and avoid relying on overly generic training.

Noisy MVTecAD: Not surprisingly, the Table 1 shows us a stability even after introducing 10% of abnormal images to corrupt the training set. This confirms the robustness of our model and of SoftPatch designed for this purpose. Moreover, this robustness is maintained for both detection and localization of anomalies. After filtering the patches that are quite isolated and therefore noisy, and then calculating a robust indicator here the geometric median allows us to cancel the effect of noise and to maintain the robustness of our method.

BTAD: We also compared RPaDiM with the 3 previously applied methods. Again, we see that here too, RPaDiM performs very well with a performance slightly below SoftPatch.

5.3 Qualitative results

In this sub-section, we present the anomaly segmentation results of our model for various classes, as shown in Figure 3 . To evaluate the model's performance, we employ the F1-score, which involves determining the threshold for each anomaly map, a common practice in previous work. Pixels with anomaly scores above the threshold are classified as anomalous and assigned a value of 1, resulting in a 0 and 1 mask that represents our predicted mask.

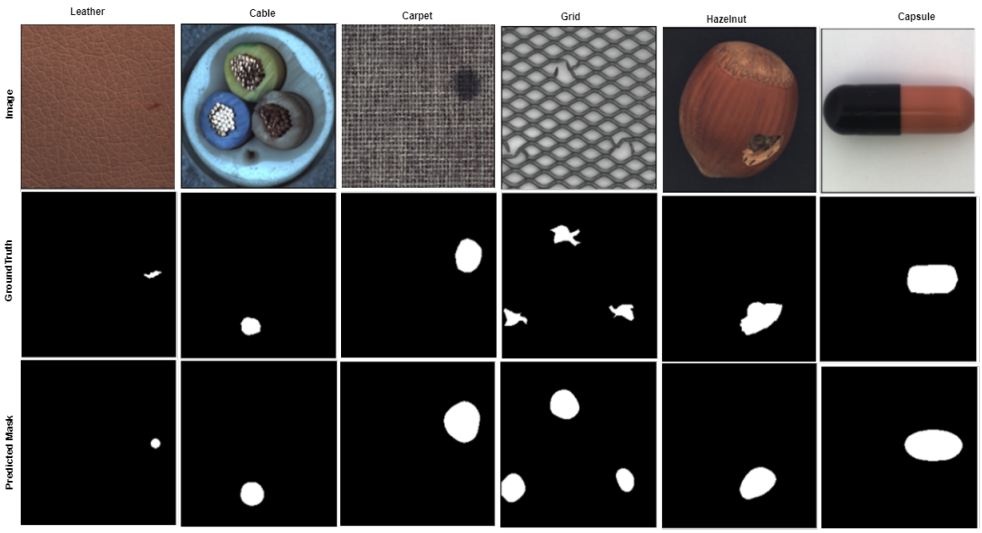

Figure 3: Examples of different anomaly localization for some classes in MVTecAD dataset. From top row to bottom row are Image, Ground Truth and the predicted mask with RPaDiM.

The Figure 3 demonstrates that our proposed method successfully localizes predicted defects, regardless of their size, shape, or position. Although some of the mask shapes appear slightly inaccurate, they still illustrate the effectiveness of our model in detecting small defects.

5.4 Ablation study

5.4.1 Effectiveness of RPaDiM

The validation of the effectiveness of RPaDiM is obvious from the Tables 1 and 2 in the sense that the added blocks that make up RPaDiM have a clear impact on the obtained performances. By removing both the locally aware patch block and the LOF block and by computing the arithmetic mean instead of the geometric median, we obtain PaDiM, and we have already seen that it suffers from robustness problems.

5.4.2 Hyper-parameters tuning

In order to implement both RPaDiM and SoftPatch, it will be necessary to first find the K and τ hyperparameters, which are essential for implementing the LOF, which is an essential component for both of these methods. Recall that PatchCore also needs a hyperparameter of its KNN algorithm, since during inference it calculates the anomaly score of a patch compared to its neighbors.

For RPaDiM, we have tested several values of K, and the results presented in the Analysis results subsection were obtained with $K = 1$ and $\tau = 0.85$. In general, the choice of K depends on the characteristics of the data and the problem to be solved, and it is recommended to experiment with different values of K to find the optimal one for a specific task.

Using K=1 in LOF means that the algorithm considers the nearest neighbor of each point as the only reference point for calculating the local density, which can be highly sensitive to noise and can lead to overfitting. However, using K=1 can be useful when the dataset has a high degree of local structure and the goal is to identify extremely isolated points as outliers. This is similar to the case of the screw category of MVTecAD (which is one of the worst performing categories in the literature), which includes misaligned images and gave us better results with $K = 1$. On the other hand, the other categories maintain their results for several values of K, as shown in the Table 3. Also, in the same Table, we have added 20% of anomalies instead of 10% to test the robustness.

Finally, it seems that the choice of K or τ has very little influence on the performance of RPaDiM, and by injecting 20% of anomalies in the training set, we maintain the performance obtained with 10%.

Table 3: The ablation study of K and τ. The performance scores are average of Image/pixel-level AUROC on MVTecAD

	K=3				K=6			
	$\tau = 0.85$		$\tau = 0.95$		$\tau = 0.85$		$\tau = 0.95$	
Category	Noisy-free	Noisy	Noisy-free	Noisy	Noisy-free	Noisy	Noisy-free	Noisy
bottle	0.989	0.989	0.989	0.99	0.988	0.988	0.988	0.989
cable	0.957	0.950	0.964	0.948	0.944	0.951	0.961	0.969
capsule	0.949	0.965	0.950	0.965	0.932	0.934	0.939	0.958
carpet	0.993	0.992	0.993	0.992	0.994	0.992	0.993	0.992
grid	0.959	0.956	0.965	0.960	0.961	0.968	0.965	0.955
hazelnut	0.971	0.962	0.977	0.945	0.957	0.948	0.978	0.937
leather	0.993	0.996	0.994	0.996	0.993	0.996	0.996	0.996
metalnut	0.981	0.955	0.981	0.954	0.981	0.981	0.981	0.957
pill	0.937	0.950	0.946	0.952	0.934	0.941	0.942	0.953
screw	0.881	0.918	0.915	0.939	0.951	0.948	0.896	0.911
tile	0.953	0.945	0.953	0.927	0.952	0.951	0.953	0.927
toothbrush	0.990	0.994	0.990	0.994	0.987	0.990	0.990	0.994
transistor	0.989	0.988	0.990	0.987	0.985	0.986	0.989	0.987
wood	0.956	0.961	0.957	0.961	0.955	0.956	0.956	0.962
zipper	0.947	0.952	0.953	0.955	0.942	0.943	0.947	0.953
Average	0.963	0.965	0.968	0.964	0.960	0.965	0.965	0.963

6 Conclusions

In this paper, we proposed a novel approach, called RPaDiM, for sensory unsupervised anomaly detection. Our method leverages the local context of image patches to capture meaningful representations. RPaDiM effectively addresses the impact of noisy data by filtering out contaminated patches and computing robust parameters for multivariate Gaussian distributions. We demonstrated the effectiveness of RPaDiM on two benchmark datasets, MVTecAD and BTAD, achieving state-of-the-art performance compared to other popular methods. In addition, we conducted extensive ablation studies and visualizations to gain insights into the behavior of our method and its ability to generalize to different anomaly types and noise levels. Our results show that RPaDiM is a powerful and flexible approach for unsupervised anomaly detection in images, with potential applications in various domains such as quality control and surveillance.

References

[1] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

[2] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020.

[3] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018.

[4] G. Bhattacharya, B. Mandal, and N. B. Puhan. Interleaved deep artifacts-aware attention mechanism for concrete structural defect classification. *IEEE Transactions on Image Processing*, 30:6957–6969, 2021.

[5] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[6] W.-H. Chu and K. M. Kitani. Neural batch sampling with reinforcement learning for semi-supervised anomaly detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 751–766. Springer, 2020.

[7] N. Cohen and Y. Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020.

[8] T. Defard, A. Setkov, A. Loesch, and R. Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*, pages 475–489. Springer, 2021.

[9] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] X. Jiang, J. Liu, J. Wang, Q. Nie, K. Wu, Y. Liu, C. Wang, and F. Zheng. Softpatch: Unsupervised anomaly detection with noisy data. *Advances in Neural Information Processing Systems*, 35:15433–15445, 2022.

[12] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021.

[13] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R. J. Radke, and O. Camps. Towards visually explaining variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8642–8651, 2020.

[14] X. Long, B. Fang, Y. Zhang, G. Luo, and F. Sun. Fabric defect detection using tactile information. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11169–11174. IEEE, 2021.

[15] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021.

[16] G. Pang, C. Yan, C. Shen, A. v. d. Hengel, and X. Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12173–12182, 2020.

[17] P. Perera, R. Nallapati, and B. Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2906, 2019.

[18] K. Pillutla, S. M. Kakade, and Z. Harchaoui. Robust Aggregation for Federated Learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

[19] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022.

[20] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[21] K. Sato, K. Hama, T. Matsubara, and K. Uehara. Predictable uncertainty-aware unsupervised deep anomaly segmentation. In *2019 international joint conference on neural networks (ijcnn)*, pages 1–7. IEEE, 2019.

[22] S. Sheynin, S. Benaim, and L. Wolf. A hierarchical transformation-discriminating generative model for few shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8495–8504, 2021.

[23] C.-C. Tsai, T.-H. Wu, and S.-H. Lai. Multi-scale patch-based representation learning for image anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3992–4000, 2022.

[24] S. Venkataramanan, K.-C. Peng, R. V. Singh, and A. Mahalanobis. Attention guided anomaly localization in images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII*, pages 485–503. Springer, 2020.

[25] Z. Zeng, B. Liu, J. Fu, and H. Chao. Reference-based defect detection network. *IEEE Transactions on Image Processing*, 30:6637–6647, 2021.

[26] C. Zhou and R. C. Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 665–674, 2017.