PaDiM: a Patch Distribution Modeling Framework for Anomaly Detection and Localization

Thomas Defard, Aleksandr Setkov, Angelique Loesch, Romaric Audigier
Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
thomas.defard@imt-atlantique.net, {aleksandr.setkov, angelique.loesch, romaric.audigier}@cea.fr

***Abstract*—We present a new framework for Patch Distribution Modeling, PaDiM, to concurrently detect and localize anomalies in images in a one-class learning setting. PaDiM makes use of a pretrained convolutional neural network (CNN) for patch embedding, and of multivariate Gaussian distributions to get a probabilistic representation of the normal class. It also exploits correlations between the different semantic levels of CNN to better localize anomalies. PaDiM outperforms current state-of-the-art approaches for both anomaly detection and localization on the MVTec AD and STC datasets. To match real-world visual industrial inspection, we extend the evaluation protocol to assess performance of anomaly localization algorithms on non-aligned dataset. The state-of-the-art performance and low complexity of PaDiM make it a good candidate for many industrial applications.**

I. Introduction

Humans are able to detect heterogeneous or unexpected patterns in a set of homogeneous natural images. This task is known as anomaly or novelty detection and has a large number of applications, among which visual industrial inspections. However, anomalies are very rare events on manufacturing lines and cumbersome to detect manually. Therefore, anomaly detection automation would enable a constant quality control by avoiding reduced attention span and facilitating human operator work. In this paper, we focus on anomaly detection and, in particular, on anomaly localization, mainly in an industrial inspection context. In computer vision, anomaly detection consists in giving an anomaly score to images. Anomaly localization is a more complex task which assigns each pixel, or each patch of pixels, an anomaly score to output an anomaly map. Thus, anomaly localization yields more precise and interpretable results. Examples of anomaly maps produced by our method to localize anomalies in images from the MVTec Anomaly Detection (MVTec AD) dataset [1] are displayed in Figure 1.

Anomaly detection is a binary classification between the normal and the anomalous classes. However, it is not possible to train a model with full supervision for this task because we frequently lack anomalous examples, and, what is more, anomalies can have unexpected patterns. Hence, anomaly detection models are often estimated in a one-class learning setting, *i.e.*, when the training dataset contains only images from the normal class and anomalous examples are not available during the training. At test time, examples that differ from the normal training dataset are classified as anomalous.

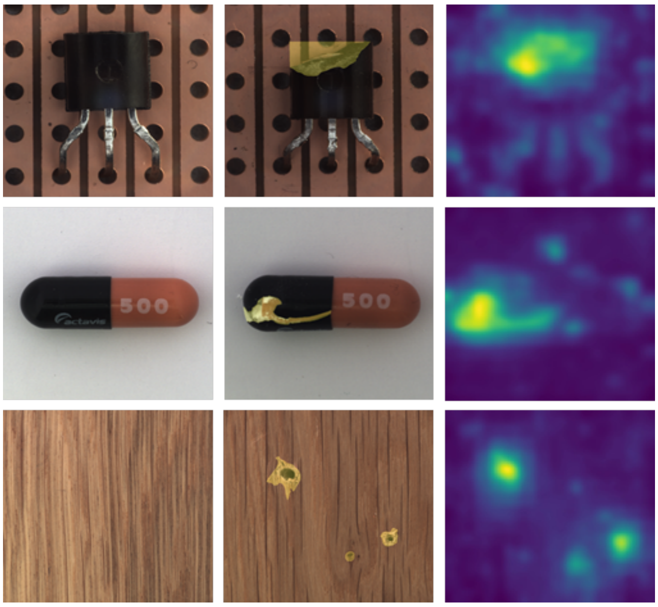


Fig. 1. Image samples from the MVTec AD [1]. *Left column*: normal images of Transistor, Capsule and Wood classes. *Middle column*: images of the same classes with the ground truth anomalies highlighted in yellow. *Right column*: anomaly heatmaps obtained by our PaDiM model. Yellow areas correspond to the detected anomalies, whereas the blue areas indicate the normality zones. Best viewed in color.

Recently, several methods have been proposed to combine anomaly localization and detection tasks in a one-class learning setting [2]–[5]. However, either they require deep neural network training [3], [6] which might be cumbersome, or they use a K-nearest-neighbor (K-NN) algorithm [7] on the entire training dataset at test time [4], [5]. The linear complexity of the KNN algorithm increases the time and space complexity as the size of the training dataset grows. These two scalability issues may hinder the deployment of anomaly localization algorithms in industrial context.

To mitigate the aforementioned issues, we propose a new anomaly detection and localization approach, named PaDiM for Patch Distribution Modeling. It makes use of a pretrained convolutional neural network (CNN) for embedding extraction and has the two following properties:

- Each patch position is described by a multivariate Gaussian distribution;
- PaDiM takes into account the correlations between dif-

PaDiM：一种用于异常检测与定位的补丁分布建模框架

托马斯·德法尔、亚历山大·塞特科夫、安吉莉克·勒施、罗马里克·奥迪吉耶
Université Paris-Saclay, CEA, List，法国帕莱索 F-91120，thomas.defard@imt-atlantique.net，{aleksandr.setkov、angelique.loesch、romaric.audigier}@cea.fr

Abstract—我们提出了一种新的补丁分布建模框架PaDiM，用于在单类学习设置中同时检测和定位图像中的异常。PaDiM利用预训练的卷积神经网络（CNN）进行补丁嵌入，并采用多元高斯分布来获得正常类别的概率表示。它还利用CNN不同语义层级之间的相关性，以更好地定位异常。在MVTec AD和STC数据集上，PaDiM在异常检测和定位方面均优于当前最先进的方法。为匹配现实世界中的视觉工业检测需求，我们扩展了评估协议，以评估异常定位算法在非对齐数据集上的性能。PaDiM的先进性能和低复杂度使其成为许多工业应用的理想选择。

一、引言

人类能够在一组同质的自然图像中检测出异质或意外的模式。这项任务被称为异常或新颖性检测，并拥有大量应用，其中视觉工业检测便是其中之一。然而，在生产线中异常事件极为罕见，且手动检测繁琐费力。因此，通过避免注意力下降并减轻操作员的工作负担，异常检测自动化将实现持续的质量控制。本文聚焦于异常检测，特别是异常定位，主要针对工业检测场景。在计算机视觉中，异常检测旨在为图像分配异常分数。异常定位则是一项更复杂的任务，它为每个像素或像素块分配异常分数，从而输出异常图。因此，异常定位能提供更精确且可解释的结果。图1展示了我们方法在MVTec异常检测（MVTec AD）数据集[1]图像中定位异常所生成的异常图示例。

异常检测是正常类与异常类之间的二元分类。然而，由于我们常常缺乏异常样本，且异常可能表现出不可预见的模式，因此无法通过完全监督的方式训练模型来完成此任务。因此，异常检测模型通常在单类学习设置中进行估计，{v*}，即训练数据集中仅包含正常类图像，且训练期间无法获得异常样本。在测试阶段，与正常训练数据集存在差异的样本会被归类为异常。

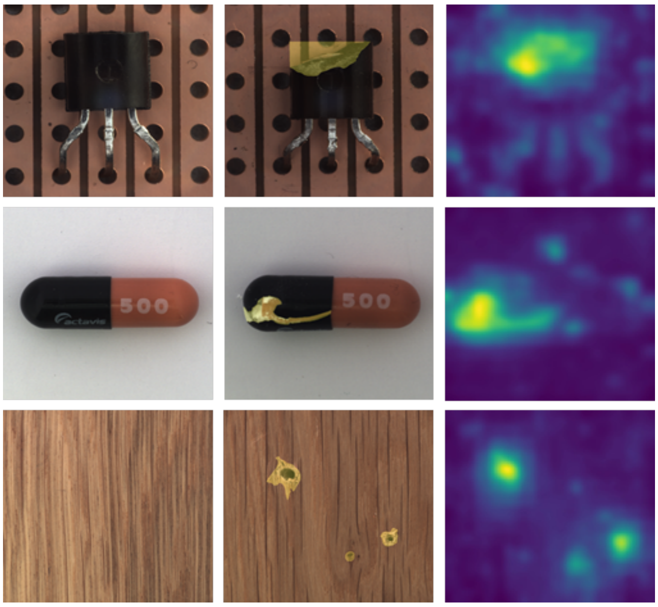


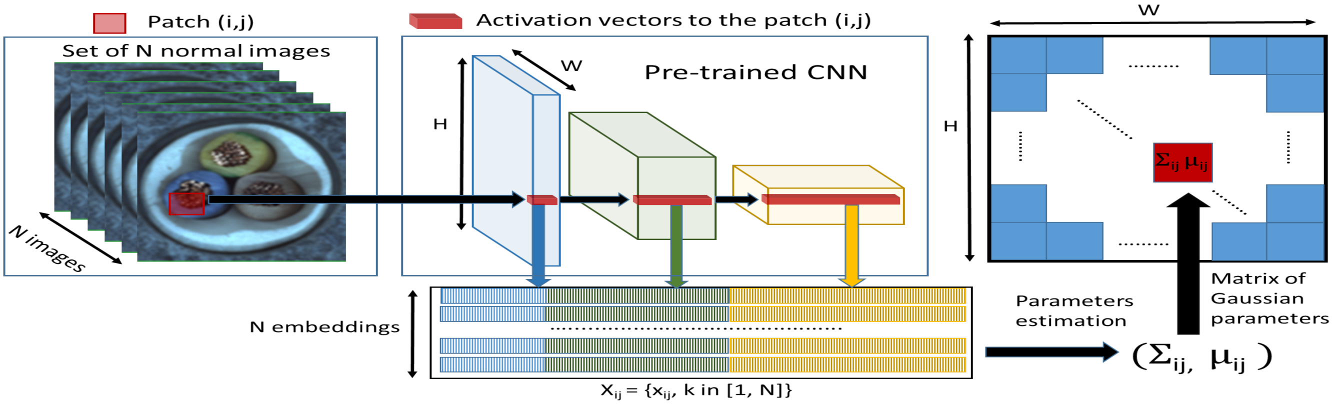
图1. MVTec AD [1] 数据集中的图像样本。*Left column*：晶体管、胶囊和木材类别的正常图像。*Middle column*：相同类别的图像，其中真实异常区域已用黄色高亮标出。*Right column*：由我们的PaDiM模型生成的异常热力图。黄色区域对应检测到的异常，而蓝色区域表示正常区域。建议彩色查看。

最近，已有多种方法被提出，用于在单类学习设置中将异常定位与检测任务相结合[2]–[5]。然而，这些方法要么需要深度神经网络训练[3]、[6]（这可能较为繁琐），要么在测试时对整个训练数据集使用K近邻（K-NN）算法[7][4]、[5]。KNN算法的线性复杂度会随着训练数据集规模的增大而增加时间和空间复杂度。这两类可扩展性问题可能阻碍异常定位算法在工业场景中的部署。

为了缓解上述问题，我们提出了一种新的异常检测与定位方法，命名为PaDiM（Patch Distribution Modeling）。该方法利用预训练的卷积神经网络（CNN）进行嵌入提取，并具备以下两个特性：

- 每个补丁位置由一个多元高斯分布描述；
- PaDiM考虑了不同特征之间的相关性，

Fig. 2. For each image patch corresponding to position (i, j) in the largest CNN feature map, PaDiM learns the Gaussian parameters (μ_{ij}, Σ_{ij}) from the set of N training embedding vectors $X_{ij} = \{x_{ij}^k, k \in [\![1, N]\!]\}$, computed from N different training images and three different pretrained CNN layers.



Patch (i,j)
Activation vectors to the patch (i,j)
Set of N normal images
N images
Pre-trained CNN
W
H
N embeddings
$X_{ij} = \{x_{ij}, k \text{ in } [1, N]\}$
W
H
$\Sigma_{ij}\ \mu_{ij}$
Parameters estimation
Matrix of Gaussian parameters
$(\Sigma_{ij},\ \mu_{ij})$

ferent semantic levels of a pretrained CNN.

With this new and efficient approach, PaDiM outperforms the existing state-of-the-art methods for anomaly localization and detection on the MVTec AD [1] and the ShanghaiTech Campus (STC) [8] datasets. Besides, at test time, it has a low time and space complexity, independent of the dataset training size which is an asset for industrial applications. We also extend the evaluation protocol to assess model performance in more realistic conditions, *i.e.*, on a non-aligned dataset.

II. Related work

Anomaly detection and localization methods can be categorized as either reconstruction-based or embedding similarity-based methods.

Reconstruction-based methods are widely-used for anomaly detection and localization. Neural network architectures like autoencoders (AE) [1], [9]–[11], variational autoencoders (VAE) [3], [12]–[14] or generative adversarial networks (GAN) [15]–[17] are trained to reconstruct normal training images only. Therefore, anomalous images can be spotted as they are not well reconstructed. At the image level, the simplest approach is to take the reconstructed error as an anomaly score [10] but additional information from the latent space [16], [18], intermediate activations [19] or a discriminator [17], [20] can help to better recognize anomalous images. Yet to localize anomalies, reconstruction-based methods can take the pixel-wise reconstruction error as the anomaly score [1] or the structural similarity [9]. Alternatively, the anomaly map can be a visual attention map generated from the latent space [3], [14]. Although reconstruction-based methods are very intuitive and interpretable, their performance is limited by the fact that AE can sometimes yield good reconstruction results for anomalous images too [21].

Embedding similarity-based methods use deep neural networks to extract meaningful vectors describing an entire image for anomaly detection [6], [22]–[24] or an image patch for anomaly localization [2], [4], [5], [25]. Still, embedding similarity-based methods that only perform anomaly detection give promising results but often lack interpretability as it is not possible to know which part of an anomalous images is responsible for a high anomaly score. The anomaly score is in this case the distance between embedding vectors of a test image and reference vectors representing normality from the training dataset. The normal reference can be the center of a n-sphere containing embeddings from normal images [4], [22], parameters of Gaussian distributions [23], [26] or the entire set of normal embedding vectors [5], [24]. The last option is used by SPADE [5] which has the best reported results for anomaly localization. However, it runs a K-NN algorithm on a set of normal embedding vectors at test time, so the inference complexity scales linearly to the dataset training size. This may hinder industrial deployment of the method.

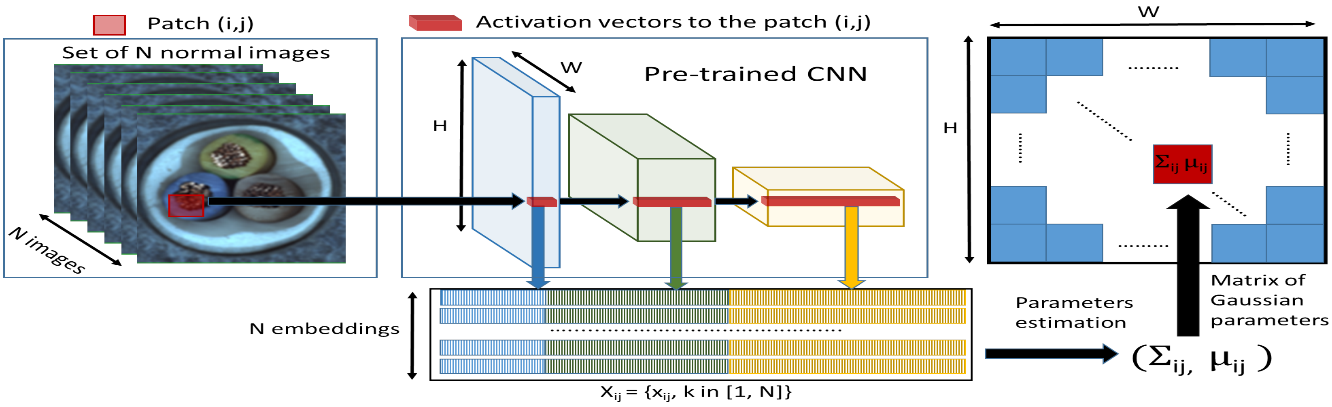
Our method, PaDiM, generates patch embeddings for anomaly localization, similar to the aforementioned approaches. However, the normal class in PaDiM is described through a set of Gaussian distributions that also model correlations between semantic levels of the used pretrained CNN model. Inspired by [5], [23], we choose as pretrained networks a ResNet [27], a Wide-ResNet [28] or an EfficientNet [29]. Thanks to this modelisation, PaDiM outperforms the current state-of-the-art methods. Moreover, its time complexity is low and independent of the training dataset size at the prediction stage.

III. Patch Distribution Modeling

A. *Embedding extraction*

Pretrained CNNs are able to output relevant features for anomaly detection [24]. Therefore, we choose to avoid ponderous neural network optimization by only using a pretrained CNN to generate patch embedding vectors. The patch embedding process in PaDiM is similar to one from SPADE [5] and illustrated in Figure 2. During the training phase, each patch of the normal images is associated to its spatially corresponding activation vectors in the pretrained CNN activation maps. Activation vectors from different layers are then concatenated to get embedding vectors carrying information from different semantic levels and resolutions, in order to encode fine-grained and global contexts. As activation maps have a lower

图 2。 对于最大CNN特征图中每个对应位置(i,j)的图像块，PaDiM从这些块中学习高斯参数(μ_{ij},Σ_{ij})。N 的集合训练嵌入向量 $X_{ij}=\{x_{ij}^{k}, k\in[[1,N]]\}$，由 N 张不同的训练图像和三个不同的预训练 CNN 层计算得出。


Patch (i,j)
Activation vectors to the patch (i,j)
Set of N normal images
N images
Pre-trained CNN
W
H
W
H
Σ_{ij} μ_{ij}
N embeddings
X_{ij} = {x_{ij}, k in [1, N]}
Parameters estimation
Matrix of Gaussian parameters
(Σ_{ij}, μ_{ij})

预训练CNN的不同语义层次。

通过这种新颖高效的方法，PaDiM在MVTec AD [1] 和上海科技大学校园（STC）[8] 数据集上的异常定位与检测任务中，超越了现有的先进方法。此外，在测试阶段，该方法具有较低的时间和空间复杂度，且与训练数据集规模无关，这一特性使其在工业应用中具备显著优势。我们同时扩展了评估协议，以在非对齐数据集上*i.e.*更真实地评估模型性能。

二、相关工作

异常检测与定位方法可分为基于重建的方法和基于嵌入相似度的方法。

基于重构的方法被广泛应用于异常检测与定位。仅使用正常训练图像来训练自编码器（AE）[1]、[9]–[11]、变分自编码器（VAE）[3]、[12]–[14]或生成对抗网络（GAN）[15]–[17]等神经网络架构。因此，异常图像因无法被良好重构而被识别。在图像层面，最简单的方法是将重构误差作为异常分数[10]，但来自潜在空间[16]、[18]、中间激活层[19]或判别器[17]、[20]的额外信息有助于更好地识别异常图像。然而，在定位异常时，基于重构的方法可将像素级重构误差作为异常分数[1]或结构相似度[9]。或者，异常图可以是基于潜在空间生成的视觉注意力图[3]、[14]。尽管基于重构的方法非常直观且可解释，但其性能受限于自编码器有时也能对异常图像产生良好的重构结果[21]。

基于嵌入相似性的方法使用深度神经网络提取描述整个图像的有意义向量以进行异常检测[6], [22]–[24]，或提取图像块以进行异常定位[2], [4], [5], [25]。然而，仅执行异常检测的基于嵌入相似性的方法虽能给出有希望的结果，但通常缺乏可解释性，因为其无法确定异常图像的哪一部分导致了高异常分数。在这种情况下，异常分数是测试图像的嵌入向量与代表训练数据集中正常性的参考向量之间的距离。正常参考可以是包含正常图像嵌入的n维球体中心[4]、[22]，高斯分布的参数[23]、[26]，或全部正常嵌入向量集合[5]、[24]。SPADE[5]采用了最后一种方案，并报告了最佳的异常定位结果。然而，该方法在测试时需对一组正常嵌入向量运行K-NN算法，因此推理复杂度与训练数据集规模呈线性增长。这可能阻碍该方法的工业部署。

我们的方法PaDiM为异常定位生成补丁嵌入，类似于上述方法。然而，PaDiM中的正常类别通过一组高斯分布来描述，这些分布还建模了所使用的预训练CNN模型语义级别之间的相关性。受[5]、[23]启发，我们选择ResNet[27]、Wide-ResNet[28]或EfficientNet[29]作为预训练网络。得益于这种建模方式，PaDiM超越了当前最先进的方法。此外，其时间复杂度较低，且在预测阶段与训练数据集大小无关。

三、补丁分布建模

A. Embedding extraction

预训练的CNN能够输出用于异常检测的相关特征[24]。因此，我们选择避免繁琐的神经网络优化，仅使用预训练的CNN生成图像块嵌入向量。PaDiM中的图像块嵌入过程与SPADE[5]的方法相似，如图2所示。在训练阶段，正常图像的每个图像块都与预训练CNN激活图中空间位置对应的激活向量相关联。随后将来自不同层的激活向量拼接起来，获得承载不同语义层次和分辨率信息的嵌入向量，从而编码细粒度与全局上下文信息。由于激活图具有较低

resolution than the input image, many pixels have the same embeddings and then form pixel patches with no overlap in the original image resolution. Hence, an input image can be divided in a grid of $(i, j) \in [1, W] \times [1, H]$ positions where WxH is the resolution of the largest activation map used to generate embeddings. Finally, each patch position (i, j) in this grid is associated to an embedding vector x_{ij} computed as described above.

The generated patch embedding vectors may carry redundant information, therefore we experimentally study the possibility to reduce their size (Section V-A). We noticed that randomly selecting few dimensions is more efficient that a classic Principal Component Analysis (PCA) algorithm [30]. This simple random dimensionality reduction significantly decreases the complexity of our model for both training and testing time while maintaining the state-of-the-art performance. Finally, patch embedding vectors from test images are used to output an anomaly map with the help of the learned parametric representation of the normal class described in the next subsection.

B. Learning of the normality

To learn the normal image characteristics at position (i, j), we first compute the set of patch embedding vectors at (i, j), $X_{ij} = \{x^k_{ij}, k \in [\![1, N]\!]\}$ from the N normal training images as shown on Figure 2. To sum up the information carried by this set we make the assumption that X_{ij} is generated by a multivariate Gaussian distribution $\mathcal{N}(\mu_{ij}, \boldsymbol{\Sigma_{ij}})$ where μ_{ij} is the sample mean of X_{ij} and the sample covariance Σ_{ij} is estimated as follows :

$$\Sigma_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} (\mathbf{x^k_{ij}} - \mu_{\mathbf{ij}})(\mathbf{x^k_{ij}} - \mu_{\mathbf{ij}})^{\mathrm{T}} + \boldsymbol{\epsilon} I \qquad (1)$$

where the regularisation term $\boldsymbol{\epsilon} I$ makes the sample covariance matrix Σ_{ij} full rank and invertible. Finally, each possible patch position is associated with a multivariate Gaussian distribution as shown in Figure 2 by the matrix of Gaussian parameters.

Our patch embedding vectors carry information from different semantic levels. Hence, each estimated multivariate Gaussian distribution $\mathcal{N}(\mu_{ij}, \boldsymbol{\Sigma_{ij}})$ captures information from different levels too and $\boldsymbol{\Sigma_{ij}}$ contains the inter-level correlations. We experimentally show (Section V-A) that modeling these relationships between the different semantic levels of the pretrained CNN helps to increase anomaly localization performance.

C. Inference : computation of the anomaly map

Inspired by [23], [26], we use the Mahalanobis distance [31] $M(x_{ij})$ to give an anomaly score to the patch in position (i, j) of a test image. $M(x_{ij})$ can be interpreted as the distance between the test patch embedding x_{ij} and learned distribution $\mathcal{N}(\mu_{ij}, \boldsymbol{\Sigma_{ij}})$, where $M(x_{ij})$ is computed as follows:

$$M(x_{ij}) = \sqrt{(x_{ij} - \mu_{ij})^T \Sigma^{-1}_{ij} (x_{ij} - \mu_{ij})} \qquad (2)$$

Hence, the matrix of Mahalanobis distances $M = (M(x_{ij}))_{1<i<W, 1<j<H}$ that forms an anomaly map can be computed. High scores in this map indicate the anomalous areas. The final anomaly score of the entire image is the maximum of anomaly map M. Finally, at test time, our method does not have the scalability issue of the K-NN based methods [4]–[6], [25] as we do not have to compute and sort a large amount of distance values to get the anomaly score of a patch.

IV. Experiments

A. Datasets and metrics

Metrics. To assess the localization performance we compute two threshold independent metrics. We use the Area Under the Receiver Operating Characteristic curve (AUROC) where the true positive rate is the percentage of pixels correctly classified as anomalous. Since the AUROC is biased in favor of large anomalies we also employ the per-region-overlap score (PRO-score) [2]. It consists in plotting, for each connected component, a curve of the mean values of the correctly classified pixel rates as a function of the false positive rate between 0 and 0.3. The PRO-score is the normalized integral of this curve. A high PRO-score means that both large and small anomalies are well-localized.

Datasets. We first evaluate our models on the MVTec AD [1] designed to test anomaly localization algorithms for industrial quality control and in a one-class learning setting. It contains 15 classes of approximately 240 images. The original image resolution is between 700x700 and 1024x1024. There are 10 object and 5 texture classes. Objects are always well-centered and aligned in the same way across the dataset as we can see in Figure 1 for classes Transistor and Capsule. In addition to the original dataset, to assess performance of anomaly localization models in a more realistic context, we create a modified version of the MVTec AD, referred as Rd-MVTec AD, where we apply random rotation (-10, +10) and random crop (from 256x256 to 224x224) to both the train and test sets. This modified version of the MVTec AD may better describe real use cases of anomaly localization for quality control where objects of interest are not always centered and aligned in the image.

For further evaluation, we also test PaDiM on the Shanghai Tech Campus (STC) Dataset [8] that simulates video surveillance from a static camera. It contains 274 515 training and 42 883 testing frames divided in 13 scenes. The original image resolution is 856x480. The training videos are composed of normal sequences and test videos have anomalies like the presence of vehicles in pedestrian areas or people fighting.

B. Experimental setups

We train PaDiM with different backbones, a ResNet18 (R18) [27], a Wide ResNet-50-2 (WR50) [28] and an EfficientNet-B5 [29], all pretrained on ImageNet [32]. Like in [5], patch embedding vectors are extracted from the first three layers when the backbone is a ResNet, in order to combine

由于分辨率高于输入图像，许多像素具有相同的嵌入，进而在原始图像分辨率中形成无重叠的像素块。因此，输入图像可被划分为一个由$(i,j)\in[1,W]\times[1,H]$个位置组成的网格，其中W xH是用于生成嵌入的最大激活图的分辨率。最终，该网格中的每个块位置(i,j)都与一个嵌入向量x_{ij}相关联，该向量按上述方式计算得出。

生成的补丁嵌入向量可能携带冗余信息，因此我们通过实验研究减小其尺寸的可能性（第V-A节）。我们发现，随机选择少量维度比经典的主成分分析（PCA）算法[30]更高效。这种简单的随机降维方法显著降低了我们模型在训练和测试阶段的复杂度，同时保持了最先进的性能。最后，测试图像的补丁嵌入向量借助下一小节描述的正常类参数化表示，用于输出异常图。

B. Learning of the normality

为了学习位置(i,j)的正常图像特征，我们首先从N张正常训练图像中计算(i,j)处的图像块嵌入向量集合$X_{ij}=\{x_{ij}^{k},k\in[[1,N]]\}$，如图2所示。为了总结该集合所承载的信息，我们假设X_{ij}由多元高斯分布$\mathcal{N}(\mu_{ij},\boldsymbol{\Sigma_{ij}})$生成，其中$\mu_{ij}$是$X_{ij}$的样本均值，样本协方差$\Sigma_{ij}$的估算方式如下：

$$\Sigma_{ij}=\frac{1}{N-1}\sum_{k=1}^{N}(\mathbf{x_{ij}^{k}}-\mu_{\mathbf{ij}})(\mathbf{x_{ij}^{k}}-\mu_{\mathbf{ij}})^{\mathrm{T}}+\boldsymbol{\epsilon}I \qquad (1)$$

其中正则化项$\boldsymbol{\epsilon}I$使样本协方差矩阵Σ_{ij}满秩且可逆。最终，每个可能的图像块位置都关联一个多元高斯分布，如图2中高斯参数矩阵所示。

我们的补丁嵌入向量携带来自不同语义层次的信息。因此，每个估计的多元高斯分布 {v*} 也捕获了来自不同层次的信息，而 {v*} 包含了层次间的相关性。我们通过实验证明（第 V-A 节），对预训练 CNN 不同语义层次间的这些关系进行建模，有助于提升异常定位性能。

C. Inference : computation of the anomaly map

受[23]、[26]的启发，我们使用马氏距离[31]$M(x_{ij})$为测试图像中位置(i,j)的图块赋予异常分数。$M(x_{ij})$可解释为测试图块嵌入x_{ij}与学习分布$\mathcal{N}(\mu_{ij},\boldsymbol{\Sigma_{ij}})$之间的距离，其中$M(x_{ij})$的计算方式如下：

$$M(x_{ij})=\sqrt{(x_{ij}-\mu_{ij})^{T}\Sigma_{ij}^{-1}(x_{ij}-\mu_{ij})} \qquad (2)$$

因此，可以计算出构成异常图的马氏距离矩阵$M\ =\ (M(x_{ij}))_{1<i<W,1<j<H}$。该图中的高分区域指示异常区域。整个图像的最终异常分数是异常图M的最大值。最后，在测试阶段，我们的方法不存在基于K-NN的方法[4]–[6]、[25]的可扩展性问题，因为我们无需计算和排序大量距离值来获取图像块的异常分数。

四、实验

A. Datasets and metrics

指标。为了评估定位性能，我们计算了两个独立于阈值的指标。我们使用受试者工作特征曲线下面积（AUROC），其中真正例率是被正确分类为异常像素的百分比。由于AUROC偏向于大型异常，我们还采用了逐区域重叠分数（PRO-score）[2]。该方法针对每个连通分量，绘制正确分类像素率的平均值随假正例率（0到0.3之间）变化的曲线。PRO-score是该曲线的归一化积分值。较高的PRO-score意味着无论大小异常都能被准确定位。

数据集。我们首先在MVTec AD [1]上评估我们的模型，该数据集专为测试工业质量控制中的异常定位算法而设计，并采用单类学习设置。它包含15个类别，约240张图像。原始图像分辨率介于700x700至1024x1024之间，包括10个物体类别和5个纹理类别。物体始终在数据集中居中并对齐，如图1中晶体管（Transistor）和胶囊（Capsule）类别所示。除了原始数据集外，为了在更真实的场景中评估异常定位模型的性能，我们创建了MVTec AD的修改版本，称为Rd-MVTec AD，其中对训练集和测试集均应用了随机旋转（-10, +10）和随机裁剪（从256x256至224x224）。这一修改后的MVTec AD版本能更好地描述质量控制中异常定位的实际应用场景，因为目标物体在图像中并非总是居中且对齐的。

为进一步评估，我们还在上海科技大学校园（STC）数据集[8]上测试了PaDiM，该数据集模拟了静态摄像头的视频监控。它包含274,515帧训练图像和42,883帧测试图像，分为13个场景。原始图像分辨率为856x480。训练视频由正常序列组成，而测试视频则包含异常情况，例如行人区域出现车辆或人员打斗。

B. Experimental setups

我们使用不同的骨干网络训练PaDiM，包括ResNet18（R18）[27]、Wide ResNet-50-2（WR50）[28]和EfficientNet-B5 [29]，所有模型均在ImageNet [32]上进行了预训练。与[5]类似，当骨干网络为ResNet时，从前三层提取补丁嵌入向量，以便结合

information from different semantic levels, while keeping a high enough resolution for the localization task. Following this idea, we extract patch embedding vectors from layers 7 (level 2), 20 (level 4), and 26 (level 5), if an EfficientNet-B5 is used. We also apply a random dimensionality reduction (Rd) (see Sections III-A and V-A). Our model names indicate the backbone and the dimensionality reduction method used, if any. For example, PaDiM-R18-Rd100 is a PaDiM model with a ResNet18 backbone using 100 randomly selected dimensions for the patch embedding vectors. By default we use $\epsilon = 0.01$ for the ϵ from Equation 1.

We reproduce the model SPADE [5] as described in the original publication with a Wide ResNet-50-2 (WR50) [28] as backbone. For SPADE and PaDiM we apply the same prepocessing as in [5]. We resize the images from the MVTec AD to 256x256 and center crop them to 224x224. For the images from the STC we use a 256x256 resize only. We resize the images and the localization maps using bicubic interpolation and we use a Gaussian filter on the anomaly maps with parameter $\sigma = 4$ like in [5].

We also implement our own VAE as a reconstruction-based baseline implemented with a ResNet18 as encoder and a 8x8 convolutional latent variable. It is trained on each MVTec AD class with 10 000 images using the following data augmentations operations: random rotation ($-2°$, $+2°$), 292x292 resize, random crop to 282x282, and finally center crop to 256x256. The training is performed during 100 epochs with the Adam optimizer [12] with an initial learning rate of 10^{-4} and a batch size of 32 images. The anomaly map for the localization corresponds to the pixel-wise L2 error for reconstruction.

V. Results

A. Ablative studies

First, we evaluate the impact of modeling correlations between semantic levels in PaDiM and explore the possibility to simplify our method through dimensionality reduction.

Inter-layer correlation. The combination of Gaussian modeling and the Mahalanobis distance has already been employed in previous works to detect adversarial attacks [26] and for anomaly detection [23] at the image level. However those methods do not model correlations between different CNN's semantic levels as we do in PaDiM. In Table I we show the anomaly localization performance on the MVTec AD of PaDiM with a ResNet18 backbone when using only one of the first three layers (Layer 1, Layer 2, or Layer 3) and when summing the outputs of these 3 models to form an ensemble method that takes into account the first three layers but not the correlations between them (Layers 1+2+3). The last row of Table I (PaDiM-R18) is our proposed version of PaDiM where each patch location is described by one Gaussian distribution taking into account the first three ResNet18 layers and correlations between them. It can be observed that using Layer 3 produces the best results in terms of AUROC among the three layers. It is due to the fact that Layer 3 carries higher semantic level information which helps to better describe normality. However, Layer 3 has a slightly worse PRO-score than Layer 2 that can be explained by the lower resolution of Layer 2 which affects the accuracy of anomaly localization. As we see in the two last rows of Table I, aggregating information from different layers can solve the trade-off issue between high semantic information and high resolution. Unlike model Layer 1+2+3 that simply sums the outputs, our model PaDiM-R18 takes into account correlations between semantic levels. As a result, it outperforms Layer 1+2+3 by 1.1p.p (percent point) for AUROC and 1.8p.p for PRO-score. It confirms the relevance of modeling correlation between semantic levels.

TABLE I
Study of the anomaly localization performance using different semantic-level CNN layers. Results are displayed as tuples (AUROC%, PRO-score%) on the MVTec AD.

Layer used	all texture classes	all object classes	all classes
Layer 1	(93.1, 87.1)	(95.6, 86.5)	(94.8, 86.8)
Layer 2	(95.0, 89.7)	(96.1, 87.9)	(95.7, 88.5)
Layer 3	(94.8, 89.6)	(97.1, 87.7)	(95.7, 88.3)
Layer 1+2+3	(95.4, 90.7)	(96.3, 88.1)	(96.0, 89.0)
PaDiM-R18	(**96.3**, **92.3**)	(**97.5**, **90.1**)	(**97.1**, **90.8**)

TABLE II
Study of the anomaly localization performance with a dimensionality reduction from 448 to 100 and 200 using PCA or random feature selection (Rd). Results are displayed as tuples (AUROC%, PRO-score%) on the MVTec AD.

	all texture classes	all object classes	all classes
Rd 100	(95.7, 91.3)	(97.2, 89.4)	(96.7, 90.5)
PCA 100	(93.7, 88.9)	(93.5, 84.1)	(93.5, 85.7)
Rd 200	(96.1, 92.0)	(97.5, 89.8)	(97.0, 90.5)
PCA 200	(95.1, 91.8)	(96.0, 88.1)	(95.7, 89.3)
all (448)	(**96.3**, **92.3**)	(**97.5**, **90.1**)	(**97.1**, **90.8**)

Dimensionality reduction. PaDiM-R18 estimates multivariate Gaussian distributions from sets of patch embeddings vectors of 448 dimensions each. Decreasing the embedding vector size would reduce the computational and memory complexity of our model. We study two different dimensionality reduction methods. The first one consists in applying a Principal Component Analysis (PCA) algorithm to reduce the vector size to 100 or 200 dimensions. The second method is a random feature selection where we randomly select features before the training. In this case, we train 10 different models and take the average scores. Still the randomness does not change the results between different seeds as the standard error mean (SEM) for the average AUROC is always between 10^{-4} and 10^{-7}.

From Table II we can notice that for the same number of dimensions, the random dimensionality reduction (Rd) outperforms the PCA on all the MVTec AD classes by at least 1.3p.p in the AUROC and 1.2p.p in the PRO-score. It can be explained by the fact that PCA selects the dimensions with the highest variance which may not be the ones that help to discriminate the normal class from the anomalous one [23].

从不同语义层次提取信息，同时为定位任务保持足够高的分辨率。基于这一思路，若使用EfficientNet-B5，我们从第7层（层级2）、第20层（层级4）和第26层（层级5）提取图像块嵌入向量。我们还应用了随机降维（Rd）（参见第三-A节和第五-A节）。我们的模型名称会标注所使用的骨干网络和降维方法（若使用）。例如，PaDiM-R18-Rd100表示使用ResNet18骨干网络，并为图像块嵌入向量随机选取100个维度的PaDiM模型。默认情况下，我们使用公式1中的$\epsilon = 0.01$作为ϵ。

我们按照原始出版物中的描述，复现了SPADE模型[5]，并使用Wide ResNet-50-2（WR50）[28]作为骨干网络。对于SPADE和PaDiM，我们采用与[5]中相同的预处理方法。我们将MVTec AD中的图像调整为256x256，并进行中心裁剪至224x224。对于STC中的图像，我们仅将其调整为256x256。我们使用双三次插值法调整图像和定位图的大小，并在异常图上使用参数为$\sigma = 4$的高斯滤波器，如[5]所述。

我们还实现了自己的变分自编码器（VAE）作为基于重建的基线模型，其编码器采用ResNet18，并配备8x8卷积潜变量。该模型在MVTec AD的每个类别上使用10,000张图像进行训练，数据增强操作包括：随机旋转（$-2^{\circ}, +2^{\circ}$）、调整尺寸至292x292、随机裁剪至282x282，最后中心裁剪至256x256。训练共进行100个周期，使用Adam优化器[12]，初始学习率为10^{-4}，批处理大小为32张图像。用于定位的异常图对应于重建的逐像素L2误差。

五、结果

A. Ablative studies

首先，我们评估了在PaDiM中建模语义层级间相关性的影响，并探讨了通过降维简化方法的可能性。

层间相关性。高斯建模与马氏距离的结合已在先前的研究中被用于检测对抗性攻击[26]及图像级别的异常检测[23]。然而这些方法并未像我们在PaDiM中那样建模不同CNN语义层级间的相关性。在表I中，我们展示了以ResNet18为骨干网络的PaDiM在MVTec AD数据集上的异常定位性能：当仅使用前三层中的单层（第1层、第2层或第3层）时，以及将这三个模型的输出求和形成集成方法（该集成考虑前三层但忽略层间相关性，记为Layers 1+2+3）时的结果。表I最后一行（PaDiM-R18）是我们提出的PaDiM版本，其中每个图像块位置通过一个综合考虑前三层及其相关性的高斯分布进行描述。可以观察到，在三层中使用第3层能在AUROC指标上取得最佳结果，这是因为第3层携带更高语义级别的信息，有助于更准确地描述特征。正常性。然而，Layer 3 的 PRO 分数略低于 Layer 2，这可以解释为 Layer 2 的分辨率较低，影响了异常定位的准确性。正如我们在表 I 的最后两行所见，聚合来自不同层的信息可以解决高语义信息与高分辨率之间的权衡问题。与简单求和输出的模型 Layer 1+2+3 不同，我们的模型 PaDiM-R18 考虑了语义级别之间的相关性。因此，它在 AUROC 上比 Layer 1+2+3 高出 1.1 个百分点，在 PRO 分数上高出 1.8 个百分点。这证实了建模语义级别之间相关性的重要性。

表一 使用不同语义级CNN层进行异常定位性能的研究。结果以MVTEC AD上的元组（AUROC%, PRO-SCORE%）形式展示。

Layer used	all texture classes	all object classes	all classes
Layer 1	(93.1, 87.1)	(95.6, 86.5)	(94.8, 86.8)
Layer 2	(95.0, 89.7)	(96.1, 87.9)	(95.7, 88.5)
Layer 3	(94.8, 89.6)	(97.1, 87.7)	(95.7, 88.3)
Layer 1+2+3	(95.4, 90.7)	(96.3, 88.1)	(96.0, 89.0)
PaDiM-R18	(**96.3**, **92.3**)	(**97.5**, **90.1**)	(**97.1**, **90.8**)

表II 使用PCA或随机特征选择(RD)将维度从448降至100和200时异常定位性能的研究。结果以MVTEC AD上的元组(AUROC%, PRO-SCORE%)形式展示。

	all texture classes	all object classes	all classes
Rd 100	(95.7, 91.3)	(97.2, 89.4)	(96.7, 90.5)
PCA 100	(93.7, 88.9)	(93.5, 84.1)	(93.5, 85.7)
Rd 200	(96.1, 92.0)	(97.5, 89.8)	(97.0, 90.5)
PCA 200	(95.1, 91.8)	(96.0, 88.1)	(95.7, 89.3)
all (448)	(**96.3**, **92.3**)	(**97.5**, **90.1**)	(**97.1**, **90.8**)

降维。PaDiM-R18从每组448维的补丁嵌入向量集中估计多元高斯分布。降低嵌入向量尺寸将减少我们模型的计算和内存复杂度。我们研究了两种不同的降维方法。第一种是应用主成分分析（PCA）算法将向量尺寸降至100或200维。第二种方法是随机特征选择，即在训练前随机选取特征。在此情况下，我们训练10个不同的模型并取平均分数。尽管如此，随机性并未在不同随机种子间改变结果，因为平均AUROC的标准误差均值（SEM）始终在10^{-4}至10^{-7}之间。

从表II中我们可以注意到，在相同维度数的情况下，随机降维（Rd）在所有MVTec AD类别上的AUROC至少高出PCA 1.3个百分点，PRO分数至少高出1.2个百分点。这可以解释为PCA选择方差最大的维度，但这些维度可能并非有助于区分正常类别与异常类别的关键维度[23]。

TABLE III
COMPARISON OF OUR PADIM MODELS WITH THE STATE-OF-THE-ART FOR THE ANOMALY LOCALIZATION ON THE MVTEC AD. RESULTS ARE DISPLAYED AS TUPLES (AUROC%, PRO-SCORE%)

Type	Reconstruction-based methods			Embedding similarity based methods			Our methods	
Model	AE simm [1], [2], [9]	AE L2 [1], [2]	VAE	Student [2]	Patch SVDD [4]	SPADE [5]	PaDiM-R18-Rd100	PaDiM-WR50-Rd550
Carpet	(87, 64.7)	(59, 45.6)	(59.7, 61.9)	(-, 69.5)	(92.6, -)	(97.5, 94.7)	(98.9, 96.0)	(**99.1**, **96.2**)
Grid	(94, 84.9)	(90, 58.2)	(61.2, 40.8)	(-, 81.9)	(96.2, -)	(93.7, 86.7)	(94.9, 90.9)	(**97.3**, **94.6**)
Leather	(78, 56.1)	(75, 81.9)	(67.1, 64.9)	(-, 81.9)	(97.4, -)	(97.6, 97.2)	(99.1, 97.9)	(**99.2**, **97.8**)
Tile	(59, 17.5)	(51, 89.7)	(51.3, 24.2)	(-, 91.2)	(91.4, -)	(87.4, 75.9)	(91.2, 81.6)	(**94.1**, **86.0**)
Wood	(73, 60.5)	(73, 72.7)	(66.6, 57.8)	(-, 72.5)	(90.8, -)	(88.5, 87.4)	(93.6, 90.3)	(**94.9**, **91.1**)
All texture classes	(78, 56.7)	(70, 69.6)	(61.2, 49.9)	(-, 79.4)	(93.7, -)	(92.9, 88.4)	(95.6, 91.3)	(**96.9**, **93.2**)
Bottle	(93, 83.4)	(86, 91.0)	(83.1, 70.5)	(-, 91.8)	(98.1, -)	(**98.4**, **95.5**)	(98.1, 93.9)	(98.3, 94.8)
Cable	(82, 47.8)	(86, 82.5)	(83.1, 77.9)	(-, 86.5)	(96.8, -)	(**97.2**, **90.9**)	(95.8, 86.2)	(96.7, 88.8)
Capsule	(94, 86.0)	(88, 86.2)	(81.7, 77.9)	(-, 91.6)	(95.8, -)	(**99.0** ,**93.7**)	(98.3, 91.9)	(98.5, 93.5)
Hazelnut	(97, 91.6)	(95, 91.7)	(87.7, 77.0)	(-, 93.7)	(97.5, -)	(**99.1**, **95.4**)	(97.7, 91.4)	(98.2, 92.6)
Metal Nut	(89, 60.3)	(86, 83.0)	(78.7, 57.6)	(-, 89.5)	(98.0, -)	(**98.1**, **94.4**)	(96.7, 81.9)	(97.2, 85.6)
Pill	(91, 83.0)	(85, 89.3)	(81.3, 79.3)	(-, 93.5)	(95.1, -)	(**96.5**, **94.6**)	(94.7, 90.6)	(95.7, 92.7)
Screw	(96, 88.7)	(96, 75.4)	(75.3, 66.4)	(-, 92.8)	(95.7, -)	(**98.9**, **96.0**)	(97.4, 91.3)	(98.5, 94.4)
Toothbrush	(92, 78.4)	(93, 82.2)	(91.9, 85.4)	(-, 86.3)	(98.1, -)	(97.9, **93.5**)	(98.7, 92.3)	(**98.8**, 93.1)
Transistor	(90, 72.5)	(86, 72.8)	(75.4, 61.0)	(-, 70.1)	(97.0, -)	(94.1, **87.4**)	(97.2, 80.2)	(**97.5**, 84.5)
Zipper	(88, 66.5)	(77, 83.9)	(71.6, 60.8)	(-, 93.3)	(95.1, -)	(96.5, 92.6)	(98.2, 94.7)	(**98.5**, **95.9**)
All object classes	(91, 75.8)	(88, 83.8)	(81.0, 71.4)	(-, 88.9)	(96.7, -)	(97.6, **93.4**)	(97.3, 89.4)	(**97.8**, 91.6)
All classes	(87, 69.4)	(82, 79.0)	(74.4, 64.2)	(-, 85.7)	(95.7, -)	(96.5, 91.7)	(96.7, 90.1)	(**97.5**, **92.1**)

It can also be noted from Table II that randomly reducing the embedding vector size to only 100 dimensions has a very little impact on the anomaly localization performance. The results drop only by 0.4p.p in the AUROC and 0.3p.p in the PRO-score. This simple yet effective dimensionality reduction method significantly reduces PaDiM time and space complexity as it will be shown in Section V-D.

B. Comparison with the state-of-the-art

Localization. In Table III, we show the AUROC and the PRO-score results for anomaly localization on the MVTec AD. For a fair comparison, we used a Wide ResNet-50-2 (WR50) as this backbone is used in SPADE [5]. Since the other baselines have smaller backbones, we also try a ResNet18 (R18). We randomly reduce the embedding size to 550 and 100 for PaDiM with WR50 and R18 respectively.

We first notice that PaDiM-WR50-Rd550 outperforms all the other methods in both the PRO-score and the AUROC on average for all the classes. PaDiM-R18-Rd100 which is a very light model also outperforms all models in the average AUROC on the MVTec AD classes by at least 0.2p.p. When we further analyze the PaDiM performances, we see that the gap for the object classes is small as PaDiM-WR50-Rd550 is the best only in the AUROC (+0.2p.p) but SPADE [5] is the best in the PRO-score (+1.8p.p). However, our models are particularly accurate on texture classes. PaDiM-WR50-Rd550 outperforms the second best model SPADE [5] by 4.8p.p and 4.0p.p in the PRO-score and the AUROC respectively on average on texture classes. Indeed, PaDiM learns an explicit probabilistic model of the normal classes contrary to SPADE [5] or Patch-SVDD [4]. It is particularly efficient on texture images because even if they are not aligned and centered like object images, PaDiM effectively captures their statistical similarity accross the normal train dataset.

Additionally, we evaluate our model on the STC dataset. We compare our method to the two best reported models performing anomaly localization without temporal information, CAVGA-RU [3] and SPADE [5]. As shown in Table IV, the best result (AUROC) on the STC dataset is achieved with our simplest model PaDiM-R18-Rd100 by a 2.1p.p. margin. In fact, pedestrian positions in images are highly variable in this dataset and, as shown in Section V-C, our method performs well on non-aligned datasets.

Detection. By taking the maximum score of the anomaly maps issued by our models (see Section III-C) we give anomaly scores to entire images to perform anomaly detection at the image level. We test PaDiM for anomaly detection with a Wide ResNet-50-2 (WR50) [28] used in SPADE and an EfficientNet-B5 [29]. The Table V shows that our model PaDiM-WR50-Rd550 outperforms every method except MahalanobisAD [23] with their best reported backbone, an EfficientNet-B4. Still our PaDiM-EfficientNet-B5 outperforms every model by at least 2.6p.p on average on all the classes in the AUROC. Besides, contrary to the second best method for anomaly detection, MahalanobisAD [23], our model also performs anomaly segmentation which characterizes more precisely the anomalous areas in the images.

C. Anomaly localization on a non-aligned dataset

To estimate the robustness of anomaly localization methods, we train and evaluate the performance of PaDiM and several

TABLE IV
COMPARISON OF OUR PADIM MODEL WITH THE STATE-OF-THE-ART FOR THE ANOMALY LOCALIZATION ON THE STC IN THE AUROC%.

Model	CAVGA-RU [3]	SPADE [5]	PaDiM-R18-Rd100
AUROC score%	85	89.9	**91.2**

表三 我们的PADIM模型与MVTEC AD异常定位领域最先进技术的比较。结果以元组形式展示（AUROC%, PRO-SCORE%）

Type	Reconstruction-based methods			Embedding similarity based methods			Our methods	
Model	AE simm [1], [2], [9]	AE L2 [1], [2]	VAE	Student [2]	Patch SVDD [4]	SPADE [5]	PaDiM-R18-Rd100	PaDiM-WR50-Rd550
Carpet	(87, 64.7)	(59, 45.6)	(59.7, 61.9)	(-, 69.5)	(92.6, -)	(97.5, 94.7)	(98.9, 96.0)	(**99.1**, **96.2**)
Grid	(94, 84.9)	(90, 58.2)	(61.2, 40.8)	(-, 81.9)	(96.2, -)	(93.7, 86.7)	(94.9, 90.9)	(**97.3**, **94.6**)
Leather	(78, 56.1)	(75, 81.9)	(67.1, 64.9)	(-, 81.9)	(97.4, -)	(97.6, 97.2)	(99.1, 97.9)	(**99.2**, **97.8**)
Tile	(59, 17.5)	(51, 89.7)	(51.3, 24.2)	(-, 91.2)	(91.4, -)	(87.4, 75.9)	(91.2, 81.6)	(**94.1**, **86.0**)
Wood	(73, 60.5)	(73, 72.7)	(66.6, 57.8)	(-, 72.5)	(90.8, -)	(88.5, 87.4)	(93.6, 90.3)	(**94.9**, **91.1**)
All texture classes	(78, 56.7)	(70, 69.6)	(61.2, 49.9)	(-, 79.4)	(93.7, -)	(92.9, 88.4)	(95.6, 91.3)	(**96.9**, **93.2**)
Bottle	(93, 83.4)	(86, 91.0)	(83.1, 70.5)	(-, 91.8)	(98.1, -)	(**98.4**, **95.5**)	(98.1, 93.9)	(98.3, 94.8)
Cable	(82, 47.8)	(86, 82.5)	(83.1, 77.9)	(-, 86.5)	(96.8, -)	(**97.2**, **90.9**)	(95.8, 86.2)	(96.7, 88.8)
Capsule	(94, 86.0)	(88, 86.2)	(81.7, 77.9)	(-, 91.6)	(95.8, -)	(**99.0** ,**93.7**)	(98.3, 91.9)	(98.5, 93.5)
Hazelnut	(97, 91.6)	(95, 91.7)	(87.7, 77.0)	(-, 93.7)	(97.5, -)	(**99.1**, **95.4**)	(97.7, 91.4)	(98.2, 92.6)
Metal Nut	(89, 60.3)	(86, 83.0)	(78.7, 57.6)	(-, 89.5)	(98.0, -)	(**98.1**, **94.4**)	(96.7, 81.9)	(97.2, 85.6)
Pill	(91, 83.0)	(85, 89.3)	(81.3, 79.3)	(-, 93.5)	(95.1, -)	(**96.5**, **94.6**)	(94.7, 90.6)	(95.7, 92.7)
Screw	(96, 88.7)	(96, 75.4)	(75.3, 66.4)	(-, 92.8)	(95.7, -)	(**98.9**, **96.0**)	(97.4, 91.3)	(98.5, 94.4)
Toothbrush	(92, 78.4)	(93, 82.2)	(91.9, 85.4)	(-, 86.3)	(98.1, -)	(97.9, **93.5**)	(98.7, 92.3)	(**98.8**, 93.1)
Transistor	(90, 72.5)	(86, 72.8)	(75.4, 61.0)	(-, 70.1)	(97.0, -)	(94.1, **87.4**)	(97.2, 80.2)	(**97.5**, 84.5)
Zipper	(88, 66.5)	(77, 83.9)	(71.6, 60.8)	(-, 93.3)	(95.1, -)	(96.5, 92.6)	(98.2, 94.7)	(**98.5**, **95.9**)
All object classes	(91, 75.8)	(88, 83.8)	(81.0, 71.4)	(-, 88.9)	(96.7, -)	(97.6, **93.4**)	(97.3, 89.4)	(**97.8**, 91.6)
All classes	(87, 69.4)	(82, 79.0)	(74.4, 64.2)	(-, 85.7)	(95.7, -)	(96.5, 91.7)	(96.7, 90.1)	(**97.5**, **92.1**)

从表II中还可以注意到，将嵌入向量维度随机降低至仅100维对异常定位性能的影响微乎其微。AUROC指标仅下降0.4个百分点，PRO分数仅下降0.3个百分点。这种简单而有效的降维方法显著降低了PaDiM的时间与空间复杂度，具体将在第V-D节中展示。

B. Comparison with the state-of-the-art

定位。在表III中，我们展示了在MVTec AD上进行异常定位的AUROC和PRO-score结果。为了公平比较，我们使用了Wide ResNet-50-2（WR50）作为主干网络，因为SPADE [5] 也采用了这一主干网络。由于其他基线模型使用较小的主干网络，我们还尝试了ResNet18（R18）。我们随机将PaDiM的嵌入维度分别降低至550（WR50）和100（R18）。

我们首先注意到，PaDiM-WR50-Rd550 在所有类别的平均 PRO 分数和 AUROC 上均优于其他所有方法。PaDiM-R18-Rd100 作为一个非常轻量的模型，其在 MVTec AD 类别上的平均 AUROC 也至少超出所有模型 0.2 个百分点。进一步分析 PaDiM 的表现时，我们发现其在物体类别上的优势较小：PaDiM-WR50-Rd550 仅在 AUROC 上表现最佳（+0.2 个百分点），而 SPADE [5] 在 PRO 分数上最优（+1.8 个百分点）。然而，我们的模型在纹理类别上表现尤为出色。在纹理类别的平均 PRO 分数和 AUROC 上，PaDiM-WR50-Rd550 分别比第二名 SPADE [5] 高出 4.8 个百分点和 4.0 个百分点。事实上，与 SPADE [5] 或 Patch-SVDD [4] 不同，PaDiM 学习了正常类别的显式概率模型。这种方法对纹理图像特别有效，因为即使纹理图像不像物体图像那样经过对齐和居中处理，PaDiM 仍能有效捕捉它们在正常训练数据集中的统计相似性。

此外，我们在STC数据集上评估了我们的模型。我们将我们的方法与两个最佳报告模型进行比较，这两个模型在没有时间信息的情况下执行异常定位，即CAVGA-RU [3]和SPADE [5]。如表IV所示，我们的最简单模型PaDiM-R18-Rd100在STC数据集上取得了最佳结果（AUROC），领先2.1个百分点。实际上，该数据集中图像中的行人位置变化很大，并且如第V-C节所示，我们的方法在非对齐数据集上表现良好。

检测。通过取我们模型生成的异常图（见第III-C节）中的最高分数，我们为整个图像分配异常分数，以在图像级别执行异常检测。我们使用SPADE中采用的Wide ResNet-50-2（WR50）[28]和EfficientNet-B5[29]测试了PaDiM的异常检测性能。表V显示，除了MahalanobisAD[23]使用其最佳报告骨干网络EfficientNet-B4外，我们的PaDiM-WR50-Rd550模型在所有方法中表现最优。尽管如此，我们的PaDiM-EfficientNet-B5模型在AUROC上所有类别的平均表现仍至少超出其他模型2.6个百分点。此外，与异常检测中表现第二好的方法MahalanobisAD[23]不同，我们的模型还能执行异常分割，从而更精确地定位图像中的异常区域。

C. Anomaly localization on a non-aligned dataset

为了评估异常定位方法的鲁棒性，我们训练并评估了PaDiM及多种方法的性能。

表四 我们的PADIM模型与STC异常定位最先进技术在AUROC%上的比较。

Model	CAVGA-RU [3]	SPADE [5]	PaDiM-R18-Rd100
AUROC score%	85	89.9	**91.2**

TABLE V
ANOMALY DETECTION RESULTS (AT THE IMAGE LEVEL) ON THE MVTEC AD USING AUROC%.

Model	GANomaly [20]	ITAE [11]	Patch SVDD [4]	SPADE [5] (WR50)	MahalanobisAD [23] (EfficientNet-B4)	PaDiM-WR50-Rd550	PaDiM EfficientNet-B5
all textures classes	-	-	94.6	-	97.2	98.8	**99.0**
all objects classes	-	-	90.9	-	94.8	93.6	**97.2**
all classes	76.2	83.9	92.1	85.5	95.8	95.3	**97.9**

state-of-the-art methods (SPADE [5], VAE) on a modified version of the MVTec AD, Rd-MVTec AD, described in Section IV-A. Results of this experiment are displayed in Table VI. For each test configuration we run 5 times data preprocessing on the MVTec AD with random seeds to obtain 5 different versions of the dataset, denoted as Rd-MVTec AD. Then, we average the obtained results and report them in Table VI. According to the presented results, PaDiM-WR50-Rd550 outperforms the other models on both texture and object classes in the PRO-score and the AUROC. Besides, the SPADE [5] and VAE performances on the Rd-MVTec AD decrease more than the performance of PaDiM-WR50-Rd550 when comparing to the results obtained on the normal MVTec AD (refer to Table III). The AUROC results decrease by 5.3p.p for PaDiM-WR50-Rd550 against 12.2p.p and 8.8p.p decline for VAE and SPADE respectively. Thus, we can conclude that our method seems to be more robust to non-aligned images than the other existing and tested works.

TABLE VI
ANOMALY LOCALIZATION RESULTS ON THE NON-ALIGNED RD-MVTEC AD. RESULTS ARE DISPLAYED AS TUPLES (AUROC%, PRO-SCORE%)

Model	VAE (R18)	SPADE (WR50)	PaDiM-WR50-Rd550
all texture classes	(54.7, 23.1)	(84.6, 75.6)	(**92.4**, **77.9**)
all object classes	(65.8, 30.2)	(88.2, 65.8)	(**92.1**, **70.8**)
all classes	(62.1, 27.8)	(87.2, 69.0)	(**92.2**, **73.1**)

D. Scalability gain

Time complexity. In PaDiM, the training time complexity scales linearly with the dataset size because the Gaussian parameters are estimated using the entire training dataset. However, contrary to the methods that require to train deep neural networks, PaDiM uses a pretrained CNN, and, thus, no deep learning training is required which is often a complex procedure. Hence, it is very fast and easy to train it on small datasets like MVTec AD. For our most complex model PaDiM-WR50-Rd550, the training on a CPU (Intel CPU 6154 3GHz 72th) with a serial implementation takes on average 150 seconds on the MVTec AD classes and 1500 seconds on average on the STC video scenes. These training procedures could be further accelerated using GPU hardware for the forward pass and the covariance estimation. In contrast, training the VAE with 10 000 images per class on the MVTec AD following the procedure described in Section IV-B takes 2h40 per class using one GPU NVIDIA P5000. Conversely, SPADE [5] requires no training as there are no parameters to learn. Still, it computes and stores in the memory before testing all the embedding vectors of the normal training images. Those vectors are the inputs of a K-NN algorithm which makes SPADE's inference very slow as shown in Table VII.

TABLE VII
AVERAGE INFERENCE TIME OF ANOMALY LOCALIZATION IN SECONDS ON THE MVTEC AD WITH A CPU INTEL I7-4710HQ @ 2.50GHZ.

Model	SPADE (WR50)	VAE (R18)	PaDiM R18-Rd100	PaDiM-WR50-Rd550
Inference time (sec.)	7.10	0.21	0.23	0.95

In Table VII, we measure the model inference time using a mainstream CPU (Intel i7-4710HQ CPU @ 2.50GHz) with a serial implementation. On the MVTec AD, the inference time of SPADE is around seven times slower than our PaDiM model with equivalent backbone because of the computationally expensive NN search. Our VAE implementation, which is similar to most reconstruction-based models, is the fastest model but our simple model PaDiM-R18-Rd100 has the same order of magnitude for the inference time. While having similar complexity, PaDiM largely outperfoms the VAE methods (see Section V-B).

Memory complexity. Unlike SPADE [5] and Patch SVDD [4], the space complexity of our model is independent of the dataset training size and depends only on the image resolution. PaDiM keeps in the memory only the pretrained CNN and the Gaussian parameters associated with each patch. In Table VIII we show the memory requirement of SPADE, our VAE implementation, and PaDiM, assuming that parameters are encoded in float32. Using equivalent backbone, SPADE has a lower memory consumption than PaDiM on the MVTec AD. However, when using SPADE on a larger dataset like the STC, its memory consumption becomes intractable, whereas PaDiM-WR50-Rd550 requires seven times less memory. The PaDiM space complexity increases from the MVTec AD to the STC only because the input image resolution is higher in the latter dataset as described in Section IV-B. Finally, one of the advantages of our framework PaDiM is that the user can easily adapt the method by choosing the backbone and the embedding size to fit its inference time requirements, resource limits, or expected performance.

VI. CONCLUSION

We have presented a framework called PaDiM for anomaly detection and localization in one-class learning setting which is based on distribution modeling. It achieves state-of-the-art performance on MVTec AD and STC datasets. Moreover, we extend the evaluation protocol to non-aligned data and the first

表五 MVTEC AD数据集上（图像级别）基于AUROC%的异常检测结果。

Model	GANomaly [20]	ITAE [11]	Patch SVDD [4]	SPADE [5] (WR50)	MahalanobisAD [23] (EfficientNet-B4)	PaDiM-WR50-Rd550	PaDiM EfficientNet-B5
all textures classes	-	-	94.6	-	97.2	98.8	**99.0**
all objects classes	-	-	90.9	-	94.8	93.6	**97.2**
all classes	76.2	83.9	92.1	85.5	95.8	95.3	**97.9**

在IV-A节描述的Rd-MVTec AD（MVTec AD的修改版本）上，我们评估了先进方法（SPADE [5]、VAE）的性能。该实验结果展示在表VI中。针对每种测试配置，我们在MVTec AD上使用随机种子运行5次数据预处理，获得5个不同版本的数据集（记为Rd-MVTec AD），随后对所得结果取平均值并记录于表VI。根据呈现的结果，PaDiM-WR50-Rd550在纹理类和物体类别的PRO分数与AUROC指标上均优于其他模型。此外，与正常MVTec AD上的结果（参见表III）相比，SPADE [5]和VAE在Rd-MVTec AD上的性能下降幅度大于PaDiM-WR50-Rd550：PaDiM-WR50-Rd550的AUROC结果下降5.3个百分点，而VAE和SPADE分别下降12.2和8.8个百分点。因此我们可以得出结论：相较于其他现有及已测试的方法，我们的方法对非对齐图像表现出更强的鲁棒性。

表六 非对齐RD-MVTEC AD上的异常定位结果。结果以元组形式显示（AUROC%，PRO-SCORE%）

Model	VAE (R18)	SPADE (WR50)	PaDiM-WR50-Rd550
all texture classes	(54.7, 23.1)	(84.6, 75.6)	(**92.4**, **77.9**)
all object classes	(65.8, 30.2)	(88.2, 65.8)	(**92.1**, **70.8**)
all classes	(62.1, 27.8)	(87.2, 69.0)	(**92.2**, **73.1**)

D. Scalability gain

时间复杂度。在PaDiM中，训练时间复杂度随数据集大小线性增长，因为高斯参数是使用整个训练数据集进行估计的。然而，与需要训练深度神经网络的方法不同，PaDiM使用了预训练的CNN，因此无需进行通常复杂的深度学习训练过程。因此，在像MVTec AD这样的小型数据集上训练非常快速且简便。对于我们最复杂的模型PaDiM-WR50-Rd550，在CPU（Intel CPU 6154 3GHz 72线程）上使用串行实现进行训练时，MVTec AD各类别的平均训练时间为150秒，总计1500在STC视频场景上平均耗时秒级。利用GPU硬件进行前向传播和协方差估计，可以进一步加速这些训练过程。相比之下，按照第IV-B节所述流程在MVTec AD数据集上训练VAE（每类使用10,000张图像）时，使用单张NVIDIA P5000 GPU每类需耗时2小时40分钟。相反，SPADE [5] 无需训练，因为其没有需要学习的参数。但在测试前，它仍需计算并存储所有正常训练图像的嵌入向量至内存中。这些向量作为K-NN算法的输入，导致SPADE的推理速度非常缓慢，如表VII所示。

表 VII 在配备英特尔酷睿 i7-4710HQ @ 2.50GHz CPU 的 MVTEC AD 数据集上进行异常定位的平均推理时间（单位：秒）。

Model	SPADE (WR50)	VAE (R18)	PaDiM R18-Rd100	PaDiM-WR50-Rd550
Inference time (sec.)	7.10	0.21	0.23	0.95

在表VII中，我们使用主流CPU（Intel i7-4710HQ CPU @ 2.50GHz）通过串行实现测量了模型推理时间。在MVTec AD数据集上，由于计算成本高昂的最近邻搜索，SPADE的推理时间比我们使用同等骨干网络的PaDiM模型慢约七倍。我们实现的VAE模型（与大多数基于重建的模型类似）是最快的，但我们简单的PaDiM-R18-Rd100模型在推理时间上仍保持同一数量级。在复杂度相近的情况下，PaDiM大幅优于VAE方法（参见第V-B节）。

内存复杂度。与SPADE [5]和Patch SVDD [4]不同，我们模型的空间复杂度与数据集训练规模无关，仅取决于图像分辨率。PaDiM在内存中仅保留预训练的CNN和与每个图像块关联的高斯参数。在表VIII中，我们展示了SPADE、我们实现的VAE以及PaDiM的内存需求（假设参数以float32格式编码）。在使用相同骨干网络时，SPADE在MVTec AD数据集上的内存消耗低于PaDiM。然而，当在STC等更大规模数据集上使用SPADE时，其内存消耗将变得难以处理，而PaDiM-WR50-Rd550所需内存减少了七倍。PaDiM的空间复杂度从MVTec AD到STC的增加仅源于后者的输入图像分辨率更高（如第IV-B节所述）。最后，我们框架PaDiM的优势之一在于：用户可通过选择骨干网络和嵌入维度来灵活调整方法，以适应其推理时间要求、资源限制或预期性能。

六、结论

我们提出了一个名为PaDiM的框架，用于基于分布建模的单类学习设置中的异常检测与定位。该框架在MVTec AD和STC数据集上实现了最先进的性能。此外，我们将评估协议扩展至非对齐数据，并首次

TABLE VIII
MEMORY REQUIREMENT IN GB OF THE ANOMALY LOCALIZATION METHODS TRAINED ON THE MVTEC AD AND THE STC DATASET.

model	SPADE (WR50)	VAE (R18)	PaDiM R18-Rd100	PaDiM-WR50-Rd550
MVTec AD	1.4	0.09	0.17	3.8
STC	37.0	-	0.21	5.2

results show that PaDiM can be robust on these more realistic data. PaDiM low memory and time consumption and its ease of use make it suitable for various applications, such as visual industrial control.

REFERENCES

[1] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection," in *CVPR*, 2019.

[2] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," in *CVPR*, 2020.

[3] S. Venkataramanan, K.-C. Peng, R. V. Singh, and A. Mahalanobis, "Attention guided anomaly localization in images," in *arXiv, 1911.08616*, 2019.

[4] J. Yi and S. Yoon, "Patch svdd: Patch-level svdd for anomaly detection and segmentation," in *arXiv, 2006.16067*, 2020.

[5] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," in *arXiv, 2005.02357*, 2020.

[6] L. Bergman and Y. Hoshen, "Classification-based anomaly detection for general data," in *ICLR*, 2020.

[7] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[8] W. Liu, D. L. W. Luo, and S. Gao, "Future frame prediction for anomaly detection – a new baseline," in *CVPR*, 2018.

[9] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," in *VISIGRAPP*, 2019.

[10] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. van den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *ICCV*, 2019.

[11] C. Huang, F. Ye, J. Cao, M. Li, Y. Zhang, and C. Lu, "Attribute restoration framework for anomaly detection," in *arXiv, 1911.10676*, 2019.

[12] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[13] K. Sato, K. Hama, T. Matsubara, and K. Uehara, "Predictable uncertainty-aware unsupervised deep anomaly segmentation," in *IJCNN*, 2019.

[14] W. Liu, R. Li, M. Zheng, S. Karanam, Z.Wu, B. Bhanu, R. J. R., and O. Camps, "Towards visually explaining variational autoencoders," in *CVPR*, 2020.

[15] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *CVPR*, 2018.

[16] S. Pidhorskyi, R. Almohsen, D. A. Adjeroh, and G. Doretto, "Generative probabilistic novelty detection with adversarial autoencoders," in *NIPS*, 2018.

[17] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," *ACCV*, 2018.

[18] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *CVPR*, 2019.

[19] K. H. Kim, S. Shim, Y. Lim, J. Jeon, J. Choi, B. Kim, and A. S. Yoon, "Rapp: Novelty detection with reconstruction along projection pathway," in *ICLR*, 2020.

[20] S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon, "Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection," in *IJCNN*, 2019.

[21] P. Perera, R. Nallapati, and B. Xiang, "OCGAN: one-class novelty detection using gans with constrained latent representations," in *CVPR*, 2019.

[22] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *ICLM*, 2018.

[23] O. Rippel, P. Mertens, and D. Merhof, "Modeling the distribution of normal data in pre-trained deep features for anomaly detection," in *arXiv, 2005.14140*, 2020.

[24] L. Bergman, N. Cohen, and Y. Hoshen, "Deep nearest neighbor anomaly detection," in *arXiv, 2002.10445*, 2020.

[25] S. R. Napoletano P, Piccoli F, "Anomaly detection in nanofibrous materials by cnn-based self-similarity," in *Sensors.*, vol. 18, no. 1, 2018, p. 209.

[26] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *NIPS*, 2018.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *ICML*, 2016.

[28] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *BMVC*, 2016.

[29] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019.

[30] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[31] P. Mahalanobis, "On the generalized distance in statistics," in *National Institute of Science of India*, 1936.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009.

表VIII 在MVTEC AD和STC数据集上训练的异常定位方法的内存需求（单位：GB）。

model	SPADE (WR50)	VAE (R18)	PaDiM R18-Rd100	PaDiM-WR50-Rd550
MVTec AD	1.4	0.09	0.17	3.8
STC	37.0	-	0.21	5.2

结果表明，PaDiM在这些更现实的数据上能够保持鲁棒性。PaDiM低内存与时间消耗以及易用性使其适用于多种应用，例如视觉工业控制。

参考文献

[1] P. Bergmann, M. Fauser, D. Sattlegger, 与 C. Steger, “Mvtec ad – 一个用于无监督异常检测的综合真实世界数据集，” 发表于 *CVPR*, 2019.[2] P. Bergmann, M. Fauser, D. Sattlegger, 与 C. Steger, “无信息学生：基于判别性潜在嵌入的学生-教师异常检测，” 发表于 *CVPR*, 2020.[3] S. Venkataramanan, K.-C. Peng, R. V. Singh, 与 A. Mahalanobis, “图像中注意力引导的异常定位，” 发表于 *arXiv, 1911.08616*, 2019.[4] J. Yi 与 S. Yoon, “Patch svdd：用于异常检测与分割的块级支持向量数据描述，” 发表于 *arXiv, 2006.16067*, 2020.[5] N. Cohen 与 Y. Hoshen, “基于深度金字塔对应关系的子图像异常检测，” 发表于 *arXiv, 2005.02357*, 2020.[6] L. Bergman 与 Y. Hoshen, “基于分类的通用数据异常检测，” 发表于 *ICLR*, 2020.[7] T. Cover 与 P. Hart, “最近邻模式分类，” *IEEE Transactions on Information Theory*, 卷 13, 期 1, 页 21–27, 1967.[8] W. Liu, D. L. W. Luo, 与 S. Gao, “用于异常检测的未来帧预测 – 一个新的基线，” 发表于 *CVPR*, 2018.[9] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, 与 C. Steger, “通过将结构相似性应用于自编码器来改进无监督缺陷分割，” 发表于 *VISIGRAPP*, 2019.[10] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, 与 A. van den Hengel, “记忆常态以检测异常：用于无监督异常检测的记忆增强深度自编码器，” 发表于 *ICCV*, 2019.[11] C. Huang, F. Ye, J. Cao, M. Li, Y. Zhang, 与 C. Lu, “用于异常检测的属性恢复框架，” 发表于 *arXiv, 1911.10676*, 2019.[12] D. P. Kingma 与 M. Welling, “自动编码变分贝叶斯，” 发表于 *ICLR*, 2014.[13] K. Sato, K. Hama, T. Matsubara, 与 K. Uehara, “可预测的、感知不确定性的无监督深度异常分割，” 发表于 *IJCNN*, 2019.[14] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R. J. R., 与 O. Camps, “面向视觉解释变分自编码器，” 发表于 *CVPR*, 2020.[15] M. Sabokrou, M. Khalooei, M. Fathy, 与 E. Adeli, “用于新颖性检测的对抗性学习单类分类器，” 发表于 *CVPR*, 2018.[16] S. Pidhorskyi, R. Almohsen, D. A. Adjeroh, 与 G. Doretto, “使用对抗自编码器的生成概率新颖性检测，” 发表于 *NIPS*, 2018.[17] S. Akcay, A. Atapour-Abarghouei, 与 T. P. Breckon, “Ganomaly：通过对抗训练进行半监督异常检测，” *ACCV*, 2018.[18] D. Abati, A. Porrello, S. Calderara, 与 R. Cucchiara, “用于新颖性检测的潜在空间自回归，” 发表于 *CVPR*, 2019.[19] K. H. Kim, S. Shim, Y. Lim, J. Jeon, J. Choi, B. Kim, 与 A. S. Yoon, “Rapp：沿投影路径重建的新颖性检测，” 发表于 *ICLR*, 2020.[20] S. Akçay, A. Atapour-Abarghouei, 与 T. P. Breckon, “Skip-ganomaly：跳跃连接与对抗训练的编码器-解码器异常检测，” 发表于 *IJCNN*, 2019.[21] P. Perera, R. Nallapati, 与 B. Xiang, “OCGAN：使用具有约束潜在表示的生成对抗网络进行单类新颖性检测，” 发表于 *CVPR*, 2019.[22] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, 和 M. Kloft, “深度单类分类,” 发表于 *ICLM*, 2018.[23] O. Rippel, P. Mertens, 和 D. Merhof, “在预训练深度特征中建模正常数据分布以进行异常检测,” 发表于 *arXiv, 2005.14140*, 2020.[24] L. Bergman, N. Cohen, 和 Y. Hoshen, “深度最近邻异常检测,” 发表于 *arXiv, 2002.10445*, 2020.[25] S. R. Napoletano P, Piccoli F, “基于CNN自相似性的纳米纤维材料异常检测,” 发表于 *Sensors.*, 第18卷, 第1期, 2018, 第209页.[26] K. Lee, K. Lee, H. Lee, 和 J. Shin, “一个用于检测分布外样本和对抗攻击的简单统一框架,” 发表于 *NIPS*, 2018.[27] K. He, X. Zhang, S. Ren, 和 J. Sun, “用于图像识别的深度残差学习,” 发表于 *ICML*, 2016.[28] S. Zagoruyko 和 N. Komodakis, “宽残差网络,” 发表于 *BMVC*, 2016.[29] M. Tan 和 Q. V. Le, “EfficientNet: 重新思考卷积神经网络的模型缩放,” 发表于 *ICML*, 2019.[30] K. Pearson, “论空间点系的最适直线与平面,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 第2卷, 第11期, 第559–572页, 1901.[31] P. Mahalanobis, “论统计学中的广义距离,” 发表于 *National Institute of Science of India*, 1936.[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, 和 L. Fei-Fei, “ImageNet: 一个大规模分层图像数据库,” 发表于 *CVPR*, 2009.