Student-Teacher Feature Pyramid Matching for Anomaly Detection

Guodong Wang*[1,2]
wanggd@buaa.edu.cn
Shumin Han*[3]
hanshumin@baidu.com
Errui Ding[3]
dingerrui@baidu.com
Di Huang†[1,2]
dhuang@buaa.edu.cn

[1] State Key Laboratory of Software Development Environment
Beihang University
Beijing, China

[2] School of Computer Science and Engineering
Beihang University
Beijing, China

[3] Department of Computer Vision Technology
Baidu, Inc.
Beijing, China

Abstract

Anomaly detection is a challenging task and usually formulated as an one-class learning problem for the unexpectedness of anomalies. This paper proposes a simple yet powerful approach to this issue, which is implemented in the student-teacher framework for its advantages but substantially extends it in terms of both accuracy and efficiency. Given a strong model pre-trained on image classification as the teacher, we distill the knowledge into a single student network with the identical architecture to learn the distribution of anomaly-free images and this one-step transfer preserves the crucial clues as much as possible. Moreover, we integrate the multi-scale feature matching strategy into the framework, and this hierarchical feature matching enables the student network to receive a mixture of multi-level knowledge from the feature pyramid under better supervision, thus allowing to detect anomalies of various sizes. The difference between feature pyramids generated by the two networks serves as a scoring function indicating the probability of anomaly occurring. Due to such operations, our approach achieves accurate and fast pixel-level anomaly detection. Very competitive results are delivered on the MVTec anomaly detection dataset, superior to the state of the art ones.

1 Introduction

Anomaly detection is generally referred to as identifying samples that are atypical with respect to regular patterns in the data set and has shown great potential in various real-world applications such as video surveillance [1, 31], product quality control [7, 8, 27] and medical

* Equal contribution.
† Corresponding author.

用于异常检测的学生-教师特征金字塔匹配

王国栋[1]
wanggd@buaa.edu.cn
韩树民[3]
hanshumin@baidu.com
丁二锐[3]
dingerrui@baidu.com
黄迪[1,2]
dhuang@buaa.edu.cn

[1] 北京航空航天大学软件开发环境国家重点实验室 中国北京

[2] 北京航空航天大学计算机科学与工程学院 中国北京

[3] 百度公司计算机视觉技术部 中国北京

摘要

异常检测是一项具有挑战性的任务，通常被视为一类学习问题，因为异常的不可预测性。本文提出了一种简单而强大的方法来解决这个问题，该方法在学生-教师框架中实现，利用了其优势，但在准确性和效率方面都有实质性的扩展。给定一个在图像分类上预训练的强大模型作为教师，我们将知识蒸馏到一个具有相同架构的单一学生网络中，以学习无异常图像的分布，这种一步转移尽可能地保留了关键线索。此外，我们将多尺度特征匹配策略整合到框架中，这种分层特征匹配使学生网络能够在更好的监督下接收来自特征金字塔的多层次知识混合，从而能够检测各种大小的异常。两个网络生成的特征金字塔之间的差异作为评分函数，指示异常发生的概率。由于这些操作，我们的方法实现了准确和快速的像素级异常检测。在 MVTec 异常检测数据集上取得了非常有竞争力的结果，优于最先进的方法。

1 引言

异常检测通常指识别数据集中相对于常规模式而言不典型的样本，并在视频监控[1, 31]、产品质量控制[7, 8, 27]和医疗等各种实际应用中显示出巨大潜力

* 贡献相同。
† 通讯作者。

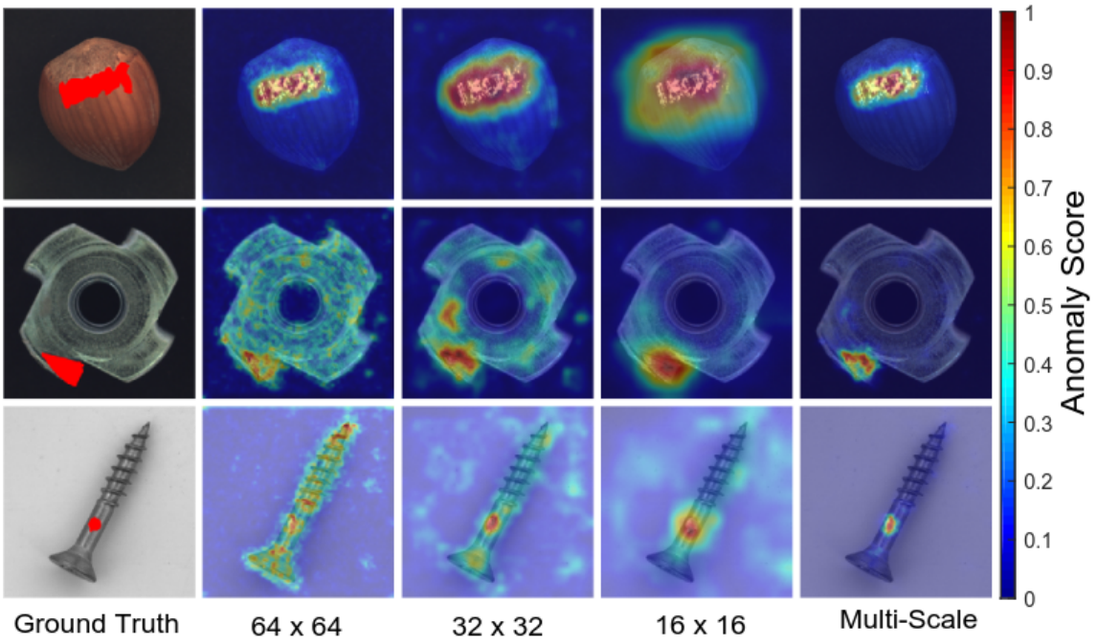


Figure 1: Visual results of our method on three defective images from the MVTec AD dataset. ResNet-18 is used as backbone and the three bottom blocks (*i.e.*, conv2_x, conv3_x, conv4_x) are selected as feature extractors. Columns from left to right correspond to input images with defects (ground truth regions in red), anomaly maps of the three blocks, and the resulting anomaly maps respectively.

diagnosis [34, 35, 40]. Its key challenge lies in the unexpectedness of anomalies which is very difficult to deal with in a supervised way, as labeling all types of anomalous instances seems unrealistic.

Previous studies address this challenge in the form of one-class learning paradigm [25]. They approximate the decision boundary for a binary classification problem by searching a feature space where the distribution of normal data is accurately modeled. Deep learning, in particular convolutional neural networks (CNNs) [20] and residual networks (ResNets) [16], provides a powerful alternative to automatically build comprehensive representations at multiple levels. Such deep features prove very effective in capturing the intrinsic characteristics of the normal data manifold [3, 10, 24, 32, 47]. Despite the promising results in their respective fields, all these methods simply predict anomalies at the image-level without spatial localization.

The pixel-level methods advance anomaly detection by means of pixel-wise comparison of image patches and their reconstructions [6, 34, 35] or per-pixel estimation of probability density on entire images [1, 37], among which Auto-encoders, Generative Adversarial Networks (GANs), and their variants are dominating models. However, their performance is prone to serious degradation when images are poorly reconstructed [30] or likelihoods are inaccurately calibrated [26].

Some recent attempts transfer the knowledge from other well-studied computer vision tasks. They directly apply the networks pre-trained on image classification and show that they are sufficiently generic to image-level detection [4, 9, 14]. Cohen and Hoshen [11] investigate this idea in pixel-level detection and delivers performance gain; unfortunately, it has the time bottleneck due to per-pixel comparison. Bergmann *et al.* [8] utilize the pre-trained model in a more efficient way by implicitly learning the distribution of normal features with a student-teacher framework and reach decent results. The difference between the outputs of the students and teacher along with the uncertainty among students' predictions serves as

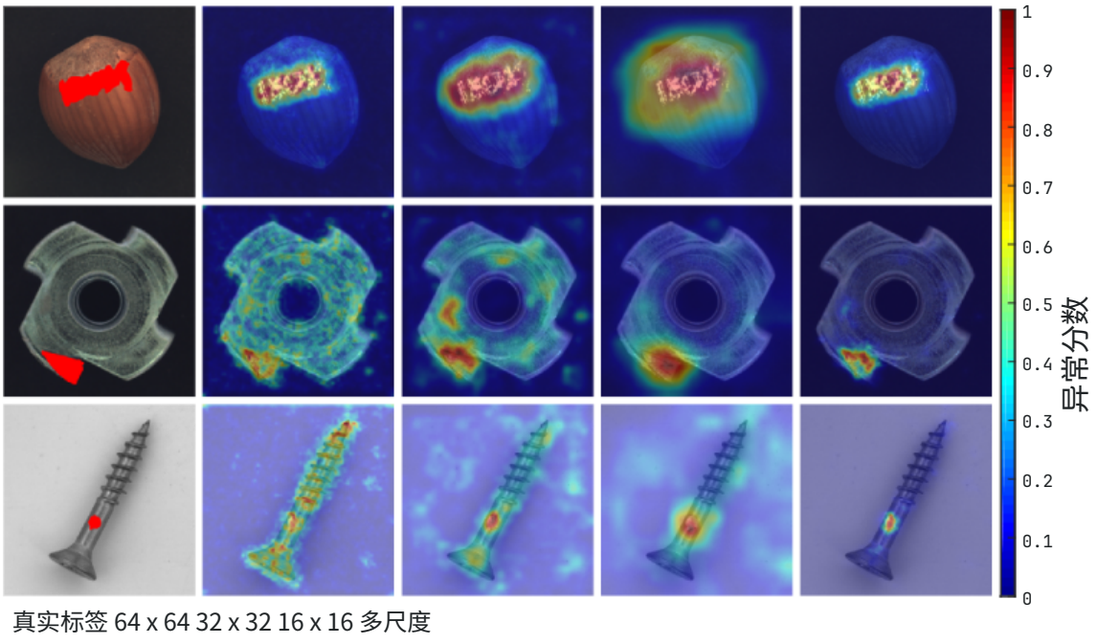


图 1：我们的方法在 MVTec AD 数据集中三张有缺陷图像上的可视化结果。使用 ResNet-18 作为骨干网络，选择底部三个块（即 conv2_x、conv3_x、conv4_x）作为特征提取器。从左到右的列分别对应带有缺陷的输入图像（红色区域为真实标签）、三个块的异常图以及最终的异常图。

诊断[34, 35, 40]。其关键挑战在于异常的不可预测性，这在监督学习中很难处理，因为标记所有类型的异常实例似乎是不现实的。

先前的研究以单类学习范式的形式解决这一挑战[25]。
他们通过搜索一个能够准确建模正常数据分布的特征空间来近似二元分类问题的决策边界。深度学习，特别是卷积神经网络（CNNs）[20]和残差网络（ResNets）[16]，提供了一种强大的替代方法，可以自动构建多层次的综合表示。这些深度特征在捕捉正常数据流形的内在特征方面非常有效[3, 10, 24, 32, 47]。尽管在各自领域取得了令人鼓舞的结果，但所有这些方法仅在图像级别预测异常，而没有空间定位。

像素级方法通过对图像块及其重建进行像素级比较[6, 34, 35]或对整个图像进行逐像素概率密度估计[1, 37]来推进异常检测，其中自编码器、生成对抗网络（GANs）及其变体是主导模型。然而，当图像重建效果不佳[30]或似然估计不准确[26]时，它们的性能容易严重下降。

一些最近的尝试从其他深入研究的计算机视觉任务中转移知识。它们直接应用在图像分类上预训练的网络,并表明这些网络对于图像级检测足够通用[4, 9, 14]。Cohen 和 Hoshen [11]在像素级检测中研究了这一想法并取得了性能提升;不幸的是,由于逐像素比较,它存在时间瓶颈。Bergmann 等人[8]以更高效的方式利用预训练模型,通过学生-教师框架隐式学习正常特征的分布,并达到了不错的结果。学生和教师输出之间的差异以及学生预测之间的不确定性作为

the anomaly scoring function. Nevertheless, two major drawbacks still remain: *i.e.*, the incompleteness of transferred knowledge and complexity of handling scaling. For the former, since knowledge is distilled from a ResNet-18 [16] into a lightweight teacher network, the big gap between their model capacities [42] tends to incur loss of important information. For the latter, multiple student-teacher ensemble pairs are required to be separately trained, each for a specific respective field, to achieve scale invariance, which leads to the inconvenience in computation. Both the facts leave much room for improvement.

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. It reports very competitive results on the MVTec anomaly detection dataset, and more results on ShanghaiTech Campus (STC) [23] and CIFAR-10 [18] are presented in the supplementary material.

2 Related Work

2.1 Image-level Anomaly Detection

Image-level techniques manifest anomalies in images of unseen categories. They can be coarsely divided into: reconstruction-based, distribution-based and classification-based.

The first group of approaches reconstruct the training images to capture the normal data manifold. An anomalous image is very likely to possess a high reconstruction error during inference, as it is drawn from a different distribution. The main weakness of these approaches comes from the excellent generalization ability of the deep models, including variational autoencoder [3], robust autoencoder [47], conditional GAN [2], and bi-directional GAN [46], which probably allows anomalous images to be faithfully reconstructed.

Distribution-based approaches model the probabilistic distribution of the normal images. The images that have low probability density values are designated as anomalous. Recent algorithms such as anomaly detection GAN (ADGAN) [12] and deep autoencoding Gaussian mixture model (DAGMM) [48] learn a deep projection that maps high-dimensional images into a low-dimensional latent space. Nevertheless, these methods have high sample com-

异常评分函数。然而，仍然存在两个主要缺点：即转移知识的不完整性和处理缩放的复杂性。对于前者，由于知识是从 ResNet-18 [16]提炼到轻量级教师网络中，它们的模型容量之间的巨大差距[42]往往会导致重要信息的丢失。对于后者，需要单独训练多个学生-教师集成对，每个对应一个特定领域，以实现尺度不变性，这导致计算上的不便。这两个事实都留下了很大的改进空间。

在本文中，我们提出了一种简单而强大的异常检测方法，该方法遵循学生-教师框架的优点，但在准确性和效率方面都有实质性的扩展。具体来说，给定一个在图像分类上预训练的强大网络作为教师，我们将知识提炼到一个具有相同架构的单一学生网络中。在这种情况下，学生网络通过将其特征与预训练网络的对应特征匹配来学习无异常图像的分布，这种一步转移尽可能地保留了关键信息。此外，为了增强尺度鲁棒性，我们将多尺度特征匹配嵌入到网络中，这种分层特征匹配策略使学生网络能够在更强的监督下接收来自特征金字塔的多层次知识混合，从而能够检测各种大小的异常（见图 1 可视化）。教师和学生网络的特征金字塔被用于比较预测，其中较大的差异表示异常发生的概率更高。

与之前的工作相比，特别是初步的学生-教师模型，我们的方法有两个优点。首先，有用的知识通过一步蒸馏很好地从预训练网络转移到学生网络，因为它们共享相同的结构。其次，由于网络的层次结构，通过提出的特征金字塔匹配方案可以方便地实现多尺度异常检测。由于这些优势，我们的方法可以进行准确和快速的像素级异常检测。它在 MVTec 异常检测数据集上报告了非常有竞争力的结果，更多关于上海科技大学校园（STC）[23]和 CIFAR-10 [18]的结果在补充材料中呈现。

2 相关工作

2.1 图像级异常检测

图像级技术可以显示未见类别图像中的异常。它们可以粗略地分为：基于重建、基于分布和基于分类的方法。

第一组方法通过重建训练图像来捕捉正常数据流形。在推理过程中，异常图像很可能具有较高的重建误差，因为它来自不同的分布。这些方法的主要弱点来自深度模型的出色泛化能力，包括变分自编码器[3]、鲁棒自编码器[47]、条件生成对抗网络[2]和双向生成对抗网络[46]，这些模型可能允许异常图像被忠实地重建。

基于分布的方法对正常图像的概率分布进行建模。
具有低概率密度值的图像被指定为异常。最近的算法，如异常检测生成对抗网络（ADGAN）[12]和深度自编码高斯混合模型（DAGMM）[48]，学习将高维图像映射到低维潜在空间的深度投影。然而，这些方法具有高样本复杂度。

plexity and demand large training data.

Classification-based approaches have dominated anomaly detection in the last decade. One useful paradigm is to feed the deep features extracted by deep generative models [9] or transferred from pre-trained networks [4, 14] into a separate shallow classification model like one-class support vector machine (OC-SVM) [36]. Another line of research depends on self-supervised learning. Geom [15] creates a dataset by applying dozens of geometric transformations to the normal images and trains a multi-class neural network over the self-labeled dataset to discriminate such transformations. At test time, anomalies are expected to be assigned with less confidence in discriminating the transformations.

2.2 Pixel-level Anomaly Detection

Pixel-level techniques are particularly designed for anomaly localization. They aim to precisely segment anomalous regions in images, which is more complicated than binary classification.

The expressive power of deep neural networks inspires a series of studies that explore how to transfer the benefits of the networks pre-trained on image classification tasks to anomaly detection. Napoletano *et al.* [27] exploit a pre-trained ResNet-18 to embed cropped training image patches into a feature space, reduce the dimension of feature vectors by PCA, and model their distribution using K-means clustering. This method requires a large number of overlapping patches to obtain a spatial anomaly map at inference time, which results in coarse-grained maps and may become a performance bottleneck.

To avoid cropping image patches and accelerate feature extraction, Sabokrou *et al.* [33] build descriptors from early feature maps of a pre-trained fully convolutional network (FCN) and adopt a unimodal Gaussian distribution to fit feature vectors of the anomaly-free images. However, the unimodel Gaussian distribution fails to characterize the training feature distribution as the problem complexity increases. More recently, a convolutional adversarial variational autoencoder with guided attention (CAVGA) [41] incorporates Grad-CAM [38] into a variational autoencoder with an attention expansion loss to encourage the deep model itself to focus on all normal regions in the image. Simliar to typical autoencoders (AE) [7, 30] and variational autoencoders (VAE) [22], CAVGA also suffers from the strong generalization ability which allows good reconstruction for anomalous images.

3 Method

3.1 Framework

We make use of the student-teacher learning framework to implicitly model the feature distribution of the normal training images. The teacher is a powerful network pre-trained on the image classification task (*e.g.*, a ResNet-18 pre-trained on ImageNet). To reduce information loss, the student shares the same architecture with the teacher. This is in essence one case of feature-based knowledge distillation [42].

Here, we need to consider a key factor, *i.e.*, position of distillation. Deep neural networks generate a pyramid of features for each input image. Bottom layers result in higher-resolution features encoding low-level information such as textures, edges and colors. By contrast, top layers yield low-resolution features that contain context information. The features created by bottom layers are often generic enough and they can be shared by various vision tasks [29,

最近的算法，如异常检测生成对抗网络（ADGAN）[12]和深度自编码高斯混合模型（DAGMM）[48]在过去十年中，基于分类的高维图像异常检测领域。学习了一种将高维图像映射到低维潜在空间，复杂性高且需要大量训练数据。一种有用的范式是将深度生成模型[9]提取的深度特征或从预训练网络[4, 14]转移的特征输入到单独的浅层分类模型中，如一类支持向量机（OC-SVM）[36]。另一条研究路线依赖于自监督学习。Geom [15]通过对正常图像应用数十种几何变换来创建数据集，并在自标记数据集上训练多类神经网络以区分这些变换。在测试时，预期异常样本在区分这些变换时会被赋予较低的置信度。

2.2 像素级异常检测

像素级技术专门用于异常定位。它们旨在精确地分割图像中的异常区域，这比二元分类更为复杂。

深度神经网络的表达能力激发了一系列研究，探索如何将预训练于图像分类任务的网络的优势转移到异常检测中。Napoletano 等人[27]利用预训练的 ResNet-18 将裁剪的训练图像块嵌入到特征空间中，通过 PCA 降低特征向量的维度，并使用 K-means 聚类对其分布进行建模。这种方法在推理时需要大量重叠的图像块来获得空间异常图，这导致了粗粒度的图，可能成为性能瓶颈。

为了避免裁剪图像块并加速特征提取，Sabokrou 等人[33]从预训练的全卷积网络（FCN）的早期特征图构建描述符，并采用单峰高斯分布来拟合无异常图像的特征向量。然而，随着问题复杂度的增加，单峰高斯分布无法很好地表征训练特征分布。最近，一种具有引导注意力的卷积对抗变分自编码器（CAVGA）[41]将 Grad-CAM [38]整合到变分自编码器中，并使用注意力扩展损失来鼓励深度模型本身关注图像中的所有正常区域。与典型的自编码器（AE）[7, 30]和变分自编码器（VAE）[22]类似，CAVGA 也存在强泛化能力的问题，这使得它能够很好地重构异常图像。

3 方法

3.1 框架

我们利用学生-教师学习框架来隐式建模正常训练图像的特征分布。教师是一个在图像分类任务上预训练的强大网络（例如，在 ImageNet 上预训练的 ResNet-18）。为了减少信息损失，学生与教师共享相同的架构。这本质上是基于特征的知识蒸馏的一种情况[42]。

在这里，我们需要考虑一个关键因素，即蒸馏的位置。深度神经网络为每个输入图像生成一个特征金字塔。底层产生更高分辨率的特征，编码低级信息，如纹理、边缘和颜色。相比之下，顶层产生低分辨率的特征，包含上下文信息。底层创建的特征通常具有足够的通用性，可以被各种视觉任务共享[29,

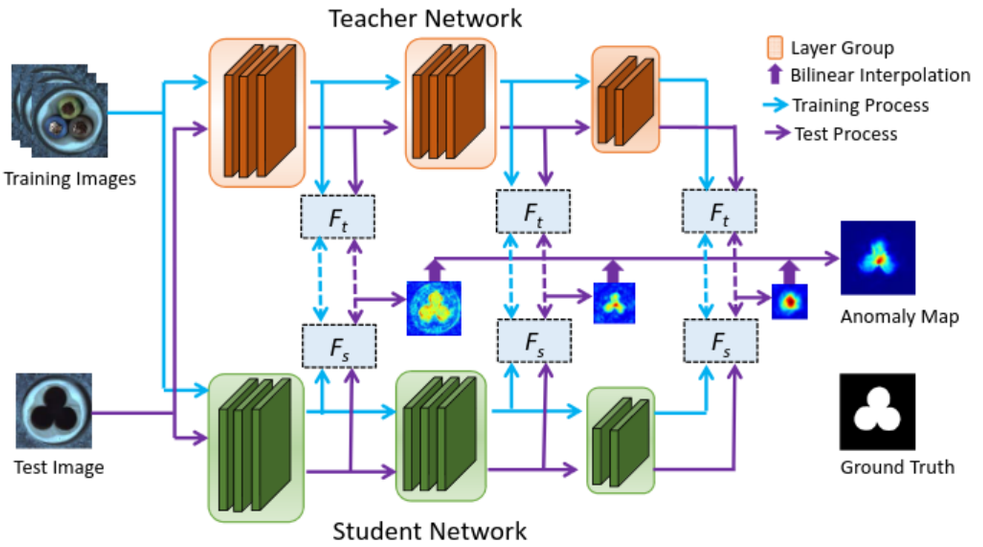


Figure 2: Schematic overview of our method. The feature pyramid of a student network is trained to match with the counterpart of a pre-trained teacher network. A test image (or pixel) has a high anomaly score if its features from the two models differ significantly. The feature pyramid matching enables our method to detect anomalies of various sizes with a single forward pass.

45]. This motivates us to integrate low-level and high-level features in a complementary way. As different layers in deep neural networks correspond to distinct receptive fields, we select the features extracted by a few successive bottom layer groups (*e.g.*, blocks in ResNet-18) of the teacher to guide the student's learning. This hierarchical feature matching allows our method to detect anomalies of various sizes.

Figure 2 gives a sketch of our method with the images from the MVTec AD dataset [8] as examples. The training and test processes are formally provided as follows.

3.2 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 $\mathcal{D} = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_n\}$, 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 $\mathbf{I}_k \in \mathbb{R}^{w \times h \times c}$, where h is the height, w is the width and c is the number of the color channels, the lth bottom layer group of the teacher and student outputs a feature map $F_t^l(\mathbf{I}_k) \in \mathbb{R}^{w_l \times h_l \times d_l}$ and $F_s^l(\mathbf{I}_k) \in \mathbb{R}^{w_l \times h_l \times d_l}$, where w_l, h_l and d_l denote the width, height and channel number of the feature map, respectively. 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. Note that $F_t^l(\mathbf{I}_k)_{ij} \in \mathbb{R}^{d_l}$ and $F_s^l(\mathbf{I}_k)_{ij} \in \mathbb{R}^{d_l}$ are feature vectors at position (i, j) in the feature maps from the teacher and student, respectively. We define the loss at position (i, j) as ℓ_2-distance between the ℓ_2-normalized feature vectors, namely,

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

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

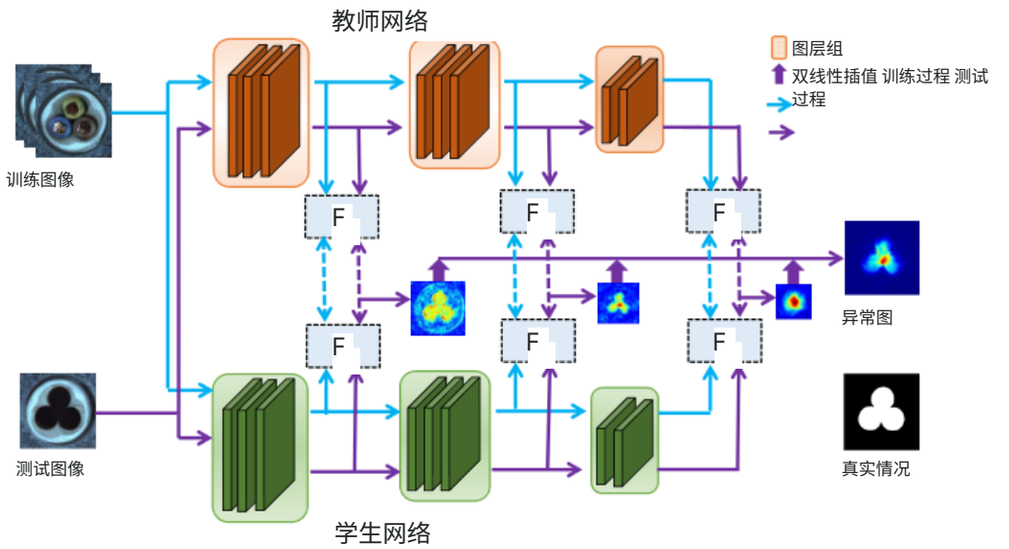


图 2：我们方法的示意概览。学生网络的特征金字塔被训练以匹配预先训练的教师网络的对应部分。如果测试图像（或像素）的特征在两个模型之间存在显著差异，则其具有高异常分数。特征金字塔匹配使我们的方法能够通过单次前向传递检测各种大小的异常。

45]。这激发我们以互补的方式整合低级和高级特征。由于深度神经网络中不同层对应不同的感受野，我们选择教师网络几个连续底层组（例如，ResNet-18 中的块）提取的特征来指导学生的学习。这种分层特征匹配使我们的方法能够检测各种大小的异常。

图 2 以 MVTec AD 数据集[8]中的图像为例，给出了我们方法的示意图。训练和测试过程正式描述如下。

3.2 训练过程

训练阶段旨在获得一个优秀的学生模型，能够在正常图像上完美模仿固定教师模型的输出。形式上，给定一个无异常图像的训练数据集 D = {I, I, . . . , I}，我们的目标是通过匹配教师模型底部 L 层组提取的特征与学生模型相应部分来捕捉正常数据流形。对于输入图像 I∈ R，其中 h 是高度，w 是宽度，c 是颜色通道数，教师和学生模型的第 l 个底层组输出特征图 F(I) ∈ R 和 F(I) ∈ R，其中 w、h 和 d 分别表示特征图的宽度、高度和通道数。由于没有关于物体外观和位置的先验知识，我们简单假设训练集中所有图像区域都是无异常的。注意，F(I)∈ R 和 F(I)∈ R 是教师和学生模型特征图中位置 (i, j) 的特征向量。我们将位置 (i, j) 的损失定义为 `-归一化特征向量之间的 `-距离，即，

$$`(I)= \frac{1}{2}\left\| \hat{F}(I)- \hat{F}(I) \right\|_{`}^{2}, \tag{1}$$

$$\hat{F}(I)= \frac{F(I)}{\|F(I)\|}, \hat{F}(I)= \frac{F(I)}{\|F(I)\|}.$$

It is worth noting that the ℓ_2 distance used in (Eq. 1) is proportional to the cosine distance as $F_t^l(\mathbf{I}_k)$ and $F_s^l(\mathbf{I}_k)$ are ℓ_2-normalized vectors. Thus the loss $\ell^l(\mathbf{I}_k)_{ij} \in (0,1)$. 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,

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

where α_l depicts the impact of the lth feature scale on anomaly detection. We simply set $\alpha_l = 1, l = 1, \ldots, L$ in all our experiments. Given a minibatch $\mathcal{B}$ sampled from the training dataset $\mathcal{D}$, we update the student by minimizing the loss $\ell_{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{k \in \mathcal{B}} \ell(\mathbf{I}_k)$. Note that we only update the student while keeping the teacher fixed throughout the training phase.

3.3 Test Process

In the test phase, we aim to obtain an anomaly map Ω 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 lth 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 $\mathbf{J}$.

4 Experiments

4.1 Dataset

We conduct experiments on the MVTec Anomaly Detection (MVTec AD) [7] 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 two popular metrics: AUC-ROC and Per-Region-Overlap (PRO) [8]. Supplementary material provides more results on ShanghaiTech Campus (STC) [23] and CIFAR-10 [18].

值得注意的是，(Eq. 1)中使用的距离与余弦距离成正比，因为 F(I)和 F(I)是`-归一化向量。因此损失`(I)∈(0, 1)。整个图像 I 的损失给出为每个位置损失的平均值，

$$`(I) = \frac{1}{wh} \sum_{=1}^{w} \sum_{=1}^{h} 我， \tag{2}$$

总损失是不同金字塔尺度下损失的加权平均

$$(我)= \sum_{=1}^{L} \alpha`(I), 满足 \alpha \geqslant 0, \tag{3}$$

其中α表示第 l 个特征尺度对异常检测的影响。在我们所有的实验中，我们简单地设置α= 1, l = 1, . . . , L。给定从训练数据集 D 中采样的小批量 B，我们通过最小化损失`= Σ`(I)来更新学生模型。请注意，我们在整个训练阶段只更新学生模型，而保持教师模型固定不变。

3.3 测试流程

在测试阶段，我们旨在获得一个大小为 w × h 的异常图Ω，用于测试图像 J ∈ R。分数Ω∈ [0, 1]表示位置(i, j)处的像素偏离训练数据流形的程度。我们将测试图像 J 输入教师和学生模型。让 F(J)和 F(J)分别表示教师和学生的第 l 个底层组生成的特征图。我们可以计算出大小为 w × h 的异常图Ω(J)，其中元素Ω(J)是位置(i, j)处的损失（等式 1）。异常图Ω(J)通过双线性插值上采样到大小 w × h。最终的异常图被定义为 L 个相同大小的上采样异常图的逐元素乘积，

$$Ω(J) = \prod_{=1}^{L} 上采样\ Ω(J)。 \tag{4}$$

如果测试图像中的任何像素是异常的，则该测试图像被指定为异常。因此，我们简单地选择异常图中的最大值，即 max(Ω(J))作为测试图像 J 的异常分数。

4 实验

4.1 数据集

我们在 MVTec 异常检测(MVTec AD)[7]数据集上进行实验，同时考虑图像级和像素级异常检测任务。该数据集专门用于对异常定位算法进行基准测试。它收集了超过 5,000 张工业产品的高分辨率图像，涵盖 15 个不同类别。对于每个类别，训练集仅包含无缺陷图像，而测试集则包含无缺陷图像和不同类型的有缺陷图像。性能通过两个常用指标来衡量：AUC-ROC 和每区域重叠(PRO)[8]。补充材料提供了在 ShanghaiTech Campus (STC)[23]和 CIFAR10[18]上的更多结果。

4.2 Implementation Details

For all the experiments, we choose the first three blocks (*i.e.*, 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.

4.3 Results

We begin with the task of finding anomalous images. As defective regions usually occupy a small proportion of the whole image, the test anomalies differ in a subtle way from the training images. This makes the MVTec AD dataset more challenging than those previously used in the literature (*e.g.*, MNIST and CIFAR-10) where the images from the other categories are regarded as anomalous to the selected one. Table 2 compares our method to state-of-the-art approaches: Geom [15], GANomaly [2], $ł_2$-AE [5], ITAE [17], Cut-Paste [21] Patch-SVDD [43], PaDiM [13] and SPADE [11]. We clearly see that our approach outperforms all the other methods. In particular, the performance is improved up to 11.7% compared with SPADE [11], which also leverages multi-scale features from a pre-trained model. It validates the superiority of the student-teacher learning framework.

We then consider the task of pixel-level anomaly detection and compare our method with the counterparts including Patch-SVDD [43], PaMiD [13], *etc*. Table 1 reports the performance in terms of the AUC-ROC and PRO metrics. We notice two trends to achieve performance gains: (1) by pre-trained models, with a Wide-ResNet50×2 network [44], SPADE reports very competitive scores; (2) by self-training techniques, Cut-Paste [21] and Patch-SVDD [43] show this potential through designing proper pretext tasks for feature learning. As our approach assumes that anomaly detection is fulfilled via the heterogeneity of the student and teacher networks, *i.e.* different network parameters learned from individual data, we employ a pre-trained model built on generic images rather than self-supervised learning on the small scale anomaly detection dataset. As Table 1 displays, our approach delivers better performance than the others. It should be noted although STAD [8] adopts the student-teacher learning framework, its performance is always inferior to that of our method. This gap can be attributed to the information loss in its two-step and single-scale knowledge transfer process. This validates our improvement in feature learning. When equipped with the same backbone as SPADE [11], our method further boosts the results, *i.e.* 0.973 and 0.923 in AUC-ROC and PRO, respectively.

5 Ablation Studies and Discussions

We first perform feature visualization to investigate what the student learns from its teacher and also conduct ablation studies on the MVTec AD dataset to answer the following three questions. Is feature pyramid matching superior to single feature matching? Is the teacher pre-trained on other datasets still useful? Is our method applicable to small training dataset?

4.2 实施细节

对于所有实验，我们选择 ResNet-18 的前三个块（即 conv2_x、conv3_x、conv4_x）作为教师网络和学生网络的金字塔特征提取器。教师网络的参数从在 ImageNet 上预训练的 ResNet-18 复制而来，而学生网络的参数则随机初始化。我们使用随机梯度下降（SGD）训练网络，学习率为 0.4，训练 100 个 epoch。批量大小为 32。训练集和测试集中的所有图像都调整为 256×256 大小。对于每个类别，我们使用 80%的训练图像来构建学生网络，保留剩余 20%用于验证。我们选择验证误差（公式 1）最低的检查点来执行异常检测。

4.3 结果

我们首先开始寻找异常图像的任务。由于缺陷区域通常只占整个图像的一小部分，测试异常与训练图像的差异很微妙。这使得 MVTec AD 数据集比文献中先前使用的数据集（如 MNIST 和 CIFAR-10）更具挑战性，在那些数据集中，其他类别的图像被视为选定类别的异常。表 2 将我们的方法与最先进的方法进行了比较：Geom [15]、GANomaly [2]、ℓ_1-AE [5]、ITAE [17]、Cut-Paste [21]、Patch-SVDD [43]、PaDiM [13]和 SPADE [11]。我们清楚地看到，我们的方法优于所有其他方法。特别是，与同样利用预训练模型多尺度特征的 SPADE [11]相比，性能提高了高达 11.7%。这验证了学生-教师学习框架的优越性。

然后，我们考虑像素级异常检测任务，并将我们的方法与包括 Patch-SVDD [43]、PaMiD [13]等对应方法进行比较。表 1 报告了 AUC-ROC 和 PRO 指标的性能。我们注意到两个趋势可以实现性能提升：(1)通过预训练模型，使用 Wide-ResNet50×2 网络[44]，SPADE 报告了非常有竞争力的分数；(2)通过自训练技术，Cut-Paste [21]和 PatchSVDD [43]通过设计适当的预训练任务来进行特征学习，展示了这种潜力。由于我们的方法假设异常检测是通过学生和教师网络的异质性来实现的，即从个别数据中学习到的不同网络参数，我们采用了基于通用图像构建的预训练模型，而不是在小规模异常检测数据集上进行自监督学习。如表 1 所示，我们的方法比其他方法表现更好。应该注意的是，尽管 STAD [8]采用了学生-教师学习框架，但其性能始终低于我们的方法。这种差距可以归因于其两步和单尺度知识转移过程中的信息损失。这验证了我们在特征学习方面的改进。当配备与 SPADE [11]相同的骨干网络时，我们的方法进一步提升了结果，即 AUC-ROC 和 PRO 分别达到 0.973 和 0.923。

5 消融研究和讨论

我们首先进行特征可视化，以研究学生从教师那里学到了什么，并在 MVTec AD 数据集上进行消融研究，以回答以下三个问题：特征金字塔匹配是否优于单一特征匹配？在其他数据集上预训练的教师是否仍然有用？我们的方法是否适用于小型训练数据集？

	Category	SSIM-AE	AnoGAN	CNN-Dict*	STAD*	Cut-Paste	Patch-SVDD	PaDiM-R18*	SPADE*	Ours*
Textures	Carpet	0.65	0.20	0.47	0.695	-	-	**0.960**	0.947	0.958
		0.87	0.54	0.72	-	0.983	0.926	**0.989**	0.975	0.988
	Grid	0.85	0.23	0.18	0.819	-	-	0.909	0.867	**0.966**
		0.94	0.58	0.59	-	0.975	0.962	0.949	0.937	**0.990**
	Leather	0.56	0.38	0.64	0.819	-	-	0.979	0.972	**0.980**
		0.78	0.64	0.87	-	**0.995**	0.974	0.991	0.976	0.993
	Tile	0.18	0.18	0.80	0.912	-	-	0.816	0.759	**0.921**
		0.59	0.50	0.93	-	0.905	0.914	0.912	0.874	**0.974**
	Wood	0.61	0.39	0.62	0.725	-	-	0.903	0.874	**0.936**
		0.73	0.62	0.91	-	0.955	0.908	0.936	0.885	**0.972**
Objects	Bottle	0.83	0.62	0.74	0.918	-	-	0.939	**0.955**	0.951
		0.93	0.86	0.78	-	0.976	0.981	0.981	0.984	**0.988**
	Cable	0.48	0.38	0.56	0.865	-	-	0.862	**0.909**	0.877
		0.82	0.78	0.79	-	0.900	0.968	0.958	**0.972**	0.955
	Capsule	0.86	0.31	0.31	0.916	-	-	0.919	**0.937**	0.922
		0.94	0.84	0.84	-	0.974	0.958	0.983	**0.990**	0.983
	Hazelnut	0.92	0.70	0.84	0.937	-	-	0.914	**0.954**	0.943
		0.97	0.87	0.72	-	0.973	0.975	0.977	**0.991**	0.985
	Metal nut	0.60	0.32	0.36	0.895	-	-	0.819	0.944	**0.945**
		0.89	0.76	0.82	-	0.931	0.980	0.967	**0.981**	0.976
	Pill	0.83	0.78	0.46	0.935	-	-	0.906	0.946	**0.965**
		0.91	0.87	0.68	-	0.957	0.951	0.947	0.965	**0.978**
	Screw	0.89	0.47	0.28	0.928	-	-	0.913	**0.960**	0.930
		0.96	0.80	0.87	-	0.967	0.957	0.974	**0.989**	0.983
	Toothbrush	0.78	0.75	0.15	0.863	-	-	0.923	**0.935**	0.922
		0.92	0.93	0.90	-	0.981	0.981	0.987	0.979	**0.989**
	Transistor	0.73	0.55	0.63	0.701	-	-	0.802	**0.874**	0.695
		0.90	0.86	0.66	-	0.930	0.970	**0.972**	0.941	0.825
	Zipper	0.67	0.47	0.70	0.933	-	-	0.947	0.926	**0.952**
		0.88	0.78	0.76	-	**0.993**	0.951	0.982	0.965	0.985
	Mean	0.69	0.44	0.52	0.857	-	-	0.901	0.917	**0.921**
		0.87	0.74	0.78	-	0.960	0.957	0.967	0.965	**0.970**

* denotes extra dataset pre-trained model used.

Table 1: Pixel-level anomaly detection. For each dataset category, PRO (top row) and AUC-ROC (bottom row) scores are given.

Geom	GANomaly	ł_2-AE	ITAE	Cut-Paste	Patch-SVDD	PaDiM-WR50*	SPADE*	Ours
0.672	0.762	0.754	0.839	0.952	0.921	0.953	0.855	**0.955**

* denotes extra dataset pre-trained model used.

Table 2: Image-level anomaly detection. The performance is measured by average AUC-ROC across 15 categories.

5.1 Feature Visualization

Figure 3 shows t-SNE visualization [39] of learned features from the student and teacher. Obviously, the features from the student and teacher on normal regions distribute closer (even overlapped) than the ones on anomalous regions. It suggests that the student learns to match the teacher's output on normal images. It also shows that the student well captures the distribution of normal patterns under the supervision of a good teacher.

5.2 Feature Matching

We first minutely investigate the effectiveness of feature extraction by each individual block of ResNet-18. Considering that the first block is a simple convolutional layer, we exclude it from comparison. We train the student by matching features extracted by its second, third, fourth and fifth blocks with the counterparts of the teacher respectively. As shown in Table 3, feature matching conducted at the end of the third and fourth blocks can achieve better performance. This is in good agreement with the previous discovery that the middle-level features play a more important role in knowledge transfer [29].

	类别	SSIM-AE	AnoGAN	CNN-Dict	STAD	Cut-Paste	Patch-SVDD	PaDiM-R18	SPADE	我们的方法
纹理	地毯	0.65	0.20	0.47	0.695	-	-	0.960	0.947	0.958
		0.87	0.54	0.72	-	0.983	0.926	0.989	0.975	0.988
	Grid	0.85	0.23	0.18	0.819	-	-	0.909	0.867	0.966
		0.94	0.58	0.59	-	0.975	0.962	0.949	0.937	0.990
	皮革	0.56	0.38	0.64	0.819	-	-	0.979	0.972	0.980
		0.78	0.64	0.87	-	0.995	0.974	0.991	0.976	0.993
	Tile	0.18	0.18	0.80	0.912	-	-	0.816	0.759	0.921
		0.59	0.50	0.93	-	0.905	0.914	0.912	0.874	0.974
	Wood	0.61	0.39	0.62	0.725	-	-	0.903	0.874	0.936
		0.73	0.62	0.91	-	0.955	0.908	0.936	0.885	0.972
对象	瓶子	0.83	0.62	0.74	0.918	-	-	0.939	0.955	0.951
		0.93	0.86	0.78	-	0.976	0.981	0.981	0.984	0.988
	电缆	0.48	0.38	0.56	0.865	-	-	0.862	0.909	0.877
		0.82	0.78	0.79	-	0.900	0.968	0.958	0.972	0.955
	胶囊	0.86	0.31	0.31	0.916	-	-	0.919	0.937	0.922
		0.94	0.84	0.84	-	0.974	0.958	0.983	0.990	0.983
	榛子	0.92	0.70	0.84	0.937	-	-	0.914	0.954	0.943
		0.97	0.87	0.72	-	0.973	0.975	0.977	0.991	0.985
	金属螺母	0.60	0.32	0.36	0.895	-	-	0.819	0.944	0.945
		0.89	0.76	0.82	-	0.931	0.980	0.967	0.981	0.976
	Pill	0.83	0.78	0.46	0.935	-	-	0.906	0.946	0.965
		0.91	0.87	0.68	-	0.957	0.951	0.947	0.965	0.978
	螺丝	0.89	0.47	0.28	0.928	-	-	0.913	0.960	0.930
		0.96	0.80	0.87	-	0.967	0.957	0.974	0.989	0.983
	牙刷	0.78	0.75	0.15	0.863	-	-	0.923	0.935	0.922
		0.92	0.93	0.90	-	0.981	0.981	0.987	0.979	0.989
	晶体管	0.73	0.55	0.63	0.701	-	-	0.802	0.874	0.695
		0.90	0.86	0.66	-	0.930	0.970	0.972	0.941	0.825
	拉链	0.67	0.47	0.70	0.933	-	-	0.947	0.926	0.952
		0.88	0.78	0.76	-	0.993	0.951	0.982	0.965	0.985
	Mean	0.69	0.44	0.52	0.857	-	-	0.901	0.917	0.921
		0.87	0.74	0.78	-	0.960	0.957	0.967	0.965	0.970

* 表示使用了额外数据集预训练的模型。

表 1: 像素级异常检测。对于每个数据集类别，给出了 PRO（上行）和 AUCROC（下行）分数。

Geom	GANomaly	ɨ-AE	ITAE	Cut-Paste	Patch-SVDD	PaDiM-WR50	SPADE	我们的方法
0.672	0.762	0.754	0.839	0.952	0.921	0.953	0.855	0.955

* 表示使用了额外的数据集预训练模型。

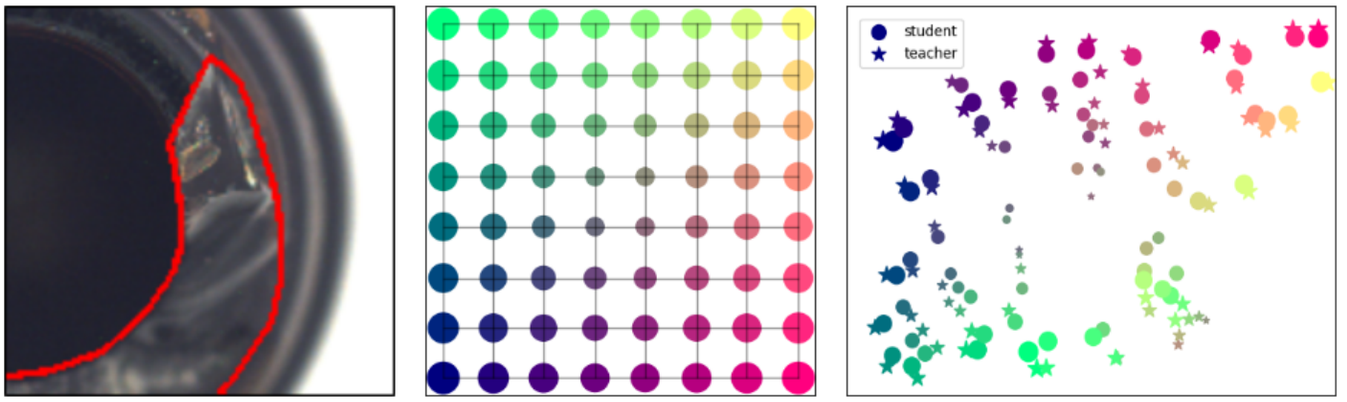
表 2：图像级异常检测。性能通过 15 个类别的平均 AUCROC 进行衡量。

5.1 特征可视化

图 3 显示了学生和教师学习到的特征的 t-SNE 可视化[39]。显然,在正常区域上,学生和教师的特征分布更接近(甚至重叠),而在异常区域上则不然。这表明学生学会了在正常图像上匹配教师的输出。它还表明,在一个优秀教师的监督下,学生很好地捕捉到了正常模式的分布。

5.2 特征匹配

我们首先详细研究了 ResNet-18 的每个单独模块进行特征提取的有效性。考虑到第一个模块是一个简单的卷积层，我们将其排除在比较之外。我们通过将学生模型的第二、第三、第四和第五个模块提取的特征分别与教师模型的对应部分进行匹配来训练学生模型。如表 3 所示，在第三和第四个模块末端进行的特征匹配可以获得更好的性能。这与之前发现的中层特征在知识迁移中发挥更重要作用的结论相一致[29]。



(a) Bottle (broken_large) (b) Points at various positions (c) t-SNE

Figure 3: t-SNE visualization [39] of learned features from the student and teacher. (a) an example test image with defects contoured by a red line. (b) point map in which different positions are encoded by different sizes and colors. (c) t-SNE visualizations for features from the student (circle) and the teacher (star) with (a) as input. Zoomed in for better display.

Metric \ # Block	2	3	4	5	[2, 3]	[2, 3, 4]	[2, 3, 4, 5]
AR_I	0.808	0.917	0.934	0.819	0.849	**0.955**	0.949
AR_P	0.915	0.953	0.957	0.860	0.950	**0.970**	0.969
PRO	0.815	0.897	0.835	0.504	0.886	**0.921**	0.886

Table 3: Ablation studies for feature matching. The performance is measured by the average image-level AUC-ROC (AR_I), average pixel-level AUC-ROC (AR_P) and average PRO across 15 categories.

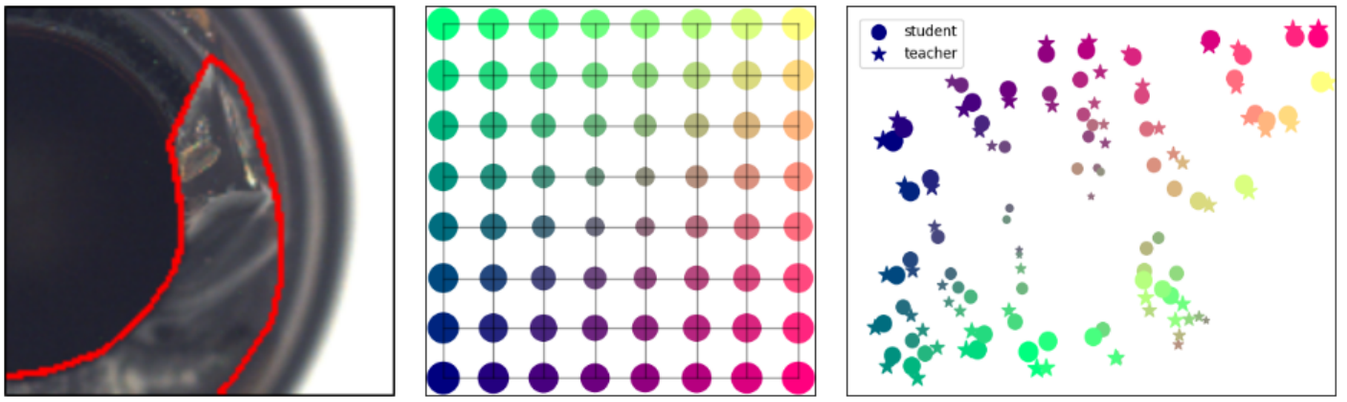
Metric \ Dataset	ImageNet	MNIST	CIFAR-10	CIFAR-100	SVHN
AR_I	**0.955**	0.619	0.826	0.835	0.796
AR_P	**0.970**	0.759	0.931	0.937	0.902
PRO	**0.921**	0.528	0.863	0.842	0.742

Table 4: Ablation studies for pre-trained datasets. The performance is measured by the average image-level AUC-ROC (AR_I), average pixel-level AUC-ROC (AR_P) and average PRO across 15 categories.

We then test three different combinations of the consecutive blocks of ResNet-18. Likewise, we match the features extracted from the corresponding compound blocks of the teacher and the student. Table 3 shows that the mixture of the second, third and fourth blocks outperforms other combinations as well as the single components. It implies that feature pyramid matching is a better way for feature learning. This finding is also validated in Figure 1. Anomaly maps generated by low-level features are more suitable for precise anomaly localization, but they are likely to include background noise. By contrast, anomaly maps generated by high-level features are able to segment big anomalous regions. The aggregation of anomaly maps at different scales contributes to accurate detection of anomalies of various sizes.

5.3 Pre-trained Datasets

To answer the second question, we pre-train the teacher on a couple of image classification benchmarks, including MNIST [19], CIFAR-10 [18], CIFAR-100 [18], and SVHN [28].



(a) 瓶子（大破损） (b) 各个位置的点 (c) t-SNE

图 3：学生和教师学习特征的 t-SNE 可视化[39]。(a) 一个缺陷被红线轮廓标出的测试图像示例。(b) 点图，其中不同位置用不同大小和颜色编码。(c) 以(a)为输入的学生（圆圈）和教师（星星）特征的 t-SNE 可视化。放大以便更好地显示。

指标 \ # 块	2	3	4	5	[2, 3]	[2, 3, 4]	[2, 3, 4, 5]
AR	0.808	0.917	0.934	0.819	0.849	0.955	0.949
AR	0.915	0.953	0.957	0.860	0.950	0.970	0.969
PRO	0.815	0.897	0.835	0.504	0.886	0.921	0.886

表 3：特征匹配的消融研究。性能通过 15 个类别的平均图像级 AUC-ROC（AR）、平均像素级 AUC-ROC（AR）和平均 PRO 来衡量。

指标 \ 数据集	ImageNet	MNIST	CIFAR-10	CIFAR-100	SVHN
AR	0.955	0.619	0.826	0.835	0.796
AR	0.970	0.759	0.931	0.937	0.902
PRO	0.921	0.528	0.863	0.842	0.742

表 4：预训练数据集的消融研究。性能通过 15 个类别的平均图像级 AUC-ROC（AR）、平均像素级 AUC-ROC（AR）和平均 PRO 来衡量。

然后，我们测试了 ResNet-18 连续块的三种不同组合。同样，我们匹配了从教师和学生相应复合块提取的特征。表 3 显示，第二、第三和第四块的混合优于其他组合以及单个组件。这意味着特征金字塔匹配是一种更好的特征学习方式。这一发现在图 1 中也得到了验证。由低级特征生成的异常图更适合精确的异常定位，但它们可能包含背景噪声。相比之下，由高级特征生成的异常图能够分割大的异常区域。不同尺度异常图的聚合有助于准确检测各种大小的异常。

5.3 预训练数据集

为了回答第二个问题，我们在几个图像分类基准数据集上预训练教师模型，包括 MNIST [19]、CIFAR-10 [18]、CIFAR-100 [18]和 SVHN [28]。

	5%		10%	
Metric	Ours	SPADE	Ours	SPADE
AR_I	**0.871**	0.782	**0.907**	0.797
AR_P	**0.961**	0.932	**0.967**	0.955
PRO	**0.892**	0.842	**0.913**	0.890

Table 5: Performance in terms of the number of training samples. The performance is measured by the average image-level AUC-ROC (AR_I), average pixel-level AUC-ROC (AR_P) and average PRO across 15 categories.

These pre-trained teachers are individually exploited to guide the student training. The MNIST and SVHN datasets simply contain digital numbers from 0 to 9. We see from Table 4 that the teacher networks pre-trained on these two datasets yield worse results. It indicates that the features learned from these two pre-trained models generalize poorly on the MVTec AD dataset. By contrast, the features extracted from the teacher networks pre-trained on CIFAR-10 and CIFAR-100 exhibit better generalization, as they contain more natural images. Note that the performance of these two pre-trained teachers is still inferior to that of the teacher pre-trained on ImageNet. This is because that the ImageNet dataset consists of a huge number of high-resolution natural images, which is crucial to learning more discriminating features.

5.4 Number of Training Samples

We investigate the effect of the training set size in this experiment. Only 5% and 10% anomaly-free images are used to train our model. It can be seen in Table 5 that our model still reaches a satisfactory level even if only a few training images are available. By contrast, SPADE suffers a serious performance degradation. This is caused by the missing of the tailored feature learning. Our model profits from this strategy and can capture the feature distribution of anomaly-free images in the few-shot scenario. Furthermore, our method uses only 10% training samples to outperform the preliminary student-teacher framework [8]. It validates the effectiveness of our feature pyramid matching technique.

6 Conclusion

We present a new feature pyramid matching technique and incorporate it into the student-teacher anomaly detection framework. Given a powerful network pre-trained on image classification as the teacher, we use its different levels of features to guide a student network with the same structure to learn the distribution of anomaly-free images. On account of the hierarchical feature matching, our method is capable of detecting anomalies of various sizes with only a single forward pass. Experimental results on the MVTec AD dataset show that our method achieves superior performance to the state-of-the-art.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (62022011), the Research Program of State Key Laboratory of Software Development Environment (SKLSDE-2021ZX-04), and the Fundamental Research Funds for the Central Universities.

10% 指标 我们的方法 SPADE 我们的方法 5% SPADE AR0.871 0.782 0.907 0.797 AR0.961 0.932 0.967 0.955 PRO 0.892 0.842 0.913 0.890 表 5：根据训练样本数量的性能。性能是通过 15 个类别的平均图像级 AUC-ROC（AR）、平均像素级 AUC-ROC（AR）和平均 PRO 来衡量的。

这些预训练的教师模型被单独用于指导学生模型的训练。MNIST 和 SVHN 数据集仅包含 0 到 9 的数字。从表 4 中我们可以看到，在这两个数据集上预训练的教师网络产生了较差的结果。这表明从这两个预训练模型中学到的特征在 MVTec AD 数据集上泛化能力较差。相比之下，从在 CIFAR-10 和 CIFAR-100 上预训练的教师网络中提取的特征表现出更好的泛化能力，因为它们包含更多的自然图像。请注意，这两个预训练教师的性能仍然不如在 ImageNet 上预训练的教师。这是因为 ImageNet 数据集包含大量高分辨率的自然图像，这对于学习更具辨别力的特征至关重要。

5.4 训练样本数量

在这个实验中，我们研究了训练集大小的影响。只使用 5%和 10%的无异常图像来训练我们的模型。从表 5 可以看出，即使只有少量训练图像可用，我们的模型仍然达到了令人满意的水平。相比之下，SPADE 遭受了严重的性能下降。这是由于缺少定制的特征学习造成的。我们的模型受益于这种策略，能够在少样本场景中捕捉无异常图像的特征分布。此外，我们的方法仅使用 10%的训练样本就超过了初步的学生-教师框架[8]。这验证了我们的特征金字塔匹配技术的有效性。

6 结论

我们提出了一种新的特征金字塔匹配技术，并将其纳入学生-教师异常检测框架中。给定一个在图像分类任务上预训练的强大网络作为教师，我们使用其不同层级的特征来指导具有相同结构的学生网络学习无异常图像的分布。由于采用了分层特征匹配，我们的方法能够在单次前向传播中检测各种大小的异常。在 MVTec AD 数据集上的实验结果表明，我们的方法达到了优于现有最先进技术的性能。

致谢

本研究得到了国家自然科学基金（62022011）、软件开发环境国家重点实验室研究计划（SKLSDE2021ZX-04）以及中央高校基本科研业务费的支持。

References

[1] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *CVPR*, 2019.

[2] Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. GANomaly: Semi-supervised anomaly detection via adversarial training. In *ACCV*, 2018.

[3] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probabiliy. Technical report, SNU Data Mining Center, 2015.

[4] Jerone T. A. Andrews, Thomas Tanay, Edward J. Morton, and Lewis D. Griffin. Transfer representation-learning for anomaly detection. In *ICML Workshops*, 2016.

[5] Caglar Aytekin, Xingyang Ni, Francesco Cricri, and Emre Aksu. Clustering and unsupervised anomaly detection with l_2 normalized deep auto-encoder representations. In *IJCNN*, 2018.

[6] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *MICCAI Workshops*, 2018.

[7] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec AD - A comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 2019.

[8] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *CVPR*, 2020.

[9] Philippe Burlina, Neil Joshi, and I-Jeng Wang. Where's wally now? deep generative and discriminative embeddings for novelty detection. In *CVPR*, 2019.

[10] Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. Anomaly detection using one-class neural networks. *arXiv:1802.06360*, 2018.

[11] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv:2005.02357*, 2020.

[12] Lucas Deecke, Robert Vandermeulen, Lukas RuffStephan Mandt, and Marius Kloft. Image anomaly detection with generative adversarial networks. In *ECML-PKDD*, pages 3–17, 2018.

[13] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In *ICPR*, 2021.

[14] Sarah M. Erfani, Sutharshan Rajasegarar, Shanika Karunasekera, and Christopher Leckie. High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recognit.*, 58:121–134, 2016.

[15] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *NeurIPS*, 2018.

参考文献

[1] Davide Abati, Angelo Porrello, Simone Calderara 和 Rita Cucchiara. 用于新颖性检测的潜在空间自回归. 发表于 CVPR, 2019.

[2] Samet Akcay, Amir Atapour-Abarghouei 和 Toby P. Breckon. GANomaly：通过对抗训练进行半监督异常检测。发表于 ACCV, 2018。

[3] Jinwon An 和 Sungzoon Cho. 基于重构概率的变分自编码器异常检测方法. 技术报告, SNU 数据挖掘中心, 2015.

[4] Jerone T. A. Andrews, Thomas Tanay, Edward J. Morton 和 Lewis D. Griffin. 用于异常检测的迁移表示学习. 发表于 ICML 研讨会, 2016.

[5] Caglar Aytekin, Xingyang Ni, Francesco Cricri 和 Emre Aksu. 使用 l 归一化深度自编码器表示进行聚类和无监督异常检测. 发表于 IJCNN, 2018.

[6] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni 和 Nassir Navab. 用于脑部 MR 图像中无监督异常分割的深度自编码模型. 发表于 MICCAI 研讨会, 2018.

[7] Paul Bergmann, Michael Fauser, David Sattlegger 和 Carsten Steger. Mvtec AD - 一个用于无监督异常检测的全面真实世界数据集。发表于 CVPR, 2019。

[8] Paul Bergmann, Michael Fauser, David Sattlegger 和 Carsten Steger. 无知学生：具有判别性潜在嵌入的学生-教师异常检测. 发表于 CVPR, 2020.

[9] Philippe Burlina, Neil Joshi, 和 I-Jeng Wang. 威利现在在哪里？用于新颖性检测的深度生成和判别嵌入。发表于 CVPR, 2019。

[10] Raghavendra Chalapathy, Aditya Krishna Menon 和 Sanjay Chawla. 使用单类神经网络进行异常检测. arXiv:1802.06360, 2018.

[11] Niv Cohen 和 Yedid Hoshen. 使用深度金字塔对应关系进行子图像异常检测. arXiv:2005.02357, 2020.

[12] Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt 和 Marius Kloft。使用生成对抗网络进行图像异常检测。发表于 2018 年 ECML-PKDD 会议论文集,第 3-17 页。

[13] Thomas Defard, Aleksandr Setkov, Angelique Loesch 和 Romaric Audigier. Padim: 一种用于异常检测和定位的补丁分布建模框架。发表于 ICPR, 2021。

[14] Sarah M. Erfani, Sutharshan Rajasegarar, Shanika Karunasekera, 和 Christopher Leckie. 使用线性一类 SVM 和深度学习进行高维和大规模异常检测。模式识别，58:121–134，2016。

[15] Izhak Golan 和 Ran El-Yaniv. 使用几何变换的深度异常检测. 发表于 NeurIPS, 2018.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[17] Chaoqin Huang, Fei Ye, Jinkun Cao, Maosen Li, Ya Zhang, and Cewu Lu. Attribute restoration framework for anomaly detection. *arXiv:1911.10676*, 2020.

[18] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[19] Yann LeCun and Corinna Cortes. Mnist handwritten digit database. 2010.

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

[21] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, 2021.

[22] Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J Radke, and Octavia Camps. Towards visually explaining variational autoencoders. In *CVPR*, 2020.

[23] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *ICCV*, 2017.

[24] Marc Masana, Idoia Ruiz, Joan Serrat, Van De Weijer Joost, and Antonio M Lopez. Metric learning for novelty and anomaly detection. In *BMVC*, 2018.

[25] M. M. Moya, M. W. Koch, and L. D. Hostetler. One-class classifier networks for target recognition applications. In *WCCI*, 1993.

[26] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *ICLR*, 2019.

[27] Paolo Napoletano, Flavio Piccoli, and Raimondo Schettini. Anomaly detection in nanofibrous materials by CNN-based self-similarity. *Sensors*, 18(2):209, 2018.

[28] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshops*, 2011.

[29] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.

[30] Michael Fauser David Sattlegger Paul Bergmann, Sindy Löwe and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *VISIGRAPP*, 2019.

[31] Alina Roitberg, Ziad Al-Halah, and Rainer Stiefelhagen. Informed democracy: Voting-based novelty detection for action recognition. In *BMVC*, 2018.

[32] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib A. Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *ICML*, 2018.

[16] 何恺明、张翔宇、任少卿和孙剑。深度残差学习用于图像识别。发表于 CVPR，2016 年。

[17] 黄超勤、叶飞、曹金坤、李茂森、张雅和陆策吾。用于异常检测的属性恢复框架。arXiv:1911.10676，2020 年。

[18] 亚历克斯·克里热夫斯基和杰弗里·辛顿。从微小图像中学习多层特征。技术报告，多伦多大学，2009 年。

[19] 扬·勒昆和科琳娜·科尔特斯。MNIST 手写数字数据库。2010 年。

[20] Yann LeCun, Léon Bottou, Yoshua Bengio 和 Patrick Haffner。基于梯度的学习应用于文档识别。IEEE 会刊，86(11):2278–2324，1998。

[21] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon 和 Tomas Pfister。CutPaste：用于异常检测和定位的自监督学习。发表于 CVPR，2021。

[22] Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J Radke 和 Octavia Camps。朝向视觉解释变分自编码器。发表于 CVPR，2020。

[23] Weixin Luo, Wen Liu, 和 Shenghua Gao. 在堆叠 RNN 框架中重新审视基于稀疏编码的异常检测。发表于 ICCV，2017 年。

[24] Marc Masana, Idoia Ruiz, Joan Serrat, Van De Weijer Joost, 和 Antonio M Lopez。用于新颖性和异常检测的度量学习。发表于 BMVC，2018 年。

[25] M. M. Moya, M. W. Koch, 和 L. D. Hostetler. 用于目标识别应用的单类分类器网络。发表于 WCCI，1993 年。

[26] Eric Nalisnick、Akihiro Matsukawa、Yee Whye Teh、Dilan Gorur 和 Balaji Lakshminarayanan。深度生成模型知道它们不知道什么吗？发表于 ICLR，2019 年。

[27] Paolo Napoletano、Flavio Piccoli 和 Raimondo Schettini。基于 CNN 自相似性的纳米纤维材料异常检测。传感器，18(2):209，2018 年。

[28] Yuval Netzer、Tao Wang、Adam Coates、Alessandro Bissacco、Bo Wu 和 Andrew Y. Ng。使用无监督特征学习读取自然图像中的数字。发表于 NeurIPS 研讨会，2011 年。

[29] Maxime Oquab、Léon Bottou、Ivan Laptev 和 Josef Sivic。使用卷积神经网络学习和转移中层图像表示。发表于 CVPR，2014 年。

[30] Michael Fauser David Sattlegger Paul Bergmann、Sindy Löwe 和 Carsten Steger。改进自编码器的无监督缺陷分割，通过应用结构相似性。发表于 VISIGRAPP，2019 年。

[31] Alina Roitberg、Ziad Al-Halah 和 Rainer Stiefelhagen。知情民主：基于投票的动作识别新颖性检测。发表于 BMVC，2018 年。

[32] Lukas Ruff、Robert A. Vandermeulen、Nico Görnitz、Lucas Deecke、Shoaib A. Siddiqui、Alexander Binder、Emmanuel Müller 和 Marius Kloft。深度单类分类。发表于 ICML，2018 年。

[33] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, Zahra Moayed, and Reinhard Klette. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *CVIU*, 172, 2018.

[34] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *IPMI*, 2017.

[35] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *MED IMAGE ANAL*, 54:30–44, 2019.

[36] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *NEURAL COMPUT*, 13(7), 2001.

[37] Philipp Seeböck, Sebastian Waldstein, Sophie Klimscha, Bianca S. Gerendas René Donner, Thomas Schlegl, Ursula Schmidt-Erfurth, and Georg Langs. Identifying and categorizing anomalies in retinal imaging data. *arXiv:1612.00686*, 2016.

[38] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, and Devi Parikh. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[39] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9 (11), 2008.

[40] Aleksei Vasilev, Vladimir Golkov, Ilona Lipp, Eleonora Sgarlata, Valentina Tomassini, Derek K. Jones, and Daniel Cremers. q-Space novelty detection with variational autoencoders. *arXiv:1806.02997*, 2018.

[41] Shashanka Venkataramanan, Kuan-Chuan Peng, Rajat Vikram Singh, and Abhijit Mahalanobis. Attention guided anomaly localization in images. In *ECCV*, 2020.

[42] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *arXiv preprint arXiv:2004.05937*, 2020.

[43] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *ACCV*, 2020.

[44] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[45] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

[46] Houssam Zenati, Manon Romain, Chuan-Sheng Foo, Bruno Lecouat, and Vijay Chandrasekhar. Adversarially learned anomaly detection. In *ICDM*, 2018.

[47] Chong Zhou and Randy C. Paffenroth. Anomaly detection with robust deep autoencoders. In *KDD*, 2017.

[33] 穆罕默德·萨博克鲁、莫森·法亚兹、马哈茂德·法西、扎赫拉·莫耶德和赖因哈德克莱特。Deep-anomaly：用于人群场景快速异常检测的全卷积神经网络。CVIU，172，2018。

[34] 托马斯·施莱格尔、菲利普·西博克、塞巴斯蒂安·M·瓦尔德斯坦、乌尔苏拉·施密特-埃尔富特和格奥尔格·朗斯。使用生成对抗网络进行无监督异常检测以指导标记发现。在 IPMI，2017。

[35] Thomas Schlegl、Philipp Seeböck、Sebastian M. Waldstein、Georg Langs 和 Ursula Schmidt-Erfurth。f-AnoGAN：使用生成对抗网络的快速无监督异常检测。医学图像分析，54:30-44，2019。

[36] Bernhard Schölkopf、John C. Platt、John Shawe-Taylor、Alex J. Smola 和 Robert C. Williamson。估计高维分布的支持。神经计算，13(7)，2001。

[37] Philipp Seeböck、Sebastian Waldstein、Sophie Klimscha、Bianca S. Gerendas René Donner、Thomas Schlegl、Ursula Schmidt-Erfurth 和 Georg Langs。识别和分类视网膜成像数据中的异常。arXiv:1612.00686，2016 年。

[38] Ramprasaath R. Selvaraju、Michael Cogswell、Abhishek Das、Ramakrishna Vedantam、和 Devi Parikh。Grad-CAM：通过基于梯度的定位从深度网络获得视觉解释。发表于 ICCV，2017 年。

[39] Laurens Van der Maaten 和 Geoffrey Hinton。使用 t-SNE 可视化数据。JMLR，9(11)，2008。

[40] Aleksei Vasilev、Vladimir Golkov、Ilona Lipp、Eleonora Sgarlata、Valentina Tomassini、Derek K. Jones 和 Daniel Cremers。使用变分自编码器进行 q 空间新颖性检测。arXiv:1806.02997，2018。

[41] Shashanka Venkataramanan、Kuan-Chuan Peng、Rajat Vikram Singh 和 Abhijit Mahalanobis。图像中的注意力引导异常定位。在 ECCV，2020。

[42] 林王和国金允。知识蒸馏和学生-教师学习用于

视觉智能：综述和新展望。arXiv 预印本 arXiv:2004.05937，2020 年。

[43] 易智勋和尹成洛。Patch SVDD：用于异常检测和分割的补丁级 SVDD。发表于 ACCV，2020 年。

[44] 谢尔盖·扎戈鲁伊科和尼科斯·科莫达基斯。宽残差网络。arXiv 预印本 arXiv:1605.07146，2016 年。

[45] Matthew D. Zeiler 和 Rob Fergus。可视化和理解卷积网络。发表于 ECCV，2014 年。

[46] Houssam Zenati、Manon Romain、Chuan-Sheng Foo、Bruno Lecouat 和 Vijay Chandrasekhar。对抗性学习的异常检测。发表于 ICDM，2018 年。

[47] Chong Zhou 和 Randy C. Paffenroth。使用鲁棒深度自编码器进行异常检测。发表于 KDD，2017 年。

[48] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *ICLR*, 2018.

[48] 宗博、宋琦、闵仁强、程伟、卢梅扎努・克里斯蒂安、
赵大基和陈海峰。用于无监督异常检测的深度自编码高斯混合模型。发表于 ICLR，2018 年。