

通过单类嵌入的反向蒸馏进行异常检测

韩秋 邓星宇 李 阿尔伯塔大学电气与计算机工程
系 {hanqiu1,xingyu}@ualberta.ca

摘要

Knowledge distillation (KD) achieves promising results on the challenging problem of unsupervised anomaly detection (AD). The representation discrepancy of anomalies in the teacher-student (T-S) model provides essential evidence for AD. However, using similar or identical architectures to build the teacher and student models in previous studies hinders the diversity of anomalous representations. To tackle this problem, we propose a novel T-S model consisting of a teacher encoder and a student decoder and introduce a simple yet effective "reverse distillation" paradigm accordingly. Instead of receiving raw images directly, the student network takes teacher model's one-class embedding as input and targets to restore the teacher's multi-scale representations. Inherently, knowledge distillation in this study starts from abstract, high-level presentations to low-level features. In addition, we introduce a trainable one-class bottleneck embedding (OCBE) module in our T-S model. The obtained compact embedding effectively preserves essential information on normal patterns, but abandons anomaly perturbations. Extensive experimentation on AD and one-class novelty detection benchmarks shows that our method surpasses SOTA performance, demonstrating our proposed approach's effectiveness and generalizability.

1. 引言

异常检测 (AD) 指在有限甚至无异常先验知识的情况下识别和定位异常。其广泛应用，如工业缺陷检测[3]、医学分布外检测[50]和视频监控[24]，使其成为一项关键任务并备受关注。在无监督异常检测的背景下，无法获得关于异常的先前信息，而是提供一组正常样本作为参考。为解决此问题，先前研究尝试在这些无异常样本上构建各种自监督任务，包括但不限于样本重建[2,5,11,16,26,34,38,48]、伪异常增广等。

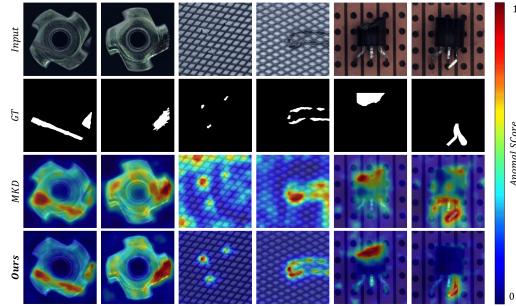


图1. MVTec上的异常检测示例[3]。多分辨率知识蒸馏（MKD）[33]采用了图2(a)中的传统KD架构。我们的反向蒸馏方法能够精确定位各类异常。

表示法 [23, 42, 46]、知识蒸馏 [4, 33, 39] 等。

在本研究中，我们从基于知识蒸馏的角度解决无监督异常检测问题。在知识蒸馏 (KD) [6, 15]中，知识在教师-学生 (T-S) 对之间传递。在无监督异常检测的背景下，由于学生在训练过程中仅接触正常样本，当查询样本异常时，学生很可能生成与教师模型存在差异的表征。这一假设构成了基于知识蒸馏的异常检测方法的基础。然而，由于以下两个原因，这一假设在实践中并不总是成立：(1) 教师网络与学生网络结构相同或相似（即存在无法区分的滤波器[33]）；(2) 知识传递/蒸馏过程中T-S模型的数据流相同。尽管使用更小的学生网络能在一定程度上缓解此问题[33, 39]，但浅层架构较弱的表征能力会阻碍模型精确检测和定位异常。

为全面解决上述问题，我们提出了一种新的知识蒸馏范式，即*Reverse Distillation*，用于异常检测。我们通过图2中的简单示意图来强调传统知识蒸馏与所提出的反向蒸馏之间的系统性差异。首先，与传统知识蒸馏框架中教师和学生均采用编码器结构不同，我们的T-S模型中

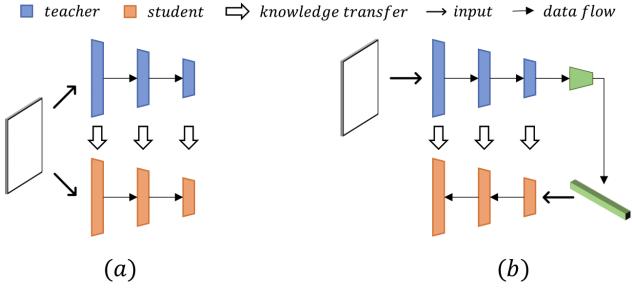


图2. (a) 传统知识蒸馏框架[6, 33]与(b) 我们Reverse Distillation范式中的师生模型及数据流。

反向蒸馏由异构架构组成：一个教师编码器和一个学生解码器。其次，学生解码器并非将原始数据直接同时输入T-S模型，而是以低维嵌入作为输入，旨在通过重建教师模型在不同尺度上的表征来模仿教师的行为。从回归的角度来看，我们的反向蒸馏利用学生网络来预测教师模型的表征。因此，此处的“反向”既指教师编码器与学生解码器在结构上的反向形态，也指知识蒸馏顺序的逆转：即先蒸馏高级表征，再处理低级特征。值得注意的是，我们的反向蒸馏具有两大显著优势：*i) Non-similarity*结构。在所提出的T-S模型中，可将教师编码器视为下采样滤波器，学生解码器视为上采样滤波器。这种“反向结构”避免了上文讨论的因滤波器未区分而导致的混淆问题[33]。*ii) Compactness embedding*。输入学生解码器的低维嵌入作为正常模式重建的信息瓶颈。若将异常特征视作对正常模式的扰动，则紧凑的嵌入有助于阻止此类异常扰动向学生模型传播，从而增强T-S模型对异常的表征差异。值得注意的是，传统的基于自编码器的方法[5, 11, 16, 26]利用像素差异检测异常，而我们的方法则通过密集描述性特征进行判别。作为区域感知描述符的深度特征，相比图像中的逐像素信息能提供更有效的判别依据。

此外，由于瓶颈嵌入的紧凑性对异常检测至关重要（如上所述），我们引入了一个单类瓶颈嵌入（OCBE）模块，以进一步压缩特征编码。我们的OCBE模块包含一个多尺度特征融合（MFF）块和一个单类嵌入（OCE）块，两者均与学生解码器联合优化。值得注意的是，前者聚合了低层和高层特征，以构建一个丰富的嵌入用于正常模式重建。后者旨在

保留对学生解码教师回应有利的关键信息。

我们在公开基准上进行了大量实验。实验结果表明，我们的逆向蒸馏范式取得了与现有技术相当的性能。所提出的OCBE模块进一步将性能提升至新的最先进水平。我们的主要贡献总结如下：

- 我们引入了一种简单而有效的*Reverse Distillation*异常检测范式。编码器-解码器结构与反向知识蒸馏策略共同解决了传统知识蒸馏模型中滤波器区分度不足的问题，从而提升了T-S模型对异常情况的判别能力。
- 我们提出一种*one-class bottleneck embedding module*方法，将教师模型的高维特征投影至紧凑的单类嵌入空间。这一创新有助于保留丰富而紧凑的编码，使学生模型能够重建无异常的表示。
- 我们进行了大量实验，结果表明我们的方法实现了新的SOTA性能。

2. 相关工作

本节简要回顾了无监督异常检测的先前研究。我们将重点阐述所提方法与现有技术之间的相似性和差异。

经典的异常检测方法侧重于利用正常支持向量定义一个紧凑的封闭单类分布。开创性研究包括单类支持向量机（OC-SVM）[35]和支持向量数据描述（SVDD）[36]。为处理高维数据，DeepSVDD[31]和PatchSVD D[43]通过深度网络估计数据表示。

另一种无监督异常检测原型是利用生成模型进行样本重建，例如自动编码器（AE）[19]和生成对抗网络（GAN）[12]。这些方法基于一个假设：仅使用正常样本训练的生成模型能够成功重建无异常区域，但会在异常区域失效[2, 5, 34]。然而，近期研究表明深度模型的泛化能力极强，即使异常区域也能被较好地还原[46]。为解决这一问题，基于重建的方法中引入了记忆机制[11, 16, 26]、图像掩码策略[42, 46]以及伪异常生成技术[28, 45]。但这些方法在现实异常检测中仍缺乏强大的判别能力[3, 5]。最近，Metaformer（MF）[40]提出利用元学习[9]来弥合基于重建方法的模型适应与重建差距。值得注意的是，所提出的反向知识蒸馏同样采用编码器-解码器架构，但它

与基于构建的方法在两方面存在差异。首先，生成模型中的编码器是与解码器联合训练的，而我们的反向蒸馏方法将预训练模型固定为教师模型。其次，该方法并非基于像素级重建误差，而是在语义特征空间上进行异常检测。

数据增强策略也被广泛使用。通过在提供的无异常样本中添加伪异常，将无监督任务转化为有监督学习任务[23, 42, 46]。然而，这些方法容易偏向伪异常点，无法检测多种异常类型。例如，CutPaste[23]通过在正常图像上添加小块来生成伪异常，并训练模型检测这些异常区域。由于该模型专注于检测边缘不连续和纹理扰动等局部特征，如图6所示，它无法检测和定位大型缺陷及全局结构异常。

最近，在大规模数据集上预训练的网络被证明能够提取用于异常检测的判别性特征[7,8,23,25,29,30]。借助预训练模型，记忆其无异常特征有助于识别异常样本[7,29]。研究[8,30]表明，使用马氏距离衡量异常与无异常特征之间的相似性可实现精确的异常检测。由于这些方法需要记忆训练样本的所有特征，其计算成本较高。

从预训练模型中进行知识蒸馏是异常检测的另一种潜在解决方案。在无监督异常检测的背景下，由于学生模型在知识蒸馏过程中仅接触无异常样本，T-S模型预期能在推理时对异常生成差异化的特征[4,33,39]。为提升T-S模型对各类异常的判别能力，研究者引入了不同策略。例如，为捕捉多尺度异常，US[4]集成了多个在不同尺度正常数据上训练的模型，MKD[33]则提出使用多层次特征对齐方法。需注意的是，尽管所提方法同样基于知识蒸馏，我们的逆向蒸馏首次采用编码器-解码器结构构建T-S模型。教师网络与学生网络的异构性，以及知识蒸馏中反向数据流的运用，使我们的方法与现有技术形成显著区别。

3. 我们的方法

问题表述：令 $\mathcal{I}^t = \{I_1^t, \dots, I_n^t\}$ 为一组可用的无异常图像， $\mathcal{I}^q = \{I_1^q, \dots, I_m^q\}$ 为包含正常与异常样本的查询集。目标是训练一个模型以识别并定位查询集中的异常。在异常检测设定中， \mathcal{I}^t 和 \mathcal{I}^q 中的正常样本均遵循相同分布——

分布外的样本被视为异常。

系统概述：图3展示了所提出的用于异常检测的逆向蒸馏框架。我们的逆向蒸馏框架包含三个模块：一个固定的预训练教师编码器 E 、一个可训练的单类瓶颈嵌入模块和一个学生解码器 D 。给定输入样本 $I \in \mathcal{I}^t$ ，教师模型 E 提取多尺度表征。我们提出训练一个学生模型 D ，以从瓶颈嵌入中恢复特征。在测试/推理阶段，教师模型 E 提取的表征能够捕捉异常样本中的异常、分布外特征。然而，学生解码器 D 无法从相应的嵌入中重建这些异常特征。所提出的T-S模型中异常表征的低相似性表明其异常分数较高。我们认为，异构的编码器与解码器结构以及逆向知识蒸馏顺序，对异常表征的差异性贡献显著。此外，可训练的OCBE模块进一步将多尺度模式压缩至极低维空间，以供下游正常表征重建。这进一步增强了我们T-S模型中异常特征的表征差异，因为教师模型生成的异常表征很可能被OCBE模块舍弃。在本节剩余部分，我们首先详述逆向蒸馏范式，接着阐述OCBE模块，最后描述使用逆向蒸馏进行异常检测与定位的方法。

3.1. 反向蒸馏

在传统的知识蒸馏（KD）中，学生网络采用与教师模型相似或相同的神经网络结构，接收原始数据/图像作为输入，并旨在使其特征激活与教师的特征激活相匹配[4, 33]。在无监督异常检测（AD）的单类蒸馏背景下，当查询样本为异常时，期望学生模型能生成与教师模型高度不同的表示[11, 26]。然而，异常样本上的激活差异有时会消失，导致异常检测失败。我们认为，这一问题源于教师网络与学生网络架构的相似性，以及在师生知识传递过程中数据流的一致性。为了提高师生模型对未知、分布外样本的表示多样性，我们提出了一种新颖的逆向蒸馏范式，其中师生模型采用编码器-解码器架构，知识从教师的深层传递至其早期层，即首先将高层次语义知识传递给学生。为进一步促进单类蒸馏，我们设计了一个可训练的单类嵌入桥接（OCEB）模块来连接教师和学生模型（见第3.2节）。

在反向蒸馏范式中，教师编码器 E 旨在提取全面的表征。我们遵循先前的工作，使用在ImageNet上预训练的编码器。

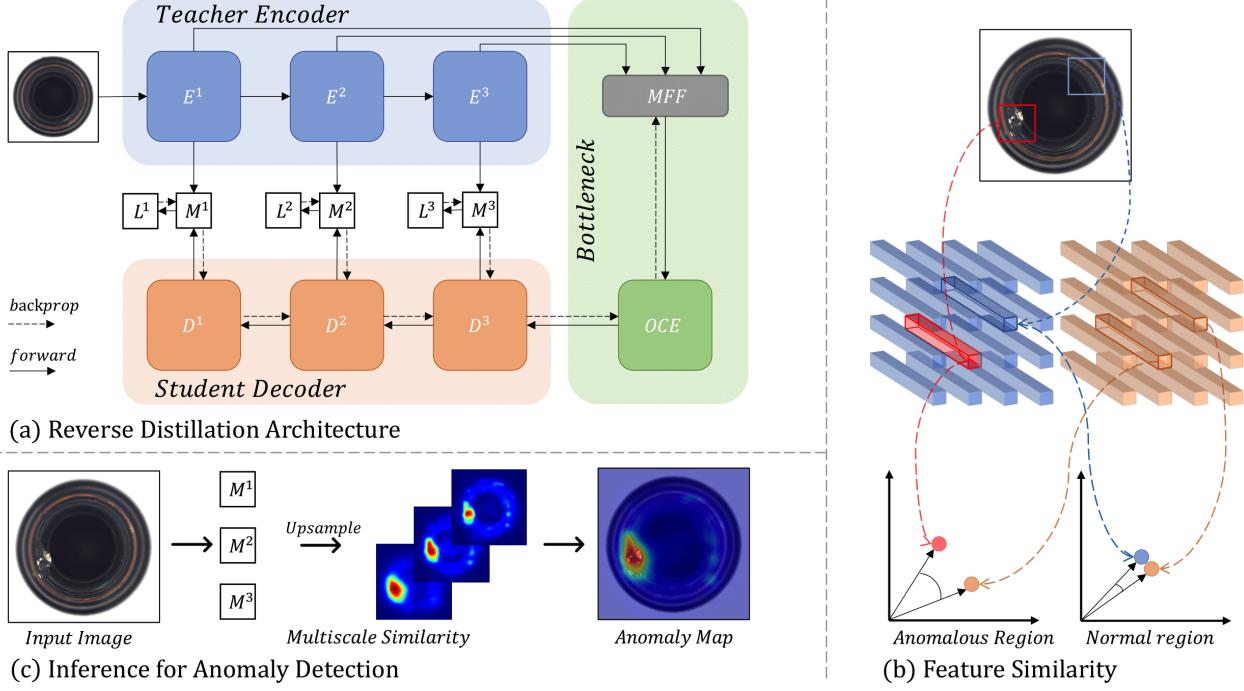


图3. 我们用于异常检测与定位的逆向蒸馏框架概览。(a) 模型包含预训练的教师编码器 E 、可训练的单类瓶颈嵌入模块(OCBE)以及学生解码器 D 。我们采用多尺度特征融合(MFF)模块整合来自 E 的低层与高层特征，并通过单类嵌入(OCE)模块将其映射为紧凑编码。训练过程中，学生网络 D 通过最小化相似度损失 \mathcal{L} 来学习模仿 E 的行为。(b) 推理阶段， E 如实提取特征，而 D 输出无异常特征。若 E 与 D 对应位置的特征向量相似度较低，则暗示存在异常。(c) 最终预测通过多尺度相似度图谱 M 的累加计算得出。

geNet [21] 作为我们的骨干网络 E 。为避免 T-S 模型收敛至平凡解，在知识蒸馏过程中教师模型 E 的所有参数均被冻结。我们在消融实验中证明，ResNet [14] 和 WideResNet [44] 都是良好的候选架构，因为它们能够从图像中提取丰富的特征 [4, 8, 23, 29]。

为了匹配 E 的中间表示，学生解码器 D 的架构与 E 对称但反向。这种反向设计有助于消除学生网络对异常情况的响应，而对称性则使其能够与教师网络具有相同的表示维度。例如，当我们以ResNet作为教师模型时，学生网络 D 由多个残差式解码块组成以实现镜像对称。具体来说，ResNet中的下采样通过核大小为1、步长为2的卷积层实现[14]。学生网络 D 中对应的解码块则采用核大小为2、步长为2的反卷积层[47]。关于学生解码器设计的更多细节见Supplementary Material。

在我们的反向蒸馏中，学生解码器 D 的目标是在训练过程中模仿教师编码器 E 的行为。在这项工作中，我们探索了基于多尺度特征的异常检测蒸馏方法。其背后的动机在于

神经网络浅层提取用于低级信息（如颜色、边缘、纹理等）的局部描述符，而深层则具有更广的感受野，能够表征区域/全局的语义与结构信息。也就是说，T-S 模型中低层与高层特征的低相似度分别暗示着局部异常和区域/全局结构异常。

数学上，令 ϕ 表示从原始数据 I 到单类瓶颈嵌入空间的投影，我们T-S模型中的配对激活对应关系为 $\{f_E^k = E^k(I), f_D^k = D^k(\phi)\}$ ，其中 E^k 和 D^k 分别代表教师模型和学生模型中的 k^{th} 编码和解码块。 $f_E^k, f_D^k \in \mathbb{R}^{C_k \times H_k \times W_k}$ ，其中 C_k 、 H_k 和 W_k 表示 k^{th} 层激活张量的通道数、高度和宽度。对于T-S模型中的知识迁移，采用余弦相似度作为KD损失，因为它能更精确地捕捉高维和低维信息中的关系[37, 49]。具体而言，对于特征张量 f_E^k 和 f_D^k ，我们沿通道轴计算它们的向量级余弦相似度损失，并得到一个二维异常图 $M^k \in \mathbb{R}^{H_k \times W_k}$ ：

$$M^k(h, w) = 1 - \frac{(f_E^k(h, w))^T \cdot f_D^k(h, w)}{\|f_E^k(h, w)\| \|f_D^k(h, w)\|}. \quad (1)$$

M^k 中的较大值表示该位置存在高度异常。考虑到多尺度知识蒸馏，通过累加多尺度异常图得到学生优化的标量损失函数：

$$\mathcal{L}_{KD} = \sum_{k=1}^K \left\{ \frac{1}{H_k W_k} \sum_{h=1}^{H_k} \sum_{w=1}^{W_k} M^k(h, w) \right\}, \quad (2)$$

其中 K 表示实验中使用的特征层数量。

3.2. 单类瓶颈嵌入

由于学生模型 D 试图在我们的反向知识蒸馏范式中恢复教师模型 E 的表示，可以直接将骨干网络中最后一个编码块的激活输出馈送至 D 。然而，这种简单连接存在两个缺陷。首先，知识蒸馏中的教师模型通常具有高容量。虽然高容量模型有助于提取丰富特征，但获得的高维描述符很可能存在显著冗余。表征的高自由度与冗余性会妨碍学生模型解码本质的无异常特征。其次，骨干网络最后一个编码块的激活通常表征输入数据的语义与结构信息。由于知识蒸馏的反向顺序，直接将这种高层表征馈送至学生解码器会给低层特征重建带来挑战。以往数据重建的研究通常引入跳跃路径连接编码器与解码器，但这种方法在知识蒸馏中并不适用，因为跳跃路径会在推理过程中向学生模型泄露异常信息。

为了解决单类蒸馏中的第一个不足，我们引入了一个可训练的单类嵌入块，将教师模型的高维表示投影到低维空间。我们将异常特征形式化为正常模式上的扰动。紧凑的嵌入块充当信息瓶颈，有助于阻止异常扰动向学生模型的传播，从而增强T-S模型在异常上的表示差异。在本研究中，我们采用ResNet[14]的第四个残差块作为单类嵌入块。

为解决解码器 D 在低层特征恢复上的问题，MFF块在单类别嵌入前对多尺度表征进行拼接。为实现特征拼接中的表征对齐，我们通过一个或多个步长为2的 3×3 卷积层对浅层特征进行下采样，随后接批量归一化与ReLU激活函数。接着采用步长为1的 1×1 卷积层配合批量归一化及ReLU激活，以获取丰富而紧凑的特征。

我们在图4中展示了OCBE模块，其中MFF聚合了低层与高层特征，以构建丰富的嵌入——

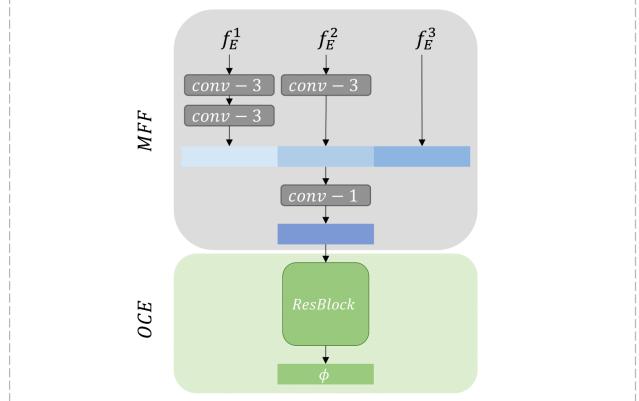


图4. 我们的一类瓶颈嵌入模块由可训练的MFF和OCE模块组成。MFF负责对齐来自教师模型 E 的多尺度特征，而OCE则将获得的丰富特征压缩为紧凑的瓶颈编码 ϕ 。

用于正常模式重建和OCE目标的基座，以保留有利于学生解码出教师响应的关键信息。图4中灰色的卷积层和绿色的ResBlock是可训练的，并在正常样本的知识蒸馏过程中与学生模型 D 联合优化。

3.3. 异常评分

在推理阶段，我们首先考虑对*anomaly localization* (AL)进行像素级异常分数的测量。当查询样本异常时，教师模型能够在其特征中反映异常性。然而，学生模型很可能无法恢复异常特征，因为学生解码器仅学习从知识蒸馏中的紧凑单类嵌入中恢复无异常表示。换言之，当查询异常时，学生 D 生成的表示会与教师模型产生差异。根据公式(1)，我们从T-S表示对中获得一组异常图，其中图 M_k 中的值反映了 k^{th} 特征张量的逐点异常性。为了定位查询图像中的异常，我们将 M^k 上采样至图像尺寸。令 Ψ 表示本研究中使用的双线性上采样操作。随后通过累加所有异常图的逐像素值，得到精确的分数图 S_{AL} ：

$$S_{AL} = \sum_{i=1}^L \Psi(M^i). \quad (3)$$

为了去除分数图中的噪声，我们通过高斯滤波器对 S_{AL} 进行平滑处理。

对于*anomaly detection*，在分数图 S_{AL} 中平均所有值对于异常区域较小的样本是不公平的。无论异常区域大小如何，都存在一个最敏感的点。

Image Size		128		256								
Category/Method		MKD [33]	Ours	GT [10]	GN [2]	US [4]	PSVDD [43]	DAAD [16]	MF [40]	PaDiM [8]	CutPaste [23]	Ours
s e r u l x e T	Carpet	79.3	99.2	43.7	69.9	91.6	92.9	86.6	94.0	99.8	93.9	98.9
	Grid	78.0	95.7	61.9	70.8	81.0	94.6	95.7	85.9	96.7	100	100
	Leather	95.1	100	84.1	84.2	88.2	90.9	86.2	99.2	100	100	100
	Tile	91.6	99.4	41.7	79.4	99.1	97.8	88.2	99.0	98.1	94.6	99.3
	Wood	94.3	98.8	61.1	83.4	97.7	96.5	98.2	99.2	99.2	99.1	99.2
	<i>Average</i>	87.7	98.6	58.5	77.5	91.5	94.5	91.0	95.5	98.8	97.5	99.5
s e r u l x e b O	Bottle	99.4	100	74.4	89.2	99.0	98.6	97.6	99.1	99.9	98.2	100
	Cable	89.2	97.1	78.3	75.7	86.2	90.3	84.4	97.1	92.7	81.2	95.0
	Capsule	80.5	89.5	67.0	73.2	86.1	76.7	76.7	87.5	91.3	98.2	96.3
	Hazelnut	98.4	99.8	35.9	78.5	93.1	92.0	92.1	99.4	92.0	98.3	99.9
	Metal Nut	73.6	99.2	81.3	70.0	82.0	94.0	75.8	96.2	98.7	99.9	100
	Pill	82.7	93.3	63.0	74.3	87.9	86.1	90.0	90.1	93.3	94.9	96.6
	Screw	83.3	91.1	50.0	74.6	54.9	81.3	98.7	97.5	85.8	88.7	97.0
	Toothbrush	92.2	90.3	97.2	65.3	95.3	100	99.2	100	96.1	99.4	99.5
	Transistor	85.6	99.5	86.9	79.2	81.8	91.5	87.6	94.4	97.4	96.1	96.7
	Zipper	93.2	94.3	82.0	74.5	91.9	97.9	85.9	98.6	90.3	99.9	98.5
<i>Average</i>		87.8	95.4	71.6	75.5	85.8	90.8	88.8	96.0	93.8	95.5	98.0
<i>Total Average</i>		87.8	96.5	67.2	76.2	87.7	92.1	89.5	95.8	95.5	96.1	98.5

表1. MVTec [3] 数据集上的 *Anomaly Detection* 结果。对于每个包含 256×256 分辨率图像的类别，AUROC (%) 排名前两位的方法已用粗体标出。根据纹理、物体及整体平均得分，我们的方法均位列第一。

异常区域。因此，我们将 S_{AL} 中的最大值定义为样本级异常分数 S_{AD} 。其原理在于，正常样本的异常分数图中不存在显著响应。

4. 实验与讨论

实证评估在MVTec异常检测与定位基准以及无监督单类新颖性检测数据集上均得以实施。此外，我们在MVTec基准上进行了消融研究，探究不同模块/区块对最终结果的影响。

4.1. 异常检测与定位

数据集。MVTec [3] 包含15个用于 *anomaly detection* 的真实世界数据集，涵盖5类纹理和10类物体。训练集共包含3,629张无异常图像。测试集则同时包含异常与无异常图像，总计1,725张。每个类别均提供多种缺陷用于测试。此外，测试数据集中还提供像素级标注，以支持 *anomaly localization* 评估。

实验设置。MVTec中的所有图像都被调整到特定分辨率（例如 128×128 、 256×256 等）。遵循先前工作的惯例，异常检测和定位每次仅在一个类别上进行。本实验中，我们在T-S模型中采用WideResNet50作为骨干网络。我们还在消融研究中报告了使用ResNet18和ResNet50的异常检测结果。为训练我们的逆向蒸馏模型，我们采用Adam优化器[18]，其参数 $\beta_1=0.5$ 、 $\beta_2=0.999$ 。学习率设置为0.005。我们训练200个周期，批次大小为16。高斯滤波器

使用 $\sigma = 4$ 来平滑异常得分图（如第3.3节所述）。

对于 *Anomaly detection*，我们采用接收者操作特征曲线下面积 (AUROC) 作为评估指标。本次实验中我们纳入了多种现有方法，包括MKD [33]、GT [10]、GANomaly (GN) [2]、Uninformed Student (US) [4]、PSVDD [43]、DAAD [16]、MetaFormer (MF) [40]、PaDiM (WResNet50) [8]以及CutPaste [23]。

对于 *Anomaly Localization*，我们同时报告了AUROC 和逐区域重叠度(RO) [4]。与用于逐像素度量的AUROC不同，PRO评分对任意大小的异常区域给予同等对待。比较基线包括MKD[33]、US[4]、MF[40]、SPADE(WResNet50)[7,29]、PaDiM(WResNet50)[8]、RIAD[46]和CutPaste[23]。

实验结果与讨论。MVTec上的异常检测结果如表1所示。平均结果显示，我们的方法超出SOTA 2.5%。在纹理和物体类别上，我们的模型分别达到了99.5%和98.0% AUROC的新SOTA。异常分数的统计分布如图5所示。正常样本（蓝色）与异常样本（红色）的非重叠分布表明我们的T-S模型具备强大的异常检测能力。

异常定位的定量结果总结于表2。在AUROC和PRO两项指标上，我们的方法在所有类别的平均得分分别达到97.8%和93.9%，均超越了现有最优水平。为探究方法对不同异常类型的鲁棒性，我们将缺陷分为两类：大型缺陷（结构异常）与微小缺陷（不明显缺陷），并通过图6与图7的可视化结果进行定性评估。相较于表1中第二名方法（即CutPaste[23]），我们的方法对整个 $\{v^*\}$ 区域产生了显著响应。

Image Size		128		256						
Category/Method		MKD [33]	Ours	US [4]	MF [40]	SPADE [7]	PaDiM [8]	RIAD [46]	CutPaste [23]	Ours
s c u x T	Carpet	95.6/-	98.1/95.3	-/87.9	-/87.8	97.5/94.7	99.1 /96.2	96.3/-	98.3/-	98.9/ 97.0
	Grid	91.8/-	97.3/92.6	-/95.2	-/86.5	93.7/86.7	97.3/94.6	98.8/-	97.5/-	99.3/97.6
	Leather	98.1/-	99.0/98.6	-/94.5	-/95.9	97.6/97.2	99.2/97.8	99.4/-	99.5/-	99.4/ 99.1
	Tile	82.8/-	92.6/84.8	-/94.6	-/88.1	87.4/75.9	94.1/86.0	89.1/-	90.5/-	95.6/90.6
	Wood	84.8/-	92.1/82.3	-/91.1	-/84.8	88.5/87.4	94.9/ 91.1	85.8/-	95.5/-	95.3/90.9
<i>Average</i>		90.6/-	95.8/90.7	-/92.7	-/88.6	92.9/88.4	96.9/93.2	93.9/-	96.3/-	97.7/95.0
s c e b O	Bottle	96.3/-	98.2/94.7	-/93.1	-/88.8	98.4/95.5	98.3/94.8	98.4/-	97.6/-	98.7/96.6
	Cable	82.4/-	97.8/90.5	-/81.8	-/93.7	97.2/90.9	96.7/88.8	84.2/-	90.0/-	97.4/91.0
	Capsule	95.9/-	96.5/87.2	-/96.8	-/87.9	99.0 /93.7	98.5/93.5	92.8/-	97.4/-	98.7/95.8
	Hazelnut	94.6/-	98.8/89.2	-/96.5	-/88.6	99.1 /95.4	98.2/92.6	96.1/-	97.3/-	98.9/95.5
	Metal Nut	86.4/-	96.6/84.1	-/94.2	-/86.9	98.1/94.4	97.2/85.6	92.5/-	93.1/-	97.3/92.3
	Pill	89.6/-	97.0/90.0	-/96.1	-/93.0	96.5/94.6	95.7/92.7	95.7/-	95.7/-	98.2/96.4
	Screw	96.0/-	98.3/94.4	-/94.2	-/95.4	98.9/96.0	98.5/94.4	98.8/-	96.7/-	99.6/98.2
	Toothbrush	96.1/-	98.2/86.7	-/93.3	-/87.7	97.9/93.5	98.8/93.1	98.9/-	98.1/-	99.1/94.5
	Transistor	76.5/-	97.6/85.2	-/66.6	-/92.6	94.1/87.4	97.5 /84.5	87.7/-	93.0/-	92.5/78.0
	Zipper	93.9/-	97.0/92.3	-/95.1	-/93.6	96.5/92.6	98.5/95.9	97.8/-	99.3/-	98.2/95.4
<i>Average</i>		90.8/-	97.6/89.4	-/90.8	-/90.8	97.6/93.4	97.8/91.6	94.3/-	95.8/-	97.9/93.4
<i>Total Average</i>		90.7/-	97.0/89.9	-/91.4	-/90.1	96.5/91.7	97.5/92.1	94.2/-	96.0-	97.8/93.9

表2. MVTec [3] 上使用AUROC和PRO指标的*Anomaly Localization*结果。AUROC代表像素级比较，而PRO则关注基于区域的行为。我们将AUROC和PRO的最佳结果以粗体标出。值得注意的是，我们的方法具有鲁棒性，且在两项指标下均代表了最先进的性能。

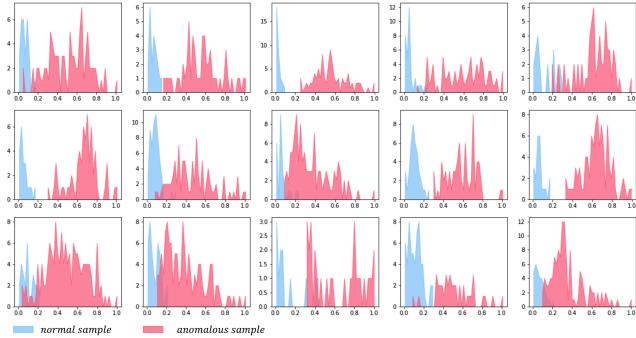


图5. MVTec [3] 所有类别的异常分数直方图（x轴：0到1的异常分数，y轴：计数）。

异常区域。

复杂度分析。近期基于预训练模型的方法通过从无异常样本中提取特征作为度量标准，取得了令人瞩目的性能[7, 8]。然而，存储特征模型会导致较大的内存消耗。相比之下，我们的方法仅依赖一个额外的CNN模型就实现了更优的性能。如表3所示，我们的模型以较低的时间和内存复杂度获得了性能提升。

Methods	Infer. time	Memory	Performance
SPADE (WResNet50)	1.40	1400	85.5/96.5/91.7
PaDiM (WResNet50)	0.95	3800	95.5/97.5/92.1
Ours (WResNet50)	0.31	352	98.5/97.8/93.9

表3. 基于预训练方法在MVTec [3]上的推理时间（Intel i7平台秒数）、内存使用量（MB）以及性能（AD-AUROC/AL-AUROC/AL-PRO）对比。

局限性。我们注意到，尽管在异常检测方面表现良好，但在*transistor*数据集上的定位性能相对较弱。这种性能下降是由于预测与标注之间的误解造成的。如图6所示，我们的方法定位了错位区域，而真实标注则同时覆盖了错位区域和原始区域。缓解此问题需要将更多特征与上下文关系相关联。我们通过实验发现，具有更广感知域的高层级特征层能够提升性能。例如，使用第二和第三层特征进行异常检测可获得94.5%的AUROC，而仅使用第三层特征可将性能提升至97.3%。此外，将图像分辨率降低至 128×128 也能达到97.6%的AUROC。我们在*supplementary material*中展示了更多异常检测与定位的案例，包括正面和负面结果。

4.2. 单类新颖性检测

为评估所提方法的泛化能力，我们在3个语义数据集上进行了*one-class novelty detection*实验，包括MNIST[22]、FashionMNIST[41]和CIFAR10[20]。MNIST是包含0-9手写数字的数据集。FashionMNIST由10个时尚产品类别的图像组成。这两个数据集均包含6万张训练样本和1万张测试样本，分辨率均为 28×28 。CIFAR10因其包含多样化的自然物体而成为新颖性检测中具有挑战性的数据集，它包含5万张训练图像和1万张测试图像，尺度为 32×32 ，共分10个类别。

按照[27]中提到的协议，我们使用单一类别的样本训练模型并检测新样本。注意，新颖性分数定义为总和

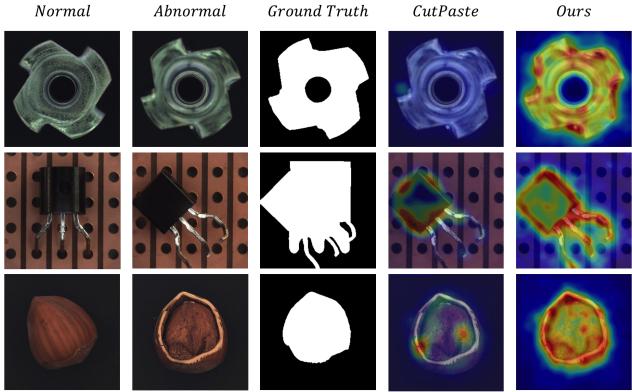


图6. 自上而下的异常情况：“metal nut”上的“flip”、“transistor”上的“misplaced”以及“hazelnut”上的“crack”。正常样本作为参考提供。

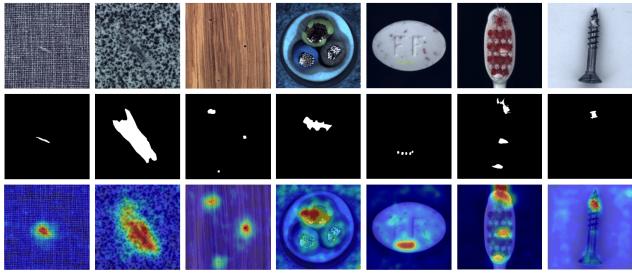


图7. 微小或不显眼异常的可视化。从左至右：carpet、tile、wood、cable、pill、toothbrush和screw。

Method	MNIST	F-MNIST	CIFAR10	Caltech-256
LSA [1]	97.5	92.2	64.1	-
OCGAN [27]	97.3	87.8	65.7	-
HRN [17]	97.6	92.8	71.3	-
DAAD [16]	99.0	-	75.3	-
MKD [33]	98.7	94.5	84.5	-
G2D [28]	-	-	-	95.7
OiG [45]	-	-	-	98.2
Ours	99.3	95.0	86.5	99.9

表4. 单类新颖性检测的AUROC(%)结果。最佳结果以**粗体**标出。

相似度图中的分数。本实验中的基线包括LSA [1]、OCGAN [27]、HRN [17]、DAAD [16] 和 MKD [33]。我们还在Caltech-256 [13] 数据集上纳入了与OiG [45] 和 G2D [28] 的比较。

表4总结了三个数据集的定量结果。值得注意的是，我们的方法取得了优异的结果。实验细节及各类别对比结果详见Supplementary Material。

4.3. 消融分析

我们研究了OCE和MFF模块在AD上的有效性，并在表5中报告了数值结果。我们采用

以预训练的残差块[14]作为基线。预训练残差块的嵌入可能包含异常特征，这会降低T-S模型的表示差异。我们可训练的OCE块压缩特征编码，MFM块将丰富特征融合进嵌入，从而实现更精确的异常检测与定位。

Metric	Pre	Pre+OCE	Pre+OCE+MFM
AUROC _{AD}	96.0	97.9	98.5
AUROC _{AL}	96.9	97.4	97.8
RPO	91.2	92.4	93.9

表5. 关于预训练瓶颈、OCE和MFF的消融研究。

表6展示了不同骨干网络作为教师模型的定性比较。直观而言，更深更宽的网络通常具备更强的表征能力，这有助于精确检测异常。值得注意的是，即使采用如ResNet18这样的小型神经网络，我们的反向蒸馏方法仍能取得优异的性能。

Backbone	ResNet18	ResNet50	WResNet50
AUROC _{AD}	97.9	98.4	98.5
AUROC _{AL}	97.1	97.7	97.8
RPO	91.2	93.1	93.9

表6. 与不同骨干网络的定量比较。

此外，我们还探讨了不同网络层对异常检测的影响，并在表7中展示了结果。对于单层特征， M^2 取得了最佳效果，因为它平衡了局部纹理和全局结构信息。多尺度特征融合有助于覆盖更多类型的异常。

Score Map	M^1	M^2	M^3	$M^{2,3}$	$M^{1,2,3}$
AUROC _{AD}	90.1	97.5	97.2	98.0	98.5
AUROC _{AL}	94.0	96.9	96.9	97.6	97.8
RPO	88.6	92.6	89.5	93.2	93.9

表7. 多尺度特征蒸馏的消融研究。

5. 结论

我们提出了一种新颖的知识蒸馏范式——反向蒸馏，用于异常检测。它全面解决了先前基于知识蒸馏的异常检测方法中的问题，并增强了教师-学生模型对异常的反应。此外，我们在反向蒸馏中引入了可训练的单类嵌入和多尺度特征融合模块，以改进单类知识迁移。实验表明，我们的方法在异常检测、异常定位和新颖性检测方面显著优于先前技术。

参考文献

[1] Davide Abati, Angelo Porrello, Simone Calderara, 与 Rita Cucchiara。基于隐空间自回归的新颖性检测。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 第481–490页, 2019年。

8[2] Samet Akcay, Amir Atapour-Abarghouei, 与 Toby P. Breckon。Ganomaly: 通过对抗训练进行半监督异常检测。载于 C. V. Jawahar, Hongdong Li, Greg Mori, 与 Konrad Schindler 编辑的 *Computer Vision – ACCV 2018*, 第622–637页, Cham, 2019年。Springer International Publishing。

1, 2, 6[3] Paul Bergmann, Michael Fauser, David Sattlegger, 与 Carsten Steger。Mvtac ad——一个用于无监督异常检测的综合真实世界数据集。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019年6月。

1, 2, 6, 7[4] Paul Bergmann, Michael Fauser, David Sattlegger, 与 Carsten Steger。无信息学生: 基于判别性隐嵌入的学生-教师异常检测。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020年6月。

1, 3, 4, 6, 7[5] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, 与 Carsten Steger。通过将结构相似性应用于自编码器改进无监督缺陷分割, 2018年。

1, 2[6] Pengguang Chen, Shu Liu, Hengshuang Zhao, 与 Jiaya Jia。通过知识回顾进行知识蒸馏。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 第5008–5017页, 2021年。

1, 2[7] Niv Cohen 与 Yedid Hoshen。基于深度金字塔对应的子图像异常检测, 2020年。

3, 6, 7[8] Thomas Defard, Aleksandr Setkov, Angeline Loesch, 与 Romaric Audigier。Padim: 一种用于异常检测与定位的补丁分布建模框架。收录于 *International Conference on Pattern Recognition*, 第475–489页。Springer, 2021年。

3, 4, 6, 7[9] Chelsea Finn, Pieter Abbeel, 与 Sergey Levine。用于深度网络快速适应的模型无关元学习。收录于 *International Conference on Machine Learning*, 第1126–1135页。PMLR, 2017年。

2[10] Izhak Golan 与 Ran El-Yaniv。使用几何变换的深度异常检测。收录于 *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 第9781–9791页, 美国纽约州红钩市, 2018年。Curran Associates Inc.。

6[11] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, 与 Anton 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。收录于 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019年10月。

1, 2, 3[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, 与 Yoshua Bengio。生成对抗网络。收录于 *Proceedings*

of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS' 14, 第2672–2680页, 美国马萨诸塞州剑桥, 2014年。MIT Press。

2 [13] Gregory Griffin, Alex Holub和Pietro Perona。Caltech-256 物体类别数据集。2007年。

8 [14] 何恺明、张祥雨、任少卿和孙剑。深度残差学习用于图像识别。于 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016年6月。

4, 5, 8 [15] Geoffrey Hinton、Oriol Vinyals和Jeffrey Dean。蒸馏神经网络中的知识。于 *NIPS Deep Learning and Representation Learning Workshop*, 2015年。

1 [16] 侯金磊、张莹莹、钟乔勇、谢迪、蒲世良和周宏。分而治之: 学习块状记忆用于无监督异常检测。于 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 第8791–8800页, 2021年10月。

1, 2, 6, 8 [17] 胡文鹏、王梦宇、秦琦、马金文和刘冰。HRN: 一种整体的一类学习方法。于 H. Larochelle、M. Ranzato、R. Hadsell、M. F. Balcan和H. Lin编辑, *Advances in Neural Information Processing Systems*, 第33卷, 第19111–19124页。Curran Associates, Inc., 2020年。

8 [18] Diederik P Kingma和Jimmy Ba。Adam: 一种随机优化方法。arXiv preprint arXiv:1412.6980, 2014年。

6 [19] Diederik P Kingma和Max Welling。自动编码变分贝叶斯, 2013年。

2 [20] Alex Krizhevsky。从微小图像中学习多层特征, 2009年。

7 [21] Alex Krizhevsky、Ilya Sutskever和Geoffrey E. Hinton。使用深度卷积神经网络进行ImageNet分类。于 *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS' 12, 第1097–1105页, 美国纽约州红钩, 2012年。Curran Associates Inc.。

4 [22] Yann LeCun。手写数字MNIST数据库, 1998年。

7 [23] 李春良、Kihyuk Sohn、Jinsung Yoon和Thomas Pfister。CutPaste: 用于异常检测和定位的自监督学习。于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 第9664–9674页, 2021年6月。

1, 3, 4, 6, 7 [24] W. Liu、D. Lian W. Luo和S. Gao。用于异常检测的未来帧预测——一个新的基线。于 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018年。

1 [25] 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。*Sensors*, 18(1), 2018年。

3 [26] Hyunjong Park、Jongyoun Noh和Bumsu Ham。学习记忆引导的正常性用于异常检测。于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020年6月。

1, 2, 3 [27] Pramuditha Perera、Ramesh Nallapati和Bing Xiang。OCGAN: 使用带约束的GAN进行一类新颖性检测

潜在表示。在*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 第2898–2906页, 2019年。7, 8 [28] Masoud Pourreza, Bahram Mohammadi, Mostafa Khaki, Samir Bouindour, Hichem Snoussi, 和 Mohammad Sabokrou。G2d: 生成以检测异常。在*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 第2003–2012页, 2021年。2, 8 [29] Tal Reiss, Niv Cohen, Liron Bergman, 和 Yedid Hoshen。Panda: 为异常检测与分割适配预训练特征。在*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 第2806–2814页, 2021年6月。3, 4, 6 [30] Oliver Rippel, Patrick Mertens, 和 Dorit Merhof。在预训练深度特征中对正常数据分布建模以进行异常检测。在*2020 25th International Conference on Pattern Recognition (ICPR)*, 第6726–6733页, 2021年。3 [31] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, 和 Marius Kloft。深度单类分类。在 Jennifer Dy 和 Andreas Krause 编辑的*Proceedings of the 35th International Conference on Machine Learning*, 第80卷*Proceedings of Machine Learning Research*, 第4393–4402页。PMLR, 2018年7月10–15日。2 [32] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, 和 Mohammad Sabokrou。关于异常、新颖性、开放集和分布外检测的统一综述: 解决方案与未来挑战。*arXiv preprint arXiv:2110.14051*, 2021年。7 [33] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, 和 Hamid R. Rabiee。用于异常检测的多分辨率知识蒸馏。在*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 第14902–14912页, 2021年6月。1, 2, 3, 6, 7, 8 [34] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, 和 Ursula Schmidt-Erfurth。f-anogan: 基于生成对抗网络的快速无监督异常检测。*Medical Image Analysis*, 54:30–44, 2019年。1, 2 [35] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, 和 Robert C Williamson。估计高维分布的支持域。*Neural computation*, 13(7):1443–1471, 2001年。2 [36] David MJ Tax 和 Robert PW Duin。支持向量数据描述。*Machine learning*, 54(1):45–66, 2004年。2 [37] Frederick Tung 和 Greg Mori。保持相似性的知识蒸馏。在*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019年10月。4 [38] Shashanka Venkataraman, Kuan-Chuan Peng, Rajat V.ikram Singh, 和 Abhijit Mahalanobis。图像中注意力引导的异常定位。在*European Conference on Computer Vision*, 第485–503页。Springer, 2020年。1 [39] Guodong Wang, Shumin Han, Errui Ding, 和 Di Huang。用于异常检测的学生-教师特征金字塔匹配, 2021年。1, 3

[40] 吴志强、陈鼎杰、傅邱山、刘定庐。学习无监督元变换器用于异常检测。发表于*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 第4369–4378页, 2021年10月。2, 6, 7[41] Han Xiao, Kashif Rasul, Roland Vollgraf。Fashion-MNIST: 一个用于基准测试机器学习算法的新型图像数据集, 2017年。7[42] 严旭东、张淮东、徐雪森、胡小伟、衡鹏安。从正常样本中学习语义上下文用于无监督异常检测。*Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3110–3118, 2021年5月。1, 2, 3[43] Jihun Yi, Sungroh Yoon。Patch SVDD: 用于异常检测与分割的块级SVDD。发表于*Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020年11月。2, 6[44] Sergey Zagoruyko, Nikos Komodakis。宽残差网络。载于Edwin R. Hancock、Richard C. Wilson、William A. P. Smith编, *Proceedings of the British Machine Vision Conference (BMVC)*, 第87.1–87.12页。BMVA Press, 2016年9月。4[45] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, Seung-Ik Lee。旧即是金: 重新定义对抗性学习单类分类器训练范式。发表于*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 第14183–14193页, 2020年。2, 8[46] Vitan Zavrtanik, Matej Kristan, Danijel Skočaj。通过修复进行重建的视觉异常检测。*Pattern Recognition*, 112:107706, 2021年。1, 2, 3, 6, 7[47] Matthew D. Zeiler, Dilip Krishnan, Graham W. Taylor, 和 Rob Fergus。解卷积网络。发表于*2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 第2528–2535页, 2010年。4[48] 周康、肖玉婷、杨建龙、程俊、刘文、罗伟新、顾再旺、刘江、高升华。利用P-Net编码结构-纹理关系用于视网膜图像异常检测。载于*Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI*, 第360–377页。Springer, 2020年。1[49] 朱金国、唐世翔、陈大鹏、余世杰、刘亚坤、戎明哲、杨爱军、王晓华。互补关系对比蒸馏。发表于*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 第9260–9269页, 2021年6月。4[50] David Zimmerer, Jens Petersen, Gregor Köhler, Paul Jäger, Peter Full, Tobias Roß, Tim Adler, Annika Reinke, Lena Maier-Hein, Klaus Maier-Hein。医学分布外分析挑战赛2021, 2021年3月。1