基于双向约束蒸馏的无监督图像异常检测

李 波,李泽超*,邢 鹏,唐金辉
(南京理工大学计算机科学与工程学院,江苏南京 210094)

摘 要: 异常检测是一项重要的计算机视觉任务,它的目标是检测异常样本同时定位异常区域. 近期,主流的无监督异常检测方案通常基于蒸馏方法和重构方法. 然而,它们仍存在相似的局限. 在基于蒸馏方法的异常检测中,学生网络通常能学习到教师网络相似的表征能力,无法针对某些异常区域产生与教师网络有明显差异的表征. 在重构模型中,编码-解码结构容易学习到简单的复原捷径,导致复原图像与输入相似,无法有效地检测异常. 为了解决上述挑战,本文提出基于双向约束蒸馏的无监督图像异常检测方法$\mathcal{N}$-Net,它通过双向蒸馏模块和多级过滤模块缓解了上述局限. 具体地,在教师学生网络中,本文首先提出蒸馏适应域特征而非原始域特征,它通过双向蒸馏分支保证了正常适应域特征的高效对齐. 然后,本文提出多级过滤模块,通过查询和压缩的方式过滤异常特征,进一步增强学习正常语义特征分布的能力,提升异常检测性能. 最后,本文在两个基准异常检测数据集 MVTec 和 VisA 上进行了大量实验,结果表明所提方法在异常检测和定位任务上取得了先进的性能.

关键词: 异常检测;双向蒸馏;特征映射;多级过滤;特征压缩
基金项目: 国家自然科学基金(No.U20B2064,No.U21B2043)
中图分类号: TP37;TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2025)03-0895-15
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240733

Unsupervised Image Anomaly Detection Based on Constrained BidiRectional Distillation

LI Bo, LI Ze-chao*, XING Peng, TANG Jin-hui
(*School of Computer Science and Engineering*, *Nanjing University of Science and Technology*, *Nanjing*, *Jiangsu* 210094, *China*)

Abstract: Anomaly detection has been widely studied and applied to various visual scenes. Recently, the mainstream unsupervised anomaly detection schemes are usually based on distillation methods and reconstruction methods. However, they still have some limitations. In distillation model, the student network can usually learn the strong representation ability of the teacher network, thus can not represent differently for the abnormal regions. In reconstruction model, the encoder-decoder model can easily learn a restoration shortcut and recover features indiscriminately. To address the above challenges, we propose $\mathcal{N}$-Net, which integrates the advantages of above two methods and alleviates limitations through the bidirectional distillation module and the multistage filtration mechanism. Specifically, in the teacher-student network, this paper first proposes distilling adaptive domain features instead of original domain features, which ensures efficient alignment of normal adaptive domain features through bidirectional distillation branches. Then, we propose a multilevel filtering module to filter abnormal features through query and compression to further enhance the ability to learn normal semantic feature distribution and improve the anomaly detection performance. Finally, a large number of experiments are carried out on two benchmark anomaly detection datasets, MVTec and VisA. The results show that the proposed method achieves advanced performance in anomaly detection and location tasks.

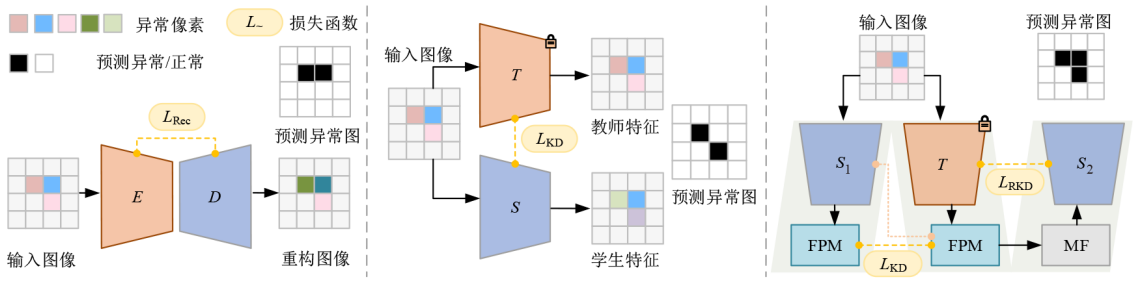
Key words: anomaly detection; bidirectional distillation; feature projection; multistage filtration; feature compact
Foundation Item(s): National Natural Science Foundation of China (No.U20B2064, No.U21B2043)

收稿日期:2024-08-07;修回日期:2024-10-22;责任编辑:朱梅玉
*通讯作者:李泽超

1 引言

图像异常检测是一项经典的计算机视觉任务，包含图像级异常检测和像素级异常检测，分别是检测异常图像和定位图像中的异常区域[1]. 异常也被称为离群值，是指不符合预期规定的模式或数据，例如工业产品中的缺陷样本，包括形变，污染和缺失等. 正常样本则是符合预期的数据，正常样本之间的特征差异很小. 异常检测在产品质检[2~4]、视频监控[5,6]和医学病理检测[7]等应用中发挥着重要的作用. 在现实场景中，正常图像通常大量存在且易于获得，而异常图像相对正常图像是罕见的，同时还呈现出多样性[5]. 因此，收集全面、多样和充足的异常样本训练有监督的检测或分割模型是极具挑战性的[3]. 目前，异常检测任务通常依赖无监督方法[8~11]（或称为自监督方法[9]），仅使用正常样本训练模型，期望能够检测和定位异常区域[12]. 其中，基于重构的方法[12~19]和基于蒸馏的方法[20~25]受到了广泛的关注.

基于自监督重构模型的方法通常假设模型仅能高质量的重构正常图像，而难以重构异常图像特征. 如图1(a)所示，重构模型通常由编码器和解码器组成. 在训练阶段，编码器提取正常图像的高级语义特征，解码器根据高级语义特征学会重构正常图像. 在推理阶段，异常图像通过重构模型后重构为正常图像，即异常图像重构前后的误差理论上更大[9]. 因此，重构模型可以根据重构前后的差异检测图像是否异常. 然而，重构模型的泛化能力强大，很容易学习到“复原捷径”，即编码器E的输出特征可能包含解码器D重构所需的所有特征，它们几乎可以良好地重构输入图像[10]. 即使一些方法探索了在编码和解码结构中加入限制模块[9,11]，抑制输入解码器的异常特征. 然而，这些可学习结构并不能完全过滤编码器提取的异常特征. 因此，在一些异常图像场景下，“复原捷径”的问题仍然会导致重构模型的异常检测效果不佳.



(a) 基于重构的模型 (b) 基于蒸馏的模型 (c) 本文所提模型

图1 无监督异常检测方法的对比图

另外一种主流的无监督异常检测方法基于知识蒸馏范式. 知识蒸馏是一种模型压缩的方法，旨在将复杂模型（称为教师模型）的“知识”转移到简单模型（称为学生模型）中. 在异常检测任务中，如图1(b)所示，它利用特征蒸馏的方式，“教会”无先验知识的学生模型直接模拟预训练教师模型对正常图像的特征表示[24]. 由于学生网络训练阶段仅学习对正常图像的表征，理论上，师生网络针对异常图像得到的表征差异大于针对正常图像的表征差异[21]. 因此，这类方法可以根据师生模型输出特征的差异推理图像是否异常. 然而，预训练的教师模型的先验知识与特定异常检测场景存在差异，在原始特征域的蒸馏可能导致师生模型无法关注下游任务的异常[20]. 此外，学生网络可能无差别地学习教师模型的知识，对于存在细微异常的图像，学生与教师模型的表征差异可能同样很小，从而导致错误检测.

基于对上述方法的深入分析，本文发现它们均存在一些局限：(1)在重构模型中，可学习参数的编码-解码结构的重构模型可能无法完全过滤编码器提取的正常特征，导致重构模型成功重构某些异常区域. (2)在蒸馏模型中，学生模型直接蒸馏教师原始域的特征，可能导致学生模型过多学习教师模型的图像表征能力，进一步导致学生模型与教师模型对异常图像的输出特征相似，造成模型的检测能力不强. 在解决异常检测任务时，它们具有相似性：基于蒸馏或重构模型的方法都期望学生模型/解码器针对正常图像的表征差异/复原误差小，同时针对异常图像的表征差异/复原误差大. 因此，本文探索了一种重构和蒸馏模型互补的异常检测方法，缓解上述限制.

本文提出一个基于双向约束蒸馏的无监督图像异常检测方法$\mathcal{N}$-Net，它的结构如图1(c)所示. 本文所提出的$\mathcal{N}$-Net主要从以下三个方面融合和优化了传统重构和蒸馏方法. 首先，本文提出带特征映射模块（Feature Projection Module，FPM）的正向蒸馏分支. 为了避免学生网络过度学习教师网络提取原始域特征，本文提出特征映射模块将教师特征映射至适应域空间，并约束正向学生网络S_1的输出与教师网络T的输出在适应特征空间对齐. 其次，本文提出逆向蒸馏分支，它首先作为特征约束模块，约束正向蒸馏中适应域空间的

特征保持原始语义,从而避免正向蒸馏中的特征映射模块收敛至平凡解. 此外,它作为可学习的解码器,与预训练的编码器构成编码-解码器结构,进一步学习正常语义特征的分布. 最后,为了进一步抑制逆向学生网络S_2复原教师异常特征的能力,本文提出多级过滤模块(Multistage Filtration, MF),分别通过查询过滤和压缩过滤的方式去除潜在的异常信息,从而保证逆向学生网络S_2有针对性地复原正常特征. 本文在图像异常检测基准数据集MVTec和VisA上进行了大量的定性和定量分析,验证了本文的$\mathcal{N}$-Net在异常检测任务上具有优越的性能. 大量的消融实验验证了本文所提模块的有效性.

2　图像异常检测相关工作

图像异常检测是计算机视觉中的一项重要任务,其目标是通过现有的正常图像构建模型以检测可能出现的异常图像. 当前的图像异常检测主要分为基于特征建模、基于重构和基于预训练模型的方法.

2.1　基于特征建模的异常检测方法

基于特征建模的方法把异常检测看作一个单分类[26]问题. Schölkopf等人[27]在训练阶段建模一个超球面,期望所有的正常样本在超球面内,并根据样本的特征与超球的位置关系判断图像是否异常. 为了处理更高维数据, Ruff等人[28]将整幅图像的特征映射到一个超球中,然后根据测试样本和球心的距离判定异常. 而Yi等人[29]将图像分割为若干的块,将空间相近的块映射到一起,解决整幅图像输入导致数据维度过高的问题. Yao等人[30]提出了一种清晰边界引导的对比学习机制区分正常与异常特征的边界,增强模型的可辨别性. 这类方法通常对异常样本的分布情况比较敏感,如果正常样本分布与异常样本的分布比较接近,那么极有可能导致错误检测.

2.2　基于重构的异常检测方法

基于重构的方法通过正常样本训练重构模型,依据重构图像和原始图像的表征差异进行异常检测[8],包括基于生成的方法和基于自编码器(AutoEncoder, AE)的方法[31]. Schlegl等人[12]首次将GAN[32]用于异常检测,它建模生成器学习正常数据的生成,由异常图像的隐变量生成无异常图像,根据图像级的差异定位异常. GANomaly[13]在特征空间计算表征差异,使用编码器提取生成的图像特征,通过比较生成图像和原始图像的特征潜在空间的数据差异推断异常. 为了提高正常图像的重构质量, Akcay等人[14]在生成器的编码器和解码器之间加入了跳跃连接层,通过跨层特征重建,改善图像重构质量. 然而,跳跃连接同时也会导致异常特征泄露到解码器,导致异常图像成功重构.

基于自编码器[31]重构的方法通过最小化重构误差学习正常特征的表征. 由于模型未学习过异常样本的重构,碰到异常样本时会产生较大的重构误差. 然而,这种假设并不总是成立,有时候自编码器的泛化能力很好,能够较好地恢复异常特征,导致图像重构前后特征差异不显著[15]. 为了缓解这类问题, Gong等人[9]在自编码器中提出存储模块,通过存储正常特征并作为解码器的输入,避免解码器重构出异常特征. Hou等人[11]通过改变特征图上的划分粒度调整自编码器对正常样本和异常样本的重建能力. Xing等人[16]提出了一种新的分区机制和查询生成方法来保存特征的上下文信息,提高了存储模块的学习能力. Zavrtanik等人[33]通过随机去除部分图像区域并根据部分图像块复原正常图像,缓解了自编码方法无差别还原异常特征的缺点. 此外,为了解决异常样本较少的问题,有的方法[34]通过生成伪造异常样本训练去噪模型,但是伪造的异常特征与真实异常仍可能有较大差异. 这些基于自编码器的方法仍存在重构图像模糊和无法通过重构完全消除异常特征的问题[20]. 因此,本文在特征重构结构中引入更精细的多级过滤模块,缓解上述限制.

2.3　基于预训练模型的异常检测方法

基于预训练模型的方法使用在外部数据集上预训练好的网络作为编码器,提取丰富的特征用于后续的异常检测. 这种方法通常也可以分为两类,其中一类直接利用预训练编码器进行特征提取,根据测试样本和训练样本的特征匹配度进行异常检测. 另一类是基于蒸馏的方法,这类方法将预训练模型视作教师模型,蒸馏学生模型,测试阶段根据师生网络的表征差异检测异常.

在基于预训练编码器的图像异常检测方法中, SPADE[35]利用预训练的网络对正常样本进行特征提取,并存储所有模板的特征. 在测试阶段,模型通过K近邻(K-Nearest Neighbor, KNN)算法[36]在存储的特征中检索K个最近的模板特征,得到图像中每个位置的异常得分. 然而,这种方法中存储的模板特征越多时,模型的计算复杂度就越高. PaDiM[37]提取图像每个位置上的多层级特征,为每个位置估计一个多元高斯分布. 模型在测试时根据图像每个位置的特征和原始分布的马氏距离判断是否异常. 由于训练图像中的物体不是完全对齐的,这种方法单独估计每个位置的特征分布时,存在较大的特征差异,不利于异常的判定. PatchCore[38]通过在局部邻域上进行特征聚合的方式来提取特征,并通过贪心策略减少存储块的数量从而降低推理阶段的计算量. Liu等人[39]认为缺陷在图像空间没有太多的共性,结合预训练模型和判别器,对预训练模型提取的特征在特征空间合成异常后,训练一个二分

类判别器识别异常. 此外,一些基于标准化流(Normalization Flow, NF)的方法[40,41]也在异常检测领域取得不错的效果. 这类方法利用预训练网络直接从正常图像中提取特征,在训练阶段将特征分布转换为高斯分布. 测试阶段,异常图像通过NF后的特征会偏离训练阶段的高斯分布,这是NF方法判断异常的依据. 然而,这类直接使用预训练模型的方法未对目标数据集的分布进行迁移和适配,模型对真实场景的适用性不足,可能导致某些异常区域难以定位.

另一种基于预训练模型的方法则采用教师学生框架,根据教师学生网络提取的特征差异检测异常区域. Bergmann等人[10]首次将教师学生网络引入异常检测领域,利用预训练的教师网络训练学生网络模拟教师网络的输出. 当模型中的师生网络的输出差异较大时,模型就能根据师生表征差异推理异常区域. STPM[21]通过多尺度特征蒸馏的方式,将教师网络的丰富"知识"传递给学生网络,并根据教师学生模型的多尺度特征差异生成异常定位图. 然而,这些方法[10,21]的教师学生模型的结构相同且数据流相似,因此针对异常输入时,教师学生网络难以产生较大的表征差异. Yamada等人[22]在知识蒸馏框架中引入了编码解码的结构,使用额外的预训练教师网络引导学生解码器还原无异常特征. 先前的蒸馏方法[10,21]中,师生网络都采用相同或相似的编码器,容易产生相似的数据流,导致师生模型对异常区域表征差异不显著. 为了缓解这个问题,Deng等人[25]使用一对教师编码器和学生解码器进行逆向的知识蒸馏,并增加了瓶颈嵌入模块以解决编码器解码器结构难以重建浅层特征的问题. 然而,简单的瓶颈嵌入模块难以完全过滤教师编码器提取的异常特征,导致学生解码器同样能够还原异常特征. Tien等人[42]认为文献[25]中仅依赖知识蒸馏任务和瓶颈嵌入模块无法为学生网络提供紧凑的特征表示,利用伪异常机制和传递损失确保正常样本投影的特征彼此接近,从而避免异常特征流入学生解码器. 为了解决学生网络无差别复制教师网络表征能力, Xing等人[43]提出了一种基于非对称蒸馏的方法,将不同形式的相同图像作为师生网络的输入,驱动学生网络学习异常区域的区别表示. 然而,这些基于知识蒸馏的方法均是直接蒸馏原始域特征,仍可能导致学生模型完全学习教师模型的先验知识,无法针对异常样本输出有差异的特征[20]. 因此,本文提出双向蒸馏模块约束学生网络对齐教师网络的特征表示,提升图像异常检测的精度.

3 方法

本节详细介绍提出的异常检测方法$\mathcal{N}$-Net,包含双向蒸馏(Bidirectional Distillation, BD)模块和多级过滤模块.

3.1 方法概述

本文所提出的$\mathcal{N}$-Net整体结构如图2所示,它包含两个主要部分:双向蒸馏模块和多级过滤模块. 双向蒸馏模块基于所提出的双向知识蒸馏范式. 其中,教师网络T和正向学生网络S_1构成正向蒸馏分支,教师网络T和逆向学生网络S_2(即学生解码器)构成逆向蒸馏分支. 多级过滤模块进一步过滤逆向蒸馏分支中的潜在异常特征,保证正常特征高精度复原,同时保证异常特征无法被成功复原.

训练阶段,教师网络T提取正常图像I的原始域特征$\boldsymbol{F}_i$,通过正向特征蒸馏方式蒸馏随机初始化参数的正向学生网络S_1,使其学会正常图像的特征表达. 为了避免正向分支中学生网络过度学习教师原始域特征,导致师生网络输出总是一致,本文将原始域特征映射到适应目标数据集的特征空间(称为适应特征空间),并在适应特征空间进行特征蒸馏. 此外,本文引入了逆向蒸馏分支,其中的多级过滤模块保存正常特征和压缩冗余特征,将教师多尺度特征映射为无异常的特征$\boldsymbol{F}_C$. 其中,学生网络S_1和S_2都通过知识蒸馏损失优化.

测试阶段,输入测试图像I',教师T提取I'的多尺度特征$\boldsymbol{F}_i$并映射为$\boldsymbol{P}_i$,正向学生网络S_1提取I'的多尺度特征$\boldsymbol{SF}_i$并映射为$\boldsymbol{SP}_i$. 多级过滤模块将教师特征$\boldsymbol{P}_i$映射为无异常的特征$\boldsymbol{F}_C$,逆向学生网络S_2根据$\boldsymbol{F}_C$输出复原特征$\boldsymbol{D}_i$. 由于学生网络S_1和S_2未学习过教师T对异常样本的表征,遇到异常图像时,师生网络会产生有差异的表征. 因此,本文根据多尺度特征$\boldsymbol{SP}_i$与$\boldsymbol{P}_i$,$\boldsymbol{D}_i$与$\boldsymbol{F}_i$的相似程度计算异常评分图M_1~M_6,并得到最终的像素级异常评分和图像级异常评分.

3.2 双向知识蒸馏模块

在传统的基于蒸馏方法的异常检测中,教师网络的输出通常直接被作为学习目标,学生网络可能完全学习教师模型的知识,弱化了师生模型的异常判别能力[22]. 为了解决上述问题,本文首先提出特征映射模块,将原始域特征映射到适应目标数据集的特征空间(称为适应特征空间),并使用教师网络的适应特征蒸馏正向学生网络S_1. 其中,适应特征空间被定义为原始域特征被映射模块调整后的输出特征. 为了加快收敛和保证特征映射模块的作用,本文仅在训练的初期将学生原始特征$\boldsymbol{SF}_i$对齐到教师映射特征$\boldsymbol{P}_i$. 同时,为了避免特征映射模块参数在训练过程收敛到平凡解,本文进一步引入逆向学生网络,将教师适应特征复原为原始域特征,保证教师适应特征的语义完整性. 因此,特征映射模块被逆向学生网络S_2约束,逆向学生网络S_2将映射特征还原为原始特征,这限制了特征映射模

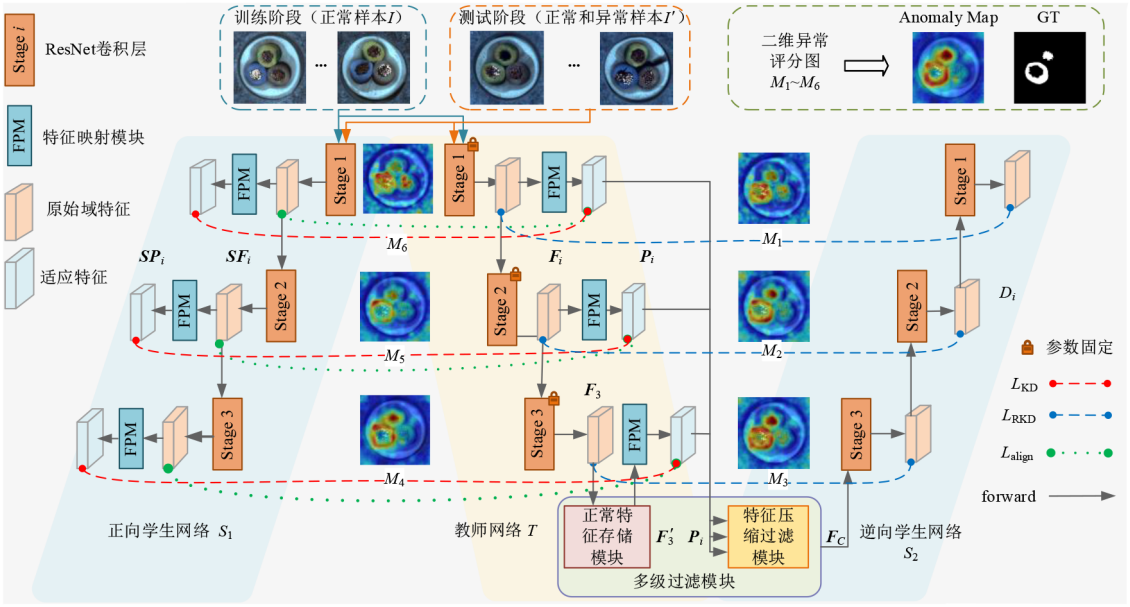


图2　本文提出的$\mathcal{N}$-Net总体架构图

块关注正常特征. 教师网络T、正向学生网络S_1和逆向学生网络S_2构成了本文的双向蒸馏模块,其中包含正向蒸馏分支(T-S_1)和逆向蒸馏分支(T-S_2).

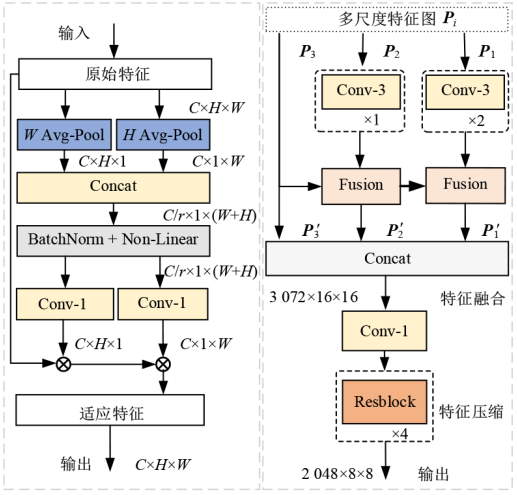
正向蒸馏分支(T-S_1)如图2所示,在本文提出的模型中,教师网络T和正向学生网络S_1都采用了WideResNet-50[44]前4个卷积层,其中选取的3个阶段特征图分别是WideResNet-50[44]的第2~4卷积层的输出特征图. 教师网络T使用预训练的参数并固定,正向学生网络S_1使用随机初始化的参数. 给定图像I,教师网络T和正向学生网络S_1分别提取多尺度的原始域特征$\boldsymbol{F}_i$和$\boldsymbol{SF}_i$. 在正向蒸馏分支中,特征蒸馏损失作用在特征映射模块转化过的适应特征$\boldsymbol{P}_i$和$\boldsymbol{SP}_i$上. 具体地,本文构造了如图3(a)所示的特征映射模块,给定教师网络提取的原始域特征$\boldsymbol{F}_i \in \mathbb{R}^{C\times H\times W}$,特征映射模块分别使用$(H,1)$和$(1,W)$的池化核沿着特征$\boldsymbol{F}_i$的水平和垂直方向聚合特征,这有助于感知正常特征中不同方向的整体信息. 这对特征图在空间维度上拼接到一起后,再在通道维度上进行缩减比为r的降维,得到特征$\boldsymbol{F}_i \in \mathbb{R}^{C/r\times 1\times (W+H)}$. 然后,特征图$F_i$被分割成尺度为$C/r\times H\times 1$和$C/r\times 1\times W$的两部分,并分别被调整通道数至初始大小$C$. 最后,两个具有方向感知能力的特征图通过Sigmoid激活函数,分别得到原始特征图分别沿着h和w方向的注意力权重. 教师网络最终的适应特征通过原始域特征与注意力权重相乘得到. 本文通过特征蒸馏的方式训练正向学生网络S_1的适应特征对齐$\boldsymbol{SP}_i$和教师网络T的适应特征$\boldsymbol{P}_i$. 该部分的正向蒸馏损失为

$$L_{\mathrm{KD}}=L\left(\boldsymbol{P}_i,\boldsymbol{SP}_i\right)=\sum_{i=1}^{T}\left(\frac{1}{h_i w_i}\sum_{h=1}^{h_i}\sum_{w=1}^{w_i}M_i^{(\boldsymbol{P}_i,\boldsymbol{SP}_i)}(h,w)\right) \tag{1}$$

其中,L_{KD}是正向蒸馏损失,i表示师生模型提取的特征层数,h_i和w_i分别是第i层特征图$\boldsymbol{P}_i$和$\boldsymbol{SP}_i$的高度和宽度. $M_i^{(\boldsymbol{P}_i,\boldsymbol{SP}_i)}(h,w)$表示特征图$\boldsymbol{P}_i$和$\boldsymbol{SP}_i$在$(h,w)$处沿着通道方向向量的余弦相似度,它被表示为

$$M_i^{(\boldsymbol{P}_i,\boldsymbol{SP}_i)}(h,w)=1-\frac{\left(\boldsymbol{P}_i(h,w)\right)^{\mathrm{T}}\cdot\boldsymbol{SP}_i(h,w)}{\left\|\boldsymbol{P}_i(h,w)\right\|\left\|\boldsymbol{SP}_i(h,w)\right\|} \tag{2}$$

其中,$\boldsymbol{P}_i(h,w)$和$\boldsymbol{SP}_i(h,w)$表示特征图中(h,w)处沿着通道方向的向量.



(a) 特征映射模块　　(b) 特征融合压缩模块

图3　子模块结构图

如图2所示,在教师网络T通过适应特征$\boldsymbol{P}_i$蒸馏正向学生网络S_1的适应特征$\boldsymbol{SP}_i$时,由于正向学生网络的原始特征$\boldsymbol{SF}_i$未被约束,正向蒸馏难以达到理想效果. 因此本文在训练的初期阶段中额外使用教师映射特征

$\boldsymbol{P}_i$初步对齐正向学生网络原始特征$\boldsymbol{SF}_i$，如图2所示．正向学生网络S_1直接学习教师映射特征$\boldsymbol{P}_i$，能够产生和教师网络T不同的数据流，从而扩大师生网络对异常的表征差异[25]．这是因为当网络结构不同或数据流不相同时，教师学生网络对于未蒸馏过的异常图像是天然不对齐的，这阻碍了学生网络过度学习教师网络的表征能力．具体地，模型在训练的前20个epoch中对齐损失的权重β设置为1，后期的权重β为0．本文通过特征蒸馏约束正向学生网络S_1提取的原始特征$\boldsymbol{SF}_i$对齐教师网络T的适应特征$\boldsymbol{P}_i$，对齐损失表示为

$$L_{\text{align}}=\beta\cdot L\left(\boldsymbol{SF}_i,\boldsymbol{P}_i\right) \tag{3}$$

逆向蒸馏分支（T-S_2）．为了约束教师适应特征$\boldsymbol{P}_i$，避免教师网络T的特征映射模块的参数收敛到平凡解，本文引入逆向学生网络S_2约束教师网络T的适应特征复原为原始输入特征．如图2所示，教师网络T与逆向学生网络S_2组成逆向蒸馏分支，逆向学生网络S_2采用WideResNet-50[44]的第2～4卷积层作为解码块．为了调整逆向学生网络S_2输出的特征图大小与教师原始特征相同，本文采用卷积核大小为2，步距为2的反卷积进行上采样．形式上，教师网络T首先提取输入图像I的原始域多尺度特征$\boldsymbol{F}_i(i=1,2,3)$，其中深层特征$\boldsymbol{F}_3$经过正常特征存储模块（Normal Feature Storage module，NFS）初步过滤异常得到$\boldsymbol{F}_3'$．多尺度特征$\boldsymbol{F}_1$、$\boldsymbol{F}_2$、$\boldsymbol{F}_3'$经过特征映射模块得到适应特征$\boldsymbol{P}_i(i=1,2,3)$．为了保证逆向学生网络能够有效复原浅层特征，本文融合多尺度适应特征$\boldsymbol{P}_i$作为逆向学生网络S_2的输入$\boldsymbol{F}_C$．在训练过程中，逆向学生网络S_2模仿教师网络T的多尺度特征$\boldsymbol{F}_i$，解码高维特征$\boldsymbol{F}_C$得到复原特征$\boldsymbol{D}_i$．本文通过逆向蒸馏损失L_{RKD}约束教师网络T的原始域特征$\boldsymbol{F}_i$和逆向学生网络S_2的复原特征$\boldsymbol{D}_i$．逆向蒸馏损失函数表示为

$$L_{\text{RKD}}=L\left(\boldsymbol{F}_i,\boldsymbol{D}_i\right) \tag{4}$$

3.3 多级过滤模块

在上述双向结构中，预训练的教师网络充当了逆向蒸馏分支的编码器，这可以缓解复原捷径的产生．为了进一步扩大教师网络T和学生网络S_2对异常特征的表征差异，本文提出了多级过滤模块过滤输入逆向学生网络S_2的潜在异常特征，避免异常特征精确复原．这样的优势是保证正常特征可以被复原获得较低的复原误差，而异常特征无法被复原从而产生较大的复原误差．

多级过滤模块包括正常特征存储块和特征融合压缩模块（Feature Fusion Compression module，FFC）．它们分别通过查询过滤方式和压缩过滤方式去除教师适应特征可能存在的异常信息．其中，NFS是一个具有可学习参数模块的矩阵，保存若干正常特征．当使用输入特征查询该矩阵时，它会计算查询特征与矩阵中特征向量的相似程度，以此权重与对应的正常特征向量进行加权求和．最后，正常特征存储模块输出对应的召回向量．召回向量理论上是正常特征存储模块中的正常向量的线性组合，难以在解码阶段复原异常向量．因此，特征存储模块能够起到过滤异常的作用．另外，FFC将原始特征压缩为更少的特征并以此输入学生网络S_2复原原始特征．基于蒸馏的异常检测目的是使未学习的异常图像差异超过正常图像的差异．经过压缩得到的紧凑特征嵌入有助于阻止异常扰动向学生模型传播，从而提高师生模型在异常图像上的表示差异．由于训练阶段的教师提取的正常特征本身是有冗余的，经过大量的正常样本训练，正常特征压缩模块能够学会使用更少的相关正常特征复原原始特征．在测试阶段，即使正常特征被压缩了，也能够正常地还原，而未学习的异常特征经过压缩后更难以复原．

正常特征存储模块．教师网络T提取的特征输入逆向学生网络时，输入部分可能包含复原所需的全部特征，而导致正常和异常特征都能精确复原．为了避免产生上述现象，本文使用正常特征存储模块保存训练过程中的正常特征，并通过查询-召回的方法召回目标特征$\boldsymbol{F}_3'\in\mathbb{R}^{C\times H\times W}$．如图2所示，正常特征存储模块本质上是一个可学习参数矩阵$\boldsymbol{M}\in\mathbb{R}^{N\times C}$，其中$N$是存储模块中存储项的个数，$\boldsymbol{m}_i$表示$\boldsymbol{M}$中的存储项，$i\in[1,N]$．矩阵$\boldsymbol{M}$在训练阶段优化可学习参数，使其中每个存储项$\boldsymbol{m}_i$存储正常语义信息．给定特征$\boldsymbol{F}_3$通道方向的向量作为查询向量$\boldsymbol{q}\in\mathbb{R}^C$．NFS计算$\boldsymbol{q}$与存储模块中各项$\boldsymbol{m}_i$的余弦相似度$w_i$，并将$w_i$中较小项收缩为0，再重新归一化得到$\hat{w}_i$，得到召回向量为$\hat{\boldsymbol{q}}=\sum_{i=1}^{N}\hat{w}_i m_i$．本文使用熵损失[9]作为损失函数，约束正常特征存储模块对正常特征的学习，查询损失表示为

$$L_{\text{query}}=\sum_{i=1}^{N}-\hat{w}_i\log\hat{w}_i \tag{5}$$

特征融合压缩模块．上述正常特征存储模块（NFS）仅对教师原始域特征$\boldsymbol{F}_3$进行了初步的异常过滤，为了进一步去除教师特征中存在的异常信息，本文提出了特征融合压缩模块．如图3（b）所示，在逆向蒸馏分支T-S_2中，教师原始域特征$\boldsymbol{F}_1$、$\boldsymbol{F}_2$和过滤特征$\boldsymbol{F}_3'$映射为多尺度适应特征$\boldsymbol{P}_1$、$\boldsymbol{P}_2$和$\boldsymbol{P}_3$．为了便于后续的特征融合，它们分别经过步长为2的3×3卷积、BN层和ReLU层调整浅层特征的尺度与深层适应特征$\boldsymbol{P}_3$相同，得到特征$\boldsymbol{P}_1'$、$\boldsymbol{P}_2'$和$\boldsymbol{P}_3'$．为了压缩浅层特征中的异常，本文通过逐层融合的方式融合特征$\boldsymbol{P}_i'$，$i=1,2,3$，该过程为

$$\boldsymbol{P}_{i-1}'=\boldsymbol{P}_i'\oplus\boldsymbol{P}_{i-1}' \tag{6}$$

其中，i的取值为3和2，$\boldsymbol{P}_3'=\boldsymbol{P}_3$．本文的特征融合操作$\oplus$表示特征图的逐点相加．为了保证逆向学生网络$S_2$能

够较好地复原教师浅层特征,本文沿着通道维度拼接初步融合的三个阶段的特征图 $\boldsymbol{P}_i'(i=1,2,3)$,并使用卷积核 1×1、步距为 1 的卷积实现通道间的特征融合,从而得到空间尺度为 $3\,072\times16\times16$ 的高维特征 $\boldsymbol{F}_E$.

然而,融合多层适应特征后的高维特征 $\boldsymbol{F}_E$ 空间尺度较大且存在冗余. 为了避免逆向学生网络 S_2 从冗余特征中恢复异常,本文通过 ResNet 残差块[44]进一步压缩特征 $\boldsymbol{F}_E$ 的空间尺度得到压缩特征 $\boldsymbol{F}_C$. 教师原始域特征 $\boldsymbol{F}_i$ 经过正常特征存储模块,特征融合压缩模块去除异常信息后,能够有效避免逆向学生网络 S_2 复原教师网络 T 提取的原始域异常特征.

3.4　损失函数与异常图推理

本文所提方法的总损失,包括正向和逆向蒸馏损失,对齐损失和查询损失,计算公式如下:

$$\mathrm{Loss}=L_{\mathrm{RKD}}+\alpha L_{\mathrm{KD}}+L_{\mathrm{align}}+\delta L_{\mathrm{query}} \tag{7}$$

其中,逆向蒸馏损失的权重为 1,正向蒸馏损失的权重为 α,查询损失的权重为 δ. 对齐损失权重 β 如式(3)所示,主要在训练的初期阶段约束学生原始特征.

在推理阶段,本文根据正向蒸馏分支的师生特征差异和逆向蒸馏分支的师生特征差异推理异常. 本文依据式(2)计算得到多尺度二维异常评分图 $M_1\sim M_6$. 异常评分图表示测试图像的每个区域为异常区域的概率. 为了得到最终的异常评分图,本文使用双线性插值将多尺度的异常评分图 M_i 上采样到 64×64 像素大小. 最终的逐像素的异常评分图 Score_P 即是测试图像的异常分割结果,它的计算过程为

$$\mathrm{Score}_P=\left(M_1+M_6\right)\times\left(M_2+M_5\right)\times\left(M_3+M_4\right) \tag{8}$$

为了去除异常图中的噪声,本文使用高斯滤波器平滑处理[25]异常评分图 Score_P. 高斯滤波器的超参 σ 表示平滑处理时的模糊程度. 样本层面的异常评分 Score_I 由二维异常评分图 Score_P 中最大的异常值决定.

4　实验结果与分析

4.1　实验细节

(1)数据集. 本文在异常检测数据集 MVTec[2]和 VisA[3]上进行了异常检测和定位的实验. MVTec 数据集是一个工业图像异常检测领域的基准数据集,包括工业生产中 15 种类别的产品,其中有 10 种物体类和 5 种纹理类. 数据集中一共有 5 354 张图像,包含大量的正常图像和多种异常图像如划痕、污染和凹痕等缺陷. 训练集有 3 629 张无异常的样本,测试集中包含剩余的正常样本和所有异常样本,共 1 725 张图像. VisA 数据集中有 12 种物品类别,共计 10 821 张图像,包括正常图像 9 621 个,异常图像 1 200 个. 其中,训练集包括 90% 的正常图像,测试集包括 10% 的正常图像和所有的异常图像. VisA 数据集中的物品在整幅图片中的占比更小,同时异常区域也不明显,因此目前的方法在 VisA 数据集上的异常检测效果都不佳. 此外, MVTec 数据集和 VisA 数据集都为有缺陷的图像提供了像素级标签.

(2)实验设置. 本文的 $\mathcal{N}$-Net 采用了预训练的 WideResNet-50[44]作为教师网络 T,随机参数初始化的正向 WideResNet-50 作为正向学生网络 S_1,随机参数初始化的逆向 WideResNet-50 作为逆向学生网络 S_2. 本文将输入图像调整到 256×256 的大小,每个批次处理的图像数为 16. Adam[45]优化器的参数设置为(0.5,0.999),学习率 lr 为 0.05,高斯滤波器的参数 σ 按照惯例设置为 4,所有的类别迭代训练 200 次. 特征融合压缩模块的残差块数量 L 设置为 4. 查询损失的权重 δ 设为 0.002,正向蒸馏损失的权重设置为 1.2,对齐损失中的权重 β 在训练的前 20 个 epoch 为 1,后期设置为 0. 本文使用的 GPU 设备为单张 NVIDIA TITAN RTX.

(3)评价指标. 本文首先引入了近期的工作[37~42]中经常使用 AUROC (Area Under the Receiver Operating Characteristic)指标作为评价异常检测和定位结果的标准. 然而,工业图像数据样本中异常区域通常只占图像的小部分,假阳性率由非异常像素数量所主导,所以尽管有假阳性检测,假阳性率仍保持在较低水平. 先前常用的指标 AUROC 容易被不均衡的正负样本影响,难以全面评价模型的优势. 因此,本文额外使用 AP 指标(Average Precision)[46]和 PRO 指标(Per-Region-Overlap)处理数量不平衡的类别分类. AP 指标作为评价像素级异常区域检测的标准更合理. AP 值通过绘制精确率(Precision)和召回率(Recall)的曲线,计算曲线下的面积得到. 它通常用于不平衡数据集,其中正样本(异常样本)远少于负样本(正常样本). 对于类别不平衡的数据集, AP 可以更好地反映模型在检测少数类时的性能. 与常用的指标相比,PRO 指标能够平等地对待任何大小的异常区域,在现实应用中,人们更关心模型是否能够完全或部分定位一个实例,而不是每个单独的像素. 这些评价指标分为像素级(P)指标和图像级(I)指标,像素级指标评价异常定位结果的准确性,图像级指标评价区分图像是否异常的准确性. 其中, AUROC 指标、PRO 指标和 AP 指标都是值越大代表模型性能表现越优秀.

4.2　量化实验结果

本节比较了本文所提方法 $\mathcal{N}$-Net 和先进方法在异常检测数据集 MVTec 和 VisA 上的量化实验结果. 对比的方法包括基于重构的 FAVAE(Factorized Action Variational AutoEncoder)[18], RIAD(Reconstruction by Inpainting for visual Anomaly Detection)[33],基于归一化流的 FastFlow[41],基于预训练模型的 SPADE(Semantic Pyramid Anomaly Detection)[35], PaDiM(Patch Distribution Modeling framework)[37], PatchCore[38]和基于知识蒸馏的 US(Uninformed Students)[10], STPM

(Student-Teacher feature Pyramid Matching)[21], RD(Reverse Distillation)[25], MemKD (Memory-guided Knowledge Distillation)[1], RD++(Revisiting Reverse Distillation)[42].

表1记录了本文方法与先进方法在工业异常检测数据集MVTec[2]和VisA[3]上的整体性能对比. 对于每个方法,本文记录了该方法在对应数据集上的像素级(P)AP、像素级(P)AUROC、像素级(P)PRO和图像级(I)AUROC指标. 本文所提出的$\mathcal{N}$-Net在这两个数据集的多个评价指标上都取得了先进结果. 在MVTec数据集基准上,本文方法在3个评价标准上得到了最好的性能. 相比其他的基于知识蒸馏的方法(STPM[21], RD[25]和MemKD[1]),本文方法的像素级AP,像素级AUROC和像素级PRO指标相比RD[25]分别提高了3.6、0.3和0.4个百分点,相比最新的方法MemKD[1]分别提高了4.8、0.1和0.4个百分点. 在VisA数据集基准上,本文所提方法的像素级AP指标达到45.8%,相比RD[25]高了1.5个百分点,相比最新的MemKD[1]高了1.7个百分点. 本文所提方法$\mathcal{N}$-Net的像素级AUROC指标达到了99.1%,比RD[25]高了0.5个百分点,比PatchCore[38]高了0.3个百分点,比MemKD[1]高了0.7个百分点. 其中,像素级AP的显著提升反映了本文模型能够较好地处理正常和异常样本不均衡的问题. 像素级的PRO指标有所提升,反映了本文提出的模型对于不同大小的异常区域都有较强的检测能力. 从像素级AP指标对比结果可以看出,与当前最先进的方法MemKD[1]相比,本文方法在像素级异常检测任务上展现了显著的竞争力. 本文所提出的$\mathcal{N}$-Net优于一般的基于蒸馏和重构的方法,显著提升了模型的异常定位能力.

表1 本文方法$\mathcal{N}$-Net与先进方法在图像异常检测数据集MVTec和VisA上的结果对比 单位:%

Dataset	Metric	FastFlow[41]	FAVAE[18]	SPADE[35]	PaDiM[37]	PatchCore[38]	STPM[21]	RD[25]	MemKD[1]	$\mathcal{N}$-Net
MVTec	P-AP	39.8	30.7	47.1	55.0	61.2	51.8	60.8	59.6	**64.4**
	P-AUROC	94.5	88.9	95.5	96.6	98.1	95.4	98.0	98.2	**98.3**
	P-PRO	85.6	74.9	89.5	91.3	93.4	87.9	94.5	94.5	**94.9**
	I-AUROC	90.5	79.3	85.4	90.8	99.2	92.4	98.6	**99.6**	99.1
VisA	P-AP	15.6	21.3	21.5	30.9	40.1	16.9	44.3	44.1	**45.8**
	P-AUROC	88.2	88.0	85.6	89.5	98.8	83.4	98.6	98.4	**99.1**
	P-PRO	59.6	67.9	65.9	85.9	91.2	62.0	94.5	**94.9**	94.7
	I-AUROC	82.2	80.3	82.1	89.1	95.1	83.3	96.1	**97.6**	96.5

此外,本文对比了最近最先进的零样本、小样本和无监督异常检测方法,比较结果如表2所示. 与最先进的无监督方法相比,本文方法的像素级异常定位能力更强,在像素级的AUROC和PRO指标上都达到了最高. 其中,SimpleNet[39]更擅长图像级的异常检测,本文方法也达到了先进水平. 与最先进的小样本和零样本异常检测方法对比,零样本方法[47,48]由于未在特定异常检测数据集上训练,完全消除了数据收集和注释的成本,但是异常检测效果仍有提升空间. 小样本方法AnomalyDiffusion[49]使用了少量异常图像,利用扩散模型从大规模数据集学习到的潜在强先验信息来增强生成异常的真实性. 相比之下,本文在不使用异常图像训练的同时,也具有很强的像素级异常定位能力,尤其是像素级的PRO指标,本文方法的性能超过了现有模型.

表2 本文方法$\mathcal{N}$-Net与不同类型的最新异常检测方法在MVTec数据集上的对比结果 单位:%

Methods	Type	I-AUROC	P-AUROC	P-PRO
AnomalyCLIP[47]	零样本	91.5	91.1	81.4
MuSc[48]	零样本	97.8	97.3	93.8
AnomalyDiffusion[49]	小样本	99.2	99.1	94.0
DiAD[50]	无监督	97.2	96.8	90.7
ADPS[43]	无监督	97.4	98.1	94.4
SimpleNet[39]	无监督	99.6	97.7	91.2
本文	无监督	99.1	98.3	94.9

为了进一步评价所提方法在不同场景下适用性,本文在表3中报告了本文方法和无监督方法在异常检测数据集MVTec上各类别的像素级AP指标对比. 在纹理类中,本文方法的像素级AP指标比RD[25]高了3.6个百分点,在物体类中,本文方法的像素级AP指标比RD++[42]高了1.7个百分点. 这说明本文的方法对纹理类别和物体类别中不同大小的异常区域都具有较好的异常检测能力. 总的来看,在MVTec数据集上15个类别的异常定位任务中,本文所提的$\mathcal{N}$-Net在8个类别中的异常定位性能达到了最佳. 这表明本文方法对于多数场景都具有强大的异常定位能力. 但是针对“Capsule”和“Screw”等少数类别,$\mathcal{N}$-Net仍有不足,其可能原因是这类物品的背景区域在图像中占比较大,模型有时会把背景区域的局部变化视为异常,造成异常定位的精度不佳.

图4直观地对比了$\mathcal{N}$-Net和当前的先进的无监督方法在推理速度(Frames Per Second, FPS)、像素级异常定位指标PRO和模型参数量上的差异. 其中,圆形的面积表示模型参数量的大小. 在像素级的PRO指标上,相比当前最先进的无监督方法RD++[42],本文方法

表3　本文方法$\mathcal{N}$-Net与先进的无监督方法在MVTec数据集上各类别的像素级AP指标对比　　单位:%

Category		US[10]	RIAD[33]	PaDiM[37]	PatchCore[38]	STPM[21]	RD[25]	RD++[42]	SimpleNet[39]	$\mathcal{N}$-Net
Textures	Carpet	52.2	61.4	60.7	66.7	65.3	66.8	64.3	40.5	**69.4**
	Grid	10.1	36.4	35.7	41.0	45.4	49.8	**50.1**	32.4	50.0
	Leather	40.9	49.1	**53.5**	51.0	42.9	52.0	51.3	42.2	49.8
	Tile	65.3	52.6	52.4	59.3	61.7	53.8	54.4	60.9	**71.3**
	Wood	53.3	38.2	46.7	52.3	47.0	50.8	52.6	44.4	**59.1**
	Average	44.4	47.5	49.8	54.1	52.5	54.6	54.5	44.1	**59.9**
Objects	Bottle	74.2	76.4	77.3	80.1	80.6	79.1	79.7	71.0	**83.0**
	Cable	48.2	24.4	45.4	70.0	58.0	59.6	61.7	66.9	**70.6**
	Capsule	25.9	38.2	46.7	**48.1**	35.9	45.1	47.1	41.3	42.5
	Hazelnut	57.8	33.8	61.1	61.5	60.3	67.9	65.7	45.1	**68.2**
	Metal Nut	83.5	64.3	77.4	88.8	79.3	82.3	83.5	**89.4**	83.3
	Pill	62.0	51.6	61.2	78.7	63.3	79.4	79.8	80.0	**81.7**
	Screw	7.8	43.9	21.7	41.4	26.9	54.5	**55.6**	35.3	51.1
	Toothbrush	37.7	50.6	54.7	51.6	48.8	54.1	56.2	38.5	**70.9**
	Transistor	27.1	39.2	**72.0**	63.2	44.4	55.6	59.1	67.5	58.8
	Zipper	36.1	63.4	58.2	**64.0**	54.9	60.9	61.1	62.2	56.5
	Average	46.0	48.6	57.6	64.7	55.2	63.9	65.0	59.7	**66.7**
Total Average		45.5	48.2	55.0	61.2	54.3	60.8	61.5	54.5	**64.4**

在参数量和推理速度上有明显优势,相比当前推理速度最快的SimpleNet[39],本文方法在异常定位方面更有优势. 总的来说,本文所提方法在取得更高异常定位精度的同时兼顾了计算复杂度.

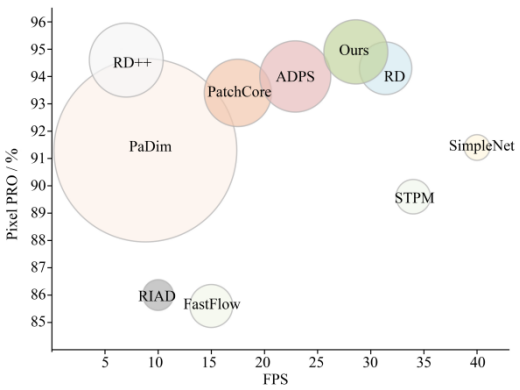


图4　本文方法与先进方法在异常定位指标、推理速度和参数量上的可视化差异

4.3　可视化分析

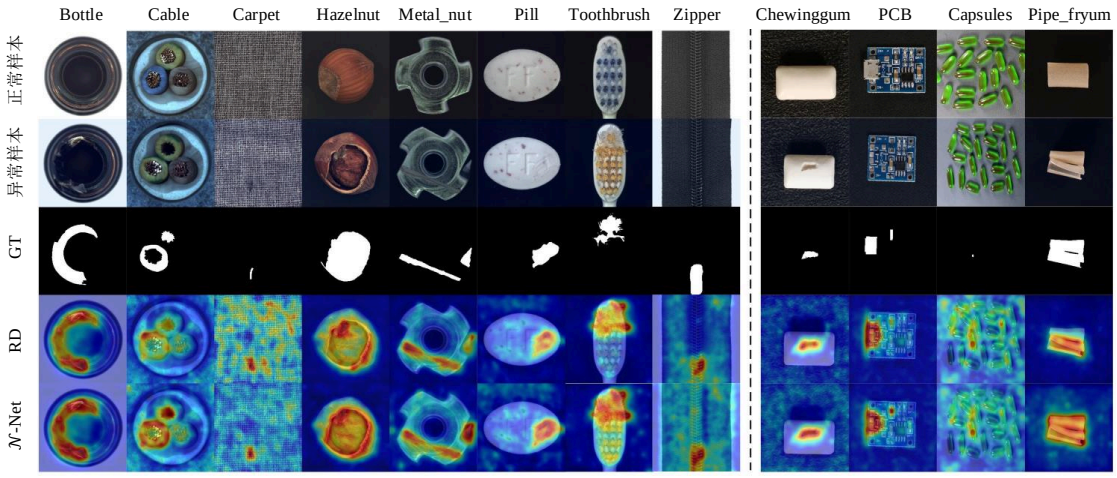
图5对比了本文的$\mathcal{N}$-Net与先进的方法RD[25]在MVTec数据集和VisA数据集上的异常定位结果. 图5中GT的白色部分表示异常区域,黑色表示正常区域,异常定位图中红色表示异常区域,蓝色表示正常区域. 从可视化结果可以看出,$\mathcal{N}$-Net的异常定位结果更精确,与GT更加接近. 例如在"Cable"和"PCB"等类别的异常定位任务中,RD[25]未能检测出图像中的所有异常区域. 对于基于蒸馏方法的RD[25],这种情况是因为学生网络过多学习了教师网络的表征能力,导致教师和学生网络对于异常表征差异很小. 本文从2个方面缓解了这种问题. (1)本文在适应特征空间对齐师生特征,避免正向学生网络过度学习教师提取的原始特征. (2)本文提出了多级过滤模块,避免异常特征通过逆向学生网络复原. $\mathcal{N}$-Net扩大了教师学生网络对异常区域的表征差异,有利于模型检测出细微和不明显的异常区域. 此外,在MVTec数据集的"Carpet"类和VisA数据集的"Capsules"类中,RD[25]除了检测出异常区域外,也对一些正常区域产生了误检测. 这是因为物品表面有轻微的变化,而RD[25]的学生网络未能完全学会这种表征,导致教师学生网络之间表征差异较大.

本文提出的2种学生网络能够充分学习教师网络的表征能力,同时联合模型多阶段的特征差异推理异常,能够缓解误检测的问题. 这些可视化对比结果进一步验证了本文方法在异常定位任务中的有效性.

4.4　消融实验

为了验证本文方法中各模块的有效性,本节首先对所提模块进行了定量或定性的消融实验,同时对各模块的作用进行了分析. 然后,分析了损失函数的变化对模型的影响,并研究了不同阶段异常预测图对最终检测性能的影响. 最后,针对各模块中的参数设置进行了消融实验.

如表4所示,本文在MVTec数据集上通过消融实验验证了特征映射模块,双向蒸馏模块和多级过滤模块的有效性. 其中,本部分实验的基线模型是一个简



(*a*) MVTec数据集上结果的可视化 (*b*) VisA数据集上结果的可视化

图5 本文方法$\mathcal{M}$-Net与RD[25]的可视化对比

表4 本文所提模块的消融实验 单位:%

序号	本文所提模块				性能指标			
	基线模型	特征映射	多级过滤	双向蒸馏	P-AP	P-AUROC	P-PRO	I-AUROC
1	√				60.78	97.96	94.45	98.62
2	√	√			61.18	97.95	94.47	98.75
3	√		√		61.41	98.11	94.59	98.70
4	√	√		√	62.77	98.23	94.67	**98.78**
5	√	√	√	√	**63.03**	**98.32**	**94.90**	98.74

单的逆向蒸馏分支,它包括教师网络T和逆向学生网络S_2,以WideResNet-50[44]为主干网络,仅使用简单的瓶颈模块代替多级过滤模块. 首先,基线模型在单独加入特征映射模块后,模型的像素级AP指标提升0.4个百分点,特征映射模块对于基线模型的作用不显著,这是因为基线方法中的逆向学生网络对齐的是教师原始域特征. 其次,基线模型单独加入所设计的多级过滤模块,像素级AP从60.7%提高到了61.41%,像素级AUROC从97.96%提高到98.11%,其余指标也得到不同程度的提升. 这说明多级过滤模块能够有效过滤基线模型中教师网络提取的异常特征,避免逆向学生网络成功复原异常特征. 然后,基线模型同时加入特征映射模块和双向蒸馏模块,模型性能的提升更加明显,像素级AP指标提升到了62.77%,像素级AUROC提升到了98.23%,图像级AUROC也达到了98.81%的最佳结果. 这验证了本文提出的双向蒸馏模块的有效性. 最后,基线模型同时加入这3个模块,模型的像素级AP和AUROC提高到了63.03%和98.32%,本文方法达到最佳的性能. 可以看出,本文所提模块有助于提升模型对图像异常区域的判别能力.

为了进一步分析多级过滤模块内部结构对模型性能的影响,本文对正常特征存储模块和特征融合压缩模块进行消融实验. 如表5所示,本文首先在基线模型(Baseline)中加入正常特征存储模块,可以发现AP指标提高了约0.6个百分点,但是其余指标的提升效果不够显著. 从这个结果可以看出,仅对教师网络T的深层特征$\boldsymbol{F}_3$使用正常特征存储模块过滤异常时,模型性能提升效果比较有限. 这是因为正常特征存储模块仅仅过滤了深层语义特征,瓶颈模块将教师网络T提取的多层次的特征融合后,逆向学生网络S_2仍然有可能恢复浅层的异常特征. 因此,本文加入多尺度的特征融合压缩模块与正常特征存储模块共同使用,模型的像素级AP和AUROC指标进一步提高至61.41%和98.11%. 可见,本文提出的多级过滤模块能够有效过滤异常,从而扩大师生网络之间对异常的表征差异.

为了直观地比较本文双向蒸馏模块相比普通蒸馏

表5 多级过滤模块中正常特征存储模块和特征融合压缩模块的消融实验 单位:%

序号	模型	P-AP	P-AUROC	P-PRO	I-AUROC
1	Baseline	60.78	97.96	94.45	98.62
2	Baseline+NFS	61.31	98.03	94.54	98.68
3	Baseline+FFC	61.26	98.07	**94.62**	98.65
4	Baseline+NFS+FFC	**61.41**	**98.11**	94.59	**98.70**

方法的提升结果,本文可视化了2种蒸馏方法多尺度的异常图.

为了公平起见,普通的蒸馏模型KD同样使用与本文相同的主干网络. 可视化结果如图6所示,本文的方法预测的尺度为64×64的异常图效果明显优于普通的蒸馏模型. 可以看到本文的方法针对局部的细微变化,明显地降低了误检测的现象. 这验证了本文双向蒸馏模块的学生网络具有更强大的异常检测能力.

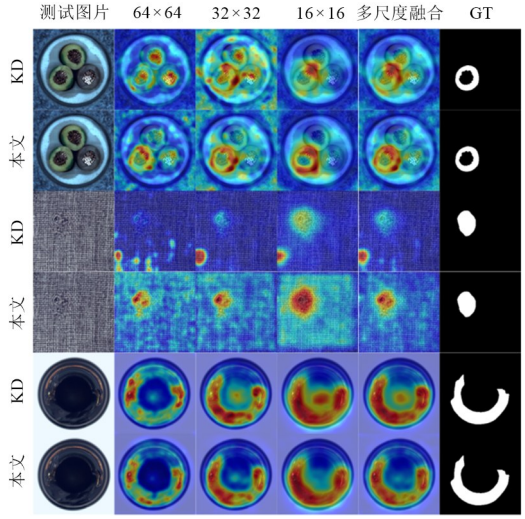
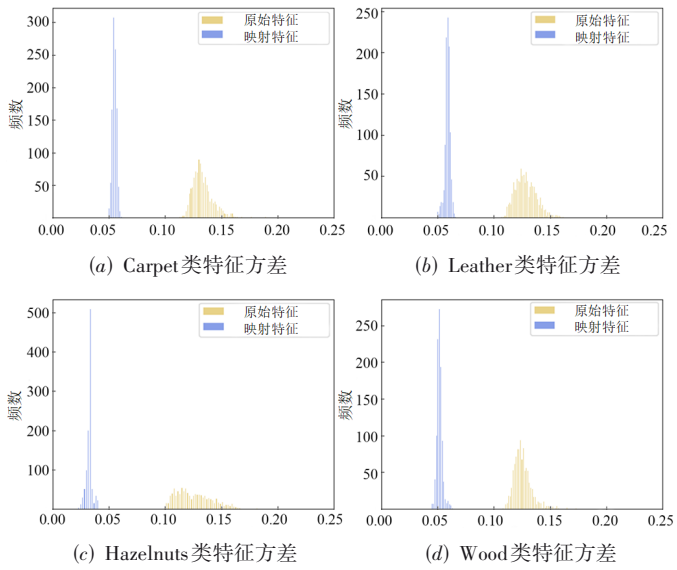


图6　本文双向蒸馏方法与正向蒸馏的可视化结果图

逆向蒸馏分支中的特征映射模块能够将教师编码器提取的特征映射到更加紧凑的空间,这有利于瓶颈模块过滤异常特征. 为了验证特征映射模块的作用,本文可视化了特征映射前后的特征沿着通道方向的方差. 如图7所示,教师编码器提取的原始特征经过特征映射模块后,原本方差较大的正常特征向量变得更加紧凑. 这有助于逆向蒸馏分支过滤异常特征,从而保证异常特征无法被复原.

为了进一步分析双向蒸馏模块中正向蒸馏分支(T-S_1)对应的损失函数对模型性能的贡献,本文对约束正向蒸馏分支的正向蒸馏损失L_{KD}和对齐损失L_{align}进行了消融实验分析. 为避免多级过滤模块的影响,实验中仅使用双向蒸馏模块和特征映射模块作为基线模型,损失函数仅使用逆向蒸馏分支中的逆向蒸馏损失L_{RKD},此时的正向蒸馏分支未被约束. 基线模型分别添加正向蒸馏损失L_{KD}和对齐损失L_{align},实验结果如表6所示. 首先,基线模型仅添加对齐损失时,由于正向学生网络S_1的特征映射模块未被约束,正向蒸馏分支的异常评分图根据正向学生网络的原始特征$\boldsymbol{SF}_i$和教师映射特征$\boldsymbol{P}_i$推理得到. 此时,模型的像素级AP指标提升到61.35%,像素级AUROC指标提升到98.10%. 这说明正向学生网络对模型的性能有一定的贡献,但是正向学生网络S_1直接学习教师网络T的适应特征$\boldsymbol{P}_i$存在困难,因此双向蒸馏模块的性能提升相对有限. 然后,基线模型添加正向蒸馏损失约束正向学生网络S_1的特征映射模块时,这时正向蒸馏分支的异常评分图根据正向学生网络的映射特征$\boldsymbol{SP}_i$和教师映射特征$\boldsymbol{P}_i$推理得到. 像素级AP指标提升到了62.43%,像素级AUROC指标提升到了98.24%. 这证明了在适应特征空间蒸馏正向学生网络S_1的特征$\boldsymbol{SP}_i$能够有效提升双向蒸馏模块的性能. 最后,当基线模型同时使用2种损失时,像素级AP指标达到了62.77%,相比基线模型提升1.59个百分点,像素级AUROC指标提升约为0.3个百分点,像素级PRO指标提升0.2个百分点. 这表明正向学生网



(a) Carpet类特征方差　(b) Leather类特征方差

(c) Hazelnuts类特征方差　(d) Wood类特征方差

图7　特征映射模块作用前后的特征沿着通道方向的方差

络初步对齐教师适应特征后，再在适应域和教师适应特征进行知识蒸馏，有利于增强双向蒸馏模块对异常的判别能力. 从上述实验可以看出，本文所提出的正向蒸馏损失 L_{KD} 和对齐损失 L_{align} 在双向蒸馏模块中起着重要的作用.

表6 双向蒸馏模块中正向蒸馏损失和对齐损失的消融实验 单位:%

序号	本文所提损失函数			性能指标			
	Base_L	L_{KD}	L_{align}	P-AP	P-AU-ROC	P-PRO	I-AU-ROC
1	√			61.18	97.95	94.47	98.75
2	√	√		62.43	**98.24**	94.58	98.84
3	√		√	61.35	98.10	94.58	**98.85**
4	√	√	√	**62.77**	98.23	**94.67**	98.81

另外，本文在不同损失函数的配置下，可视化了模型使用正向蒸馏损失 L_{KD} 和对齐损失 L_{align} 对结果的影响. 如图8所示，基线模型使用单独的损失函数时，在“Tile”和“Hazelnut”类别中出现了异常区域检测不全的问题，而在“Carpet”和“Toothbrush”类别中，可视化结果中存在误检测和定位准确性差的情况. 当使用两种损失同时约束双向蒸馏模块中正向蒸馏分支时，这些问题能被显著改善. 这表明同时使用正向蒸馏损失和对齐损失有助于约束正向学生网络在适应域对齐教师特征，提升双向蒸馏模块的检测性能.

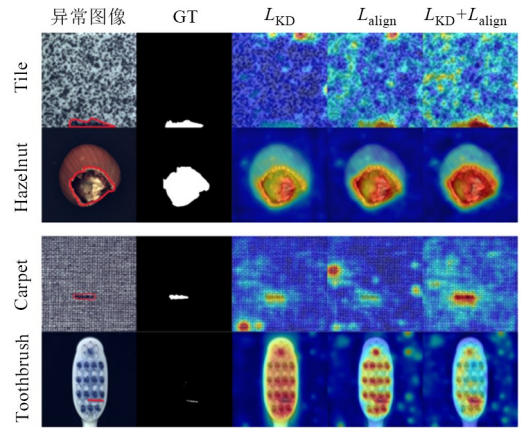


图8 双向蒸馏模块中正向蒸馏损失和对齐损失的可视化

此外，本文在MVTec数据集上实验对比了 $\mathcal{N}$-Net 使用不同类型的损失函数对检测结果的影响，其中包括常用的 L_1 损失、L_2 损失和余弦相似度损失 cosine. 图9记录了本文方法使用不同类型的损失函数进行训练的结果，其中包括了模型在MVTec数据集各类别上的像素级AUROC性能指标和平均结果. 使用余弦相似度损失训练的模型在9个类别中达到了实现最佳的检测结果，并且像素级AUROC的平均结果也明显优于使

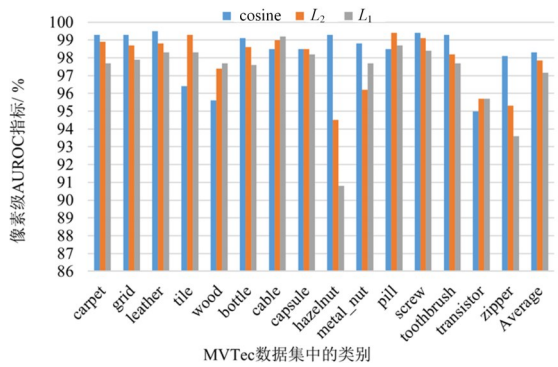


图9 不同类型损失函数的消融实验

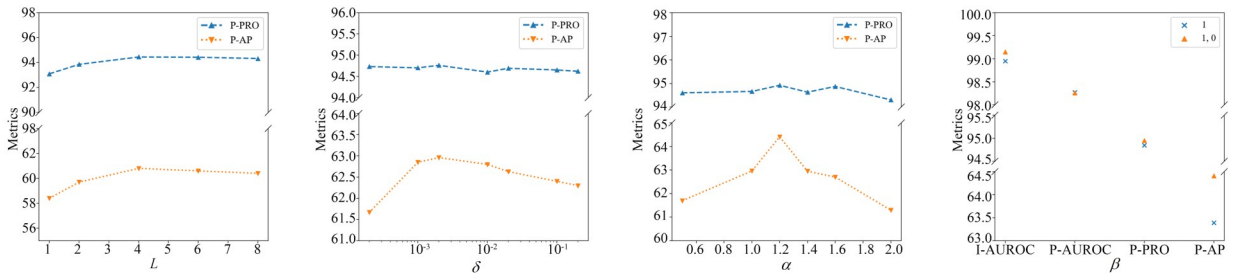
用 L_1 损失和 L_2 损失训练的结果. 这是因为高维向量间的余弦距离不会受限于向量维度的影响，而 L_1 损失和 L_2 损失都容易受向量维度的影响，损失值的范围不固定. 由于本文模型中特征图通道方向的向量维度较高，因此，本文针对这个任务选择了基于余弦相似度的损失函数.

本文针对正向蒸馏分支和逆向蒸馏分支中的异常评分图组合方式进行分析. 如表7所示，这些组合分别是仅使用逆向蒸馏分支的异常图 M_1、M_2、M_3，仅使用正向蒸馏分支的异常图 M_4、M_5、M_6 和联合异常图 $M_1 \sim M_6$. 实验结果表明，联合正向和逆向蒸馏的异常图 $M_1 \sim M_6$ 检测异常能得到最佳的结果.

表7 不同阶段异常图的组合对结果的影响 单位:%

Metrics	M_1、M_2、M_3	M_4、M_5、M_6	$M_1 \sim M_6$
P-AUROC	98.1	96.9	**98.3**
I-AUROC	**99.0**	96.6	98.8
P-AP	61.4	57.2	**63.0**

最后，本文针对网络模型中的超参数进行了消融分析，结果如图10所示. 首先，本文仅对逆向蒸馏分支进行研究，其中的FFC模块的残差块数量 L 对逆向蒸馏分支的作用结果如图10(a)所示. 当 L 增加到一定程度时，对逆向蒸馏分支的作用逐渐趋于稳定，因此本文选取数量 L=4. 然后，本文在整个模型中调整查询损失的权重 δ. 如图10(b)所示，权重 δ 仅对数量级较敏感，在查询损失权重 δ=0.002 时，模型的性能最好. 其次，本文固定超参 δ，调整正向蒸馏损失权重 α. 正向蒸馏损失和逆向蒸馏损失是相似的损失，理论上两者权重应当比较相近. 从图10(c)中可以看出，当 α=1.2 时，模型的性能达到最佳. 最后，本文固定参数 δ 和 α，比较对齐损失权重 β 恒为1和训练阶段前20个epoch为1，后期为0时对性能的影响. 从图10(d)可以看出，训练初期，学生网络对齐教师映射特征，后期训练学生映射特征对齐教师映射特征，网络能够获得较好的性能.



(*a*) 特征融合压缩模块数L的消融实验　(*b*) 损失权重δ的消融实验　(*c*) 损失权重α的消融实验　(*d*) 对齐损失权重β的消融实验

图 10　本文模型中超参数的消融实验

5　结束语

本文针对图像异常检测任务,提出了基于双向约束蒸馏的无监督图像异常检测方法$\mathcal{N}$-Net. 首先,不同于传统的蒸馏模型直接学习教师特征,本文提出了特征映射模块让正向学生网络仅学习教师适应域特征,避免了学生模型的过度学习教师表征能力. 其次,$\mathcal{N}$-Net引入了逆向蒸馏分支,约束教师网络适应域特征被精确还原为原始域特征的同时保证了语义完整性. 然后,本文提出了多级过滤模块,约束逆向蒸馏分支仅学习正常特征的复原能力,从而扩大异常特征复原前后的差异. 最后,本文根据师生模型对输入图像的表征差异推理异常区域,实现高精度的异常检测和定位. 本文在图像异常检测数据集MVTec和VisA上进行了大量实验,验证了异常检测模型$\mathcal{N}$-Net性能的优越性和所提模块的有效性. 未来,我们将会进一步探索不对称网络结构的蒸馏范式,实现高效的异常检测.

参考文献

[1] GU Z H, LIU L, CHEN X, et al. Remembering normality: Memory-guided knowledge distillation for unsupervised anomaly detection[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 16355-16363.

[2] BERGMANN P, FAUSER M, SATTLEGGER D, et al. MVTec AD: A comprehensive real-world dataset for unsupervised anomaly detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 9584-9592.

[3] ZOU Y, JEONG J, PEMULA L, et al. SPot-the-difference self-supervised pre-training for anomaly detection and segmentation[M]//Computer Vision - ECCV 2022. Cham: Springer Nature Switzerland, 2022: 392-408.

[4] 严莉, 张凯, 徐浩, 等. 基于图注意力机制和Transformer的异常检测[J]. 电子学报, 2022, 50(4): 900-908.
YAN L, ZHANG K, XU H, et al. Abnormal detection based on graph attention mechanisms and transformer[J]. Acta Electronica Sinica, 2022, 50(4): 900-908. (in Chinese)

[5] PARK H, NOH J, HAM B. Learning memory-guided normality for anomaly detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 14372-14381.

[6] ZHAO Y R, DENG B, SHEN C, et al. Spatio-temporal AutoEncoder for video anomaly detection[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM, 2017: 1933-1941.

[7] XIANG T G, ZHANG Y X, LU Y Y, et al. SQUID: Deep feature in-painting for unsupervised anomaly detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 23890-23901.

[8] 尚文利, 石贺, 赵剑明, 等. 基于SAE-LSTM的工艺数据异常检测方法[J]. 电子学报, 2021, 49(8): 1561-1568.
SHANG W L, SHI H, ZHAO J M, et al. An anomaly detection method of process data based on SAE-LSTM[J]. Acta Electronica Sinica, 2021, 49(8): 1561-1568. (in Chinese)

[9] GONG D, LIU L Q, LE V, et al. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 1705-1714.

[10] BERGMANN P, FAUSER M, SATTLEGGER D, et al. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 4182-4191.

[11] HOU J L, ZHANG Y Y, ZHONG Q Y, et al. Divide-and-assemble: Learning Block-wise memory for unsupervised anomaly detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 8771-8780.

[12] SCHLEGL T, SEEBÖCK P, WALDSTEIN S M, et al. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery[M]//Information Processing in Medical Imaging. Cham: Springer International Publishing, 2017: 146-157.

[13] AKCAY S, ATAPOUR-ABARGHOUEI A, BRECKON T P. GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training[M]//Computer Vision - ACCV 2018. Cham: Springer International Publishing, 2019: 622-637.

[14] AKÇAY S, ATAPOUR-ABARGHOUEI A, BRECKON T P. Skip-GANomaly: Skip connected and adversarially trained encoder-decoder anomaly detection[C]//2019 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE, 2019: 1-8.

[15] ZONG B, SONG Q, MIN M R, et al. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection[C]//International Conference on Learning Representations. Scottsdale: ICLR, 2018: 1-19.

[16] XING P, LI Z C. Visual anomaly detection via partition memory bank module and error estimation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 3596-3607.

[17] YAN X D, ZHANG H D, XU X M, et al. Learning semantic context from normal samples for unsupervised anomaly detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(4): 3110-3118.

[18] DEHAENE D, ELINE P. Anomaly localization by modeling perceptual features[EB/OL]. (2020-08-12) [2024-08-07]. https://arxiv.org/abs/2008.05369v1.

[19] LI Z, LI N, JIANG K, et al. Superpixel masking and inpainting for self-supervised anomaly detection[C]//British Machine Vision Conference. London: BMVC, 2020: 219-232.

[20] LIU J Q, XIE G Y, WANG J B, et al. Deep industrial image anomaly detection: A survey[J]. Machine Intelligence Research, 2024, 21(1): 104-135.

[21] WANG G, HAN S, DING E, et al. Student-teacher feature pyramid matching for anomaly detection[C]//British Machine Vision Conference. London: BMVC, 2021: 1589-1605.

[22] YAMADA S, KAMIYA S, HOTTA K. Reconstructed student-teacher and discriminative networks for anomaly detection[C]//2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE, 2022: 2725-2732.

[23] 邢鹏, 蒋鑫, 潘永华, 等. 基于特征约束蒸馏学习的视觉异常检测[J]. 软件学报, 2023, 34(9): 4378-4391.
XING P, JIANG X, PAN Y H, et al. Feature constrained restricted distillation learning for visual anomaly detection[J]. Journal of Software, 2023, 34(9): 4378-4391. (in Chinese)

[24] SALEHI M, SADJADI N, BASELIZADEH S, et al. Multiresolution knowledge distillation for anomaly detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 14902-14912.

[25] DENG H Q, LI X Y. Anomaly detection via reverse distillation from one-class embedding[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 9727-9736.

[26] MOYA M M, HUSH D R. Network constraints and multiobjective optimization for one-class classification[J]. Neural Networks, 1996, 9(3): 463-474.

[27] SCHÖLKOPF B, PLATT J C, SHAWE-TAYLOR J, et al. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2001, 13(7): 1443-1471.

[28] RUFF L, VANDERMEULEN R, GOERNITZ N, et al. Deep one-class classification[C]//International Conference on Machine Learning. Stockholm: ICML, 2018: 4393-4402.

[29] YI J H, YOON S. Patch SVDD: Patch-Level SVDD for anomaly detection and segmentation[M]//Computer Vision - ACCV 2020. Cham: Springer International Publishing, 2021: 375-390.

[30] YAO X C, LI R Q, ZHANG J, et al. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 24490-24499.

[31] KINGMA D P, WELLING M. Auto-encoding variational Bayes[J]. 2nd International Conference on Learning Representations ICLR 2014 - Conference Track Proceedings. Scottsdale: ICLR, 2014: 33.

[32] GOODFELLOW I, POUGET-Abadie J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. New York: ACM, 2014: 2672-2680.

[33] ZAVRTANIK V, KRISTAN M, SKOČAJ D. Reconstruction by inpainting for visual anomaly detection[J]. Pattern Recognition, 2021, 112: 107706.

[34] ZHANG X, LI S Y, LI X, et al. DeSTSeg: Segmentation guided denoising student-teacher for anomaly detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 3914-3923.

[35] COHEN N, HOSHEN Y. Sub-image anomaly detection with deep pyramid correspondences[EB/OL]. (2021-02-03)[2024-08-07]. https://arxiv.org/abs/2005.02357v3.

[36] COVER T, HART P. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.

[37] DEFARD T, SETKOV A, LOESCH A, et al. PaDiM: A patch distribution modeling framework for anomaly detection and localization[M]//Pattern Recognition ICPR In-

ternational Workshops and Challenges. Cham: Springer International Publishing, 2021: 475-489.
[38] ROTH K, PEMULA L, ZEPEDA J, et al. Towards total recall in industrial anomaly detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 14298-14308.
[39] LIU Z K, ZHOU Y M, XU Y S, et al. SimpleNet: A simple network for image anomaly detection and localization[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 20402-20411.
[40] RUDOLPH M, WEHRBEIN T, ROSENHAHN B, et al. Fully convolutional cross-scale-flows for image-based defect detection[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2022: 1829-1838.
[41] YU J W, ZHENG Y, WANG X, et al. FastFlow: Unsupervised anomaly detection and localization via 2D normalizing flows[EB/OL]. (2021-11-16)[2024-08-07]. https://doi.org/10.48550/arXiv.2111.07677.
[42] TIEN T D, NGUYEN A T, TRAN N H, et al. Revisiting reverse distillation for anomaly detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 24511-24520.
[43] XING P, TANG H, TANG J H, et al. ADPS: Asymmetric distillation postsegmentation for image anomaly detection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(4): 7051-7064.
[44] ZAGORUYKO S, KOMODAKIS N. Wide residual networks[EB/OL]. (2017-06-14)[2024-08-07]. https://arxiv.org/abs/1605.07146v4.
[45] KINGMA D P, BA J. Adam: A method for stochastic optimization[EB/OL]. (2017-1-30) [2024-08-07]. https://doi.org/10.48550/arXiv.1412.6980.
[46] ZAVRTANIK V, KRISTAN M, SKOČAJ D. DRÆM-A discriminatively trained reconstruction embedding for surface anomaly detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 8310-8319.
[47] ZHOU Q, PANG G, TIAN Y, et al. Anomalyclip: Objectagnostic prompt learning for zero-shot anomaly detection[C]//International Conference on Learning Representations. Scottsdale: ICLR, 2024: 3565-3583.
[48] LI X R, HUANG Z M, XUE F, et al. MuSc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images[EB/OL]. (2024-01-30)[2024-08-07]. https://arxiv.org/abs/2401.16753v1.
[49] HU T, ZHANG J N, YI R, et al. AnomalyDiffusion: Few-shot anomaly image generation with diffusion model[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(8): 8526-8534.
[50] HE H Y, ZHANG J N, CHEN H X, et al. A diffusion-based framework for multi-class anomaly detection[C]//Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. New York: ACM, 2024: 8472-8480.

作者简介

李　波　男,2000年4月出生. 现为南京理工大学计算机科学与工程学院硕士研究生. 主要研究方向为异常检测、计算机视觉.
E-mail: 122106010714@njust.edu.cn

邢　鹏　男,1998年9月出生. 现为南京理工大学计算机科学与工程学院博士研究生. 主要研究方向为异常检测、计算机视觉、AIGC.
E-mail: xingp_ng@njust.edu.cn

李泽超　男,1985年5月出生. 现为南京理工大学计算机科学与工程学院教授,副院长,博士生导师. 主要研究方向为图像视频分析、目标检测、模式识别. 中国电子学会会员编号:E190031283S.
E-mail: zechao.li@njust.edu.cn

唐金辉　男,1981年2月出生. 现为南京理工大学计算机科学与工程学院教授,院长. 主要研究方向为多媒体分析与计算机视觉相关工作. 中国电子学会会员编号:E190031289M.
E-mail: jinhuitang@njust.edu.cn