软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(9):4378−4391 [doi: 10.13328/j.cnki.jos.006643]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于特征约束蒸馏学习的视觉异常检测*

邢 鹏, 蒋 鑫, 潘永华, 唐金辉, 李泽超

(南京理工大学 计算机科学与工程学院, 江苏 南京 210094)

通信作者: 李泽超, E-mail: zechao.li@njust.edu.cn

摘 要: 针对视觉异常检测任务, 提出一种基于特征约束的蒸馏学习方法, 充分利用教师网络模型的特征来指导学生模型高效的识别异常图像. 具体地, 引入 vision transformer (ViT) 作为异常检测任务的主干网络, 并提出中心特征策略约束学生网络的输出特征. 由于教师网络的特征表达能力较强, 特征中心策略从教师网络中动态地为学生网络生成正常样本的特征表示中心, 从而提升学生网络对正常数据特征输出的描述能力, 进而扩大了学生网络和教师网络对于异常数据的特征差异; 另一方面, 为了最小化学生网络和教师网络在正常图像特征表示上的差异, 引入格拉姆 (Gram) 损失函数对学生网络编码层之间的关系进行约束. 在 3 个异常检测通用数据集和 1 个真实工业异常检测数据集上进行了实验验证, 相比当前最优方法, 所提方法取得了显著的性能提升.

关键词: 异常检测; 特征蒸馏; 异常评分; 中心分布; 一致性约束

中图法分类号: TP391

中文引用格式: 邢鹏, 蒋鑫, 潘永华, 唐金辉, 李泽超. 基于特征约束蒸馏学习的视觉异常检测. 软件学报, 2023, 34(9): 4378–4391. http://www.jos.org.cn/1000-9825/6643.htm

英文引用格式: Xing P, Jiang X, Pan YH, Tang JH, Li ZC. Feature Constrained Restricted Distillation Learning for Visual Anomaly Detection. Ruan Jian Xue Bao/Journal of Software, 2023, 34(9): 4378–4391 (in Chinese). http://www.jos.org.cn/1000-9825/6643.htm

Feature Constrained Restricted Distillation Learning for Visual Anomaly Detection

XING Peng, JIANG Xin, PAN Yong-Hua, TANG Jin-Hui, LI Ze-Chao

(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: This study proposes a new feature constrained distillation learning method for visual anomaly detection, which makes full use of the features of the teacher model to instruct the student model to efficiently identify abnormal images. Specifically, the vision transformer (ViT) model is introduced as the backbone network of anomaly detection tasks, and a central feature strategy is put forward to constrain the output features of the student network. Considering the strong feature expressiveness of the teacher network, the central feature strategy is developed to dynamically generate the feature representation centers of normal samples for the student network from the teacher network. In this way, the ability of the student network to describe the feature output of normal data is improved, and the feature difference between the student and teacher networks in abnormal data is widened. In addition, to minimize the difference between the student and teacher networks in the feature representation of normal images, the proposed method leverages the Gram loss function to constrain the relationship between the coding layers of the student network. Experiments are conducted on three general anomaly detection data sets and one real-world industrial anomaly detection data set, and the experimental results demonstrate that the proposed method significantly improves the performance of visual anomaly detection compared with the state-of-the-art methods.

Key words: anomaly detection; feature distillation; anomaly score; central distribution; consistency constraint

1 引 言

在计算机视觉领域, 异常检测 (anomaly detection) 是指从输入图像中识别出与正常图像不同的图像数据[1]. 由

* 基金项目: 国家自然科学基金 (U20B2064, U21B2043)
收稿时间: 2021-06-26; 修改时间: 2021-09-27; 采用时间: 2021-12-13; jos 在线出版时间: 2022-12-22
CNKI 网络首发时间: 2022-12-26

于异常数据的稀缺性和多样性, 异常检测与常规的分类任务区别在于训练阶段模型仅使用正常数据, 异常数据不可用[2]. 异常检测使用正常样本训练模型, 利用训练得到的模型判断新图像是否属于正常样本[3]. 异常检测通常应用于森林火灾检测, 病理图像异常检测[4], 工业数据异常检测[5]中. 以森林火灾检测任务为例, 模型在训练时几乎无法获得火灾发生时的数据, 因此常规的分类方法在这些场景下无法使用, 而异常检测算法则能够有效地处理这些任务.

针对视觉异常检测任务, 不少方案已经被提出. 例如在基于图像重建的异常检测方案中, An 等人提出通过自动编码器网络学习重建正常图像[6], 这类工作的假设是自编码器网络模型只能重建正常图像, 而无法重建未学习过的类别图像, 因此测试时可以通过比较重建图像和原始图像的差异来判断输入是否异常. 但是由于自编码器重建能力强, 重建后得到的差异并不明显[7], 这类方法仍不能很好地解决异常检测任务. 在基于知识蒸馏的异常检测方案中, 文献 [8,9] 引入了孪生网络或者教师-学生网络. 仅使用正常样本的图像数据训练学生网络, 并通过教师网络和学生网络的输出特征差异来判断是否为异常. 然而, 现有的知识蒸馏方法并未充分利用教师网络学习得到的特征表示, 从而限制了异常检测的性能. 在文献 [9] 中, 学生网络直接学习教师网络对于正常样本的中间池化层特征表示, 并利用输出特征的差异来衡量异常值的高低. 由于池化层丢失较多细节信息, 在输入高分辨率图像时, 异常检测效果不佳. 另外, 在基于孪生网络结构的方法中, 例如文献 [8], 教师网络和孪生网络的最后一层特征图的差异被用于衡量样本异常分数, 但在学习过程中模型没有对正常样本的输出特征分布进行约束. 这类方法使得学生网络仅能模拟教师网络的输出, 难以在蒸馏学习中捕获到正常样本的特征分布, 导致其无法有效度量新样本的异常程度.

为了解决以上的问题, 本文面向视觉异常检测任务提出一种基于特征约束的蒸馏学习算法, 使学生网络能够捕获正常图像的视觉特征分布, 以更好地度量图像的异常程度. 为了提取视觉特征, 所提出方法引入 vision transformer (ViT) 作为主干网络, 其在中间编码层上具有较大的感受野, 提升了各编码层特征的语义表达能力. 为了更好地捕获正常样本的特征分布, 本文提出中心特征学习策略, 利用教师网络学习正常样本的分布中心. 为了有效度量样本的异常程度, 本文提出特征约束策略, 要求学生网络的特征输出与教师网络的分布中心一致. 此外, 本文构建编码层特征的格拉姆 (Gram) 矩阵[10], 最小化教师网络和学生网络拉姆矩阵之间的差异, 以进一步约束学生网络的不同编码层特征之间的关系与教师网络保持一致, 保证学生网络的性能. 为了验证所提出方法的异常检测性能, 本文在 3 个异常检测通用图像数据集和一个真实工业异常检测图像数据集上进行了实验验证. 实验结果表明本文所提出方法相比于现有异常检测方法取得了显著的性能提升.

本文主要贡献如下.

(1) 本文提出了基于特征约束蒸馏学习的异常检测方法, 在蒸馏学习过程中约束学生网络得到的特征与教师网络得到的特征具有多重一致性, 并将学习得到的正常样本的分布中心作为判断异常的重要依据.

(2) 本文提出了中心特征学习策略, 在蒸馏学习中捕获正常样本的特征分布中心, 并约束学生网络对正常样本的输出与分布中心一致, 以提升学生网络的性能.

(3) 本文将 ViT 模型应用到异常检测任务, 并利用格拉姆矩阵的一致性约束正常样本的特征编码, 得到更好的学生网络.

本文第 2 节介绍与本文相关的异常检测方法. 第 3 节介绍本文所提出的异常检测方法. 第 4 节设计实验以及实验结果的对比分析. 第 5 节对本文工作进行总结.

2 相关工作研究

针对视觉异常检测任务已有大量的方法, 其中基于图像重建的异常检测方法, 基于生成对抗网络 (GAN)[11]的异常检测方法和基于知识蒸馏的异常检测方法研究最为广泛.

基于图像重建的异常检测工作主要使用自编码器网络重建 (或表示) 正常样本[6]. 通过学习正常样本的潜在特征表示, 期望自编码器不能像正常样本那样精确地重建 (或表示) 异常样本, 再通过比较重建后的图像和原始图像的差异 (均方误差)[12]来判别检测异常样本. 但是自编码器网络具有较强的重建能力, 在实际中能对未经过训练的

异常图像进行良好重建[13], 正常样本和异常样本经过重建后差异很小, 所以异常检测效果不佳. Gong 等人[14]和 Park 等人[15]提出在自编码器网络中引入记忆力机制来削弱网络的重建能力, 让解码网络仅从学习到正常样本的记忆模块中提取特征重建样本. 但是这导致模型无法很好地重建正常样本, 造成异常检测效果不佳. Ye 等人[16]引入旋转, 着色自监督任务训练重建网络模型用于异常检测任务, 让模型能学习到图像的语义信息并根据语义信息通过 U-Net[17]生成图像, 再计算生成后图像和原始图像的均方误差和结构相似性 SSIM[18]. 但是这类方法对异常图像也有良好的重建能力, 无法适用背景复杂的图像异常检测任务.

基于生成对抗网络的异常检测工作主要通过生成器学习正常图像的分布. 例如, AnoGAN [19]通过生成器学习生成正常图像, 测试时随机采样某向量经过生成器生成新图像, 再通过新图像和原测试图像多次迭代更新采样的向量, 最后通过比较生成的图像与原图的差异检测异常. OCGAN [20]期望将整个高维隐空间限制为正常样本的表示, 则异常样本的表示在该空间几乎不存在, 从而产生较高的重构误差. 但是此类方法在针对背景复杂任务时, 对于正常样本重构误差同样较大, 也不能很好的检测背景复杂的异常.

Hinton 等人[21]首次提出了知识蒸馏的概念, 目的是让学生网络学习到庞大教师网络的知识而不依赖大规模图像的训练. 近期, 一些工作[9,22]在异常检测任务中引入蒸馏学习. 这类工作是通过蒸馏学习正常样本在教师网络中的特征表示, 并认为学生网络和教师网络对于正常样本的特征表示类似, 而对于未学习的类别样本特征输出会产生较大的差异. 然而, 学生网络缺乏对正常样本特征分布的描述, 仅通过输出差异很难有效衡量样本的异常程度. 其次, 这类方法要么使用池化层的特征损失细节信息, 要么使用最后一层的特征图信息缺少浅层信息描述, 对噪声信息敏感. 不同于以前方法, 本文提出的方法不仅解决了基于知识蒸馏的异常检测方法中没有对正常样本空间约束的问题, 还通过多重一致性约束提升了学生网络编码能力, 提升了异常检测的性能.

3 主要方法

本文引入 ViT 网络作为骨干网络, 提出基于特征约束蒸馏学习解决视觉异常检测任务, 在蒸馏学习过程中提出了多种约束以提升学生网络的性能.

3.1 预备知识

近期 DosoViTskiy 等人[23]提出了 ViT 模型, 将 Transformer 模型[24]引入计算机视觉领域. ViT 模型首先将图片分为若干图像块, 每个图像块经过线性投影层表示为一维向量标记, 在此基础上, 引入了与单个标记相同维度的可学习的参数 cls 标记, 两类标记分别与其位置信息合并后输入 Transformer 模型训练. cls 标记经过训练后可以捕获到丰富的视觉特征信息, 故将其经过 Transformer 编码网络的特征输出直接用于图像分类. ViT 模型在大规模数据集上进行训练之后取得了超过现有的卷积神经网络模型的分类效果, 在视频分析[25], 目标检测[26,27]和多模态任务[28]中都展示出优越的性能.

异常检测的目标是训练一个能区分正常样本和异常样本的模型. 在本文中, 异常检测模型的训练集为 $D=\{d_1,d_2,d_3,\ldots,d_{n_d}\}$, n_d 表示训练样本数量, 且训练集中不含有任何异常样本. 测试集表示为 $TE_{\text{test}}=\{te_1,te_2,te_3,\ldots,te_{n_{te}}\}$, n_{te} 表测试集样本数量. 与文献 [9] 的蒸馏方法相似, 即使用 D 通过预训练的教师网络 T 在数据集 D 上训练学生网络 S. 通过教师-学生网络为 TE_{test} 数据集的每幅图像计算异常分数, 然后通过异常分数检测样本是否异常[8].

本文针对视觉异常检测任务所提出方法的框架图如图 1 所示, 每幅图像被分为若干的图像块, 每个图像块经过 ViT 模型得到一维向量标记, 和可学习的参数 cls 标记一并作为输入分别送入教师网络 T 和学生网络 S 进行蒸馏学习. 教师网络 T 与学生网络 S 结构相同, 教师网络使用 ImageNet[29]预训练的权重, 学生网络 S 则使用随机初始化的权重. 在正常样本的数据集 D 上, 蒸馏学习训练学生网络 S, 最后训练得到的模型能对正常图像和异常图像输出有差异的特征表示.

3.2 中心特征约束方法

为了让学生网络能够捕获正常样本的特征分布, 以便于更好的度量样本的异常程度, 本文提出了中心特征约束方法. 如图 2 所示, 我们首先利用教师网络对于训练数据的输出特征学习特征分布中心 (*anchor*), 再通过约束学

生网络的输出特征与教师网络的特征分布中心保持一致来学习教师网络中的知识. 最后, 如图 2(c) 所示, 经过训练后的学生网络对于正常样本的特征表示在特征空间中围绕在 *anchor* 周围. 而在学习过程中, 学生网络并未对异常样本的特征表示施加约束, 从而异常样本的特征表示在特征空间中距离 *anchor* 较远, 进一步扩大了正常样本异常样本在特征空间的表示差异.

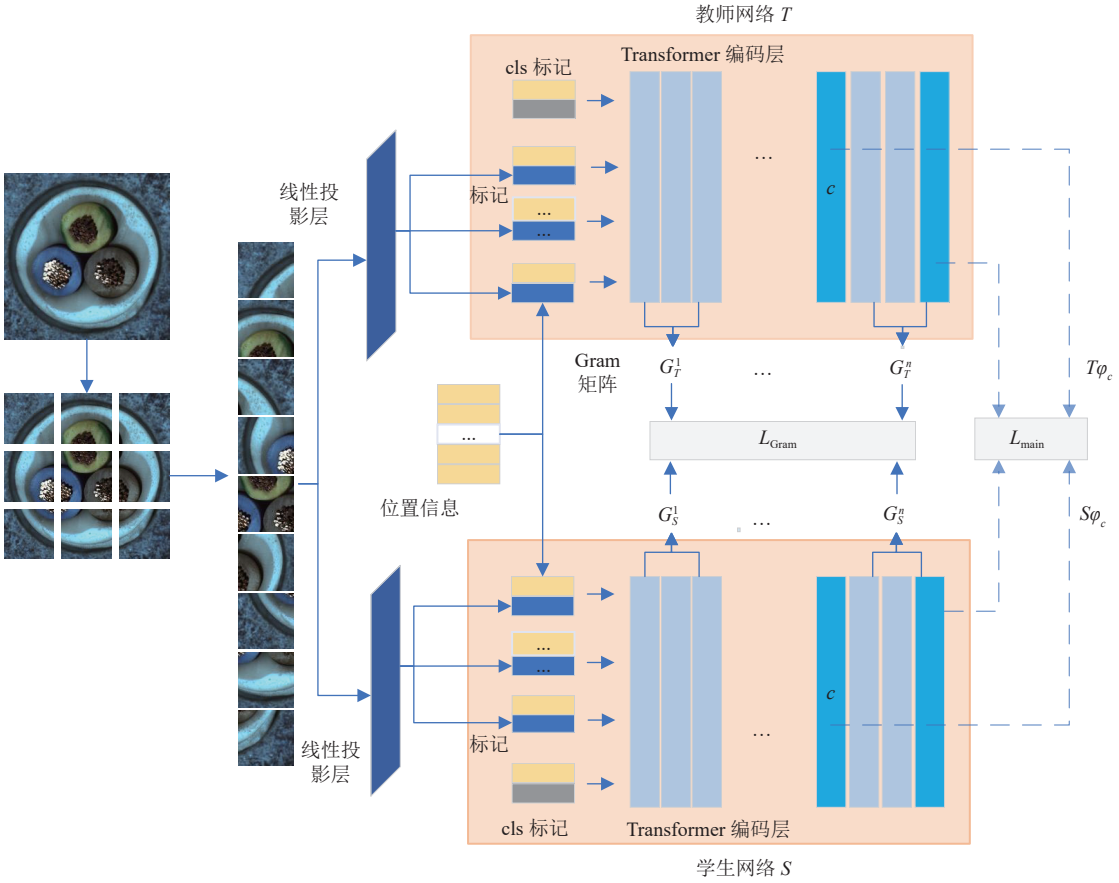


图 1 本文所提出方法框架图

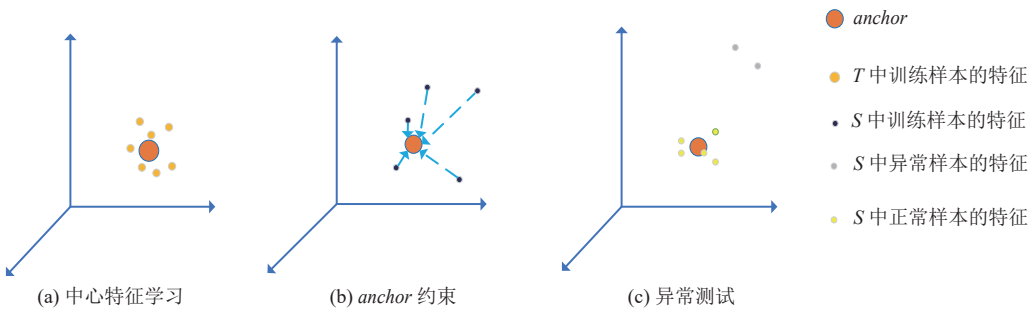


图 2 中心特征约束示意图

本文提出从教师网络中学习特征分布中心. 实际上, 特征分布中心可由多种复杂方法学习得到. 为了简化起见和更好的验证中心特征约束的有效性, 本文采用最简单的方法得到特征分布中心, 即当正常样本分块被映射为多个标记输入教师网络时, 其 cls 标记经过 Transformer 编码器的输出特征的平均值作为 *anchor*, 如图 2(a) 所示.

$$anchor = \frac{1}{n_d}\sum_{j=1}^{n_d} T\varphi_c^{d_j^{\text{cls}}} \tag{1}$$

其中, $T\varphi_c^{d_j^{\text{cls}}}$ 代表训练样本 d_j 经过教师网络 T 时, 第 c 层编码器中 cls 标记对应的特征输出. 此外, 我们使用滑动平均的方式动态学习分布中心, 从而有效地节省计算资源.

为了更好地约束学生网络, 本文提出使用基于 *anchor* 约束的损失函数 L_{anchor}:

$$L_{anchor} = \left\| S\varphi_c^{\text{cls}} - anchor \right\|_2 \tag{2}$$

其中, $S\varphi_c^{\text{cls}}$ 代表训练样本经过学生网络 S 时第 c 层编码器中 cls 标记对应的输出特征. 通过最小化上述损失函数, 学生网络能够有效捕获到正常样本的特征分布, 从而有效的区分正常样本和异常样本.

3.3 编码一致性约束

为了不丢失图像细节信息, 增强学生网络的编码能力, 本文不仅使用最后的特征层进行蒸馏学习, 还引入了中间层特征用于蒸馏学习. 本文将中间层特征的损失函数 L_{distill} 定义为网络教师-学生网络中间层特征之间的欧式距离. 但是朴素的欧式距离 (例如 l_2 距离) 无法保证学生与教师网络训练至收敛时输出的特征最为接近. 鉴于此, 我们提出利用余弦距离度量 T 和 S 网络的中间层输出特征之间的相似度, 对输出的特征施加方向上的约束. L_{distill} 表示为:

$$L_{\text{distill}} = \sum_{c\in groupc}\left(\|T\varphi_c - S\varphi_c\|_2 + \sum_p 1 - \frac{< T\varphi_{c,p}, S\varphi_{c,p} >}{\left\|T\varphi_{c,p}\right\| \times \left\|S\varphi_{c,p}\right\|} \right) \tag{3}$$

其中, *groupc* 表示蒸馏约束的中间层集合, $T\varphi_{c,p}$ 和 $S\varphi_{c,p}$ 分别表示训练样本经过教师网络 T 和学生网络 S 第 c 层编码器 p 标记的特征输出. $<\cdot,\cdot>$ 表示内积函数, $\|\cdot\|$ 表示模长函数.

为了进一步提升学生网络学习到教师网络丰富的知识, 本文进一步约束教师-学生网络的层间关系, 以保持教师网络和学生网络编码层之间关系的一致性. 为此, 本文使用格拉姆 (Gram) 矩阵表示教师网络的编码层之间的关系, 并让学生网络学习这种层间关系, 能让网络更好的拟合到教师网络的输出. Gram 关系矩阵 G 可以表示为:

$$G = \varphi_{c_1} \cdot (\varphi_{c_2})^{\text{T}} \tag{4}$$

其中, φ_{c_1} 和 φ_{c_2} 表示教师 (或学生) 网络的编码器在 c_1 和 c_2 层的输出特征, $(\cdot)^{\text{T}}$ 表示转置运算; G 表示网络中编码层 c_1 和 c_2 的关系矩阵. $c_1, c_2 \in [0, M-1]$, M 代表 Transformer 编码网络的层数. L_{Gram} 损失函数表示为:

$$L_{\text{Gram}} = \|G_T - G_S\|_2 \tag{5}$$

其中, G_T 和 G_S 分别表示教师网络和学生网络的 Gram 关系矩阵.

3.4 总体优化目标

为了使学生网络能够更好地学习教师网络特征表达以更好的度量样本的异常程度, 本文提出了中心特征约束和多重一致性约束. 将上述损失函数进行融合, 所提出方法的总体优化目标函数为:

$$L = L_{\text{distill}} + \lambda \times (L_{anchor} + L_{\text{Gram}}) \tag{6}$$

其中, λ 是模型超参数, 衡量所提出约束的重要性.

3.5 异常检测

本文检测图像的异常与否由异常分数判别. 异常分数高的样本被判断为异常图像, 异常分数低的样本被判定为正常图像. 通过多种特征约束策略, 本文提出的方案针对正常样本在教师网络第 c 层的特征输出 $T\varphi_c$ 和学生网络第 c 层的特征输出 $S\varphi_c$ 差异小, 对于异常样本, 两者的差异较大. 为了更好地度量图像的异常分数, 本文提出了两种异常分数计算策略, 并将他们进行融合以进行异常判断.

(1) 基于 *anchor* 的异常分数计算: 如图 2(c) 所示, 在学习得到的特征空间中, 正常样本的特征靠近 *anchor*, 而异常样本的特征 *anchor* 相距较远. 因此, 本文设计了基于 *anchor* 的差异评分函数:

$$S_A = \left\| T\varphi_c^{\text{cls}} - anchor \right\|_2 + \left\| S\varphi_c^{\text{cls}} - anchor \right\|_2 \tag{7}$$

其中, $T\varphi_c^{\text{cls}}$ 和 $S\varphi_c^{\text{cls}}$ 分别代表样本经过教师网络和学生网络在第 c 层中 cls 标记对应的特征输出.

(2) 基于输出特征差异的异常分数计算: 主流异常检测方法采用教师网络 T 和学生网络 S 输出特征图的差异平

均值作为异常分数. 但是, 大部分图像的异常仅存在于一些细节区域. 这种平均值的处理方式极易受噪声的影响. 为了解决这个问题, 本文提出一种新的 Top k 均值方法, 即计算同一图像块经过教师网络和学生网络后差异, 从所有图像块的差异分数中选出前 k 个最大值, 然后计算平均值, 以前 k 特征差异平均值代表整幅图像的差异, 以减少噪声的影响. 在此, $Diff_{c,p}$ 用来表示样本第 p 个标记 (不包括 cls 标记) 经过第 c 层编码网络后教师-学生网络之间的差异值:

$$Diff_{c,p} = \left\| T\varphi_{c,p} - S\varphi_{c,p} \right\|_2 + 1 - \frac{< T\varphi_{c,p}, S\varphi_{c,p} >}{\left\| T\varphi_{c,p} \right\| \times \left\| S\varphi_{c,p} \right\|} \tag{8}$$

其中, $T\varphi_{c,p}$ 和 $S\varphi_{c,p}$ 分别代表 $T\varphi_c$ 和 $S\varphi_c$ 的第 p 项. 然后将所有的 $Diff_{c,p}$ 按照降序排序, 选择出前 k 个最大值, 并计算平均值. 该平均值使用符号 $AvgTopk(Diff_c)$ 表示. 此外, 由于 cls 标记经过编码网络学习到的特征具有丰富的视觉信息, 异常判断时应充分利用该信息, 将 cls 标记经过教师-学生网络在第 c 层产生的差异记为 $Diff_{c,\mathrm{cls}}$. 因此, 图像的基于输出特征差异的异常分数可由公式 (9) 计算得到:

$$S_D = \sum_{c \in groupc} (Diff_{c,\mathrm{cls}} + AvgTopk(Diff_c)) \tag{9}$$

综上, 单个样本的异常分数表示为:

$$Score = S_D + S_A \tag{10}$$

给定异常分数阈值 δ, 若 $Score \geqslant \delta$, 则判定样本为异常样本; 若 $Score < \delta$, 则判定样本为正常样本.

4 实验验证

4.1 实验设置

为了验证所提出方法的异常检测性能, 本文在 3 个异常检测通用数据集 (CIFAR10[30], CIFAR100[30]和 Fashion-minst[31]) 和一个真实工业异常检测数据集 MVTec[32]上进行了实验验证.

● CIFAR10: 该数据集共包含 10 类图像. 每一类有 6 000 幅彩色图像, 其中 5 000 幅图像用来训练模型, 其余 1 000 幅图像用来验证性能. 所有图像的大小都是 32×32.

● CIFAR100: 该数据集的图像来自 100 个类别. 每一类具有 600 幅彩色图像, 其中 500 幅训练图像和 100 幅测试图像. 所有图像的大小都是 32×32.

● Fashion-minst: 该数据集共有来自 10 类的 60 000 幅图像. 每一类有 6 000 幅灰度图像, 其中 5 000 幅训练图像和 1 000 幅测试图像. 所有图像的大小都是 28×28.

● MVTec: 该数据集是工业异常检测数据集, 为高分辨率彩色图像, 包括 15 个类别的常见工业情景图像. 每个类别图像数目不定, 但是都有专用测试数据集. 测试数据集中包括正常和异常样本. 其中, 异常图像和正常图像相比, 仅存在细微之处的差异, 如铁丝网的断裂, 皮革的裂痕等.

在实验中, 所有图像将统一放大或缩小到 224×224 分辨率. 对于灰度图像, 我们将单通道图像扩充为三通道图像, 每个通道都是 224×224 大小的灰度图. 对于 CIFAR10 和 Fashion-minst 数据集, 训练时将某一类别作为正常样本, 在训练阶段仅用正常样本进行训练. 测试时, 在测试集中其他类别的样本被设置为异常样本. 对于 CIFAR100 数据集, 实验每次选取其中 5 个类别的图像作为正常样本, 其余 95 个类别的图像作为异常样本进行训练验证, 共有 20 次. 对于 MVTec 数据集, 每个类别单独处理, 训练时使用正常样本进行训练, 测试时将测试集的所有异常类型样本统一认为是异常样本. 在所提出蒸馏模型中, 教师-学生网络均采用文献 [23] 中设计的“ViT-Base”结构, 图像均分为大小为 16×16 的图像块, 编码网络的层数 M 设置为 12, 超参数 λ 设置为 1, 异常检测时 Top k 的参数 k 设置为 20, 集合 $groupc$ 设置为 $\{3,6,9,11\}$, 模型训练中采用随机梯度下降算法优化网络模型. 在 CIFAR10 数据集, CIFAR100 数据集和 Fashion-mnist 数据集上, 学习率 lr=0.1, batchsize=32, 而在 MVTec 数据集上, 学习率 lr=0.2, batchsize =4, 均采用 Warmup 方法优化学习率. 本文采用接受者操作特征曲线下面积 (AUROC) 作为评价指标验证异常检测方法的性能[8], 其中, AUROC 指标越接近于 100% 代表检测的效果越优秀.

4.2 比较分析

为了验证本文方法的有效性, 我们在 4 个异常检测数据集上进行了大量的实验, 并将本文方法与视觉异常检

测的代表性方法进行比较分析, 比较结果如表 1–表 4 所示. 所有比较方法的结果均来自于原文.

本文在 CIFAR10 数据集上与 ARAE[6], OCSVM[33], AnoGAN[19], DSVDD[3], CapsNetpp[34], OCGAN[20], LSA[35], DROCC[36], CAVGA-Du[13], GeoTrans[37], U-Std[8], ARFAD[16]和目前最优的 MKDAD[9]方法进行了实验对比, 实验结果如表 1 所示. 这些方法包括基于编码器重建的方法 ARAE, 基于生成对抗网络的方法 AnoGAN, 以及基于知识蒸馏的方法 MKDAD. 从对比结果可以发现, 本文提出的异常检测方法在 CIFAR10 数据集上平均 AUROC 指标达到了 96.3%, 远高于现有的方法. 此外, 本文方法比目前最优的方案 MKDAD 仍高出 9.1%. 这充分说明本文所提出的特征约束蒸馏学习的有效性, 学生网络能够更优的学习到教师网络的知识. 最后, 从单个类别的性能可以看出, 本文所提方法在所有类别上均取得了最优的检测效果, 其中在 8 个类别上的检测结果 AUROC 超过了 95%. 比如对于“Cat”类别, 现有的异常检测方法的 AUROC 指标都没超过 80%, 而本文方法达到了 92.8%, 远远优于现有方法.

表 1　不同方法在 CIFAR10 数据集上的异常检测比较结果 AUROC (%)

比较方法	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	平均值
ARAE[6]	72.2	43.1	69.0	55.0	75.2	54.3	70.1	51.0	72.2	40.0	60.2
OCSVM[33]	63.0	44.0	64.9	48.7	73.5	50.0	72.5	53.3	64.9	50.8	58.6
AnoGAN[19]	67.1	54.7	52.9	54.5	65.1	60.3	58.5	62.5	75.8	66.5	61.8
DSVDD[3]	61.7	65.9	50.8	59.1	60.9	65.7	67.7	67.3	75.9	73.1	64.8
CapsNetpp[34]	62.2	45.5	67.1	67.5	68.3	63.5	72.7	67.3	71.0	46.6	61.2
OCGAN[20]	75.7	53.1	64.0	62.0	72.3	62.0	72.3	57.5	82.0	55.4	65.7
LSA[35]	73.5	58.0	69.0	54.2	76.1	54.6	75.1	53.5	71.7	54.8	64.1
DROCC[36]	73.5	58.0	69.0	54.2	76.1	54.6	75.1	53.5	71.7	54.8	64.1
CAVGA-Du[13]	65.3	78.4	76.1	74.7	77.5	55.2	81.3	74.5	80.1	74.1	73.7
GeoTrans[37]	76.2	84.8	77.1	73.2	82.8	84.8	82	88.7	89.5	83.4	82.3
U-Std[8]	78.9	84.9	73.4	74.8	85.1	79.3	89.2	83.0	86.2	84.8	82.0
MKDAD[9]	90.5	90.4	79.6	77.0	86.7	91.4	88.9	86.8	91.5	88.9	87.2
ARFAD[16]	78.5	89.8	86.1	77.4	90.5	84.5	89.2	92.9	92.0	85.5	86.6
本文方法	**97.3**	**98.4**	**94.5**	**92.8**	**97.5**	**94.7**	**95.7**	**98.8**	**96.6**	**97.0**	**96.3**

表 2　不同方法在 CIFAR100 数据集上的异常检测比较结果 AUROC (%)

比较方法	0	1	2	3	4	5	6	7	8	9	10
DAGMM[38]	43.4	49.5	66.1	52.6	56.9	52.4	55.0	52.8	53.2	42.5	52.7
DSEBM[39]	64.0	47.9	53.7	48.4	59.7	46.6	51.7	54.8	66.7	71.2	78.3
ALOCC[40]	52.7	56.6	61.2	61.1	66.7	50.6	63.9	66.2	50.9	73.4	71.1
ADGAN[41]	63.1	54.9	41.3	50.0	40.6	42.8	51.1	55.4	59.2	62.7	79.8
GANomaly[42]	57.9	51.9	36.0	46.5	46.6	42.9	53.7	59.4	63.7	68.0	75.6
GeoTrans[37]	74.7	68.5	74.0	81.0	78.4	59.1	81.8	65.0	85.5	**90.6**	87.6
ARFAD[16]	77.5	70.0	62.4	76.2	77.7	64.0	86.9	65.6	82.7	90.2	85.9
本文方法	**92.8**	**90.4**	**83.2**	**87.0**	**92.2**	**86.5**	**88.9**	**90.3**	**88.2**	89.1	**94.4**

比较方法	11	12	13	14	15	16	17	18	19	平均值	—
DAGMM[38]	46.4	42.7	45.4	57.2	48.8	54.4	36.4	52.4	50.3	50.5	
DSEBM[39]	62.7	66.8	52.6	44.0	56.8	63.1	73.0	57.7	55.5	58.8	
ALOCC[40]	56.9	63.6	56.0	57.9	58.2	57.0	73.5	61.0	58.8	60.9	
ADGAN[41]	53.7	58.9	57.4	39.4	55.6	63.3	66.7	44.3	53.0	54.7	—
GANomaly[42]	57.6	58.7	59.9	43.9	59.9	64.4	71.8	54.9	56.8	56.5	
GeoTrans[37]	83.9	83.2	58.0	**92.1**	68.3	73.5	**93.8**	**90.7**	85.0	78.7	
ARFAD[16]	83.5	84.6	67.6	84.2	74.1	80.3	91.0	85.3	**85.4**	78.8	
本文方法	**88.7**	**88.2**	**87.1**	90.9	**87.9**	**88.0**	92.4	89.2	82.6	**88.9**	

表 3 不同方法在 Fashion-minst 数据集上的异常检测比较结果 AUROC (%)

比较方法	T-shirt/top	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot	平均值
ARAE[6]	93.7	99.1	91.1	94.4	92.3	91.4	83.6	98.9	93.9	97.9	93.6
OCSVM[33]	91.9	99.0	89.4	94.2	90.7	91.8	83.4	98.8	90.3	98.2	92.8
DAGMM[38]	30.3	31.1	47.5	48.1	49.9	41.3	42.0	37.4	51.8	37.8	41.7
DSEBM[39]	89.1	56.0	86.1	90.3	88.4	85.9	78.2	98.1	86.5	96.7	85.5
DSVDD[3]	98.2	90.3	90.7	94.2	89.4	91.8	83.4	98.8	91.9	99.0	92.8
LSA[35]	91.6	98.3	87.8	92.3	89.7	90.7	84.1	97.7	91.0	98.4	92.2
MKDAD[9]	92.5	99.2	92.5	93.8	93.0	98.2	84.9	99.0	94.3	97.5	94.5
ARAFD[16]	92.7	99.3	89.1	93.6	90.8	93.1	85.0	98.4	97.8	98.4	93.9
本文方法	93.0	99.6	92.2	94.9	92.0	97.4	84.5	98.72	97.2	98.2	94.8

表 4 不同方法在 MVTec 数据集上的异常检测比较结果 AUROC (%)

比较方法	Bottle	Hazelnut	Capsule	Metal Nut	Leather	Pill	Wood	Carpet	Tile	Grid	Cable	Transistor	Toothbrush	Screw	Zipper	平均值
AVID[43]	88	86	85	63	58	86	83	70	66	59	64	58	73	66	84	73
AE-ssim[44]	88	54	61	54	46	60	83	67	52	69	61	52	74	51	80	63
AE-12[44]	80	88	62	73	44	62	74	50	77	78	56	71	98	69	80	71
AnoGAN[19]	69	50	58	50	522	62	68	49	51	51	53	67	57	35	59	55
LSA[35]	86	80	71	67	70	85	75	74	70	54	61	50	89	75	88	73
CAVAG-du[13]	89	84	83	67	71	88	85	73	70	75	63	73	91	77	87	78
DSVDD[3]	86	71	69	50	73	77	87	54	81	59	71	65	70	64	74	72
VAE-grad[45]	92.2	97.6	91.7	90.7	92.5	93	83.8	73.5	65.4	96.1	91	91.9	98.5	94.5	86.9	89.3
GeoTrans[37]	74.3	33.3	67.8	82.4	82.5	65.2	48.2	45.9	53.9	61.9	84.7	79.8	94.0	44.6	87.4	67.1
MKDAD[9]	99.4	**98.4**	80.5	73.6	95.1	82.7	94.3	79.3	91.6	78.0	89.2	85.6	92.2	83.3	**93.2**	87.7
ARFAD[16]	94.1	85.5	68.1	66.7	86.2	78.6	92.3	70.6	73.5	**88.3**	83.2	84.3	**100.0**	**100.0**	87.6	83.9
本文方法	**100.0**	96.4	77.7	**95.1**	**97.9**	**88.1**	**96.0**	**81.3**	**99.1**	71.7	**92.5**	**88.4**	95.8	95.0	81.0	**90.4**

本文在 CIFAR100 数据集上与 DAGMM[38], DSEBM[39], ALOCC[40], ADGAN[41], GANomaly[42], ARAFD[16]和 GeoTrans[37]方法进行了实验对比, 20 种构造不同正常样本类和异常样本类的设置下的实验结果如表 2 所示. 本文所提出的方法在 CIFAR100 数据集上平均 AUROC 指标达到了 88.9%, 比目前最优的方法高出约 10%. 本文方法在大部分类别的检测中获得了最优的性能, 在“5”“7”等设置时均超出现有的方法约 20%.

为了验证本文所提出的方法对于灰度图像异常检测有效性, 在 Fashion-mnist 数据集上对比了 ARAE[6], OCSVM[33], DAGMM[38], DSEBM[39], DSVDD[3], LSA[35], MKDAD[9]和 ARAFD[16], 实验结果如表 3 所示. 本文所提出方法对于灰度图像异常检测同样有效, 也取得最优的性能.

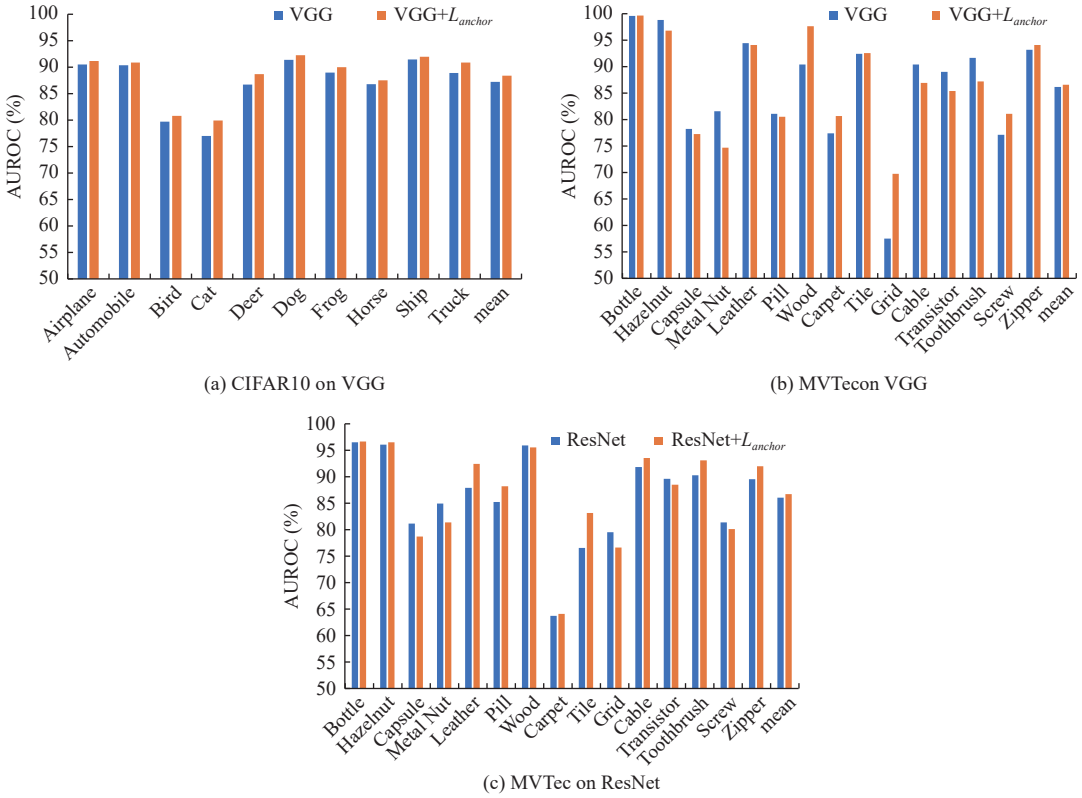
为了验证本文方法针对高分辨率工业异常场景的有效性, 本文将所提出的方法与 AVID[43], AE-ssim[44], AE-L2[44], AnoGAN[19], LSA[35], CAVAG-du[13], DSVDD[3], VAE-grad[45], GeoTrans[37], MKDAD[9], ARFAD[16]在真实的工业数据集 MVTec 进行了实验对比, 结果如表 4 所示. 本文方法在“Bottle”“Carpet”类别的异常检测 AUROC 指标可以达到 99% 以上. 对于“Metal Nut”类别, 本文提出的方法对比于 MKDAD 方法提升了约 13%, 平均的 AUROC 达到了 90.4%. 但是针对“Grid”, 本文的方法仍存在不足, 其可能的原因是“Grid”纹理图像的异常区域与纹理语义类似, 造成检测效果不佳.

综合以上实验结果, 可以验证本文所提出的方法能针对灰度图像, 彩色图像和高分辨率工业图像, 其异常检测均能达到优异的效果, 主要是因为本文方法在蒸馏学习过程中引入了多重特征约束, 有效提升了学生网络的性能.

4.3 消融实验

为了验证本文所提及的各个模块的作用, 本节针对各个模块设计了以下的实验进行分析.

为了验证中心特征约束策略的有效性进行了下列实验. 首先在 VGG 卷积神经网络框架[46]中引入了 L_{anchor} 损失函数, 在 CIFAR10 数据集和真实工业异常检测数据集 MVTec 进行了对比实验. 其中, 学习率 lr=0.001, batchsize=128, 采用随机梯度下降算法优化网络模型. VGG 教师网络采用 ImageNet 预训练 VGG16 网络[45], 学生网络同样采取随机权重. 实验结果如图 3 所示, 其中 VGG 代表直接使用 VGG16 蒸馏实现异常检测, VGG+Lanchor 代表加入了中心特征约束策略的异常检测, “mean”表示平均 AUROC 值. 从实验结果可以看出, 引入中心特征约束策略后, 在 CIFAR10 数据集上, 所有类别的异常检测效果均有一定的提升, 约提升 1%. 在 MVTec 数据集上, 异常检测总体性能有所提升, 但存在部分类别性能有所下降的现象. 比如, “Wood”和“Grid”两个类别的异常检测性能明显提升, 表明中心特征约束对于此类图像检测有潜力. 而在“Cable”类别上, 施加约束后, 异常检测效果下降, 其可能的原因是在 VGG 网络蒸馏中, 学习得到分布中心距离正常样本和异常样本太接近, 无法起到区分作用, 而对于异常和正常类别差距较大的两类, 中心特征约束能起到有效的作用. 其次, 本文同样在先进的 ResNet 网络[47]中验证了 L_{anchor} 损失函数的有效性. 如图 3(c) 所示, 为了简化起见, 我们仅在 MVTec 数据集上进行了对比实验. 可以发现, L_{anchor} 损失函数对于 ResNet 卷积神经网络异常检测十分有效, 多数类别的异常检测效果均有显著提升, 如“leather”和“pill”. 综上实验结果, 中心特征约束策略可以提升模型的异常检测的性能, 且十分有效. 在异常检测中的纹理图像中, 使用中心特征约束策略的 VGG 网络优于使用同样策略的 ResNet 网络. 从 CIFAR10 数据集上的提升结果发现当正常样本和异常样本类别不同时 (比如正常类是“DOG”, 其他类为异常), 提升效果更加明显. 直觉上, 非细粒度图像正常和异常类别的特征中心相距更远, 所以鉴别效果更优. 从 MVTec 实验结果分析, 本文的中心特征约束策略同样对 MVTec 数据集有效.



(a) CIFAR10 on VGG

(b) MVTecon VGG

(c) MVTec on ResNet

图 3 VGG 或者 ResNet 网络引入 L_{anchor} 损失在 CIFAR10 和 MVTec 数据集的表现

为了验证本文所提出方法中损失函数的作用, 在 MVTec 数据集上设计实验验证每个模块的效果. 实验结果如图 4 所示. ViT 表示仅使用 L_{distill} 损失函数的检测方法, ViT+L_{Gram} 表示添加 L_{distill} 损失函数与 L_{Gram} 损失函数的检测

方法, ViT+L_{Gram}+L_{anchor} 表示添加 L_{distill} 损失函数, L_{Gram} 损失函数和 L_{anchor} 损失函数的方法. 从在该数据集上的平均检测结果来看, 本文所提出的各个模块都有助于模型性能的提升, 平均每个模块能提升 1%. 通过比较 ViT+L_{Gram} 和 ViT 的结果, 我们可以发现所有类别的 AUROC 值在 L_{Gram} 损失函数的约束下均得到了提升. 其原因是在 ViT 的编码结构中, Gram 矩阵表示的是无损信息的层间关系矩阵, 这种层间关系约束能够增强学生网络的编码能力, 并对于蒸馏学习解决异常问题十分有效. 将 ViT+L_{Gram}+L_{anchor} 的结果与其他两个方法结果以及和图 3(b)和图 3(c) 中结果进行对比, 我们观察到两点结论, 其一是 ViT 的整体检测效果明显高于卷积神经网络表明 ViT 结合中心特征约束策略更适用于异常检测. 其二是一些如“Metal Nut”“Cable”和“Transistor”等类别的异常检测结果明显有所提升. 其可能原因有: 1) ViT 对于图像的编码能力强于 VGG 网络, 中间层特征捕获的细节信息更多, 可以学习到更能代表正常样本分布的分布中心; 2) 有效的特征中心能提升中心特征约束策略的约束效果. 这可以从 “Tile”“Screw”“Wood”等类别在逐渐增加损失后效果逐步提升的结果中得到验证. 综合上述实验结果, L_{Gram} 损失函数的引入, 约束了教师网络和学生网络在层间关系的一致性, 加强了网络的编码能力, 提升了异常检测的性能. L_{anchor} 损失函数通过中心特征学习策略, 学习了正常样本的分布中心, 拉近正常样本的特征与分布中心的距离, 使学生网络的捕获正常样本分布, 能更合理的度量样本异常程度来加强异常检测效果. 但是仍有例外, 如“Grid”类别中, 编码网络很难处理异常只在一些纹理不明显的区域, 这样对于后续异常度量过程中, 异常样本和正常样本特征接近, 异常样本可能也靠近分布中心, 造成检测效果不佳.

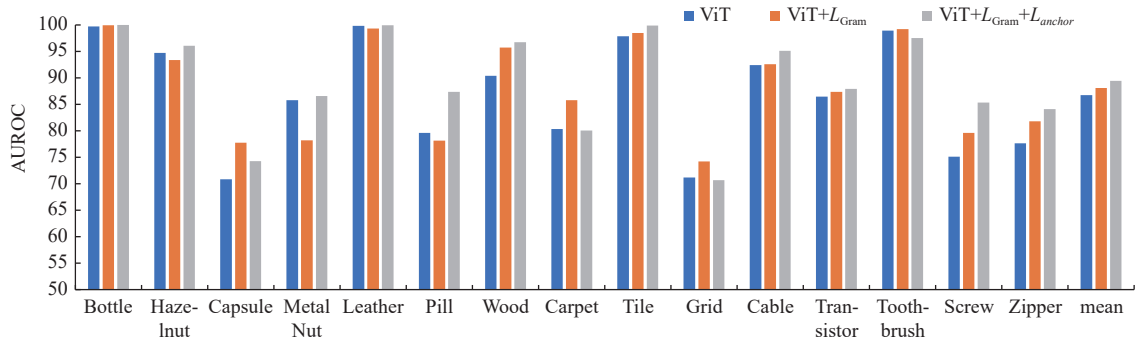


图 4 本文所提出模块在 MVTec 数据集的提升效果

为了验证 ViT 骨干网络使用 损失函数比 VGG 骨干网络和 ResNet 骨干网络更加有效, 本文将蒸馏框架中骨干网络分别设置为 VGG16 网络, ResNet18[47]和 ViT 网络, 在 MVTec 数据集进行对比实验. 图 5 展示了 VGG, ResNet 与 ViT 加入 损失函数前后在 MVTec 数据集上各个类别 AUROC 变化情况. 涨幅差值表示在骨干网络基础上加入中心特征约束策略前后, 对 MVTec 数据集各个类别的提升效果, 效果下降表示为负值. 实验结果表明在 ViT 为主干网络时, 10 个类别的异常检测效果均有所提升, 在“Pill”和“Metal Nut”类别的提升达到了 8%, 而 VGG 作为主干网络时仅有几个类别的效果提升, ResNet 的表现趋于平稳, 虽有提升, 但是提升较小. 这表明 ViT 中使用中心特征约束策略比 VGG 网络和 ResNet 网络更加有效. 造成这一现象的可能原因首先是 ViT 的输出特征对于每个类别离其本身的分布中心相较于卷积神经网络中更聚集, 分布中心也更能代表此类样本的分布特性. 其次是 ViT 网络更由于分块, 实际感受野变大, 与 VGG 和 ResNet 中使用池化层特征不同, ViT 中使用中间层无损的编码了细节信息, 编码后特征包含更丰富的语义信息同时也保留了大量的细节信息, 可加强异常检测效果.

为了进一步展示本文方法的优势, 我们展示不同方法在 MVTec 数据集中部分图像的检测效果. 如图 6 所示. 本文的方法与最新的研究方法 MKDAD 加入 L_{anchor} 损失函数 (图中的 VGG+L_{anchor}) 和以骨干网络为 ResNet 实现的检测方法加入 L_{anchor} 损失函数 (图中的 ResNet+L_{anchor}) 对比. 如图 6(a) 所示, 本文的方法在针对图像中细微的异常区域更有效. 图 6(a) 中异常为细微的“异常物体”, 从对比结果可以看出本文的方法精确的分辨出图像为异常图像, 但是其余两个方法均不能识出此类型的细微异常. 从图 6(c) 中可以发现蒸馏异常检测任务中 VGG 对于颜色敏感性不如 ResNet, 而本文的方法对于此张图像异常分数 S=8.2, 远远大于判定阈值 δ=7.62, 判定为异常图像, 检测效果良好. 通过可视化样本的检测效果, 本文发现图 6(d) 中, 以 ResNet 作为骨干网络对于边界异常区域的检测效果很差, 其可能的原因是感受野阻碍了检测效果. 而本文的方法异常检测效果十分良好. 尽管本文的方法在多数

数据集上均有一定优势, 但是仍存在一些问题. 如图 6(b) 所示, 此图像虚线框区域存在轻微褶皱, 易被模型过拟合检测为异常图像. 此 3 类方法均在此图像中检测失败. 综上结果, 本文方法在针对 MVTec 数据集异常检测有良好的表现, 且克服了卷积网络作为骨干网络对于细微区域和边界区域检测效果不好的问题.

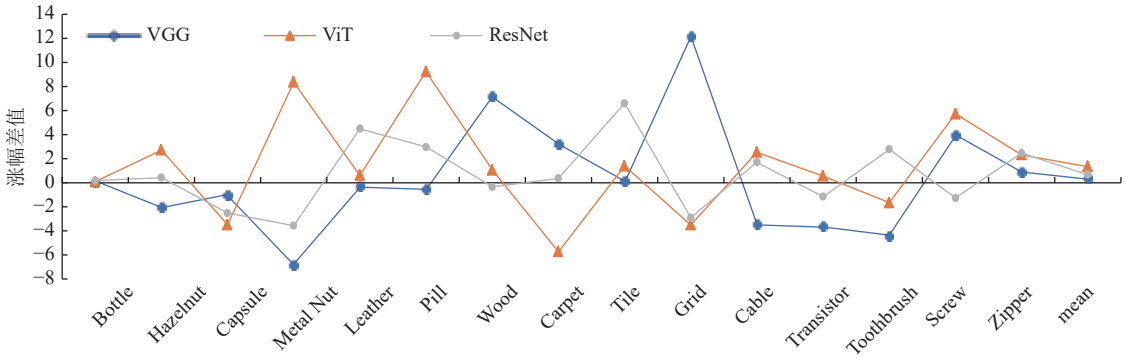
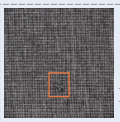
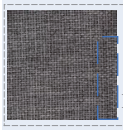


图 5 L_{anchor} 损失函数在不同主干网络的效果

示例图	方法	异常分数	阈值 δ	结果	真实标签
	OURS	7.06	6.99	异常	异常
	VGG+L_{anchor}	0.8	0.9	正常	
	ResNet+L_{anchor}	1.22	1.26	正常	
(a)					
	OURS	7.74	7.62	异常	正常
	VGG+L_{anchor}	1.12	0.9	异常	
	ResNet+L_{anchor}	1.30	1.26	异常	
(b)					

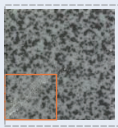
示例图	方法	异常分数	阈值 δ	结果	真实标签
	OURS	8.20	7.62	异常	异常
	VGG+L_{anchor}	0.48	0.49	正常	
	ResNet+L_{anchor}	1.70	1.62	异常	
(c)					
	OURS	8.59	7.62	异常	异常
	VGG+L_{anchor}	0.55	0.49	异常	
	ResNet+L_{anchor}	1.57	1.62	正常	
(d)					

图 6 不同骨干网络的异常检测效果

在实验过程中, 本文结合知识蒸馏与 ViT 的结构特性, 采取两种不同训练策略, 即 (1) 所有学生网络的权重均从教师网络学习, 包括位置编码权重和编码结构层的权重; (2) 将教师网络中学习到的位置权重直接赋予学生网络位置权重, 即保证两个图像块位置信息相同, 并固定位置编码权重不更新. 在工业异常数据集 MVTec 上实验的结果如图 7 所示. 结果表明策略 (2) 的效果基本高于策略 (1), 且每个类别均有一定幅度的提升. 这说明共享学习到的位置信息对学生网络的学习能力起到了促进作用. 其可能的原因是训练学习得到的位置信息类似于“基础知识”, 而赋值学习到的位置信息直接让学生网络掌握了这种“基础知识”, 对于后面的蒸馏学习过程起到了促进作用, 让学生网络更加侧重编码能力的学习.

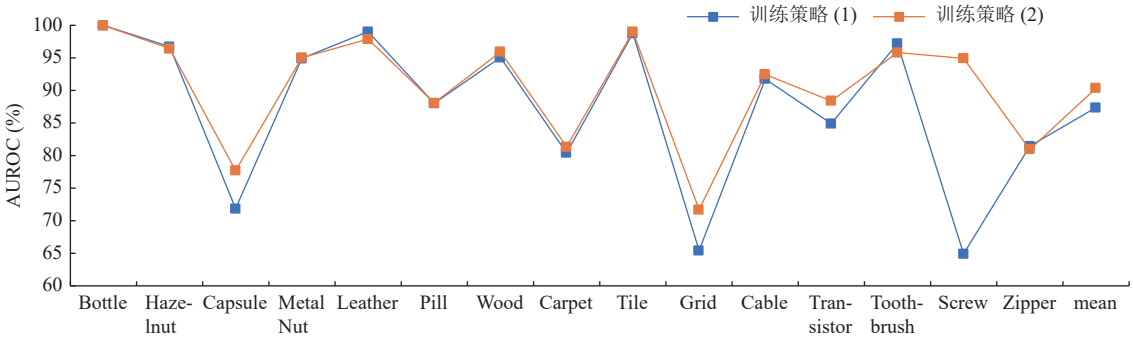


图 7 位置编码的影响

5 总 结

本文在视觉异常检测任务中首次引入了 ViT 模型作为主干网络, 并通过特征约束蒸馏学习的方法解决异常检测的问题. 以往的知识蒸馏模型没有在学生网络学习中约束正常样本的特征分布, 只是学习模拟教师网络对于正常样本的输出, 无法学习到鉴别特征. 为了解决这个问题, 本文提出了中心特征学习策略学习了正常样本的特征分布中心, 约束学生网络学习正常样本的特征分布, 以更加有效的方式度量图像的异常程度. 在此基础上, 本文方法通过让学生网络与教师网络保持层间关系的一致性, 加强了网络的编码能力, 进一步提升了异常检测性能. 通过多项消融实验验证了本文方法的有效性. 未来的研究工作将深入研究如何高效地学习中心特征, 以更有效的方式限制样本空间, 以此来解决异常检测问题.

References:

[1] Domingues R, Filippone M, Michiardi P, Zouaoui J. A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recognition, 2018, 74: 406–421. [doi: 10.1016/j.patcog.2017.09.037]

[2] Pang GS, Shen CH, Cao LB, van den Hengel A. Deep learning for anomaly detection: A review. ACM Computing Surveys, 2022, 54(2): 38. [doi: 10.1145/3439950]

[3] Ruff L, Vandermeulen R, Goernitz N, Deecke L, Siddiqui SA, Binder A, Müller E, Kloft M. Deep one-class classification. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 4393–4402.

[4] Song J, Xiao L, Lian ZC, Cai ZY, Jiang GP. Overview and prospect of deep learning for image segmentation in digital pathology. Ruan Jian Xue Bao/Journal of Software, 2021, 32(5): 1427−1460 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6205.htm [doi: 10.13328/j.cnki.jos.006205]

[5] Ding XO, Yu SJ, Wang MX, Wang HZ, Gao H, Yang DH. Anomaly detection on industrial time series based on correlation analysis. Ruan Jian Xue Bao/Journal of Software, 2020, 31(3): 726−747 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5907.htm [doi: 10.13328/j.cnki.jos.005907]

[6] An J, Cho S. Variational autoencoder based anomaly detection using reconstruction probability. Special Lecture on IE, 2015, 2(1): 1–18.

[7] Zhou C, Paffenroth RC. Anomaly detection with robust deep autoencoders. In: Proc. of the 23rd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Halifax: ACM, 2017. 665–674. [doi: 10.1145/3097983.3098052]

[8] Bergmann P, Fauser M, Sattlegger D, Steger C. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 4182–4191. [doi: 10.1109/CVPR42600.2020.00424]

[9] Salehi M, Sadjadi N, Baselizadeh S, Rohban MH, Rabiee HR. Multiresolution knowledge distillation for anomaly detection. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 14897–14907. [doi: 10.1109/CVPR46437.2021.01466]

[10] Yim J, Joo D, Bae J, Kim J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 7130–7138. [doi: 10.1109/CVPR.2017.754]

[11] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680. [doi: 10.5555/2969033.2969125]

[12] Zhao H, Gallo O, Frosio I, Kautz J. Loss functions for image restoration with neural networks. IEEE Trans. on Computational Imaging, 2017, 3(1): 47–57. [doi: 10.1109/TCI.2016.2644865]

[13] Liu W, Luo WX, Lian DZ, Gao SH. Future frame prediction for anomaly detection—A new baseline. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6536–6545. [doi: 10.1109/CVPR.2018.00684]

[14] Gong D, Liu LQ, Le V, Saha B, Mansour RM, Venkatesh S, van den Hengel A. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 1705–1714. [doi: 10.1109/ICCV.2019.00179]

[15] Park H, Noh J, Ham B. Learning memory-guided normality for anomaly detection. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 14360–14369. [doi: 10.1109/CVPR42600.2020.01438]

[16] Ye F, Huang CQ, Cao JK, Li MS, Zhang Y, Lu CW. Attribute restoration framework for anomaly detection. IEEE Trans. on Multimedia, 2020, 24: 116–127. [doi: 10.1109/TMM.2020.3046884]

[17] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of the 18th Int'l Conf.

on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241. [doi: 10.1007/978-3-319-24574-4_28]

[18] Horé A, Ziou D. Image quality metrics: PSNR vs. SSIM. In: Proc. of the 20th Int'l Conf. on Pattern Recognition. Istanbul: IEEE, 2010. 2366–2369. [doi:DOI: 10.1109/ICPR.2010.579] [doi: 10.1109/ICPR.2010.579]

[19] Schlegl T, Seeböck P, Waldstein SM, Schmidt-Erfurth U, Langs G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Proc. of the 25th Int'l Conf. on Information Processing in Medical Imaging. Boone: Springer, 2017. 146–157. [doi: 10.1007/978-3-319-59050-9_12]

[20] Perera P, Nallapati R, Xiang B. OCGAN: One-class novelty detection using GANs with constrained latent representations. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2893–2901. [doi: 10.1109/CVPR.2019.00301]

[21] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.

[22] Sabokrou M, Fayyaz M, Fathy M, Moayed Z, Klette R. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Computer Vision and Image Understanding, 2018, 172: 88–97. [doi: 10.1016/j.cviu.2018.02.006]

[23] DosoViTskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Prof. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021. 1–22.

[24] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez NA, Kaiser L, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010. [doi: 10.5555/3295222.3295349]

[25] Neimark D, Bar O, Zohar M, Asselmann D. Video transformer network. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision Workshops. Montreal: IEEE, 2021. 3156–3165. [doi: 10.1109/ICCVW54120.2021.00355]

[26] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 213–229. [doi: 10.1007/978-3-030-58452-8_13]

[27] Beal J, Kim E, Tzeng E, Park DH, Zhai A, Kislyuk D. Toward transformer-based object detection. arXiv:2012.09958, 2020.

[28] Hu RH, Singh A. UniT: Multimodal multitask learning with a unified transformer. arXiv:2102.10772, 2021.

[29] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

[30] Krizhevsky A. Learning multiple layers of features from tiny images. Technical Report, Toronto: University of Toronto, 2009.

[31] Xiao H, Rasul K, Vollgraf R. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747, 2017.

[32] Bergmann P, Fauser M, Sattlegger D, Steger C. MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9584–9592. [doi: 10.1109/CVPR.2019.00982]

[33] Chen YQ, Zhou XS, Huang TS. One-class SVM for learning in image retrieval. In: Proc of the 2001 Int'l Conf. on Image Processing. Thessaloniki: IEEE, 2001. 34–37. [doi: 10.1109/ICIP.2001.958946]

[34] Li XY, Kiringa I, Yeap T, Zhu XD, Li YF. Exploring deep anomaly detection methods based on capsule net. In: Proc. of the 33rd Canadian Conf. on Artificial Intelligence. Ottawa: Springer, 2020. 375–387. [doi: 10.1007/978-3-030-47358-7_39]

[35] Abati D, Porrello A, Calderara S, Cucchiara R. Latent space autoregression for novelty detection. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 481–490. [doi: 10.1109/CVPR.2019.00057]

[36] Goyal S, Raghunathan A, Jain M, Simhadri HV, Jain P. DROCC: Deep robust one-class classification. In: Proc. of the 37th Int'l Conf. on Machine Learning. Vienna: PMLR, 2020. 3711–3721.

[37] Golan I, El-Yaniv R. Deep anomaly detection using geometric transformations. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 9781–9791. [doi: 10.5555/3327546.3327644]

[38] Zhai SF, Cheng Y, Lu WN, Zhang ZF. Deep structured energy based models for anomaly detection. In: Prof. of the 33rd Int'l Conf. on Machine Learning. New York City: JMLR.org, 2016. 1100–1109.

[39] Zong B, Song Q, Min MR, Cheng W, Lumezanu C, Cho DK, Chen HF. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: Prof. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018. 1–19.

[40] Sabokrou M, Khalooei M, Fathy M, Adeli E. Adversarially learned one-class classifier for novelty detection. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3379–3388. [doi: 10.1109/CVPR.2018.00356]

[41] Deecke L, Vandermeulen R, Ruff L, Mandt S, Kloft M. Image anomaly detection with generative adversarial networks. In: Proc. of the

2018 European Conf. on Machine Learning and Knowledge Discovery in Databases. Dublin: Springer, 2018. 3–17. [doi: 10.1007/978-3-030-10925-7_1]

[42] Akcay S, Atapour-Abarghouei A, Breckon TP. GANomaly: Semi-supervised anomaly detection via adversarial training. In: Proc. of the 14th Asian Conf. on Computer Vision. Perth: Springer, 2018. 622–637. [doi: 10.1007/978-3-030-20893-6_39]

[43] Sabokrou M, Pourreza M, Fayyaz M, Entezari R, Fathy M, Gall J, Adeli E. AVID: Adversarial visual irregularity detection. In: Proc. of the 14th Asian Conf. on Computer Vision. Perth: Springer, 2018. 488–505. [doi: 10.1007/978-3-030-20876-9_31]

[44] Bergmann P, Löwe S, Fauser M, Sattlegger D, Steger C. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In: Prof. of the 14th Int'l Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications. Prague: SciTePress, 2019. 372–380.

[45] Dehaene D, Frigo O, Combrexelle S, Eline P. Iterative energy-based projection on a normal data manifold for anomaly localization. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020. 1–17.

[46] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2015. 1–14.

[47] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

附中文参考文献:

[4] 宋杰, 肖亮, 练智超, 蔡子羹, 蒋国平. 基于深度学习的数字病理图像分割综述与展望. 软件学报, 2021, 32(5): 1427−1460. http://www.jos.org.cn/1000-9825/6205.htm [doi: 10.13328/j.cnki.jos.006205]

[5] 丁小欧, 于晟健, 王沐贤, 王宏志, 高宏, 杨东华. 基于相关性分析的工业时序数据异常检测. 软件学报, 2020, 31(3): 726−747. http://www.jos.org.cn/1000-9825/5907.htm [doi: 10.13328/j.cnki.jos.005907]

邢鹏(1998－), 男, 博士生, 主要研究领域为异常检测, 计算机视觉.

唐金辉(1981－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为多媒体分析与检索, 社交媒体分析, 计算机视觉, 人工智能.

蒋鑫(2001－), 男, 博士生, 主要研究领域为深度学习, 计算机视觉.

李泽超(1985－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为图像视频分析, 目标检测, 模式识别.

潘永华(1990－), 男, 博士生, 主要研究领域为深度学习, 视觉问答, 机器学习.