中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2025)02-0451-16

论文引用格式:Qu H C and Lin J J. 2025. Multimodal multiscale industrial anomaly detection via flows. Journal of Image and Graphics, 30(02):0451-0466(曲海成, 林俊杰. 2025. 基于归一化流的多模态多尺度工业场景缺陷检测. 中国图象图形学报, 30(02):0451-0466)[DOI:10.11834/jig.240183]

基于归一化流的多模态多尺度工业场景缺陷检测

曲海成,林俊杰*

辽宁工程技术大学软件学院,葫芦岛　125000

摘　要:**目的**　工业缺陷检测是现代工业质量控制中至关重要的一环,针对工业多模态缺陷检测场景下,捕捉不同形状大小、在RGB图像上感知度低的缺陷,以及减少单模态原始特征空间内存在的噪声对多模态信息交互的干扰的挑战,提出了一种基于归一化流的多模态多尺度缺陷检测方法。**方法**　首先,使用Vision Transformer和Point Transformer对RGB图像和3D点云两个模态的信息提取第1、3、11块的特征构建特征金字塔,保留低层次特征的空间信息助力缺陷定位任务,并提高模型对不同形状大小缺陷的鲁棒性;其次,为了简化多模态交互,使用过点特征对齐算法将3D点云特征对齐至RGB图像所在平面,通过构建对比学习矩阵的方式实现无监督多模态特征融合,促进不同模态之间信息的交互;此外,通过设计代理任务的方式将信息瓶颈机制扩展至无监督,并在尽可能保留原始信息的同时,减少噪声干扰得到更充分有力的多模态表示;最后,使用多尺度归一化流结构捕捉不同尺度的特征信息,实现不同尺度特征之间的交互。**结果**　本文方法在MVTec-3D AD数据集上进行性能评估,实验结果显示Detection AUCROC(area under the curve of the receiver operating characteristic)指标达到93.3%,Segmentation AUPRO(area under the precision-recall overlap)指标达到96.1%,Segmentation AUCROC指标达到98.8%,优于大多数现有的多模态缺陷检测方法。**结论**　本文方法对于不同形状大小、在RGB图像上感知度低的缺陷有较好的检测效果,不但减少了原始特征空间内噪声对多模态表示的影响,并且对不同形状大小的缺陷具有一定的泛化能力,较好地满足了现代工业对于缺陷检测的要求。

关键词:多模态多尺度工业场景;缺陷检测;无监督特征融合(UFF);代理任务;归一化流

Multimodal multiscale industrial anomaly detection via flows

Qu Haicheng, Lin Junjie*

School of Software, Liaoning Technical University, Huludao　125000, China

Abstract: **Objective** Defect detection stands as a fundamental cornerstone in modern industrial quality control frameworks. As industries have advanced, the array of defect types has become increasingly diverse. Some defects present formidable challenges, as they are scarcely perceptible when examined using individual RGB images. This approach necessitates additional information from complementary modalities to aid in detection. Consequently, conventional deep learning methods, which rely solely on single modal data for defect identification, have proven inadequate for meeting the dynamic demands of contemporary industrial environments. Here, an innovative approach is proposed to address the nuanced challenges inherent in multimodal defect detection scenarios prevalent in modern industries, where defects vary considerably in

收稿日期:2024-04-07;**修回日期**:2024-06-17;**预印本日期**:2024-06-23

*通信作者:林俊杰　linjunjie2569@foxmail.com

基金项目:国家自然科学基金项目(42271409);辽宁省高等学校基本科研项目(LJKMZ20220699)

Supported by: National Natural Science Foundation of China(42271409); Scientific Research Foundation of the Higher Education Institutions of Liaoning Province, China(LJKMZ20220699)

shape and size and often exhibit low perceptibility within individual modalities. The proposed method can integrate a novel multimodal multiscale defect detection framework grounded in the principles of normalizing flows once the inherent noise interference within single modal feature spaces is addressed and the synergies between multimodal information are harnessed. **Method** The proposed method is structured into four main components: feature extraction, unsupervised feature fusion, an information bottleneck mechanism, and multiscale normalizing flow. First, in the feature extraction stage, features at different levels are assumed to contain varying degrees of spatial and semantic information. Low-level features contain more spatial information, whereas high-level features convey richer semantic information. Given the emphasis on spatial detail information in pixel-level defect localization tasks. Vision transformers and point transformers were used to extract features from RGB images and 3D point clouds, with a focus on blocks 1, 3, and 11, to obtain multimodal representations at different levels. The representations are subsequently fused and structured into a feature pyramid. This approach not only preserves spatial information from low-level features to aid in defect localization but also enhances the model's robustness to defects of varying shapes and sizes. Second, in the unsupervised feature fusion stage, the multimodal interaction was streamlined by employing the point feature alignment technique to align 3D point cloud features with the RGB image plane. Unsupervised multimodal feature fusion was achieved by constructing a contrastive learning matrix, thereby facilitating interaction between different modalities. Moreover, in the information bottleneck mechanism stage, a proxy task is designed to extend the information bottleneck mechanism to unsupervised settings. The aim is to obtain a more comprehensive and robust multimodal representation by minimizing noise interference within single-modal raw feature spaces while preserving the original information as much as possible. Finally, in the multiscale normalizing flow stage, the structure uses parallel flows to capture feature information at different scales. Through the fusion of these flows, interactions between features at various scales are realized. Additionally, an innovative approach for anomaly scoring is employed, wherein the average of the Top-K values in the anomaly score map replaces traditional methods such as those that use the mean or maximum values. This approach yields the final defect detection results. **Result** The proposed method is evaluated on the MVTec-3D AD dataset. This dataset is meticulously curated, encompassing 10 distinct categories of industrial products, with a comprehensive collection of 2 656 training samples and 1 137 testing samples. Each category is meticulously segmented into subclasses, delineated by the nature of the defects. The proposed method was experimentally validated, and the results demonstrated its exceptional performance. An AUCROC of 93.3%, a segmentation AUPRO of 96.1%, and a segmentation AUCROC of 98.8% were achieved. These metrics not only reflect the method's effectiveness but also have advantages over the majority of existing multimodal defect detection methodologies. Moreover, visualizations were conducted on selected samples, comparing the detection outcomes using only RGB images against those utilizing RGB images in conjunction with 3D point clouds. The latter combination has revealed defects that remain elusive when relying solely on RGB imagery. This empirical evidence firmly establishes the advantage of integrating data from both modalities, as posited in the hypothesis of this study. The ablation studies conducted provide additional insights into the efficacy of the proposed method. The introduction of an information bottleneck resulted in incremental improvements across all three metrics: 1.4% in the detection AUCROC, 2.1% in the segmentation AUPRO, and 3.5% in the segmentation AUCROC. The integration of a multiscale normalizing flow further enhanced the performance, with gains of 2.5%, 3.6%, and 1.6% across the respective metrics. These findings are indicative of the substantial contributions of both the information bottleneck and the multiscale normalizing flow to the overall performance of the defect detection framework used in this work. **Conclusion** The main contributions of this study are as follows: unsupervised feature fusion was employed to encourage information exchange between different modalities. An information bottleneck was introduced into the feature fusion module to mitigate the impact of noise in the original feature space of single modalities on multimodal interaction. Additionally, multimodal representations were utilized at different levels to construct feature pyramids, addressing the issue of the poor performance of previous flow-based methods in handling defects of varying scales. The proposed method demonstrates promising detection performance across defects of diverse shapes and sizes, including those with low perceptibility on RGB images. When the impact of noise within the original feature space on the multimodal representation is mitigated, the proposed approach can not only improve the robustness of the method but also enhance its ability to generalize the defects of varying characteristics. This approach effectively aligns with the stringent demands of modern industry for accurate and reli-

able defect detection methodologies.

Key words: multimodal and multi-scale industrial scene; anomaly detection; unsupervised feature fusion (UFF); pretext task; normalizing flow

0 引 言

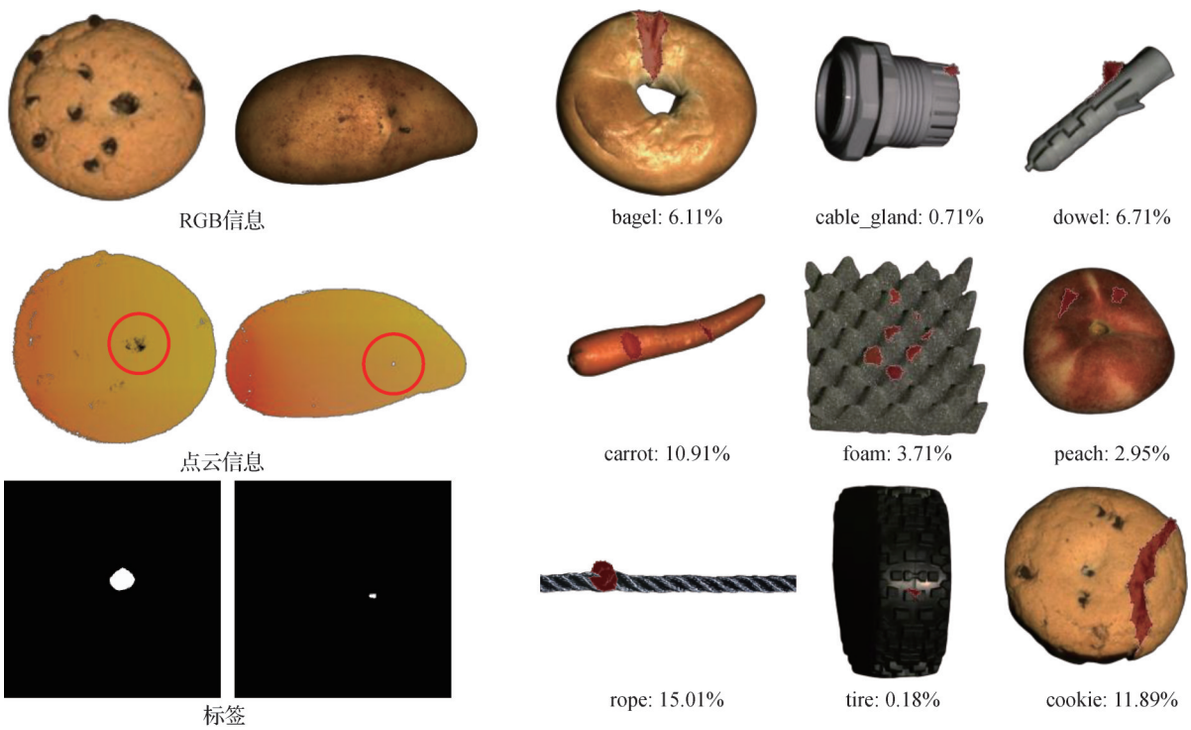
在工业生产过程中，工业缺陷检测扮演着重要的角色，旨在及时发现和定位产品或设备的缺陷，是保障生产质量和提升生产水平的重要技术之一。随着工业自动化程度的提高和生产规模的扩大，工业缺陷检测面临日益严峻的挑战，传统的人工检测方法存在主观性高、效率低和成本高等问题，难以满足大规模工业生产的需求（赵振兵 等，2021）。因此，现代工业正转向利用计算机视觉和机器学习技术，以实现工业缺陷检测的自动化和高效率化（汤勃等，2017）。

随着现代工业的发展和深度学习领域研究的不断深入，大量深度学习方法在缺陷检测领域落地应用，使得流水线上的良品率不断增加。作为附随性结果，可用于深度学习方法训练的缺陷标注样本变得稀少，令有监督方法受到了较大的限制（Pang 等，2021），而无监督方法具有对标注样本依赖性低、更贴合目前缺陷检测领域内标注样本稀缺的特点，故目前工业缺陷检测的方法主要采用无监督方法。

工业缺陷检测领域内的无监督方法根据训练使用数据的不同可以分为单模态方法和多模态方法（Pang 等，2022）。单模态的方法只利用 RGB 图像或者 3D 点云等单一模态的信息进行缺陷检测。例如基于自编码器（auto-encoder，AE）的方法（Zhou 和 Paffenroth，2017）和基于生成式对抗网络（generative adversarial network，GAN）的方法（Schlegl 等，2019），其只使用正常样本的 RGB 图像作为输入进行训练，以重建误差检测和定位缺陷。另外，基于聚类方法使用预训练模型作为特征提取器（Reiss 等，2021）提取 RGB 图像的特征，采取不同构建库策略，如通过计算输入样本与库之间的相似度（Liu 等，2023）进行缺陷检测。此外，基于知识蒸馏的方法（Deng 和 Li，2022），通过教师模型蒸馏得到正常样本的软知识，指导学生模型进行训练，以教师模型和学生模型对异常输入表现的不同判别异常样本。基于流模型的方法，使用不同的策略学习 RGB 信息原始数据分布，如使用条件正则化流根据附加条件（Gudovskiy 等，2022）以及使用 2D 归一化流（Yu 等，2021）和半监督学习方式（Rudolph 等，2021）学习更复杂的数据分布，并将分布外数据视为缺陷。单模态方法在 RGB 图像上取得了较为优越的性能表现，但由于现代工业中产品缺陷的多样性与复杂性，某些缺陷可能在单一模态上的感知度较低，如图 1（a）所示，使得单模态方法很难检测出这些缺陷。

随着 3D 传感器的发展，针对单一模态下感知度较低的缺陷，人工质量检查员往往需要使用 3D 信息辅助 RGB 图像进行缺陷检测，故 3D 信息对于精确检测和定位缺陷是较为重要的。基于上述原因，一些学者开展了一系列基于 3D 点云信息的缺陷检测研究，首先是基于自编码器以及生成式对抗网络的方法（Bergmann 和 Sattlegger，2022），但其对于 3D 信息难以定位重建位置，且丢失了大量空间结构信息，导致缺陷定位效果不佳。除此之外，基于点云局部特征描述子（fast point feature histograms，FPFH）的方法（Rani 等，2024）以及基于深度几何描述子的方法（Bergmann 和 Sattlegger，2022）由于其快速处理大规模点云数据以及适用于实时要求场景的特性，基于深度点云局部特征描述子（deep geometric descriptors，DGD）的方法在 3D 缺陷检测方向上的表现优越，但由于缺少了类似 RGB 信息中的色彩信息，所以对一些工业产品中非空间结构类型缺陷，如色彩不均等缺陷表现一般。随着第 1 个公共 3D 缺陷检测数据集 MVTec-3D AD（Bergmann 等，2021）的提出，多模态缺陷检测领域涌现出了一系列优秀的工作。其次，Rudolph 等人（2023）将基于知识蒸馏的方法从单模态扩展到多模态以及 Horwitz 和 Hoshen（2022）通过构建不同模态的记忆库将基于记忆库的方法扩展到多模态，但二者都直接拼接两个模态的特征作为模型的输入，这忽略了各模态原始特征空间中可能存在的冲突信息，导致模型难以学习两个模态之间互补的信息。为了更高效地利用两个模态的特征，并促进不同模态间信息交互，Wang 等人（2023）通过构建对比学习矩阵的方式，鼓励两个模态间信息融合，得到更有力的多模态表示，但其忽略

了对于缺陷定位任务所需要的低层特征蕴含的大量空间信息以及各模态原始特征空间内的噪声对于多模态信息交互的负面影响。Mai等人(2023)将信息瓶颈机制(information bottleneck, IB)扩展至多模态领域,消除单模态内噪声的干扰,学习一种充分无冗余的多模态表示,但其属于有监督方法,无法适应目前缺陷检测任务重标注样本稀少的特点。上述多模态缺陷检测大多只关注使用复杂的特征融合策略进行特征融合,而在某些程度上忽略了在不同缺陷在产品差异极大的占比,即多尺度信息,如图1(b)所示,不同形状大小的缺陷在不同产品上的占比可能有近百倍的差异。



(a) 不同模态对于缺陷的感知　　(b) 不同形状大小的缺陷在产品中的占比

图1　缺陷在不同模态的感知度以及不同产品上占比示意图

Fig. 1　Schematic diagram of perception of defects in different modes and proportion of defects in different products
((a) different modes' perception of defects; (b) proportion of defects of different shapes and sizes in products)

综上所述,为了解决当前工业缺陷检测领域存在的不同形状大小缺陷差异大、某些缺陷在RGB图像上的可感知度低的问题以及如何降低单模态原始特征空间内的噪声对多模态融合中存在的影响,本文提出了一个基于归一化流的多模态多尺度的缺陷检测方法。本文主要工作总结如下:1)将信息瓶颈机制扩展至无监督对去除原始特征空间内的冗余信息,并通过构建对比矩阵的方式进行无监督特征融合(unsupervised feature fusion, UFF)(Wang等,2023),得到不同层次更充分有力的多模态表示。2)使用多尺度归一化流(Zhou等,2024)结构捕捉不同尺度的缺陷信息,提高模型对于不同形状大小缺陷的泛化能力。3)在MVTec-3D AD数据集上进行大量的对比分析实验,验证了本文方法的有效性和准确性。

1　本文方法

本文方法主要由4个部分构成,分别是RGB特征提取、3D点云特征提取、无监督特征融合以及多尺度归一化流,整体结构如图2所示,其分成两个部分进行训练,首先,对图2中的(a)(b)(c)部分进行训练,对两个模态的信息进行不同层次的特征抽取;其次,将各个模态的特征通过信息瓶颈机制进行去噪,尽可能地消除原始特征空间内的噪声,再由无监督特征融合模块对去噪后的特征进行特征融合得到多模态表示;最后,训练图2(d)中的部分,接受多模态表示作为多尺度归一化流的输入,给出最后的缺陷检测和定位结果。

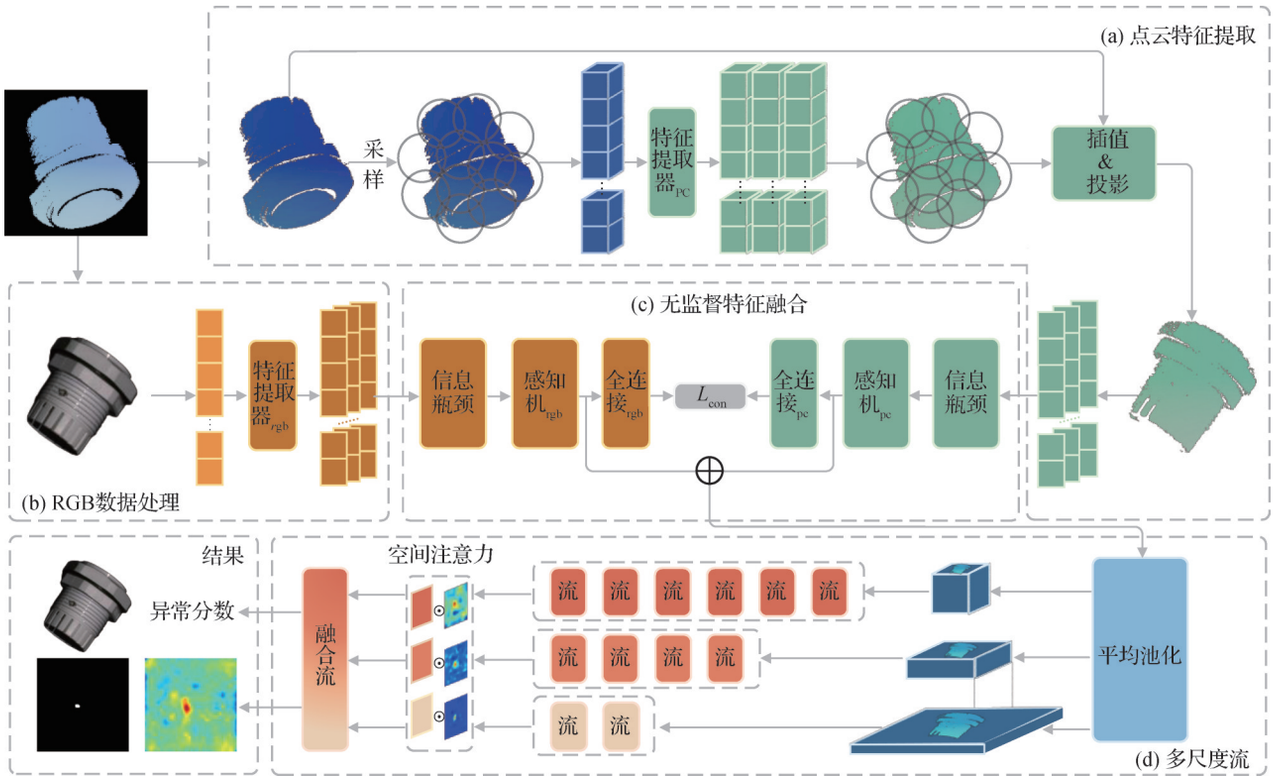


图 2　模型结构图

Fig. 2　Model structure diagram

((a)point cloud feature extraction;(b)RGB data processing;(c)unsupervised feature fusion;(d)multi-scales flow)

1.1　RGB 特征提取

本文使用在 ImageNet(Deng 等，2009)上使用 DINO(DETR with improved denoising anchor boxes)(Caron 等，2021)方式预训练的 ViT-B/8(vision transformer base with 8 pathes)(Dosovitskiy 等，2021)作为 RGB 特征提取器对 RGB 图像提取第 1、3、11 块输出的特征，以备合成不同层次的多模态表示。

1.2　3D 点云特征提取

本文使用在 ShapeNet(Chang 等，2015)上预训练的 Point Transformer (Pang 等，2022)作为 3D 特征提取器对 3D 点云提取第 1、3、11 块输出的特征以备合成不同层次的多模态表示。由于将完整的点云数据输入 Transformer 中进行处理可能带来较高的计算成本，故使用最远点采样方法(Qi 等，2017)，对由 N 个点构成的点云数据 p 进行采样，选取具有代表性的点，使得这些点尽可能覆盖空间分布，同时减少冗余的点，减少计算成本，提高点云处理的效率。将采样后的点云分成 M 组，每组 S 个点，将每组点都输入特征提取器中进行特征提取，得到输出 g，即 M 组点特征，记为 $p_n(n \in \{1,2,\cdots,N\})$。由于在采样之后，可能导致在空间上分布不均匀，使得点特征密度不平衡。所以采样后以中心点 c_m 使用反距离权重插值法将特征插值回原始点云 $p_n(n \in \{1,2,\cdots,N\})$ 维度大小，该过程可以描述为

$$\alpha_m = \frac{\dfrac{1}{\left\| c_i - p_j \right\|_2 + \delta}}{\sum_{k=1}^{M}\sum_{t=1}^{N}\dfrac{1}{\left\| c_k - p_t \right\|_2 + \delta}},\quad p'_n = \sum_{m=1}^{M}\alpha_m g_m \tag{1}$$

式中，δ 是足够小的常数。在插值后，为了简化两个模态之间的交互以及避免特征空间上的语义信息差距导致的冲突，使用原数据中的点坐标以及相机参数将 3D 点云特征 p'_n 投影至 2D 平面上，计算得到的投影点记为 $\hat{p}$。由于点云的稀疏性，将 3D 特征投影到 2D 平面时，2D 平面上可能没有对应的点。为了保证信息的完整性，将 2D 平面上没有对应点处的像素值设为 0，并将投影得到的特征图表示为 $\{\hat{p}_{x,y} \mid (x,y) \in D\}$，其中 D 代表 RGB 图像所在平面。

1.3　无监督特征融合

3D 点云是一系列三维坐标点的集合，其较好地表示了物体在空间中的形状和结构信息，弥补了

RGB图像缺少的空间结构信息，故RGB图像信息和3D点云信息在特征层面上具有一定的互补性，两个模态的融合可能会带来更强于单一模态的融合特征表示而进一步提高缺陷检测的性能。要使得两个模态的特征进行有效融合，需要考虑两个模态间的特征是否互补、是否冲突，以及如何有效地令两个模态进行信息交互，使得模型能够较好地学习到两个模态之间的内在关系。

基于上述考虑，本文通过构建3个基于块的对比矩阵的方式对3个层次的特征进行无监督特征融合（unsupervised feature fusion，UFF），并使用信息噪声对比估计（information noise contrastive estimation，InfoNCE）损失函数（van den Oord等，2019）令同一块的不同模态特征的互信息更高，而不同块的不同模态特征的互信息更低。将两个模态抽取出来的特征切分成块的形式，$\boldsymbol{f}_{\mathrm{rgb}}^{(i,j)}$和$\boldsymbol{f}_{\mathrm{pc}}^{(i,j)}$记为特征块，其中，$i$为特征的索引，$j$为特征块的索引，并使用两个多层感知机（multilayer perceptron，MLP）捕捉两个模态之间的交互信息，再通过两个全连接层将两个模态的特征映射成键特征向量或值特征向量，记为$\boldsymbol{h}_{\mathrm{rgb}}^{(i,j)}$和$\boldsymbol{h}_{\mathrm{pc}}^{(i,j)}$，以进行后续的匹配以及相似度度量，具体优化目标为

$$L_{\mathrm{con}}=\frac{\boldsymbol{h}_{\mathrm{rgb}}^{(i,j)}\cdot\boldsymbol{h}_{\mathrm{pc}}^{(i,j)\mathrm{T}}}{\sum_{t=1}^{N_b}\sum_{k=1}^{N_p}\boldsymbol{h}_{\mathrm{rgb}}^{(t,k)}\cdot\boldsymbol{h}_{\mathrm{pc}}^{(t,k)\mathrm{T}}} \tag{2}$$

式中，N_b指代批次大小，N_p指代补丁数，UFF的结构如图3所示，其中，信息瓶颈机制将在下一节中具体介绍。

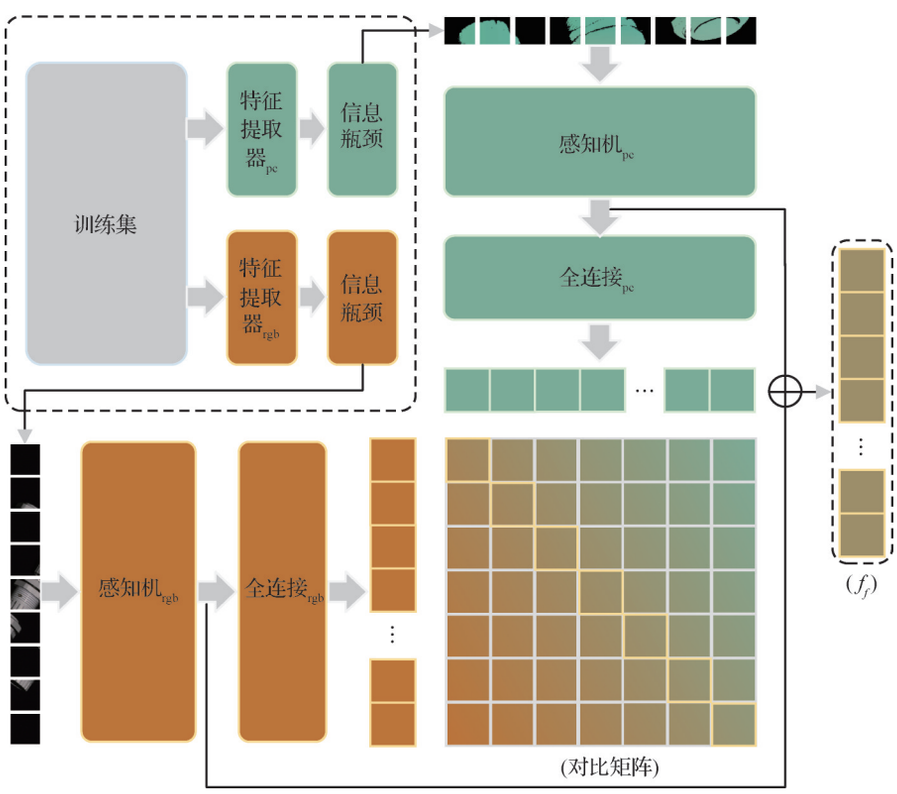


图3　无监督特征融合模块的结构

Fig. 3　The structure of unsupervised feature fusion module

1.4　无监督信息瓶颈机制

近年来，一些研究（Guo等，2023；Zhu等，2023）中将多模态特征融合时，单模态内干扰特征交互融合的冗余信息定义为噪声。除此之外，单模态特征中包含的噪声还可能在融合过程中跨模态交互引入负面影响的新噪声，妨碍模型学习模态之间的互补信息。为了尽可能地去除单模态信息中的噪声，提升模型检测和定位缺陷的性能，引入信息瓶颈来尽可能地学习到一种更充分的无冗余的多模态表示。

传统的信息瓶颈机制是一种有监督学习方法，假设给定一个输入变量X，一个压缩的表示变量T，一个以T为输入的输出变量$\hat{Y}$以及真实标签Y，旨在寻找一个充分的无冗余的压缩表示T。定义两个约束，其中，约束1使T与$\hat{Y}$的互信息（Tishby等，2000）

尽可能高；约束2使T尽可能少保留输入X的信息。在满足这两个约束的条件下，反复逼近学习，所得T则为一个条件下充分的无冗余的压缩表示，目标函数表示为

$$R_{\mathrm{IB}} = I(T;\hat{Y}) - \beta I(T;X) \tag{3}$$

式中，$I(T;\hat{Y})$为压缩表示T和输出变量$\hat{Y}$之间的互信息，$I(T;X)$为压缩表示T和输入X之间的互信息，由于互信息计算困难，故一般使用KL(Kullback-Leibler)散度近似表示为

$$I(T;X) = KL[\,p(x,y)\,|\,p(x)p(y)\,] \tag{4}$$

式中，β代表压缩率。β的值越高，压缩表示T中所含有X的原始信息越少。

传统的信息瓶颈机制使用输出变量$\hat{Y}$与真实标签Y之间的均方误差(mean square error, MSE)(Wang等，2022)作为监督信号进行有监督学习，所以其无法很好地适应缺陷检测任务标注样本稀少的特点，故本文中设计使用重建模块的重建误差作为代理任务(Albelwi，2022)代替原有的监督信号，将有监督的信息瓶颈机制扩展至无监督领域，无监督的信息瓶颈机制模块结构如图4所示。其中RE_{rgb}和RE_{pc}分别是VQ-VAE(variational quantized variational auto-encoder)(van den Oord 等，2018)和PointFlow(Yang等，2019)，是适用于RGB模态和3D点云模态(point cloud，PC)的重建模块，结构与两个生成式模块的编码器相同，并将两个生成式模块中对应的编码器结构去除。由于在消解单模态空间内噪声的同时可能会导致一些有助于缺陷检测和定位的原始信息流失，所以本模块中最后的输出不完全依赖于去噪的结果，而是将去噪后的结果与原始信息通道维度进行拼接得到最后的结果。由于VQ-VAE的离散化潜在空间可以更容易地学习到训练数据的复杂分布，捕捉更精细的细节，且重建误差相较于传统的变分自编码器(variational auto-encoder，VAE)较低，故选用VQ-VAE作为RGB模态的重建模块。考虑到RGB图像和3D点云在物理性质以及感知空间上的不同，并且点云特征中存在一些与RGB特征不同的非结构化信息，故对3D点云特征不使用与RGB特征相同的重建模块，而使用PointFlow，其结构设计更适合3D点云特征，并且其潜在空间的设计也带来了更高的灵活性，R_{rgb}和R_{pc}是两个模块重建后的结果，RE_{rgb}和RE_{pc}在推理时不参加计算。

无监督信息瓶颈机制旨在尽可能地去除单模态特征空间中存在的干扰模型学习两个模态交互信息的噪声，以及减少特征融合时产生的噪声，所以将代理任务设置为两个生成式模块的重构误差，是用于近似表示互信息$I(\cdot)$，即当两个重建模块能够在一定程度重建出X，则将T视为对于X在尽可能去除噪声的条件下，且最大限度保留原有的语义信息后的压缩表示。本文修改了传统的信息瓶颈机制中的两个约束，现给出新的约束定义，约束1：使T与X的互信息尽可能低；约束2：使T与R的互信息尽可能高，目标函数为

$$L_{\mathrm{IB}} = I(T;X) - \beta I(R;X) \tag{5}$$

通过第1个约束条件互信息下界，即

$$I(R;X) \geqslant \int \mathrm{d}x\mathrm{d}r\mathrm{d}t\, p(x)\,p(r|x)\,p(t|x)\log(r|t) \tag{6}$$

以及第2个约束条件互信息下界，即

$$I(T;X) \leqslant \int \mathrm{d}x\mathrm{d}r\mathrm{d}t\, p(t|x)\log\left(\frac{p(t|x)}{q(t)}\right) \tag{7}$$

可将式(5)进一步表示为

$$\begin{aligned} L_{\mathrm{IB}} &= I(T;X) - \beta I(R;X) \geqslant \\ &\int \mathrm{d}x\mathrm{d}r\mathrm{d}t\, p(x)\,p(r|x)\,p(t|x)\log q(r|t) - \\ &\beta \int \mathrm{d}x\mathrm{d}r\, p(x)\,p(t|x)\log\left(\frac{p(t|x)}{q(t)}\right) = \mathrm{E}(x,r)\sim p(x,r), z\sim \\ &p(t|x)\left[\log q(r|t) - \beta \times KL\left(p(t|x)\,\|\,q(t)\right)\right] = J_{\mathrm{IB}} \end{aligned} \tag{8}$$

式中，$q(r|t)$和$q(t)$是对$p(r|t)$和$p(t)$的变分近似，通过最大化J_{IB}的下界，间接对目标函数L_{IB}进行优化。最后使用蒙特卡洛近似采样将式(8)进一步表示为

$$\begin{aligned} J_{\mathrm{IB}} = \frac{1}{n}\Big[& E_{\kappa\sim p(\kappa)}[\log q(r_j|t_j)] - \\ & \beta \times KL(N(\mu_{t_j},\sigma_{t_j}^2)\,\|\,N(0,1))\Big] \end{aligned} \tag{9}$$

式中，$p(r|t)$和$p(t)$为使用神经网络学习到高斯分布的期望μ_T以及方差σ_T^2，$\kappa\sim N(0,1)$，n为采样样本数，j为样本索引，通过最大化J_{IB}，使得T学习到最充分的多模态表示的同时，消解原始数据空间中的干扰多模态信息交互的冗余信息和噪声，从而达到提高模型在缺陷检测任务上的性能。

1.5 多尺度归一化流

本文使用多尺度归一化流来捕捉不同形状大小的缺陷带来的不同尺度的缺陷特征信息。多尺度归一化流由两个部分组成，分别是用于处理不同层次

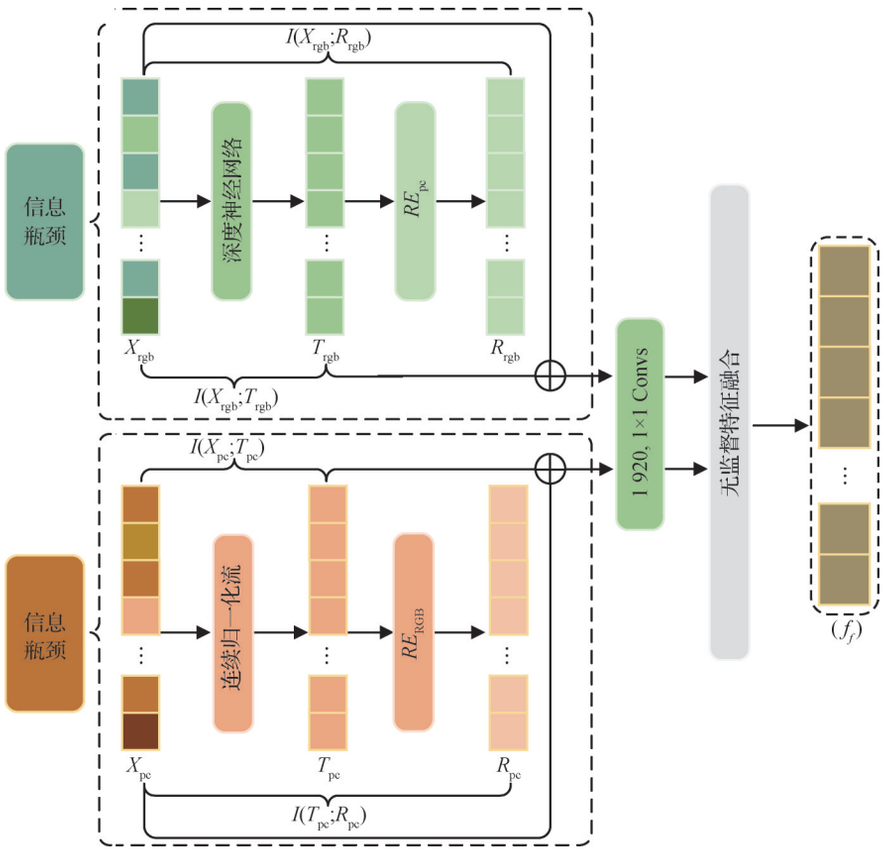


图4　无监督信息瓶颈机制的结构

Fig. 4　The structure of unsupervised information bottleneck mechanism

尺度特征的平行流，以及合并平行流处理结果实现多尺度交互并最后给出缺陷检测和定位结果的融合流（fusion flow，FF）。

归一化流的学习以及表达能力一般随着其中流块的增加而增加。不同层次尺度的特征蕴含的信息侧重点不同，RGB图像特征的低层的信息更多地包含图像的空间信息以及一些低纬信息，如边缘、颜色和纹理等，而高层特征相较于低层特征包含更多的抽象的语义信息，如图像中的对象以及场景等；3D点云特征的低层信息更倾向于表示点云的基本属性，例如点的位置、领域结构信息和局部集合特征，更具有细粒度，而高层特征可能更多地包含对象的全局性结构和语义信息，如物体的分类、形状的整体特性以及点云中不同部分的相对位置或者功能性描述。所以针对不同层次尺度的特征，应该创建不同流块的归一化流，本文方法对3个层次尺度的特征创建了3个流，流块数按特征维度的高低以及实验结果分别为2，5，12，具体实验将在第2节进行解释。为了节省计算成本，将3个层级的多模态表示记为$\{f_f^1, f_f^2, f_f^3\}$，本文方法使用960个和480个1×1的卷积核以及步长为2、池化核大小分别为2和4的池化层对f_f^1和f_f^2进行处理后，再输入平行流进行处理。

融合流负责接受平行流的输出结果，其使用了一个融合网络N_{fuse}进行特征融合，得到不同尺度的融合结果，记为$y_{fuse}^i (i = 1, 2, 3)$，最终实现融合不同尺度的信息，其结构如图5所示，融合过程可以表示为

$$\begin{aligned} &\left(y_{fuse}^1, y_{fuse}^2, y_{fuse}^3\right) = N_{fuse}\left(y_p^1, y_p^2, y_p^3\right) \\ &p_s^i = STN\left(y_{fuse}^i\right), i = 1, 2, 3 \\ &p_t^i = STN\left(y_{fuse}^i\right), i = 1, 2, 3 \end{aligned} \tag{10}$$

式中，STN为归一化流中由神经网络实现的scale$(\cdot)$和translation$(\cdot)$函数，用于学习复杂的数据分布。

融合网络N_{fuse}中，首先将由平行流处理过的所有尺度的输出通过平均池化层缩小至ms，降低计算复杂度，再将池化输出拼接在一起，初步融合各个尺度信息。再将拼接后的特征通过两个3×3的卷积层对不同尺度信息进行融合处理，实现不同尺度间信息的交互，此过程记为$fuse(\cdot)$，最后将得到的融合特征按通道维度切分为3份，再重新放大到原尺寸，得到最后的融合结果$y_{fuse}^i (i = 1, 2, 3)$，即

$$
\begin{aligned}
&ms = \min\left(size\left(y_{\mathrm{p}}^{1}\right), size\left(y_{\mathrm{p}}^{2}\right), size\left(y_{\mathrm{p}}^{3}\right)\right)\\
&y_{\mathrm{avg}}^{i} = AvgPooling_{ms}\left(y_{\mathrm{p}}^{i}\right), i = 1, 2, 3\\
&tmp_{\mathrm{concat}} = cat\left(y_{\mathrm{avg}}^{j}\right), j = 1, 2, 3\\
&fuse(\cdot) = Conv_3 \circ SiLU \circ LN \circ Conv_3\\
&y_{\mathrm{fuse}}^{k} = split\left(fuse\left(tmp_{\mathrm{concat}}\right)\right), k = 1, 2, 3
\end{aligned}
\tag{11}
$$

在得到最后缺陷检测和定位的结果前，将不同尺寸的融合结果放大到输入图像的尺寸。进行图像级别缺陷检测时使用加法聚合策略（Zhou 等，2024），将不同尺度融合流结果相加，进一步扩大每个尺度上缺陷指示的影响，捕捉所有可能的缺陷，并选取前 K 个异常分数作为整幅图像的异常分数。进行像素级别缺陷检测时使用乘法聚合策略（Zhou 等，2024），通过逐元素乘法操作以结合不同尺度的缺陷分数图，即当且仅当所有尺度的缺陷分数图同时指向缺陷时，缺陷分数才会进一步升高，得到像素级别缺陷定位结果。

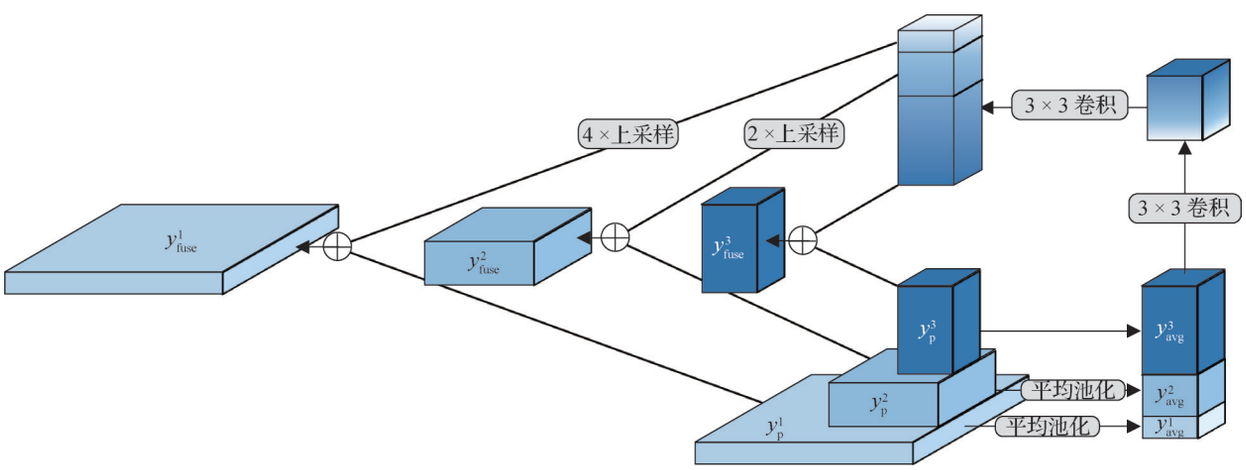


图 5　融合网络结构图

Fig. 5　Fusion network structure diagram

2　实验结果及分析

实验采用的硬件配置为 Xeon（R）Platinum 8352V 处理器，单块 GTX 4090 显卡，软件配置为 CUDA 11.6，cuDNN 8.6.0，PyTorch 1.9.1，系统版本为 Ubuntu18.04。无监督特征融合与多尺度归一化流分别训练，无监督特征融合模块使用的优化器为 AdamW，学习率为 0.000 3，共迭代 2 200 轮；多尺度归一化流模块使用的优化器为 AadmW，学习率为 0.000 1，共迭代 200 轮。Point Transformer 中组数设定为 1 024，组内大小为 128，平行流中流块数量分别是（2，5，12），信息瓶颈中的 β 参数为 0.01，计算图像异常分数的 K 值为 3%。本文方法在 MVTec-3 AD 数据集的 10 个类上使用 Detection AUCROC（area under the curve of the receiver operating characteristic），Segmentation AUPRO（area under the precision-recall overlap）以及 Segmentation AUCROC 的 3 个指标进行性能评估（下文中简称为 I-AUCROC、P-AUPRO、P-AUCROC），并与现有的一些方法进行对比，同时进行完备的消融实验以证明本文中所使用的每个模块的有效性。

2.1　数据集

实验选用的数据集为 MVTec-3 AD 数据集，该数据集于 2021 年提出，是 3D 工业缺陷检测领域中较为权威的一个数据集，其中一共有 10 个类，2 656 个训练样本以及 1 137 个测试样本，类内又将不同的缺陷划分为不同的子类，且每一个样本都拥有 RGB 图像和 3D 点云，RGB 图像由工业相机从相同的角度进行拍摄；3D 点云则由 3D 传感器进行高分辨率的扫描得到三维度的信息，3 个维度分别表示不同的位置信息，将 3 个维度的数据对应地映射成点云。

2.2　评估指标

I-AUCROC、P-AUPRO 以及 P-AUCROC 这 3 个性能指标用于评估不同模型在 MVtec-3D AD 中 10 个类上的性能指标，并且计算不同模型在 10 个类上的平均 AUCROC。图像级别缺陷检测和像素级别缺陷定位通过每一个类下的 AUCROC 曲线下的面积表示，ROC（receiver operating characteristic）展示了在不同阈值下，模型的真阳性率（true positive rate，TPR）和假

阳性率(false positive rate,FPR)之间的关系,ROC曲线下的面积越大,代表模型很好地平衡了真阳性率和假阳性率,能够较好地区分正常样本和缺陷样本。AUCROC的值越高,代表模型性能越强,当AUCROC = 1时,代表模型能完美地区分正常样本和缺陷样本,当AUCROC = 0.5时,说明模型的性能接近于随机猜测。AUPRO指标代表模型识别得到的一类样本中预测缺陷区域和真实区域重叠部分的均值。

2.3 结果分析

为了验证本文方法在缺陷检测任务上的有效性,表1—表3展示了本文方法与现有的一些缺陷检测方法于MVTec-3D AD数据集上在不同模态上的性能表现。部分产品的缺陷定位效果如图6所示。一部分模型不支持使用P-AUPRO与P- AUCROC指标进行评估,故在表中不予展示。为了验证本文方法对于检测单一模态下感知度低的某些缺陷的有效性,在图7中可视化了部分样本在单RGB图像下以及在RGB图像下并辅以3D点云进行多模态检测的效果图。可以看出,结合3D点云信息后,检测出了在RGB图像上感知度较低的缺陷。

表1 不同模型在MVTec-3D AD上的I-AUCROC结果

Table 1 I-AUCROC of different models on the MVTec-3D AD

模态	模型	类别										平均
		Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	
3D	Depth GAN	0.530	0.376	0.607	0.603	0.497	0.484	0.595	0.489	0.536	0.521	0.523
	Depth AE	0.468	0.731	0.497	0.673	0.534	0.417	0.485	0.549	0.564	0.546	0.546
	Depth VM	0.510	0.542	0.469	0.576	0.609	0.699	0.450	0.419	0.668	0.520	0.546
	Voxel GAN	0.383	0.623	0.474	0.639	0.564	0.409	0.617	0.427	0.663	0.577	0.537
	Voxel AE	0.693	0.425	0.515	0.790	0.494	0.558	0.537	0.484	0.639	0.583	0.571
	Voxel VM	0.750	**0.747**	0.613	0.738	0.823	0.693	0.679	0.652	0.609	0.690	0.699
	3D-ST	0.862	0.484	0.832	0.894	0.848	0.663	0.763	0.687	**0.958**	0.486	0.748
	FPFH	0.825	0.551	**0.952**	0.797	0.883	0.582	0.758	0.889	0.929	0.653	0.782
	本文	**0.911**	0.667	0.913	**0.912**	**0.894**	**0.744**	**0.847**	**0.902**	0.886	**0.766**	**0.844**
RGB	DifferNet	0.859	0.703	0.643	0.435	0.797	0.790	0.787	0.643	0.715	0.590	0.696
	PADiM	**0.975**	0.775	0.698	0.582	0.959	0.663	0.858	0.535	0.832	0.760	0.764
	PatchCore	0.876	0.880	0.791	0.682	0.912	0.701	0.695	0.618	0.841	0.702	0.770
	STFPM	0.930	0.847	**0.890**	0.575	0.947	0.766	0.710	0.598	0.965	0.701	0.793
	CS-Flow	0.941	**0.930**	0.827	0.795	**0.990**	0.886	0.731	0.471	**0.986**	0.745	0.830
	本文	0.961	0.927	0.889	**0.875**	0.929	**0.914**	**0.890**	**0.893**	0.963	**0.899**	**0.914**
RGB+3D	Depth GAN	0.538	0.372	0.580	0.603	0.430	0.534	0.642	0.601	0.443	0.577	0.532
	Depth AE	0.648	0.502	0.650	0.488	0.805	0.522	0.712	0.529	0.540	0.552	0.595
	Depth VM	0.513	0.551	0.477	0.581	0.617	0.716	0.450	0.421	0.598	0.623	0.555
	Voxel GAN	0.680	0.324	0.565	0.399	0.497	0.482	0.566	0.579	0.601	0.482	0.518
	Voxel AE	0.510	0.540	0.384	0.693	0.446	0.632	0.550	0.494	0.721	0.413	0.538
	Voxel VM	0.553	0.772	0.484	0.701	0.751	0.578	0.480	0.466	0.689	0.611	0.609
	3D-ST	**0.950**	0.483	**0.986**	0.921	0.905	0.632	**0.945**	**0.988**	0.976	0.542	0.833
	PatchCore+FPFH	0.918	0.748	0.967	0.883	**0.932**	0.582	0.896	0.912	0.921	0.886	0.865
	本文	0.926	**0.953**	0.931	**0.966**	0.912	**0.891**	0.931	0.927	**0.978**	**0.915**	**0.933**

注:加粗字体表示各列在不同模态条件下的最优结果。

表 2　不同模型在 MVTec-3D AD 上的 P-AUPRO 结果

Table 2　P-AUPRO of different models on the MVTec-3D AD

模态	模型	类别										平均
		Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	
3D	Depth GAN	0.111	0.072	0.212	0.174	0.160	0.128	0.003	0.042	0.446	0.075	0.142
	Depth AE	0.147	0.069	0.293	0.217	0.207	0.181	0.164	0.066	0.545	0.142	0.203
	Depth VM	0.280	0.374	0.243	0.526	0.485	0.314	0.199	0.388	0.543	0.385	0.374
	Voxel GAN	0.440	0.453	0.875	0.755	0.782	0.378	0.392	0.639	0.775	0.389	0.588
	Voxel AE	0.260	0.341	0.581	0.351	0.502	0.234	0.351	0.658	0.015	0.185	0.348
	Voxel VM	0.453	0.343	0.521	0.697	0.680	0.284	0.349	0.634	0.616	0.346	0.492
	FPFH	**0.973**	**0.879**	**0.982**	**0.906**	0.892	**0.735**	**0.977**	**0.982**	**0.956**	**0.961**	**0.924**
	本文	0.916	0.823	0.920	0.811	**0.902**	0.728	0.959	0.962	0.927	0.904	0.885
RGB	CFlow-AD	0.855	0.919	0.958	0.867	**0.969**	0.500	0.889	0.935	0.904	0.919	0.871
	PatchCore	0.901	0.949	0.928	0.877	0.892	0.563	0.904	0.932	0.908	0.906	0.876
	PADiM	**0.980**	0.944	0.945	0.925	0.961	0.792	**0.966**	**0.940**	0.932	0.912	0.930
	本文	0.968	**0.951**	**0.968**	**0.926**	0.949	**0.927**	0.914	0.922	**0.978**	**0.939**	**0.944**
RGB+3D	Depth GAN	0.421	0.422	0.778	0.696	0.494	0.252	0.285	0.362	0.402	0.631	0.474
	Depth AE	0.432	0.158	0.808	0.491	0.841	0.406	0.262	0.216	0.716	0.478	0.481
	Depth VM	0.388	0.321	0.194	0.570	0.408	0.282	0.244	0.349	0.268	0.331	0.336
	Voxel GAN	0.664	0.620	0.766	0.740	0.783	0.332	0.582	0.790	0.633	0.483	0.639
	Voxel AE	0.467	0.750	0.808	0.550	0.765	0.473	0.721	0.918	0.019	0.170	0.564
	Voxel VM	0.510	0.331	0.413	0.715	0.680	0.279	0.300	0.507	0.611	0.366	0.471
	3D-ST	0.950	0.483	0.986	0.921	0.905	0.632	0.945	**0.988**	**0.976**	0.542	0.833
	PatchCore+FPFH	**0.976**	**0.969**	0.979	**0.973**	0.933	0.888	**0.975**	0.981	0.950	0.971	0.960
	本文	0.968	0.962	**0.989**	0.942	**0.952**	**0.924**	0.974	0.962	0.957	**0.976**	**0.961**

注：加粗字体表示各列在不同模态条件下的最优结果。

表 3　不同模型在 MVTec-3D AD 上的 P-AUCROC 结果

Table 3　P-AUCROC of different models on the MVTec-3D AD

模态	模型	类别										平均
		Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	
3D	FPFH	**0.994**	**0.966**	**0.999**	**0.946**	**0.966**	**0.927**	**0.996**	**0.999**	**0.996**	**0.990**	**0.978**
	本文	0.916	0.923	0.920	0.911	0.902	0.898	0.959	0.962	0.927	0.914	0.923
RGB	PatchCore	0.983	**0.984**	**0.980**	0.974	0.972	0.849	**0.976**	**0.983**	**0.987**	**0.977**	0.967
	本文	**0.987**	0.981	0.978	**0.983**	**0.976**	**0.907**	0.964	0.972	0.973	0.970	**0.969**
RGB+3D	PatchCore+FPFH	**0.996**	0.992	**0.997**	**0.994**	0.981	0.974	**0.996**	**0.998**	0.994	**0.995**	**0.992**
	本文	0.994	0.992	0.989	0.982	**0.982**	**0.994**	0.979	0.992	**0.997**	0.987	0.989

注：加粗字体表示各列在不同模态条件下的最优结果。

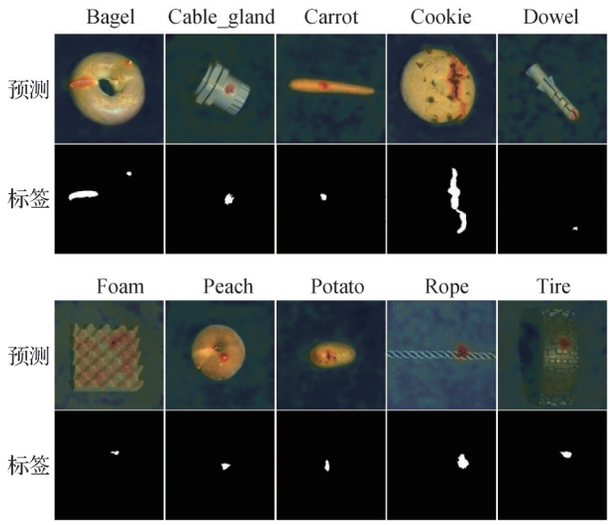


图6　部分产品的缺陷定位效果图

Fig. 6　Location effect diagram of some product defects

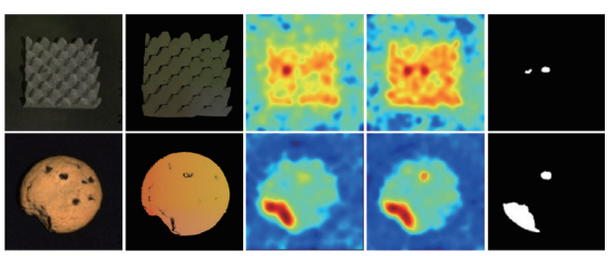


图7　部分产品不同模态下缺陷定位效果图

Fig. 7　Comparison of defect localization of some products under RGB and multimodal

从上述数据可以看出，本文方法同时使用RGB图像和3D点云进行缺陷检测时，I-AUCROC、P-AUPRO以及P-AUCROC这3个指标分别得到了93.3%、96.1%和98.9%的成绩，相较于绝大部分现有的一些缺陷检测方法展示出了优越的性能，仅在P-AUCROC指标上以98.9%的成绩略逊于Patch-Core+FPFH的99.2%，以及在部分数据集上的指标略低于3D-ST，这可能是因为二者在3D点云信息上所使用的点云局部特征描述子和深度几何描述子相对于Ponint Transformer能更好地捕捉局部空间特征（Rani等，2024）。

局部空间特征对于缺陷检测任务是较为重要的，局部空间特征能帮助检测系统精确地定位物体上的缺陷位置。例如，在金属表面或纺织品中，缺陷如划痕、凹坑或破洞往往具有独特的局部形状和纹理。通过识别这些局部特征，系统能够更准确地定位缺陷，且局部特征可以帮助模型区分缺陷和背景之间的微小差异，所以PatchCore+FPFH的定位性能要略高于本文方法。从表1—表3可以看出，二者在多模态缺陷检测任务上的表现优于本文方法时，其在3D点云信息上的表现也普遍要优于本文方法，如表1中RGB+3D部分的第7行第4列、表2中RGB+3D部分的第8行第2列与第3列等。本文方法中处理3D点云信息所使用的Point Transformer，相对于点云局部特征描述子和深度几何描述子，除了捕捉局部空间特征之外，还提取了3D点云信息中的依赖关系以及上下文信息。由于在实际缺陷检测场景中缺陷常常与相邻位置的状态有关，故依赖关系以及上下文信息可以有效降低漏检率（Albelwi，2022），PatchCore+FPFH的I-AUCROC（86.5%）远低于本文方法（93.3%），也在一定程度上佐证了：专注于局部空间特征提取的方法虽然在缺陷定位上表现亮眼，但是数据中的依赖关系和上下文信息对于降低漏检率十分关键。

2.3　消融实验

为了验证本文方法中各个模块的有效性以及对于性能的影响，设计了6组对比实验，其中前5组为消融实验，第6组为完整的性能表现。w/o表示去除某个模块，w表示保留某个模块，并对特定对比实验组进行补充说明：

1）为了探讨不同流块数量的流对于缺陷的学习能力以及多尺度流（multi-scales flow，MF）对于不同尺度特征的捕捉能力的影响，设计了第4组实验，在本组实验中不构建平行流和融合流，只使用不同的流块进行消融实验，直接将特征提取器的最高层次的特征进行融合，并按照流块的个数分为5种情况，分别是{1，2，5，12，13}，记为sf_1，sf_2，sf_5，sf_{12}，sf_{13}，最后的缺陷分数改用均值计算；

2）为了验证融合流对于多尺度特征交互融合上的有效性，设计了第5组实验，在本组实验中，构建平行流的同时，不构建融合流，直接将3个尺度的流模型的结果放大至输入图像大小，然后进行加法融合策略以及乘法融合策略得到缺陷检测结果；

3）在未使用无监督特征融合模块以及不构建平行流时，直接将两个模态的特征分别通过960个1×1的卷积核和上采样处理后按通道维度进行拼接得到维度大小为（1 920，3 136）的多模态融合表示，具体的消融实验表现情况见表4，其中MF代表多尺度流结构（multi-scales flows），FF代表融合流结构（fusion

flow）。

对比第1组实验与第6组实验的表现，验证了单一模态信息和多模态信息对于缺陷检测和定位性能的影响。在仅有RGB图像参与缺陷检测任务时，方法的性能远高于只有3D点云参加检测；而RGB图像和3D点云同时参与检测时，相较于单一模态缺陷检测的3个指标则分别得到了一定的提高。由此可以得到结论，在3D点云中可能存在一些与RGB图像互补的结构空间信息在UFF模块的作用下与RGB信息一起形成了相较于单一模态更加强大的表示，助力提升在缺陷检测任务上的性能表现，且RGB图像在缺陷检测任务中的占比更大，影响因子更高。

对比第3组实验和第1组实验，在未使用UFF模块的情况下，直接将两个模态的信息简单地拼接在一起时，得到的性能表现并不如单一的RGB图像参与检测时的性能，可以得到结论，直接拼接两个模态的信息，可能由于不同模态在原始空间上分布的不同，或者不同模态对于世界的感知机制不同导致一定的冲突，使得在简单拼接得到的融合表示中产生了一些冲突信息，而这些冲突信息妨碍模型学习到正常的数据分布，从而使得第3组的表现不如第1组。

观察第4组实验，本文设计使用原多尺度流中不同流块数量验证流块数量对于数据分布的学习能力，可以看到随着流块的增加，模型对于数据分布的学习能力得到了进一步的提升，但使用5个流块和12个流块时模型的性能表现趋于平稳，并且在使用13个流块时，缺陷定位性能下降，故继续堆叠流块数量可能会导致过拟合，从而影响性能表现。具体的验证流块数量对于多尺度流的性能表现的影响将在表5中展示。

最后，第2组实验以及第5组实验中的结果表明，信息瓶颈和融合流的加入使得本文方法的性能得到了提升，证明了这两个模块在去除单模态特征空间内噪声和实现不同尺度特征之间信息交互方面的有效性。

表4　不同模态以及模块下方法的性能

Table 4　Performance under different modals and modules

编号	方法	I-AUCROC	P-AUPRO	P-AUCROC
第1组	w/o PC	0.914	0.935	0.969
	w/o RGB	0.844	0.866	0.923
第2组	w/o IB	0.919	0.940	0.954
第3组	w/o UFF	0.921	0.942	0.974
第4组	w/o MF & w/o FF, sf_1	0.854	0.896	0.929
	w/o MF & w/o FF, sf_2	0.873	0.912	0.935
	w/o MF & w/o FF, sf_5	0.906	0.921	0.969
	w/o MF & w/o FF, sf_{12}	0.908	0.925	0.971
	w/o MF & w/o FF, sf_{13}	0.911	0.923	0.967
第5组	w MF & w/o FF	0.920	0.944	0.979
第6组	Complete	**0.933**	**0.961**	**0.989**

注：加粗字体表示各列最优结果。

2.4　超参数分析

2.4.1　平行流流块

为了讨论在平行流中如何针对不同维度以及尺度的特征选择合适的流块数量，本文设计了不同的流块数量方案，以寻求性能表现最佳的流块数量方案，具体的实验结果见表5。从表5数据可以看出，对于最高层的融合特征，在堆叠12个流块之后，性能达到最高，故本文认为12个流块已经可以很好地学习高层特征。当使用流块数量为（2，5，12）的方案8时，本文方法的I-AUCROC和P-AUCROC达到

最高，而P-AUPRO仅稍微劣于方案9的0.963。流块的堆叠往往带来较高的计算成本，故选用方案8作为实验方案。从方案9和方案10的结果可以看出，当叠加的流块过多时，可能会导致一定程度上的过拟合风险，导致性能降低；从方案1—7的结果可以看出，当叠加的流块过少时，模型的复杂度太低，无法很好地捕捉所有尺度上的特征，从而体现出较为孱弱的性能。

表5　不同流块方案下的MMF-AD性能

Table 5　Performance of MMF-AD under different flow blocks schemes

方案	流块数量	I-AUCROC	P-AUPRO	P-AUCROC
1	(1,1,12)	0.918	0.931	0.946
2	(2,1,12)	0.922	0.939	0.958
3	(3,1,12)	0.923	0.937	0.958
4	(4,1,12)	0.924	0.938	0.957
5	(2,2,12)	0.927	0.943	0.963
6	(2,3,12)	0.929	0.951	0.974
7	(2,4,12)	0.930	0.957	0.982
8	(2,5,12)	**0.933**	0.961	**0.989**
9	(2,6,12)	0.927	**0.963**	0.985
10	(2,7,12)	0.929	0.960	0.977

注：加粗字体表示各列最优结果。

2.4.2　无监督信息瓶颈机制β

在本文提出的无监督信息瓶颈机制中存在一个参数β，用于表示压缩表示与原始输入之间的压缩率，为了讨论该参数对模型性能的影响，本文设置了不同的参数值，具体实验结果见表6。

从表6可以看出，当$\beta=0.01$时，模型的性能表现最佳；且令人瞩目的是第1组的性能表现不如表4的第2组，即当$\beta=1$时，使用信息瓶颈机制时的本文方法的性能低于未使用信息瓶颈机制。对于参数β，当β越大，代表对于原输入信息的保有量越低。可将信息瓶颈定义成一个由两个玩家抽取积木塔中积木的游戏，游戏定义为若玩家抽取积木导致积木倒塌，则玩家失败，反之获胜。若将可以抽取并保持积木塔形状的积木定义为噪声，则信息瓶颈机制的优化目的可以抽象为平衡积木塔的原始结构以及抽取积木之间的关系，在尽可能抽取积木的情况下保持塔的形状，β越大，抽取的积木越多。当β过大，则使得原数据的保有量太低，在缺陷检测任务中可以视为原始空间信息减少，使得影响了模型的性能；而β过小，以至于模型几乎忽略信息压缩这一方面，模型可能会倾向于简单地学习输入到输出的复杂映射，而不进行任何有意义的信息压缩。

表6　不同β下方法的性能

Table 6　Performance under different β

方案	β	I-AUCROC	P-AUPRO	P-AUCROC
1	1	0.881	0.927	0.943
2	1E−1	0.925	0.957	0.968
3	1E−2	**0.933**	**0.961**	**0.989**
4	1E−3	0.929	0.957	0.979
5	1E−4	0.920	0.945	0.961
6	0	0.912	0.937	0.946

注：加粗字体表示各列最优结果。

2.4.3　使用Top-K计算图像异常分数中K值

本文中对于计算图像异常分数使用Top-K的方法，即以像素异常分数中前K%的分数的均值，取代以往的直接使用均值计算缺陷分数（记为均值法）和取缺陷分数最高的像素的缺陷分数作为整幅图像的缺陷分数（记为最大值法）的计算方法。为了确定一个合适的K值，此处预设了9种方案，其分别是（1，1%，2%，3%，5%，10%，25%，50%，100%），其中，1代表最大值法，100%则代表均值法，无论是Top-K、均值法或者最大值法影响的都是图像的缺陷分数，故只与I-AUCROC指标有关，现绘制关于不同K值下的I-AUCROC指标（各个类的I-AUCROC的均值）折线图如图8所示。从图8可以看出，当K值设置为3%时，随着K值的继续增大，I-AUCROC开始减少，并且100%处发生了快速下降。由表2得，本文方法的性能对于像素异常分数的预测较为准确，即缺陷定位性能较好。故I-AUCROC的突然下降可能与较高的定位性能有关，由于图像中缺陷像素被赋予了较高的缺陷分数，而其他区域则被赋予了较低的分数，当缺陷区域较小时，若使用均值法计算整张缺陷分数，缺陷区域的高分则会被低分稀释，从而导致分类性能下降。

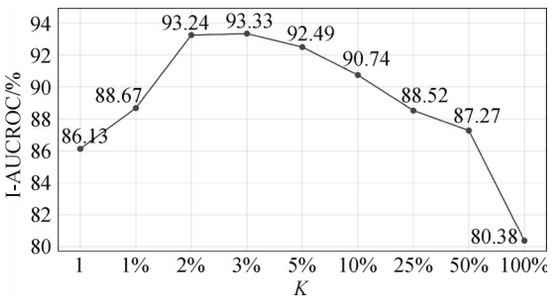


图 8　不同 K 值下模型的 I-AUCROC

Fig. 8　I-AUCROC under different K values

4　结　论

为了适应工业缺陷检测场景下的缺陷尺度不断变化，解决单一模态的某些缺陷不易感知的问题，本文提出了一种基于归一化流的多模态多尺度缺陷检测方法。采用无监督特征融合的方式融合不同模态的特征，为了减少单模态原始特征空间的噪声对于多模态交互的影响，在特征融合模块中应用信息瓶颈，并利用不同层次的特征构建特征金字塔，处理以往基于流的方法对于缺陷尺度不同而表现不佳的问题，所提出的方法在 MVTec-3D AD 数据集上的表现优于以往的大部分方法。

在消融实验中，发现 RGB 图像和点云对于多模态缺陷检测任务的影响因子不同。在未来的工作中，将进一步探索这两个模态之间的权重关系，进而能找到更合理的方案，从而更高效地利用两个模态的信息进行缺陷检测。

参考文献（References）

Albelwi S. 2022. Survey on self-supervised learning: auxiliary pretext tasks and contrastive learning methods in imaging. Entropy, 24(4): #551 [DOI: 10.3390/e24040551]

Bergmann P, Jin X, Sattlegger D and Steger C. 2021. The MVTec 3D-AD dataset for unsupervised 3D anomaly detection and localization [EB/OL]. [2024-04-07]. https://arxiv.org/pdf/2112.09045.pdf

Bergmann P and Sattlegger D. 2022. Anomaly detection in 3D point clouds using deep geometric descriptors [EB/OL]. [2024-04-07]. https://arxiv.org/pdf/2202.11660.pdf

Caron M, Touvron H, Misra I, Jegou H, Mairal J, Bojanowski P and Joulin A. 2021. Emerging properties in self-supervised vision transformers//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: EEEE: 9630-9640 [DOI: 10.1109/ICCV48922.2021.00951]

Chang A X, Funkhouser T, Guibas L, Hanrahan P, Huang Q X, Li Z M, Savarese S, Savva M, Song S R, Su H, Xiao J X, Yi L and Yu F. 2015. ShapeNet: an information-rich 3D model repository [EB/OL]. [2024-04-07]. https://arxiv.org/pdf/1512.03012.pdf

Deng H Q and Li X Y. 2022. Anomaly detection via reverse distillation from one-class embedding//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 9727-9736 [DOI: 10.1109/CVPR52688.2022.00951]

Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/CVPR. 2009.5206848]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale [EB/OL]. [2024-04-07]. https://arxiv.org/pdf/2010.11929.pdf

Gudovskiy D, Ishizaka S and Kozuka K. 2022. CFLOW-AD: real-time unsupervised anomaly detection with localization via conditional normalizing flows//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE: 1819-1828 [DOI: 10.1109/wacv51458.2022.00188]

Guo X B, Kot A and Kong A W K. 2023. Pace-adaptive and noise-resistant contrastive learning for multimodal feature fusion. IEEE Transactions on Multimedia, 25: 9437-9448 [DOI: 10.1109/TMM.2023.3252270]

Horwitz E and Hoshen Y. 2022. An empirical investigation of 3D anomaly detection and segmentation [EB/OL]. [2024-04-07]. https://arxiv.org/pdf/2203.05550v2.pdf

Zhao Z B, Jiang Z G, Li Y X, Qi Y C, Zhai Y J, Zhao W Q and Zhang K. 2021. Overview of visual defect detection for transmission line components. Journal of Image and Graphics, 26(11): 2545-2560 (赵振兵，蒋志钢，李延旭，戚银城，翟勇杰，赵文清，张珂. 2021. 输电线路部件视觉缺陷检测综述. 中国图象图形学报，26(11): 2545-2560) [DOI: 10.11834/jig.200689]

Liu H Q, Xu X D, Li E H, Zhang S C and Li X L. 2023. Anomaly detection with representative neighbors. IEEE Transactions on Neural Networks and Learning Systems, 34(6): 2831-2841 [DOI: 10.1109/tnnls.2021.3109898]

Mai S, Zeng Y and Hu H F. 2023. Multimodal information bottleneck: learning minimal sufficient unimodal and multimodal representations. IEEE Transactions on Multimedia, 25: 4121-4134 [DOI: 10.1109/TMM.2022.3171679]

Pang G S, Shen C H, Cao L B and Van Den Hengel A. 2021. Deep learning for anomaly detection: a review. ACM Computing Surveys, 54(2): #38 [DOI: 10.1145/3439950]

Pang Y T, Wang W X, Tay F E H, Liu W, Tian Y H and Yuan L. 2022. Masked autoencoders for point cloud self-supervised learning//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 604-621 [DOI: 10.1007/978-3-031-20086-1_35]

Qi C R, Yi L, Su H and Guibas L J. 2017. PointNet++: deep hierarchical feature learning on point sets in a metric space [EB/OL]. [2024-04-07]. https://arxiv.org/pdf/1706.02413.pdf

Rani A, Ortiz-Arroyo D and Durdevic P. 2024. Advancements in point cloud-based 3D defect detection and classification for industrial systems: a comprehensive survey [EB/OL]. [2024-04-07]. https://arxiv.org/pdf/2402.12923.pdf

Reiss T, Cohen N, Bergman L and Hoshen Y. 2021. PANDA: adapting pretrained features for anomaly detection and segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 2805-2813 [DOI: 10.1109/cvpr46437.2021.00283]

Rudolph M, Wandt B and Rosenhahn B. 2021. Same same but DifferNet: semi-supervised defect detection with normalizing flows//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 1906-1915 [DOI: 10.1109/wacv48630.2021.00195]

Rudolph M, Wehrbein T, Rosenhahn B and Wandt B. 2023. Asymmetric student-teacher networks for industrial anomaly detection//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2591-2601 [DOI: 10.1109/WACV56688.2023.00262]

Schlegl T, Seeböck P, Waldstein S M, Langs G and Schmidt-Erfurth U. 2019. f-AnoGAN: fast unsupervised anomaly detection with generative adversarial networks. Medical Image Analysis, 54: 30-44 [DOI: 10.1016/j.media.2019.01.010]

Tang B, Kong J Y and Wu S Q. 2017. Review of surface defect detection based on machine vision. Journal of Image and Graphics, 57(1): 47-56 (汤勃, 孔建益, 伍世虔. 2017. 机器视觉表面缺陷检测综述. 中国图象图形学报, 22(12): 1640-1663) [DOI: 10.11834/jig.160623]

Tishby N, Pereira F C and Bialek W. 2000. The information bottleneck method [EB/OL]. [2024-04-07]. https://arxiv.org/pdf/physics/0004057.pdf

Van Den Oord A, Li Y Z and Vinyals O. 2019. Representation learning with contrastive predictive coding [EB/OL]. [2024-04-07]. https://arxiv.org/pdf/1807.03748.pdf

van den Oord A, Vinyals O and Kavukcuoglu K. 2018. Neural discrete representation learning [EB/OL]. [2024-04-07]. https://arxiv.org/pdf/1711.00937.pdf

Wang Q, Ma Y, Zhao K and Tian Y J. 2022. A comprehensive survey of loss functions in machine learning. Annals of Data Science, 9(2): 187-212 [DOI: 10.1007/s40745-020-00253-5]

Wang Y, Peng J L, Zhang J N, Yi R, Wang Y B and Wang C J. 2023. Multimodal industrial anomaly detection via hybrid fusion//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 8032-8041 [DOI: 10.1109/CVPR52729.2023.00776]

Yang G D, Huang X, Hao Z K, Liu M Y, Belongie S and Hariharan B. 2019. PointFlow: 3D point cloud generation with continuous normalizing flows//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4540-4549 [DOI: 10.1109/ICCV.2019.00464]

Yu J W, Zheng Y, Wang X, Li W, Wu Y S, Zhao R and Wu L W. 2021. FastFlow: unsupervised anomaly detection and localization via 2D normalizing flows [EB/OL]. [2024-04-07]. https://arxiv.org/pdf/2111.07677.pdf

Zhou C and Paffenroth R C. 2017. Anomaly detection with robust deep autoencoders//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada: ACM: 665-674 [DOI: 10.1145/3097983.3098052]

Zhou Y X, Xu X, Song J K, Shen F M and Shen H T. 2024. MSFlow: multiscale flow-based framework for unsupervised anomaly detection. IEEE Transactions on Neural Networks and Learning Systems, 2024: 1-14 [DOI: 10.1109/TNNLS.2023.3344118]

Zhu Y, Xu R D, An H, Tao C B and Lu K. 2023. Anti-noise 3D object detection of multimodal feature attention fusion based on PV-RCNN. Sensors, 23(1): #233 [DOI: 10.3390/s23010233]

作者简介

曲海成，男，副教授，硕士生导师，主要研究方向为遥感影像高性能计算、视觉信息计算、目标检测与识别。
E-mail：quhaicheng@lntu.edu.cn
林俊杰，通信作者，男，硕士研究生，主要研究方向为工业缺陷检测。E-mail：linjunjie2569@foxmail.com