中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-15

论文引用格式:Zhao Mingming, Zhang Erhu. A fine-grained industrial defect detection method based on semantic-detail synergistic fusion[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.260260. (赵明明, 张二虎. 语义—细节协同融合的细粒度工业品缺陷检测[J/OL]. 中国图象图形学报,XXXX:1-15. DOI: 10.11834/jig.260260.)[DOI:10.11834/jig.260260]

语义—细节协同融合的细粒度工业品缺陷检测

赵明明,张二虎

西安理工大学印刷包装与数字媒体学院,西安　710048

摘　要:**目的**　针对多类别无监督工业品缺陷检测中,现有重建模型难以兼顾浅层细节特征,且易陷入“捷径学习”导致细粒度缺陷漏检的问题,提出一种语义—细节协同融合的细粒度缺陷检测方法SDSF-ViTAD(Semantic-Detail Synergistic Fusion-ViTAD)。**方法**　首先,采用基于DINO先验的纯视觉Transformer提取多层级密集先验;其次,设计语义—细节协同融合模块SDSFM(Semantic-Detail Synergistic Fusion Module),利用交叉注意力机制实现深浅层特征的对齐,并引入空间感知门控机制动态过滤背景噪声,精准强化细粒度特征;接着,在重建阶段引入基于信息瓶颈的特征扰动机制,切断特征直接复制的通路,并结合Top-K稀疏注意力优化解码过程,强制模型通过全局上下文推断正常模式,从而在推理时显著放大异常区域的重建误差;最后,引入结合难分样本挖掘的余弦距离损失函数,使得模型在训练过程中重点关注难以拟合的正常区域,增强模型对细粒度异常的判别能力。**结果**　在VisA(visual anomaly detection dataset)数据集上的实验结果表明,本文算法与基准模型ViTAD相比,在图像级平均受试者工作特征曲线下面积mAUROC(mean area under the receiver operating characteristic curve)、平均精度mAP(mean average precision)和平均F1分数mF1(mean F1-score)指标上分别提升了3.28%、3.00%和3.70%,在像素级mAP和mF1指标上分别提升了3.84%和2.98%,其中对微小缺陷高度敏感的平均区域重叠率曲线下面积mAUPRO(mean area under the per-region overlap)指标提升了3.8%,同时,在金属零件缺陷检测MPDD(Metal Parts Defect Detection)和BeanTech异常检测(BeanTech Anomaly Detection, BTAD)公开数据集上的实验进一步证明了该模型在应对不同工业场景多类别统一检测中的通用性与有效性。**结论**SDSF-ViTAD模型提升了复杂工业场景下的细粒度缺陷检测精度,优于现有主流多类别无监督算法。

关键词:多类别无监督缺陷检测;细粒度缺陷;视觉Transformer;注意力机制;捷径学习;特征扰动

A fine-grained industrial defect detection method based on semantic-detail synergistic fusion

Zhao Mingming, Zhang Erhu

Faculty of Printing, Packing and Digital Media, Xi'an University of Technology, Xi'an 710048, China

Abstract: **Objective** In the field of industrial manufacturing, automated surface defect detection is an essential component for quality control and production efficiency. Deep learning techniques have been widely applied. However, they face specific challenges in real-world scenarios. These challenges include the scarcity of defective samples, the difficulty of identifying fine-grained defects, and the necessity to evaluate multiple categories using a unified framework. Fine-grained

收稿日期:2026-05-09;**修回日期**:2026-07-18

基金项目:国家自然科基金项目(项目编号:62573345);国家科技重大专项(项目编号:2025ZD1606404)

Supported by: Project supported by the National Natural Science Foundation of China (Grant No. 62573345); the National Science and Technology Major Project (Grant No. 2025ZD1606404)

© 中国图象图形学报版权所有

defects, such as micro-cracks, scratches, and spots, exhibit low contrast and subtle variations against complex background textures. This makes them difficult for conventional models to separate from the background. Furthermore, the paradigm of Multi-class Unsupervised Anomaly Detection (MUAD) requires a single network to model the normal feature distributions of various distinct products simultaneously. In this setting, existing reconstruction-based models frequently encounter the problem of "shortcut learning". Because the network fits a massive cross-category normal feature space, local features of different products tend to overlap. Consequently, during the inference phase, the decoder utilizes normal prior knowledge from other categories to reconstruct the abnormal regions of the input. This identity mapping phenomenon causes the reconstruction residual of abnormal areas to approach zero, resulting in missed detections for fine-grained defects. To address the insufficient utilization of shallow detail features and the shortcut learning bottleneck, this study proposes a fine-grained industrial defect detection method based on semantic-detail collaborative fusion, denoted as SDSF-ViTAD. The objective is to develop a unified framework capable of localizing fine-grained defects across diverse industrial products by balancing deep semantic understanding and shallow texture details. **Method** The proposed SDSF-ViTAD framework utilizes a purely vision-based Transformer (ViT) pre-trained with DINO. This model serves as a frozen encoder to extract multi-level dense feature priors from the input images. To integrate information across different levels, the Semantic-Detail Synergistic Fusion Module (SDSFM) is designed. The deepest feature map extracted by the encoder provides global perception and semantic understanding, serving as the query vector in a cross-attention mechanism. The shallower feature maps contain rich edge and texture information. These maps are concatenated along the channel dimension, processed through average pooling and layer normalization, and then utilized as the key and value vectors. To prevent semantic mismatch and the introduction of background noise during this process, a spatial-aware gating mechanism is utilized. This mechanism applies one-dimensional convolution to the query vector to generate an adaptive gating weight. This weight dynamically filters the cross-attention output to retain defect details and suppress noise. Following the SDSFM, a feature perturbation mechanism based on the information bottleneck concept is introduced at the reconstruction stage. During training on normal samples, a dropout layer randomly discards a specific proportion of the fused activation values. This operation cuts off the direct feature copying pathway. Consequently, it forces the decoder to utilize the uncorrupted global context to infer and reconstruct the missing normal texture, thereby fitting the intrinsic manifold of the normal samples. During testing, the perturbation is disabled. Since the decoder is trained solely to reconstruct normal patterns, it cannot replicate unseen abnormal features. This inability significantly amplifies the reconstruction error in defective regions. Additionally, the standard dense self-attention in the decoder is replaced with a Top-K sparse attention mechanism. The Top-K mechanism applies a hard mask to the attention score matrix, restricting each token to interact only with the K tokens that share the highest similarity. By truncating low-similarity connections, the model compresses the information transmission pathways and limits the shortcut learning problem. Finally, the network is optimized using a multi-level Cosine Distance loss combined with hard normal mining. This mechanism guides the model to focus on hard-to-fit normal regions during the training process, thereby enhancing its discriminative capability for fine-grained anomalies. **Result** Experimental results on the VisA dataset validate the proposed algorithm. Compared with the baseline model (ViTAD), it improves the image-level mAUROC, mAP, and mF1 metrics by 3.28%, 3.00%, and 3.70%, respectively. It also improves the pixel-level mAP and mF1 metrics by 3.84% and 2.98%, respectively. Among them, the mAUPRO metric, which is highly sensitive to tiny defects, is improved by 3.8%. Furthermore, experiments on the public MPDD and BTAD datasets further demonstrate the universality and effectiveness of the proposed model in unified multi-class detection across different industrial scenarios. **Conclusion** The SDSF-ViTAD model improves the accuracy of fine-grained defect detection in complex industrial scenarios, outperforming existing mainstream multi-category unsupervised algorithms.

Key words: multi-class unsupervised defect detection; fine-grained defect; visual transformer; attention mechanism; shortcut learning; feature perturbation

论文引用格式:Zhao M M and Zhang E H. 2026. A fine-grained industrial defect detection method based on semantic-detail synergistic fusion. Journal of Image and Graphics,31(00): 0000-0000)(赵明明,张

© 中国图象图形学报版权所有

二虎. 2026. 语义—细节协同融合的细粒度工业品缺陷检测. 中国图象图形学报,31(00):0000-0000)[DOI:10. 11834/jig. 260260]

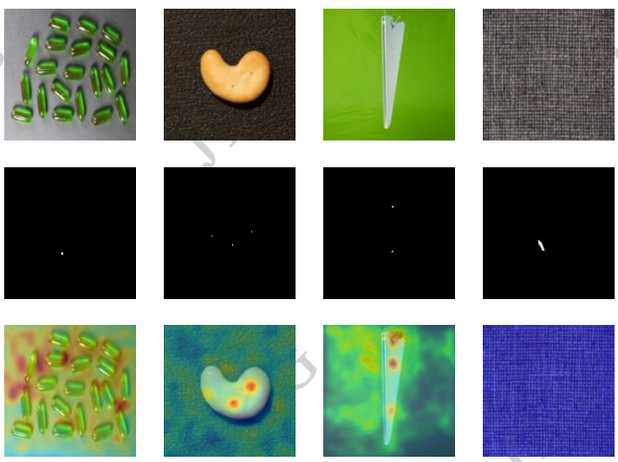
0 引 言

近年来,以卷积神经网络为代表的深度学习技术凭借其强大的特征自动学习能力,在图像分类、目标检测与语义分割等视觉任务中取得了突破性进展,并被广泛应用于工业品表面缺陷检测领域。

然而,相比于一般的目标检测任务,工业品表面缺陷检测面临着更为严峻的挑战。第一,缺陷样本匮乏。实际工业生产过程中严格把控次品率,相比于更易被获取的正常样本,缺陷样本的数量极少;同时,对缺陷进行精确标注的成本高昂(王帅,2021),远不能满足传统监督学习方法对数据量的要求。第二,细粒度缺陷的识别难题。裂纹、划痕、麻点等工业缺陷(常文娴,2022)通常极其微小且对比度低,容易受多变的异构背景干扰,呈现出显著的"类内差异大、类间差异小"的细粒度视觉表征特点(Em 等,2017)。这种特性使得常规模型难以提取到具有判别性的局部细节特征,导致其检测与定位精度大幅下滑,无法满足工业高精度质检的实际需求(吴建豪等,2017)。第三,多类别统一检测的泛化挑战。随着工业质检需求的升级,模型需要仅利用一个统一网络来同时处理多个类别的工业产品,这催生了多类别无监督异常检测(multi-class unsupervised anomaly detection, MUAD)这一极具挑战性的前沿范式。然而,在该设定下,模型极易陷入"捷径学习"的困境,当单一网络同时拟合庞大且复杂的跨类别正常特征空间时,不同产品的局部特征极易发生流形重叠,这使得解码器在处理异常样本时,容易错误地调用跨类别的正常先验知识,将异常区域完美重构。这种特征误用导致异常区域的重建残差趋近于零,大幅削弱了模型的判别能力并引发严重的漏检、误检现象。

现有方法的细粒度缺陷检测情况如图 1 所示。直观可见,主流方法在应对工业品细微瑕疵时存在明显局限。一方面,部分方法的热力图响应过于弥散,无法精准聚焦瑕疵点,边界定位十分模糊,导致微小缺陷的漏检。另一方面,部分方法容易受正常背景的干扰,在非缺陷区域错误产生异常高亮红斑,产生密集的误报。这种微小瑕疵易漏检、复杂背景易误报的现象,表明了现有方法在细粒度检测任务上的不足。

为应对缺陷样本匮乏的问题,基于正样本的无监督深度学习方法(Defard 等,2021)成为解决这一问题关键,但这类方法多是为每一类正样本图像单独训练一个模型,无法适应换产频繁的多类别工业场景。对于多类别统一检测,近年学术界涌现了一系列前沿的工作。早期的研究致力于通过干预底层特征的传递来限制异常重构。例如,UniAD(You 等,2022)利用邻域掩码注意力与特征抖动策略对输入特征进行信息隐藏与随机扰动,强制解码器调用正常先验知识来恢复局部掩码区域;OmniAL(Zhao 等,2023)则通过在特征中合成伪异常,强制破坏异常区域的恒等映射。HVQ-Trans(Lu 等,2023)引入向量量化机制以自发诱导特征差异,ReContrast(Guo 等,2023)则利用双编码器交叉重建来缓解特征重叠。这类方法虽然在抑制模型对异常的完美重构方面卓有成效,但其本质多局限于对局部特征的机械干预。在面对细粒度缺陷时,粗粒度的掩码遮挡或强制的量化压缩极易破坏原始图像的细节信息。同时,由于缺乏全局语义信息的指导,模型在复杂异构背景下难以精准定位微弱瑕疵。



(a) 误检 (b) 漏检 (c) 误检 (d) 漏检

((a)False detection;(b)Missed detection;(c)False detection;(d)Missed detection)

图 1 现有方法细粒度缺陷检测的误检与漏检示例

Fig. 1 Examples of false detections in existing methods for fine-grained defect detection

对于细粒度缺陷检测,早期模型多采用扩大感

© 中国图象图形学报版权所有

受野、多尺度特征融合及注意力机制等策略(Dong等,2022),但其缺乏高层语义指导限制了模型对细粒度缺陷检测性能的进一步提升。随着模型架构的演进,研究重点逐渐转向全局上下文建模与跨类特征的解耦。在生成式模型层面,LafitE(Yin等,2023)与DiAD(He等,2024)将扩散模型引入异常重建,并辅以语义引导机制,从而在去噪重构阶段有效维持了多类别图像的全局语义一致性。在纯视觉架构层面,统一异常检测方法MoEAD(Meng等,2024)引入了混合专家机制(mixture of experts, MoE),通过动态路由为不同类别的正常特征分配专属参数空间;MambaAD(He等,2024)则率先将状态空间模型引入多类别检测,以增强长距离上下文建模;ViTAD(Zhang等,2023)利用纯Transformer构建块,抽象出一个简洁高效的特征重建无监督异常检测框架。此外,AnomalyGPT(Gu等,2024)、IAD-GPT(Li等,2025)等基于大语言模型的新型方法,通过引入文本先验实现跨类别的语义对齐。这类方法虽然一定程度上强化了模型对多类别全局语义的认知,但在面对对比度极低的细粒度缺陷时,其表征与判别能力依然受限:MoEAD缺乏高层语义对底层特征的显式指导;MambaAD的全局序列扫描机制,以及LafitE、DiAD等扩散模型在隐空间中的去噪采样过程,往往会产生不可避免的特征平滑效应,导致底层像素级细节的模糊;而AnomalyGPT等大模型复杂的跨模态映射同样容易导致细粒度细节的丢失,微小的异常特征极易被庞大的参数网络淹没。

一些研究也探索了特征空间的细化策略。王素琴等人(2024)验证了频率域线索在增强背景与缺陷差异性方面的有效性,杨杰等人(2025)则通过引入特征扰动池增强了模型对多类别特征的判别鲁棒性。然而,频率分析方法在处理复杂异构纹理时容易引入非结构化背景噪声,而单纯的特征扰动策略虽然提升了鲁棒性,却缺乏深层语义与浅层细节的深度协同融合,导致模型在面对低对比度微小瑕疵时,依然效果受限。

综上所述,现有的多类别工业异常检测方法往往陷入“重语义而轻细节”或“重细节而轻语义”的单一化困境。要在统一网络下同时解决跨类别特征混淆与细粒度缺陷漏检这两个核心问题,关键在于融合高层抽象语义与浅层局部细节,实现两者的深度交互与优势互补。

为此,本文面向工业制造实际需求,以ViTAD为基准网络,提出了一种语义—细节协同融合的细粒度工业品缺陷检测方法(SDSF-ViTAD)。具体而言,本文贡献如下:1)提出了语义—细节协同融合模块(Semantic-Detail Synergistic Fusion Module, SDSFM)。该模块利用交叉注意力机制对齐并融合高层语义信息与浅层细节特征,并引入空间感知门控,在精准提取细粒度缺陷特征的同时,有效抑制复杂背景噪声的干扰。2)设计了一种面向解码重构的双重稀疏约束机制。其一,在特征融合模块之后引入基于信息瓶颈的特征扰动机制,通过随机丢弃部分融合激活值,强制解码器利用周围的全局上下文去推理而非复制缺失的信息,使其深度拟合正常样本的内在模式,缓解模型的捷径学习;其二,将原解码器中的标准稠密自注意力替换为Top-K稀疏注意力,限制当前Token仅与相似度最高的前K个Token进行信息交互,进一步缓解模型的捷径学习问题。3)提出了结合难分样本挖掘的余弦距离损失函数。在余弦距离损失的基础上引入难分样本挖掘机制,使模型在训练过程中重点关注难以拟合的正常区域,增强模型对细粒度异常的判别能力。

1 本文方法

本文提出的SDSF-ViTAD算法整体框架如图2所示,在训练阶段,模型仅使用正常样本。

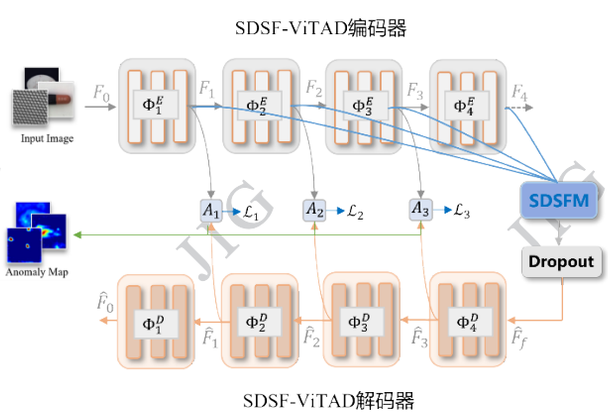


图2 SDSF-ViTAD算法网络原理图

Fig2 The schematic illustration of SDSF-ViTAD

整体流程可概括为以下几个核心环节:首先,采用基于DINO预训练的视觉Transformer(vision transformer, ViT)作为冻结编码器,提取输入图像的多层

© 中国图象图形学报版权所有

级密集特征先验。其次，设计语义—细节协同融合模块(SDSFM)，将编码器提取到的深层语义特征与浅层细节特征进行对齐与融合，生成富含细粒度信息的解码向量。接着，在融合模块之后引入基于信息瓶颈的特征扰动机制，通过随机丢弃部分激活值，强制解码器学习正常样本的内在流形，缓解模型“捷径学习”的问题。此外，解码器采用Top-K稀疏注意力替代标准稠密注意力，进一步压缩冗余信息传递路径。最后，采用结合难分样本挖掘的多层级余弦距离损失函数来指导网络优化，并计算编码器原始特征与解码器重建特征的差异，生成像素级异常得分图，实现缺陷的定位与评分。

1.1 语义—细节协同融合模块设计

在细粒度缺陷检测任务中，诸如划痕、微小斑点等缺陷往往表现为极低对比度的局部异常。ViT编码器提取的深层特征($\boldsymbol{F}_4$)虽然具备强大的全局感知和高级语义理解能力，但不可避免地丢失了对检测微小缺陷至关重要的细粒度细节，而较浅层特征($\boldsymbol{F}_1$, $\boldsymbol{F}_2$, $\boldsymbol{F}_3$)蕴含丰富的边缘和纹理信息。为了更充分地利用浅层细节信息，并实现其与深层语义的对齐，本文设计了语义—细节协同融合模块，该模块的核心思想是通过语义引导提取、动态门控过滤与特征内部强化三个递进步骤，生成高质量的解码向量，其内部结构如图3所示。

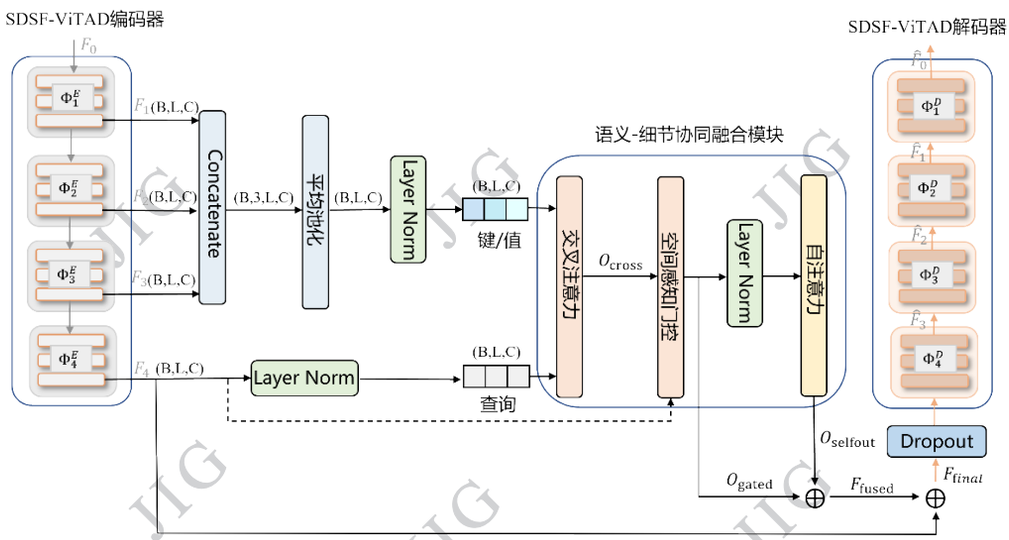


图3 语义—细节协同融合模块设计原理图

Fig3 Diagram of Semantic-Detail Synergistic Fusion Module Design

(1)基于交叉注意力的细粒度特征提取。首先，将编码器提取到的最具全局语义的最深层特征($\boldsymbol{F}_4$)作为查询向量$\boldsymbol{Q}$，旨在提供“正常模式应该是什么样”的语义指导。其次，将各浅层特征沿通道维度拼接，经过平均池化与层归一化后，构建键向量$\boldsymbol{K}$和值向量$\boldsymbol{V}$。通过计算$\boldsymbol{Q}$与$\boldsymbol{K}$之间的相似度，网络能够精准定位并提取出与深层语义相匹配的浅层细粒度特征，计算过程如式(1)所示：

$$\boldsymbol{O}_{\text{cross}} = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)V\#(1)$$

式中，d_k为键矩阵的特征维度。

(2)空间感知门控动态加权与去噪。在上述交叉注意力的提取过程中，若直接利用高层语义去查询浅层纹理容易造成语义错配，引入无关噪声。为此，本模块在交叉注意力之后引入了空间感知门控机制。该机制首先利用一维卷积对包含全局上下文信息的查询向量$\boldsymbol{Q}$进行空间特征编码，生成自适应的门控权重$\boldsymbol{G}$，计算过程如式(2)。随后，利用该权重对交叉注意力的输出$\boldsymbol{O}_{\text{cross}}$进行动态加权。这一操作严格控制了浅层信息的注入量，在保留有效缺陷细节的同时抑制噪声，计算过程如式(3)。

$$G = \sigma\left(\mathrm{Conv1D}\left(\boldsymbol{Q}\right)\right)\#(2)$$

式中，σ(·)表示Sigmoid激活函数，Conv1D表示一维卷积操作。

$$\boldsymbol{O}_{\text{gated}} = \boldsymbol{O}_{\text{cross}} \odot G\#(3)$$

式中，⊙指元素间的乘积。

(3)特征内部强化与残差聚合。最后，将加权后

© 中国图象图形学报版权所有

的特征送入自注意力层以强化关键信息。为保证网络训练的稳定性和特征梯度的平滑传递，模块引入了可学习的缩放参数γ_1和γ_2，构建了跨模块的残差连接。最终融合特征$\boldsymbol{F}_{\text{final}}$的计算过程如式(4)、(5)与(6)所示：

$$\boldsymbol{O}_{\text{selfout}} = SelfAttention\left(\text{LN}\left(\boldsymbol{O}_{\text{gated}}\right)\right)\#(4)$$

式中，SelfAttention(·)表示自注意力操作，LN(·)表示层归一化。

$$\boldsymbol{F}_{\text{fused}} = \boldsymbol{O}_{\text{gated}} + \gamma 1 \cdot \boldsymbol{O}_{\text{selfout}}\#(5)$$

$$\boldsymbol{F}_{\text{final}} = Q + \gamma 2 \cdot \boldsymbol{F}_{\text{fused}}\#(6)$$

1.2 基于信息瓶颈的特征扰动机制

在多类别无监督重建任务中，深层ViT解码器极易陷入捷径学习。为打破这种恒等映射，本文在语义—细节协同融合模块的输出端引入了基于信息瓶颈的特征扰动机制。设未受扰动前的融合特征为$\boldsymbol{F}_{\text{final}}$，则送入解码器的实际特征$\boldsymbol{F}_{\text{input}}$如式(7)所示：

$$\boldsymbol{F}_{\text{input}} = Dropout(\boldsymbol{F}_{\text{final}}, p)\#(7)$$

式中，Dropout(·)为随机丢弃操作，p为随机丢弃率。

该机制通过训练与测试阶段的行为差异，实现了对正常特征流形的强约束：

训练阶段(强制推理)：网络仅输入正常样本。Dropout层随机丢弃比例为p的激活单元，人为制造信息缺失。这切断了特征直接复制的捷径，迫使解码器利用未受损的全局上下文去推理并重构缺失的正常纹理，从而深度拟合正常样本的流形空间。

测试阶段(误差放大)：特征扰动关闭。当输入包含缺陷时，由于解码器仅掌握恢复正常模式的能力，无法对未见过的异常特征进行有效推理，从而在缺陷区域产生明显的重建偏差。

1.3 Top-K稀疏注意力重构机制

基准模型的解码器采用标准自注意力进行特征重建。然而，考虑到正常工业品图像具有强烈的局部结构规律性(如周期纹理、规则几何形状)，局部区域的特征表达应主要取决于空间邻域内具有结构相似性的少数关键Token，而非受全局冗余信息的干扰。

传统的稠密注意力矩阵，如图4(a)不仅引入了大量无效的远程依赖，还为解码器提供了过多无关的上下文信息，客观上加剧了模型的捷径学习问题：当模型处理缺陷区域时，它并不会尝试去理解图像正常的结构规律，而是倾向于在全局范围内搜索与当前缺陷区域高度相似的正常模式来“走捷径”的修复当前缺陷，使得模型即使不理解图像的内在逻辑，也能通过简单的信息复制实现极低的重构误差，从而导致异常检测失效。

为此，本文进一步探索了一种更具结构性的信息约束策略—Top-K稀疏注意力机制(Chen等人，2023)，并将其用于可训练的解码器中。标准的视觉自注意力机制中，查询矩阵$\boldsymbol{Q}$、键矩阵$\boldsymbol{K}$和值矩阵$\boldsymbol{V}$之间的注意力权重计算如式(8)：

$$Attention\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{\text{d}_{\text{k}}}}\right)V\#(8)$$

此时，注意力分数矩阵$\boldsymbol{A} = \frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{\text{d}_{\text{k}}}} \in \mathbf{R}^{N\times N}$是一个稠密矩阵($N$为序列长度)，意味着每个Token都会接收来自所有其他Token的信息。

Top-K稀疏注意力机制在计算softmax激活之前，对注意力分数矩阵$\boldsymbol{A}$施加了硬掩码操作。对于矩阵$\boldsymbol{A}$中的第i行(即第i个Token对所有Token的注意力分数)，记其前k个最大值所在的索引集合为$\boldsymbol{\Omega}_{\text{i}}$。构造一个掩码后的稀疏分数矩阵$\tilde{\boldsymbol{A}}$，其元素定义如式(9)所示：

$$\tilde{\boldsymbol{A}}_{i,j} = \begin{cases} \boldsymbol{A}_{i,j}, & if\, j \in \boldsymbol{\Omega}_{\text{i}} \\ -\infty, & otherwise \end{cases}\#(9)$$

式中，i和j分别表示注意力矩阵空间位置的行列索引，$\tilde{\boldsymbol{A}}_{i,j}$表示更新后的稀疏注意力分数，$\boldsymbol{A}_{i,j}$表示原始的注意力分数。

将未选中的注意力分数置为$-\infty$后，再经过softmax函数归一化，这些位置的权重将严格变为0。改进后的Top-K稀疏注意力输出如式(10)所示：

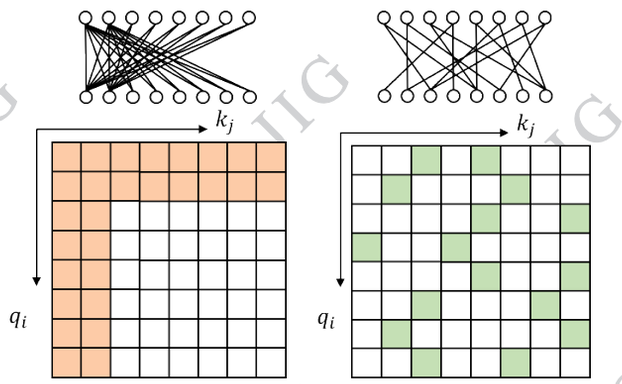
$$TopKAttention\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right) = softmax\left(\tilde{\boldsymbol{A}}\right)V\#(10)$$

通过该机制，当前Token被限制为仅与相似度最高的前k个Token进行信息交互，其余低相似度连接被直接截断。这种在解码器侧施加的稀疏约束，在不影响编码器特征提取能力的前提下，有效压缩了解码器的信息传递通路，使其专注于最相关的Token，以进行正常模式的推理与重构。

1.4 损失函数

在本文的无监督缺陷检测任务中，模型在训练阶段采用结合难分样本挖掘(Hard Normal Mining)的多层级余弦距离损失(Cosine Distance loss)作为模型优化的目标函数。相比于L_1或均方误差(Mean

© 中国图象图形学报版权所有



（a） 传统稠密注意力矩阵　（b） Top-k稀疏注意力矩阵

((a)Traditional dense attention matrix;(b)Top-k sparse attention matrix)

图4 注意力矩阵对比

Fig. 4 Attention matrix comparison

Squared Error, MSE)损失，余弦距离能够消除特征绝对幅值的干扰，强制网络专注于多层级特征在空间向量方向与语义表征上的一致性。在此基础上，通过引入难分样本挖掘机制，能够有效调节平滑背景与复杂纹理区域的梯度权重，引导解码器更精准地隐式拟合正常样本的多尺度特征分布。

具体而言，给定输入图像，设$\boldsymbol{F}_i$与$\hat{\boldsymbol{F}}_i$分别表示编码器提取的第i层原始特征与解码器重构出的对应层级特征。对于特征图上任意空间位置(h,w)，其对应的特征向量分别记为$\boldsymbol{F}_i(h,w)$与$\hat{\boldsymbol{F}}_i(h,w)$。第$i$层特征在位置$(h,w)$处的余弦距离如式(11)所示：

$$A_i(h,w)=1-\frac{\boldsymbol{F}_i(h,w)^{\mathrm{T}}\cdot\hat{\boldsymbol{F}}_i(h,w)}{\left\|\boldsymbol{F}_i(h,w)\right\|\left\|\hat{\boldsymbol{F}}_i(h,w)\right\|}\#(11)$$

式中，$A_i(h,w)$表示模型在当前像素位置的重构误差，其取值范围在0到1之间，i∈{1, 2, 3}表示编解码器中被约束的第3、6、9层特征，$\|\cdot\|$表示向量的L_2范数，公式右侧减号后面的部分为两特征向量的余弦相似度。

当重构效果较好时(如平滑背景等易分样本)，特征$\boldsymbol{F}_i(h,w)$与$\hat{\boldsymbol{F}}_i(h,w)$高度相似，余弦相似度趋近于1，重构误差$A_i(h,w)$接近于0；当重构效果较差时(如复杂边缘与纹理等难分样本)，特征差异较大，余弦相似度较小，重构误差$A_i(h,w)$接近于1，为了实现聚焦困难样本、优化结构不平衡正常模式的目标，本文对重构误差$A_i(h,w)$施加非线性指数放大，针对第i阶段特征，其引入难分样本挖掘后的重构损失L_i定义如式(12)所示：

$$L_i=\frac{1}{H_i\times W_i}\sum_{h=1}^{H_i}\sum_{w=1}^{W_i}(A_i(h,w))^{\gamma}\#(12)$$

式中，H_i和W_i分别表示第i层特征图的高和宽，γ为引入的聚焦参数(实验中设置为2)。

面对结构简单的易分背景区域，其重构损失较小，经过指数平方后，重构损失被进一步压缩，网络将减少在此类已学好区域的资源分配；而面对纹理交错的难分边界区域，其重构误差较大，经过指数运算后，相对于易分样本，其产生的梯度相对权重被放大，从而强制模型在训练阶段重点学习此类复杂的正常流形模式。

模型最终的总体损失函数L_{total}为所选用的多个层级重构损失之和，如式(13)所示：

$$L_{\text{total}}=\sum_{i=1}^{3}L_i\#(13)$$

通过最小化总体损失L_{total}，模型能够在正常样本的驱动下，深度学习并记忆多种工业品多尺度的正常模式先验。

2 实验与结果分析

2.1 实验数据集

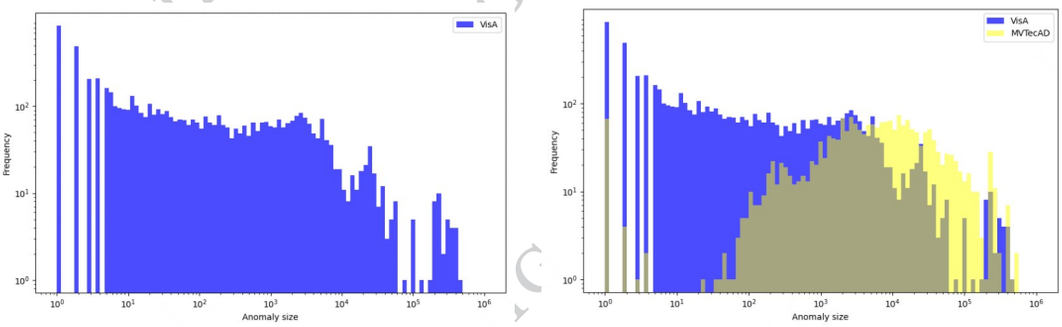
本文选用亚马逊云科技于2022年发布的VisA数据集作为实验基础。与早期的工业缺陷数据集相比，VisA数据集在缺陷的尺度以及目标的空间分布上均提出了更高的挑战，是目前评估细粒度异常检测算法性能的权威数据集之一。

VisA数据集共包含10821张高分辨率图像，其中正常样本9621张，异常样本1200张，涵盖了12个不同的子类别。根据图像中目标的物理结构与分布特点，这12个类别可划分为单实例图像、多实例图像和复杂结构图像。数据集中的异常图像包含各种类别的缺陷，包括划痕、凹痕、色斑或裂纹等表面缺陷，以及错位或缺失部件等结构缺陷。

为了直观量化VisA数据集的细粒度特性，本文遍历了该数据集中所有异常样本的真实掩码，提取了每个缺陷连通区域的像素面积，并绘制了缺陷面积的频率分布直方图。图中横坐标表示缺陷的像素面积，纵坐标表示对应面积缺陷出现的频数。为了更清晰地展现跨数量级的尺寸差异，横纵坐标均采

© 中国图象图形学报版权所有

用了对数尺度,结果如图5(a)。此外,引入了异常检测领域主流的MVTec AD(MVTec anomaly detection dataset)数据集进行对比,结果如图5(b)所示。



(a) VisA数据集缺陷像素面积频率分布　(b) VisA与MVTec AD数据集异常尺度分布对比

((a) distribution of defective pixel areas in the VisA; (b) comparison of anomaly scale distribution between VisA and MVTec AD)

图5 数据集对比

Fig. 5 Dataset Comparison Chart

从图5(b)中的对比结果可以清晰看出,VisA数据集的缺陷面积分布呈现典型的长尾效应,存在大量面积集中在10^0到10^3量级的微小缺陷。相比之下,MVTec AD数据集的缺陷面积则主要集中在10^3到10^5的区间,多表现为较大面积的异常。这一可视化结果表明,相比于传统的MVTec AD数据集,VisA数据集蕴含着更加丰富的微小异常特征。因此更适合作为本文细粒度工业缺陷检测任务的数据集。

2.2 实验环境配置

本文所有实验均在配置为32GB内存和单张NVIDIA GeForce RTX 4070 Super显卡的工作站上完成,算法网络基于PyTorch 2.1.2深度学习框架与CUDA 11.8加速库构建。在模型训练阶段,总训练轮数设置为100轮,初始学习率设定为$1\times e^{-4}$,并采用AdamW优化器进行网络参数的迭代更新,训练批量大小设定为8。

2.3 评价指标

为了全面评估模型在细粒度缺陷检测任务中的性能,本文分别从图像级别和像素级别开展定量分析。具体的评估指标包括:受试者工作特征曲线下面积(area under the receiver operating characteristic curve, AUROC)、平均精度(average precision, AP)以及F1分数(F1-score, F1)。此外,针对像素级缺陷定位,本文额外引入了区域重叠率曲线下面积(area under the per-region overlap, AUPRO)。

在实际工业场景中,正常背景与微小缺陷像素存在极端的不平衡。在上述指标中,AP能够比AUROC更客观地反映模型对极少数异常样本的检测精度,而AUPRO采用基于连通域(Region-based)的计算方式,对图像中的微小、细长缺陷极其敏感,能够最真实地验证本文算法对细粒度缺陷的精准定位能力。上述所有指标的取值范围均为0到100,数值越高代表模型性能越优。

2.4 对比实验结果及分析

将SDSF-ViTAD模型与当前无监督异常检测领域中最具代表性的6种算法进行综合对比实验。包括基于图像重构与伪影生成的DRAEM(Zavrtanik等,2021)、基于统一Transformer架构的多类别重构模型UniAD(You等,2022)、基于预训练特征提取与高斯噪声合成的SimpleNet(Liu等,2023)、结合知识蒸馏与特征分割的DeSTSeg(Zhang等,2023)、基于去噪扩散概率模型(Diffusion Model)的DiAD(He等,2024)以及基于状态空间模型的MambaAD(He等,2024),实验结果如表1所示。

从表1的数据可知,SDSF-ViTAD算法在整体检测性能上展现出明显优势。在图像级评价指标中,SDSF-ViTAD的mAUROC、mAP和mF1分别达到了93.87%、94.75%和89.89%,全面超越了所有对比算法,表明了本文算法对全局异常图像具有更精准的分类判别能力。在像素级评价指标中,SDSF-ViTAD同样展现了优秀的细粒度缺陷感知能力。其像素级mAUROC达到了最优的98.72%,在衡量细

© 中国图象图形学报版权所有

表1　不同模型在VisA数据集上的检测结果

Table 1　Detection results of different models on the VisA dataset

/%

方法/种类	图像级别			像素级别			
	mAUROC	mAP	mF1	mAUROC	mAP	mF1	mAUPRO
DRAEM	79.5	82.8	79.4	91.4	24.8	30.4	59.1
UniAD	88.8	90.8	85.8	98.3	33.7	39.0	85.5
SimpleNet	87.2	87.0	81.7	96.8	34.7	37.8	81.4
DeSTSeg	88.9	89.0	85.2	96.1	39.6	43.4	67.4
DiAD	86.8	88.3	85.1	96.0	26.1	33.0	75.2
MambaAD	93.69	94.10	89.35	98.52	39.19	43.55	**91.06**
ViTAD(baseline)	90.59	91.75	86.19	98.24	36.8	41.04	84.19
SDSF-ViTAD	**93.87**	**94.75**	**89.89**	**98.72**	**40.64**	**44.02**	87.99

粒度缺陷定位精度的mAUPRO指标上，本文算法达到了87.99%，较基准模型提升了3.8%，大幅领先于DiAD(75.2%)和SimpleNet(81.4%)，且优于性能强劲的UniAD(85.5%)。尽管MambaAD方法(91.06%)在该指标上高于本文算法，但其基于状态空间模型的序列扫描机制过度聚焦局部像素，在多类别异构背景下缺乏全局语义把控，易在复杂纹理区引入误检，导致其像素级mAP与mF1指标明显低于本文算法。

各模型在VisA数据集上的检测结果如图6所示，图中分别展示了包含气泡与反光干扰的胶囊、具有微小划痕的管材、存在微弱异色斑点的腰果与蜡烛，以及带有熔化物的印刷电路板(Printed Circuit Board, PCB)等缺陷图像。以上图像的检测存在两方面的挑战：一方面，划痕与异色斑点等缺陷特征微弱、有效面积较小，与正常背景的区分度较低；另一方面，胶囊的高频反光以及PCB板复杂的背景纹理，又与部分真实缺陷的局部特征高度相似。对于与背景十分相似的微小划痕和异色斑点缺陷，UniAD与SimpleNet等模型容易将微弱异常平滑化，均存在明显的漏检情况，而本文方法能够精准定位到该类缺陷。在胶囊气泡反光与PCB异物缺陷的检测上，MambaAD虽然对局部像素敏感，取得了部分正确的响应，但由于缺乏全局语义把控，往往会将正常的反光和复杂的交界纹理误判为异常，存在较多误检情况，而本文模型均实现了准确的定位，且检测边界更为收敛。综上所述，本文模型在微弱缺陷与复杂背景干扰缺陷的检测上有着更好的表现。

2.5　消融实验

为了进一步验证本文模型各项改进的有效性，本节按照递进关系，进行了一系列的消融实验。实验过程分为四个阶段：首先，在基准网络中引入语义—细节协同融合模块(SDSFM)，以验证该跨层特征融合机制在改善模型细粒度缺陷检测能力方面的作用；其次，在加入SDSFM的基础上，进一步引入基于信息瓶颈的特征扰动机制，探究其在缓解网络捷径学习问题、提升网络整体异常判别能力方面的实际效果，并分析不同随机丢弃率p对模型精度的具体影响；接着，在解码器内部替换Top-K稀疏注意力，以验证该内部结构性约束在进一步抑制捷径学习问题及提升细粒度缺陷检测精度方面的贡献，同时通过对比实验确定最佳的注意力保留数量k值；最后，在基准模型的基础上引入改进的损失函数，以验证该损失函数对模型最终检测精度的提升效果。实验结果表明，引入上述改进策略后，模型在各项指标上均取得了一定的性能提升。

2.5.1　SDSFM与特征扰动机制的有效性分析

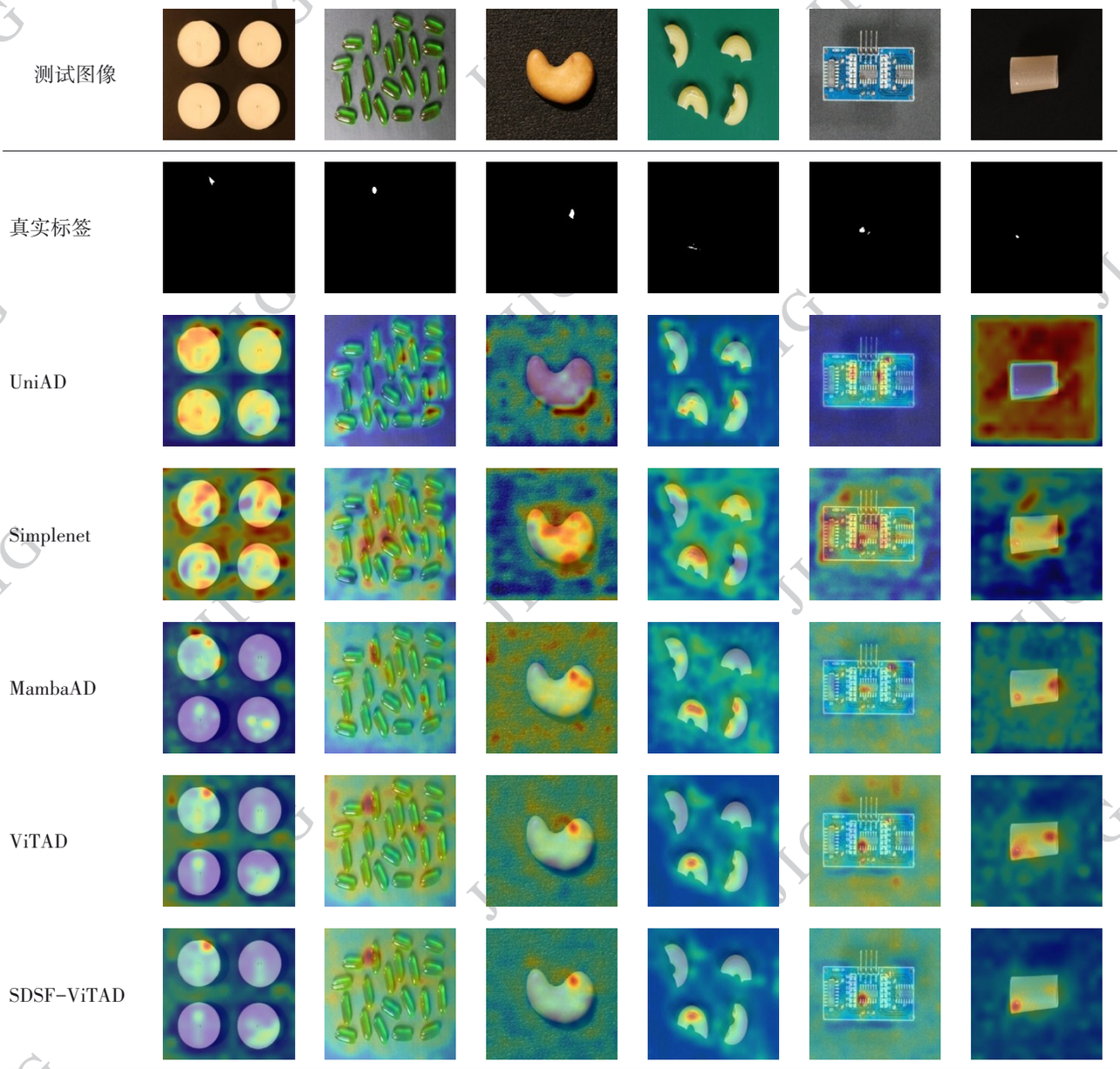
观察表2中的消融实验结果可知，逐步引入语义—细节协同融合模块与特征扰动机制对基准模型的性能产生了积极的影响。

对比表2的第1、2行数据可知，在基准模型中引入SDSFM后，模型在像素级检测指标上获得了较大提升。其中，反映细粒度定位精度的mAUPRO指标提升了2.79%，mAP和mF1也分别提升了0.76%和

© 中国图象图形学报版权所有

图6　VisA数据集上各方法检测结果的可视化图

Fig 6　Visualization of detection results of various methods on the VisA dataset



0.51%。这表明该模块通过融合浅层细节特征与深层语义信息，有效缓解了基准网络深层单尺度解码带来的局部信息丢失问题，增强了模型对细粒度缺陷的像素级别感知能力。

对比表2的第3~8行数据可知，在引入SDSFM的基础上，进一步引入特征扰动机制后，模型的图像级别检测指标得到了有效提升。当随机丢弃率p=0.5时，图像级评价指标mAUROC、mAP和mF1分别较基准模型提升了2.05%、1.97%和2.35%。这表明，特征扰动机制缓解了基准模型的捷径学习问题，模型学习到了正常样本的流形分布，能够对缺陷图像的语义信息理解的更好，从而提升了图像级检测指标的精度。

从表2中的p=0.1至p=0.6的变化趋势可以观察到，随着随机丢弃率p的增加，模型各项检测指标总体呈现先上升后轻微回落的趋势。当p值较小时，特征扰动强度不足，网络仍倾向于捷径学习；当p增加至0.5时，各项评价指标均达到峰值，表明此时的信息瓶颈约束达到了最优平衡；而当p进一步升至0.6时，模型性能开始出现轻微衰退，此时，过度的特征丢失破坏了正常语义的重建。因此，本文最终选取p=0.5作为该机制的最佳超参数。

© 中国图象图形学报版权所有

表2　消融实验结果对比

Table 2　Ablation Experiment Results Comparison

/%

	p值	图像级别			像素级别			
		mAUROC	mAP	mF1	mAUROC	mAP	mF1	mAUPRO
ViTAD	/	90.59	91.75	86.19	98.24	36.8	41.04	84.19
+SDSFM	/	90.60	91.38	86.85	98.5	37.56	41.55	86.98
+SDSFM+Dropout	p=0.1	91.64	92.92	87.46	98.37	38.32	42.10	85.88
	p=0.2	92.02	93.18	87.72	98.45	38.94	42.71	85.97
	p=0.3	92.39	93.41	88.29	98.50	39.18	42.95	86.51
	p=0.4	92.58	93.68	88.41	98.53	39.67	43.19	86.94
	p=0.5	**92.64**	**93.70**	**88.54**	**98.57**	**39.86**	**43.26**	**87.08**
	p=0.6	92.58	93.67	88.36	98.59	39.47	43.01	87.06

2.5.2　Top-K稀疏注意力的有效性分析

稀疏注意力的核心超参数k为保留的Token数量，其直接决定了信息瓶颈的强弱程度。为探究最优配置，本文开展了k∈{32, 64, 128, 196, 256}的消融实验，结果如表3所示。

观察表3中的结果可知，随着k值的增大，模型性能呈现出明显的先升后降趋势。当k值过小时，过强的信息约束导致解码器获取的上下文不足，无法有效重建正常特征，造成性能下降；而当取值偏大时，过多的交互连接引入了冗余信息并削弱了瓶颈约束，使得模型性能进一步下降；当k=256时，该机制则完全退化为无约束的标准稠密注意力基线水平。相比之下，适度稀疏实现最优平衡，当k=64时，模型图像与像素级的七项指标实现全面提升，综合性能达到最优。解码器的每个Token保留了最相关的64个交互连接，在充分压缩冗余注意力路径的同时，保留了足够的上下文信息用于正常模式的重建，有效缓解捷径学习而不损失重建能力。基于此，本文最终选取k=64作为该机制的最优超参数。

2.5.3　改进损失函数的有效性分析

本文最后验证了结合难分样本挖掘的多层级余弦重构损失的实际贡献，结果如表4所示。数据表明，引入该损失机制后，模型的各项检测指标均得到了提升，进一步证明了该优化策略的有效性。

表3　Top-K稀疏注意力k值消融实验结果

Table 3　Ablation experiment results of Top-K sparse attention k value

/%

	k值	图像级别			像素级别			
		mAUROC	mAP	mF1	mAUROC	mAP	mF1	mAUPRO
ViTAD	256	90.59	91.75	86.19	98.24	36.8	41.04	84.19
+Top-K	32	90.43	91.64	86.03	**98.26**	36.86	41.04	84.91
	64	**90.69**	**91.90**	86.39	98.25	**37.25**	**41.58**	84.56
	128	90.60	91.43	86.30	98.25	36.64	40.76	84.73
	196	90.58	91.66	**86.44**	98.25	36.45	40.84	**84.97**

© 中国图象图形学报版权所有

表4　引入基于难分样本挖掘的余弦重构损失实验结果对比

Table 4　Introduction of Experimental Result Comparison of Cosine Reconstruction Loss Based on Hard Sample Mining

/%

	图像级别			像素级别			
	mAUROC	mAP	mF1	mAUROC	mAP	mF1	mAUPRO
ViTAD	90.59	91.75	86.19	98.24	36.8	41.04	84.19
+改进损失函数	91.33	92.31	87.18	98.44	37.63	41.83	85.65

2.6　模型性能分析

为了验证本文方法在复杂工业场景下的可落地性与部署潜力，我们对模型的核心性能指标进行了评估，结果如表5所示。

表5　模型性能分析

Table 5　Model Performance Analysis

方法	参数量/M	计算量/G	Latency/ms	VRAM/M
ViTAD	38.586	9.668	3.12	222.13
SDSF-ViTAD	40.067	10.6	4.24	263.37

测试结果显示，本文提出的SDSF-ViTAD模型总参数量为40.067M，计算量为10.6G。相较于基准模型ViTAD（参数量为38.586M，计算量为9.668G），本文模型在改进各模块后，参数量与计算量的增幅分别仅为1.48M和0.93G，整体架构依然属于中等规模。

本文模型在真实物理环境下的单张图像推理延迟（Latency）为4.24毫秒，峰值显存占用（video random access memory，VRAM）为263.37MB。这一数据表明，该模型能够满足工业高速流水线的实时节拍要求（单次检测<50ms），并具备在算力与空间受限的边缘工控设备上进行部署的潜力。总体而言，SDSF-ViTAD算法在提升细粒度缺陷检测精度的同时，较好地兼顾了工业级应用的轻量化与吞吐率需求。

2.7　在其他数据集上的应用

2.7.1　金属零件缺陷数据集实验

为验证本文方法在细粒度工业品缺陷检测任务上的通用性与鲁棒性，本节采用金属零件缺陷公开数据集（Metal Parts Defect Detection，MPDD）进行对比实验。MPDD数据集共包含1346幅图像，其中训练集包含888幅正常图像，测试集包含176幅正常图像与282幅异常图像。该数据集中涵盖了黑色支架（Bracket Black）、棕色支架（Bracket Brown）、白色支架（Bracket White）、连接器（Connector）、金属板（Metal Plate）以及管材（Tubes）六类金属部件图像，其异常类型主要包括划痕、污渍、边缘缺失、变形及凹陷。

从表6的结果得知，SDSF-ViTAD模型在MPDD数据集上取得了较好的检测性能。在图像级指标上，mAUROC达到90.23%，略高于MambaAD（90.21%），并明显优于基准模型ViTAD（87.73%），在mAP和mF1指标上，本文方法分别达到93.86%和91.14%，虽略低于MambaAD（94.38%和91.71%），但仍大幅领先于其他对比方法。在像素级指标上，SDSF-ViTAD展现出更强的细粒度定位能力：其mAUROC达到98.03%，为所有方法中最高，mAP和mF1分别达到41.08%和42.22%，显著优于MambaAD（33.48%和39.25%）和ViTAD（35.35%和37.55%），最能反映细粒度定位精度的mAUPRO指标达到93.24%，同样优于所有对比方法。

考虑到本文的研究重点为细粒度缺陷检测，本节选取了MPDD数据集中缺陷尺寸较小、且易受背景干扰的两类缺陷图像进行可视化分析。从图7的异常热力图对比结果可以看出，在面对微弱的点状瑕疵与极细划痕时，UniAD倾向于在正常物体边缘产生响应并导致真实的缺陷漏检；Simplenet检测则受背景干扰严重；而MambaAD与ViTAD存在一定的误检现象，且检测定位结果与真实缺陷掩码形状不符。相比之下，本文提出的SDSF-ViTAD方法展现出了最优的检测效果，其不仅有效减少了误检与漏检现象，而且对划痕缺陷的定位更加精准，生成的高响应区域更加聚集。

2.7.2　BeanTech异常检测数据集实验

本节进一步采用BeanTech异常检测（BeanTech Anomaly Detection，BTAD）公开数据集进行对比实

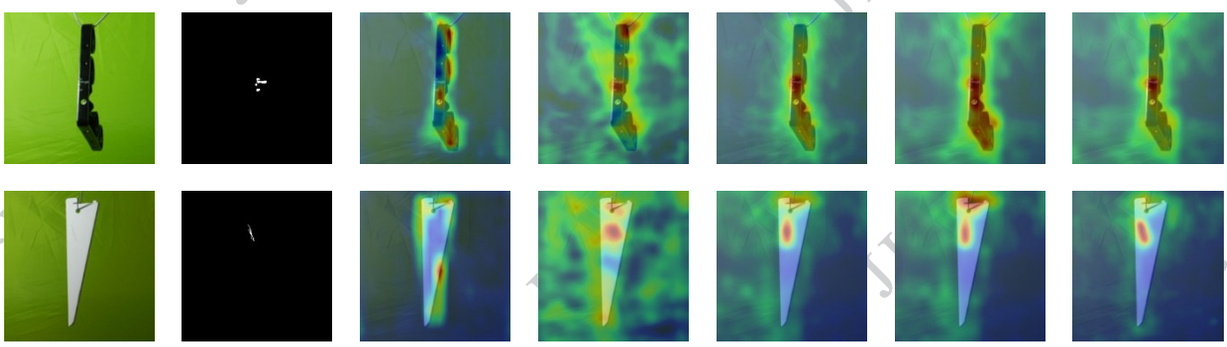
© 中国图象图形学报版权所有

表6　不同模型在MPDD数据集上的测试结果

Table 6　Test results of different models on the MPDD dataset

/%

方法/种类	图像级别			像素级别			
	mAUROC	mAP	mF1	mAUROC	mAP	mF1	mAUPRO
UniAD	50.61	61.54	73.56	88.24	8.85	13.43	59.85
SimpleNet	88.39	92.84	87.71	94.98	27.96	31.41	86.85
MambaAD	90.21	**94.38**	**91.71**	97.68	33.48	39.25	93.03
ViTAD(baseline)	87.73	90.22	87.97	97.7	35.35	37.55	92.57
SDSF-ViTAD	**90.23**	93.86	91.14	**98.03**	**41.08**	**42.22**	**93.24**



（a）测试图像　（b）真实标签　（c）UniAD　（d）Simplenet　（e）MambaAD　（f）ViTAD　（g）SDSF-ViTAD

((a) test image;(b) ground truth;(c)UniAD;(d)Simplenet;(e)MambaAD;(f)ViTAD;(g)SDSF-ViTAD)

图7　MPDD数据集上各方法检测结果的可视化图

Fig.7　Visualization of detection results of various methods on the MPDD dataset

验。BTAD数据集共包含2540幅图像，其中训练集包含1799幅正常图像，测试集包含451幅正常图像与290幅异常图像。该数据集涵盖了三类工业产品图像（产品01、产品02、产品03），其异常类型主要包括划痕、凹痕、裂纹、污染物及缺失。

表7展示了各模型在BTAD数据集上的测试结果。在像素级指标上，本文模型的mAP达到62.81%，较基准模型提升了3.49%，同时，mAUPRO（77.58%）与mF1（59.56%）也均高于其他对比方法，表明模型对细粒度缺陷的定位精度有所改善。在图像级指标方面，本文模型的mAP和mF1分别为97.52%和93.50%，高于其他对比模型。尽管图像级mAUROC（95.89%）较基准模型（95.95%）降低了0.06%，但综合整体数据可见，本文方法在保持图像级分类能力的基础上，有效提升了像素级的异常定位能力。

此外，本节在BTAD数据集中选取了背景纹理复杂且缺陷尺寸微小的两类工业产品进行可视化分析。其中，第一类产品的缺陷为微小的点状污渍，第二类产品的缺陷为极细的横向划痕。这类缺陷的共同特征是与复杂背景的对比度低，且极易被工件本身的周期性结构（如同心圆环、密集条纹）所掩盖。可视化结果如图8所示。分析结果可知，面对复杂的背景纹理，UniAD、Simplenet以及MambaAD均受到了严重的干扰，在正常的纹理区域（如正常条纹和圆环边缘）产生了大量的异常响应，未能有效聚焦于真实的微弱缺陷。基准模型ViTAD的热力图同样存在响应弥散和背景噪声。相比之下，本文模型能够有效抑制复杂背景纹理的干扰，异常热力图背景更加干净，同时，准确地定位到了微小的点状缺陷与细微划痕，其高响应区域与真实标签图像保持了较好的一致性。

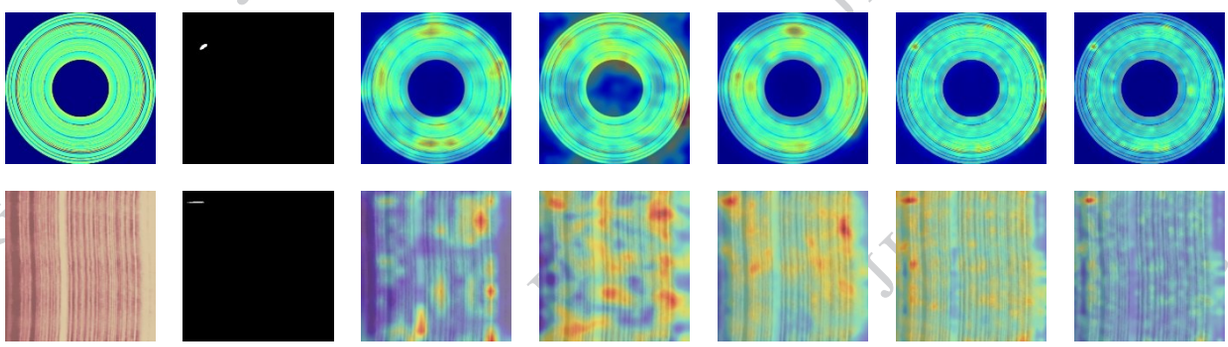
© 中国图象图形学报版权所有

表7　不同模型在BTAD数据集上的测试结果

Table 7　Test results of different models on the BTAD dataset

/%

方法/种类	图像级别			像素级别			
	mAUROC	mAP	mF1	mAUROC	mAP	mF1	mAUPRO
UniAD	66.12	64.00	65.18	76.80	14.13	20.73	33.6
SimpleNet	93.12	97.49	93.30	96.47	41.59	45.44	71.30
MambaAD	91.91	96.35	93.05	97.57	52.60	55.16	77.52
ViTAD(baseline)	**95.95**	96.58	92.48	97.70	59.32	58.17	75.59
SDSF-ViTAD	95.89	**97.52**	**93.50**	**97.92**	**62.81**	**59.56**	**77.58**



（a）测试图像　（b）真实标签　（c）UniAD　（d）Simplenet　（e）MambaAD　（f）ViTAD　（g）SDSF-ViTAD

((a) test image; (b) ground truth; (c)UniAD; (d)Simplenet; (e)MambaAD; (f)ViTAD; (g)SDSF-ViTAD)

图8　BTAD数据集上各方法检测结果的可视化图

Fig 8　Visualization of detection results of various methods on the BTAD dataset

3　结　论

针对多类别无监督场景下细粒度缺陷易被漏检且模型易陷入“捷径学习”的问题，本文提出了一种语义—细节协同融合的工业品缺陷检测方法SDSF-ViTAD。该方法通过构建语义—细节协同融合模块，在深层全局语义的引导下对齐并融合浅层细节特征，增强了模型对微弱缺陷的感知与抗背景干扰能力。同时，在解码器重构阶段引入基于信息瓶颈的特征扰动机制与Top-K稀疏注意力，并结合难分样本挖掘的余弦重构损失，有效切断了特征直接复制的通路，迫使模型深度拟合正常流形，缓解了模型的“捷径学习”现象。在VisA、MPDD和BTAD数据集上的实验结果表明，本文方法在各项图像级与像素级指标上均取得了较好的综合表现，且有效减少了细粒度缺陷的漏检与误检。模型参数量与推理延迟适中，具备良好的工业部署潜力。

但是，当前方法依赖固定的随机丢弃率与稀疏度超参数，面对不同纹理复杂度的工业产品时适应性有限。今后我们将探索轻量化网络设计与自适应特征扰动机制，以提升模型在跨领域、多类别工业场景中的泛化能力与实时检测性能。

参考文献(References)

Chang W X. 2022. Metal complex surface defect detection technology based on machine vision. Xi′an: Xi′an Technological University (常文娴. 2022. 基于机器视觉的金属复杂表面缺陷检测技术. 西安: 西安工业大学)

Defard T, Setkov A, Loesch A, et al. 2021. PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection and Localization. 25th International Conference on Pattern Recognition, 12664: 475-489

Dong H W, Song K C, He Y, et al. 2020. PGA-Net: Pyramid Feature

© 中国图象图形学报版权所有

Fusion and Global Context Attention Network for Automated Surface Defect Detection. IEEE Transaction on Industrial Informatics, 2020, 16(12):7448-7458

Em Y, Gag F, Lou Y, Wang S, Huang T and Duan L Y. 2017. Incorporating intra-class variance to fine-grained visual recognition//Proceedings of the 2017 IEEE International Conference on Multimedia and Expo. Hong Kong: IEEE: 1452-1457

Gu Z P, Zhu B K, Zhu G, Chen Y Y, Tang M and Wang J Q. 2024. Anomalygpt: Detecting industrial anomalies using large vision-language models//Proceedings of the AAAI Conference on Artificial Intelligence: 1932-1940

Guo J, Lu S, Jia L, Zhang W and Li H. 2023. Recontrast: domain specific anomaly detection via contrastive reconstruction//Advances in Neural Information Processing Systems: 10721-10740

He H Y, Bai Y, Zhang J N, He Q, Chen H, Gan Z, et al. 2024. Mambaad: exploring state space models for multi-class unsupervised anomaly detection[EB/OL]. [2026-05-08]. https://arxiv.org/pdf/2404.06564.pdf

He H Y, Zhang J N, Chen H X, Chen X H, Li Z S, Chen X, et al. 2023. Diad: A diffusion-based framework for multi-class anomaly detection[EB/OL]. [2026-05-08]. https://arxiv.org/pdf/2312.06607.pdf

Li Z W, Yu Z T, Ye Q L, Xie W C, Zhuo W and Shen L L. 2025. IAD-GPT: Advancing visual knowledge in multimodal large language model for industrial anomaly detection. IEEE Transactions on Instrumentation and Measurement, 74: 1-12

Liu Z K, Zhou Y M, Xu Y S and Wang Z L. 2023. Simplenet: A simple network for image anomaly detection and localization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE: 20402-20411

Lu R Y, Wu Y J, Tian L, Wang D S, Chen B, Liu X Y, et al. 2023. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection[EB/OL]. [2026-05-08]. https://arxiv.org/pdf/2310.14228.pdf

Luo J H and Wu J X. 2017. A survey on fine-grained image classification based on deep convolutional features. Acta Automatica Sinica, 43(8): 1306-1318 (罗建豪, 吴建鑫. 2017. 基于深度卷积特征的细粒度图像分类研究综述. 自动化学报, 43(8): 1306-1318)

Meng S Y, Meng W C, Zhou Q H, Li S Z, Hou W Y and He S B. 2024. Moead: A parameter-efficient model for multi-class anomaly detection//European Conference on Computer Vision. Cham: Springer Nature Switzerland: 345-361

Wang S. 2021. Research on key algorithms of product surface defect detection based on machine vision. Shenyang: University of Chinese Academy of Sciences (Shenyang Institute of Computing Technology, Chinese Academy of Sciences) (王帅. 2021. 基于机器视觉的产品表面缺陷检测关键算法研究. 沈阳: 中国科学院大学(中国科学院沈阳计算技术研究所))

Wang S Q, Cheng C, Shi M and Zhu D M. 2024. Defect detection method for industrial product surfaces with similar features by combining frequency and ViT. Journal of Image and Graphics, 29(10): 3074-3089 (王素琴, 程成, 石敏, 朱登明. 2024. 结合频率和ViT的工业产品表面相似特征缺陷检测方法. 中国图象图形学报, 29(10): 3074-3089)

Yang J, Hu W J and Zang Y. 2025. Feature disturbance pool-based fusion mechanism for multi-class industrial defect detection. Journal of Image and Graphics, 30(5): 1404-1418 (杨杰, 胡文军, 臧影. 2025. 特征扰动池融合机制的多类工业缺陷检测. 中国图象图形学报, 30(5): 1404-1418)

Yin H N, Jiao G L, Wu Q H, Karlsson B F, Huang B Q and Lin C Y. 2023. Lafite: Latent diffusion model with feature editing for unsupervised multi-class anomaly detection[EB/OL]. [2026-05-08]. https://arxiv.org/pdf/2307.08059.pdf

You Z, Cui L, Shen Y, Yang K, Lu X, Zheng Y, et al. 2022. A unified model for multi-class anomaly detection. Advances in Neural Information Processing Systems, 35: 4571-4584

Zavrtanik V, Kristan M and Skočaj D. 2021. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 8330-8339

Zhang J N, Chen X H, Wang Y B, Wang C J, Liu Y, Li X T, et al. 2023. Exploring plain vit reconstruction for multi-class unsupervised anomaly detection[EB/OL]. [2026-05-08]. https://arxiv.org/pdf/2312.07495.pdf

Zhang X, Li S, Li X, Huang P, Shan J L and Chen T. 2023. Destseg: Segmentation guided denoising student-teacher for anomaly detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE: 3914-3923

Zhao Y. 2023. Omnial: a unified cnn framework for unsupervised anomaly localization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE: 3924-3933

作者简介

赵明明，女，硕士研究生，主要研究方向为模式识别与智能信息处理。E-mail:18281539967@163.com

张二虎，通信作者，男，教授，主要研究方向为数字图像处理、模式识别与机器学习。 E-mail:eh-zhang@xaut.edu.cn

© 中国图象图形学报版权所有