# MVTec Dataset

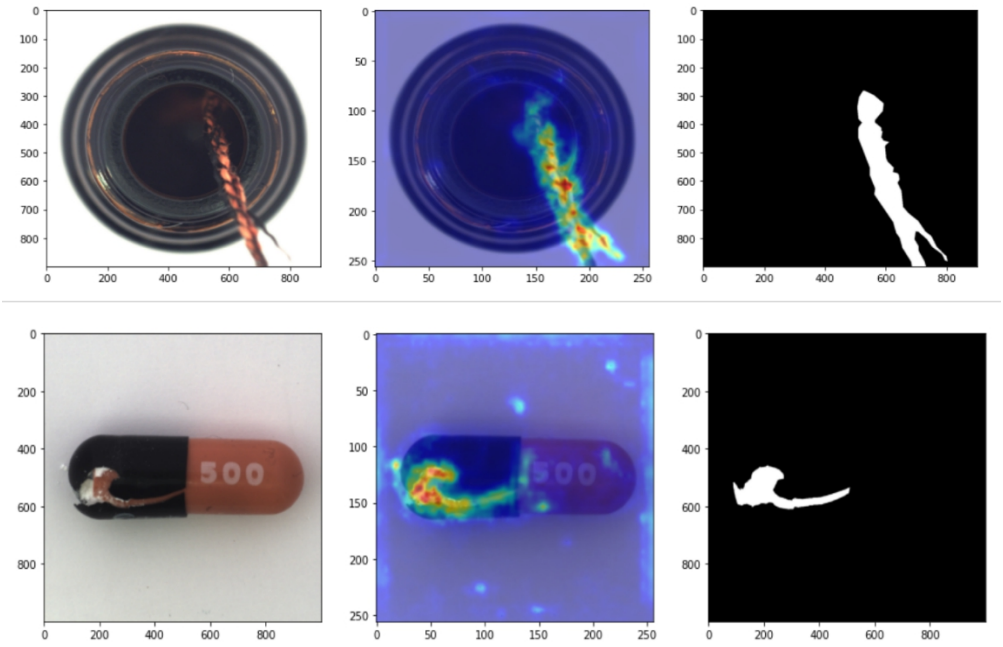

## Abstract

Anomaly detection is a difficult task that is frequently expressed as a one-class learning problem due to the unpredictable nature of anomalies. This study presents a straightforward yet effective solution to this problem, which is implemented in the student-teacher framework. For its benefits, but vastly improves it in terms of accuracy and efficiency. We distil the information from a powerful model that has been pre-trained on image classification as the instructor. To study the distribution, put all of your information into a single student network with the same architecture. To the greatest extent possible We also incorporate a multi-scale feature matching technique. This hierarchical feature matching allows the student network to get a mixture of multi-level knowledge from the feature pyramid, thus allowing to detect anomalies of various sizes.


## Introduction

Humans are very good at recognizing if an image is similar to what they have previously observed or if it is something novel or anomalous. So far, machine learning systems, however, seem to have difficulties with such tasks. There are many relevant applications that must rely on unsupervised algorithms that can detect anomalous regions. In the manufacturing industry, for example, optical inspection tasks often lack defective samples or it is unclear what kinds of defects may appear. In active learning systems, structures that are identified as anomalous might indicate the necessity of including a specific image for training. Therefore, it is not surprising that recently a significant amount of interest has been directed towards novelty detection in natural image data using modern machine learning architectures. A number of algorithms have been proposed that test whether a network is able to detect if new input data matches the distribution of the training data. Many of these algorithms, however, focus on classification settings in which the inlier and outlier distributions differ significantly. This is commonly known as outlier detection or one-class-classification. A common evaluation protocol is to arbitrarily label a number of classes from existing object classification datasets as outlier classes and use the remaining classes as inliers for training. It is then measured how well the trained algorithm can distinguish between previously unseen outlier and inlier samples. While this classification on an image level is important, it is unclear how current state-of-the-art methods perform on what we call anomaly detection tasks. The problem setting is to find novelties in images that are very close to the training data and differ only in subtle deviations in possibly very small, confined regions. Clearly, to develop machine learning models for such and other challenging scenarios we require suitable data. Curiously, there is a lack of comprehensive real-world datasets available for such scenarios.

In this paper, we propose a simple yet powerful approach to anomaly detection, which follows the student-teacher framework for the advantages but substantially extends it in terms of both accuracy and efficiency. Specifically, given a powerful network pre-trained on image classification as the teacher, we distill the knowledge into a single student network with the identical architecture. In this case, the student network learns the distribution of anomaly-free images by matching their features with the counterparts of the pre-trained network, and this one-step transfer preserves the crucial information as much as possible. Furthermore, to enhance the scale robustness, we embed multi-scale feature matching into the network, and this hierarchical feature matching strategy enables the student network to receive a mixture of multi-level knowledge from the feature pyramid under a stronger supervision and thus allows to detect anomalies of various sizes (see Figure 1 for visualization). The feature pyramids from the teacher and student networks are compared for prediction, where a larger difference indicates a higher probability of anomaly occurrence. Compared to the previous work, especially the preliminary student-teacher model, the benefits of our approach are two-fold. First, useful knowledge is well transferred from the pre-trained network to the student network within one-step distillation, as they share the same structure. Second, thanks to the hierarchical structure of the network, multi-scale anomaly detection is conveniently reached by the proposed feature pyramid matching scheme. Due to such strengths, our approach conducts accurate and fast pixel-level anomaly detection.



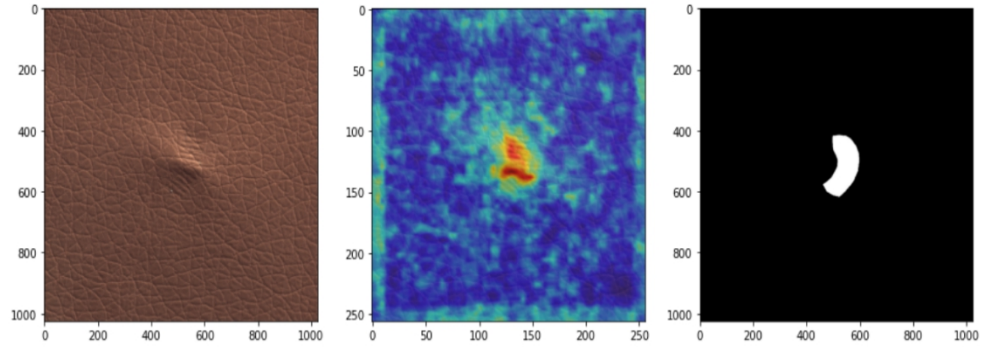


Figure 1: Visual results of our method on three defective images from the MVTec AD dataset. ResNet-18 is used as backbone. Columns from left to right correspond to input images with defects, anomaly map, and the ground truth respectively

# Related Work

## Image-level Visual Anomaly Detection Methods:

1. Density Estimation

It assumes that if the testing image or image feature does not meet the probability distribution model that is estimated with the normal image samples, it will be classified as anomaly.

Parametric methods:

1. Gaussian model
2. Gaussian mixture model

Nonparametric methods:

1. Nearest neighbor
2. Kernel density estimation

Disadvantages:

- ➢ Estimating a reasonable probability density requires a large number of training samples.
- ➢ When the feature dimension of the samples is very large (such as image data), this problem of the number of training samples becomes particularly prominent.
- ➢ The scalability of these classic models is poor.

2. One-class classification

Classifies a single class, concretely which attempts to construct decision boundary of the target class (normal images) in the feature space.

1. one-class support vector machines (OCSVM)

2. and support vector data description (SVDD)
3. one-class classification neural network (OCCNN).

Advantages:
- ➢ They do not require a large number of training samples.

Disadvantages:
- ➢ Still suffer from the problems of dimension disaster and scalability.

3. Image reconstruction:

It maps the image to a low-dimensional vector representation (latent space), and then try to find an inverse mapping or reconstruction for the original

Image. It assume that the reconstruction errors of the normal images are small, while that of abnormal images are larger.

Autoencoder
Generative adversarial network (GAN)
BiGAN
GAN disadvantages:
- ➢ This model needs to perform an iterative search process.
- ➢ Its efficiency is usually unsatisfactory in practice.

4. Self-supervised classification:

Self-supervised models can learn unique and more significant characteristics and features of normal samples. The representations learned for the target objects not only reflect their color, texture and other low-level features, but also reflect the high-level features such as the location, shape, position and direction. By learning these features only from normal samples, the model then can effectively discern abnormal samples without such characteristics.

RotateNet

## Pixel-level Visual Anomaly Detection Methods:

1. Image Reconstruction:

A typical method of image reconstruction is to compress and reconstruct the input image with the deep convolution autoencoder. It first learn to reconstruction of the normal images. Then, potential anomalies are detected by evaluating the pixel difference between the input image and the reconstructed image. It cannot reconstruct the abnormal image as well as the normal ones.

1. variational autoencoder (VAE)

2. generative adversarial network (GAN)
3. VAE-GAN

Disadvantages:

- ➢ It usually is expected to regenerate high-quality images for comparisons. However, high quality image generation itself is still a challenging task.
- ➢ The reconstruction approaches usually struggle to regenerate the sharp edges and complex texture structure for images.

2. Feature modeling:

It detects anomaly in feature space, after extracting features using Neural Networks. It models the feature distribution of normal images. For anomaly detection, if the regional feature corresponding to the local region of the test image deviate from the modeled feature distribution, this region will be labeled as abnormal.

1. sparse coding
2. Gaussian mixed model
3. Kmeans clustering

Disadvantages:

- ➢ It is very time-consuming during both training and testing, especially when the deep neural network is required to extract the deep image features.
- ➢ Because each local region of an image is evaluated independently, it may not be able to infer anomalies by leveraging the global context information of the image.
- ➢ the multi-scale strategy implicitly assumes that each scale is independent and totally ignores the relationship between different feature scales which is very important for making comprehensive detection decision.

For MVTec, many methods were evaluated based on image reconstruction and feature modeling, but found that the detection performances of these methods on different data categories are quite unstable. However, Bergmann proposed an unsupervised anomaly detection method with a student-teacher distillation framework, it leverage the transferred deep convolution features and detect the anomaly through feature regression, and achieved very promising results on MVTec AD dataset.

# Method

## Framework

The feature distribution of the normal training images is implicitly modelled using the student-teacher learning paradigm. The teacher is a powerful network that has been pre-trained on the image classification task (e.g., a ResNet-18 pre-trained on ImageNet). To prevent information loss, the student and the teacher use the same architecture. In essence, this is a situation of knowledge distillation based on features.

Here, we must consider a crucial factor: the distillation position. Deep neural networks (DNNs) create a feature pyramid for each input image. Higher-resolution results from the bottom layers Low-level information such as textures, edges, and colors are encoded via features. On the other hand, top Low-resolution features with context information are produced by layers. The characteristics that were created by Bottom layers are frequently general enough to be used in various vision tasks.

This encourages us to combine low- and high-level characteristics in a complementary manner. We choose different layers in deep neural networks because they correlate to different receptive fields. the features retrieved by a few successive bottom layer groups of the teacher to steer the student's learning (e.g., blocks in ResNet-18). Our ability to match features in a hierarchical order is enabled by this hierarchical feature matching and enables us to detect anomalies of different sizes.

## Training Process

The training phase aims to obtain a good student which can perfectly imitate the outputs of a fixed teacher on normal images. Formally, given a training dataset of anomaly-free images D = {**I1, I2, ..., In**}, our goal is to capture the normal data manifold by matching the features extracted by the L bottom layer groups of the teacher with the counterparts of the student.

For an input image **Ik**, the **l**th bottom layer group of the teacher and student outputs a feature map **Fl**

Since there is no prior knowledge regarding the appearances and locations of objects, we simply assume that all image regions are anomaly-free in the training set. We
define the loss at position (i; j) as l2-distance between the l2-normalized feature vectors,
namely,

$$\ell^l(\mathbf{I}_k)_{ij} = \frac{1}{2}\left\|\hat{F}_t^l(\mathbf{I}_k)_{ij} - \hat{F}_s^l(\mathbf{I}_k)_{ij}\right\|_{\ell_2}^2, \tag{1}$$

$$\hat{F}_t^l(\mathbf{I}_k)_{ij} = \frac{F_t^l(\mathbf{I}_k)_{ij}}{\|F_t^l(\mathbf{I}_k)_{ij}\|_{\ell_2}},\ \hat{F}_s^l(\mathbf{I}_k)_{ij} = \frac{F_s^l(\mathbf{I}_k)_{ij}}{\|F_s^l(\mathbf{I}_k)_{ij}\|_{\ell_2}}.$$

The loss for the entire image $\mathbf{I}_k$ is given as an average of the loss at each position,

$$\ell^l(\mathbf{I}_k) = \frac{1}{w_l h_l} \sum_{i=1}^{w_l} \sum_{j=1}^{h_l} \ell^l(\mathbf{I}_k)_{ij}, \tag{2}$$

and the total loss is the weighted average of the loss at different pyramid scales,

$$\ell(\mathbf{I}_k) = \sum_{l=1}^{L} \alpha_l \ell^l(\mathbf{I}_k), \quad \text{s.t. } \alpha_l \geq 0, \tag{3}$$

## Test Process

In the test phase, we aim to obtain an anomaly map $\Omega$ of size $w \times h$ regarding a test image $\mathbf{J} \in \mathbb{R}^{w \times h \times c}$. The score $\Omega_{ij} \in [0,1]$ indicates how much the pixel at position $(i,j)$ deviates from the training data manifold. We forward the test image $\mathbf{J}$ into the teacher and the student. Let $F_t^l(\mathbf{J})$ and $F_s^l(\mathbf{J})$ denote the feature maps generated by the $l$th bottom layer group of the teacher and the student, respectively. We can compute an anomaly map $\Omega^l(\mathbf{J})$ of size $w_l \times h_l$, whose element $\Omega_{ij}^l(\mathbf{J})$ is the loss (Eq. 1) at position $(i,j)$. The anomaly map $\Omega^l(\mathbf{J})$ is upsampled to size $w \times h$ by bilinear interpolation. The resulting anomaly map is defined as the element-wise product of $L$ equal-sized upsampled anomaly maps,

$$\Omega(\mathbf{J}) = \prod_{l=1}^{L} \text{Upsample}\, \Omega^l(\mathbf{J}). \tag{4}$$

A test image is designated as anomaly if any pixel in the image is anomalous. As a result, we simply choose the maximum value in the anomaly map, *i.e.*, $\max(\Omega(\mathbf{J}))$ as the anomaly score for the test image $\mathbf{J}$.

# Experiments

## Dataset

We conduct experiments on the MVTec Anomaly Detection (MVTec AD) dataset, with both the image-level and pixel-level anomaly detection tasks considered. The dataset is specifically created to benchmark algorithms for anomaly localization. It collects more than 5,000 high-resolution images of industrial products covering 15 different categories. For each category, the training set only includes defect-free images and the test set comprises both defect-free images and defective images of different types. The performance is measured by AUC-ROC metric.

## Implementation Details

For all the experiments, we choose the first three blocks (conv2_x, conv3_x, conv4_x) of ResNet-18 as the pyramid feature extractors for both the teacher and student networks. The parameters of the teacher network are copied from the ResNet-18 pre-trained on ImageNet, while those of the student network are initialized randomly. We train the network using stochastic gradient descent (SGD) with a learning rate of 0.4 for 100 epochs. The batch size is 32. All the images in the training and test sets are resized to 256×256. For each category, we use 80% of training images to build the student, keeping the remaining 20% for validation. We select the checkpoint with the lowest validation error (Eq. 1) to perform anomaly detection.

# Results:

| Category | Pixel-level auc-roc score | Image-level auc-roc score |
|---|---|---|
| bottle | 0.98 | 1.00 |
| cable | 0.93 | 0.88 |
| capsule | 0.96 | 0.94 |
| carpet | 0.98 | 0.97 |
| grid | 0.98 | 0.96 |
| hazelnut | 0.98 | 1.00 |
| leather | 0.99 | 1.00 |
| metal_nut | 0.94 | 1.00 |
| pill | 0.96 | 0.94 |
| screw | 0.97 | 0.91 |
| tile | 0.96 | 0.96 |
| toothbrush | 0.98 | 0.89 |
| transistor | 0.81 | 0.90 |
| wood | 0.95 | 1.00 |
| zipper | 0.97 | 0.95 |

# Conclusion

We introduce a new feature pyramid matching technique and include it into the framework for detecting student-teacher anomalies. We employ the multiple layers of features of a powerful network pre-trained on image classification as the teacher to guide a student network with the same structure to learn the distribution of anomaly-free images. Our approach is capable of detecting abnormalities of varied sizes due to the hierarchical feature matching.

Using only one forward pass. Experiments on the MVTec AD dataset have revealed that our solution outperforms the current state-of-the-art method.