

本文引用格式: 陈永彬,黄嘉思,黄国健.多尺度自监督特征融合的工业视觉缺陷检测方法[J].自动化与信息工程,2026,47(4):17-25.

CHEN Yongbin, HUANG Jiasi, HUANG Guojian. Multi-scale self-supervised feature fusion method for industrial visual defect detection[J]. Automation & Information Engineering, 2026,47(4):17-25.

## 多尺度自监督特征融合的工业视觉缺陷检测方法\*

陈永彬 黄嘉思 黄国健

(广东机电职业技术学院, 广东 广州 510550)

**摘要:** 针对产品品类切换后视觉缺陷检测模型需重新训练与调整参数的问题, 提出一种多尺度自监督特征融合的工业视觉缺陷检测方法。该方法以冻结的自监督视觉主干网络通过多 Hook 并行抓取浅、中、深 3 层 patch token 特征, 分层构建核心子集记忆库; 通过等权得分级融合机制将各层最近邻距离合成统一的异常图像, 从而避免模型逐类调整参数与重新训练。在 MVTec-AD 数据集全部 15 类、3 个随机种子下, 图像级 AUROC 达  $0.9874 \pm 0.0017$ , 跨种子标准差仅为同主干基线的 11%~20%, 区域级 PRO 达  $0.8163 \pm 0.0121$ 。归因消融实验结果表明: 跨种子标准差减小的主要原因是多层 Hook 的等权得分级融合机制, 全部 15 类上多层相对单层 L9 的标准差降幅约为 45%。零样本迁移至 VisA 数据集全部 12 类时, P-AUROC 达 0.9929; Real-IAD 数据集前 6 类子集的 P-AUROC 达 0.9847。在自建的发动机缸套缺陷图像数据集上, P-AUROC 达  $0.9673 \pm 0.0015$ 。在半精度模式下单图推理耗时约为 25 ms, 可满足一般中速生产线的节拍要求。

**关键词:** 工业视觉; 缺陷检测; 自监督学习; 多尺度特征; 零样本迁移

中图分类号: TP391.41

文献标志码: A

文章编号: 1674-2605(2026)04-0003-09

DOI: 10.12475/aie.20260403

开放获取

## Multi-scale Self-supervised Feature Fusion Method for Industrial Visual Defect Detection

CHEN Yongbin HUANG Jiasi HUANG Guojian

(Guangdong Mechanical and Electrical Polytechnic, Guangzhou 510550, China)

**Abstract:** Aiming at the problem that visual defect detection models require retraining and parameter tuning after product category switching, we propose an industrial visual defect detection method based on multi-scale self-supervised feature fusion. The method employs a frozen self-supervised visual backbone network to extract shallow, middle, and deep patch token features in parallel via multiple hooks, and hierarchically constructs a core subset memory bank for each layer. It then fuses the nearest neighbor distances from all layers into a unified anomaly map using an equal-weight score-level fusion mechanism, thereby avoiding category-wise parameter tuning and retraining. On all 15 categories of the MVTec-AD dataset under 3 random seeds, the image-level AUROC reaches  $0.9874 \pm 0.0017$ , with a cross-seed standard deviation that is only 11%~20% of that of the same backbone baseline, while the region-level PRO achieves  $0.8163 \pm 0.0121$ . Ablation study results attribute the reduction in cross-seed standard deviation primarily to the equal-weight score-level fusion mechanism of multiple hooks; the standard deviation reduction of multi-layer compared to single-layer L9 across all 15 categories is approximately 45%. When zero-shot transferred to all 12 categories of the VisA dataset, the P-AUROC reaches 0.9929; on the first 6 categories of the Real-IAD dataset, the P-AUROC achieves 0.9847. On a self-collected engine cylinder liner defect image dataset, the P-AUROC reaches  $0.9673 \pm 0.0015$ . In half-precision mode, the inference time per image is about 25 ms, which can meet the takt time requirements of typical medium-speed production lines.

**Keywords:** industrial visual; anomaly detection; self-supervised learning; multi-scale features; zero-shot transfer

## 0 引言

工业视觉缺陷检测是制造业质量控制的核心环节,广泛应用于汽车零部件、3C 电子、纺织品、食品包装等行业。当前主流的工业视觉缺陷检测系统以 HALCON、Cognex VisionPro、Sherlock 等规则式商业软件为主,采用边缘检测、形态学处理、Blob 分析、模板匹配等经典图像处理算法对每类缺陷手工调整参数。然而,上述工业视觉缺陷检测系统存在以下问题:1) 参数与成像条件、缺陷形态高度耦合,导致跨品类不可迁移;2) 单类缺陷检测需要调整多个参数,专家编程成本较高;3) 无法识别未见过的缺陷形态,需补充相应检测规则

近年来,基于深度学习的异常检测方法发展迅速。PatchCore<sup>[1]</sup>结合预训练卷积神经网络特征与核心子集采样记忆库,在 MVTec-AD 数据集上的图像级受试者工作特征曲线下面积 (image-level area under the receiver operating characteristic curve, I-AUROC) 接近 99%。WinCLIP<sup>[2]</sup>、AnomalyCLIP<sup>[3]</sup>、MVFA<sup>[4]</sup>等模型借助对比语言-图像预训练 (contrastive language-image pre-training, CLIP) 跨模态特征,可实现零样本或少样本检测。DINOv2<sup>[5]</sup>等自监督视觉模型在异常检测任务中表现出较强的泛化能力。主流基于深度学习的异常检测方法虽然检测精度较高,但大多数方法仍依赖“按类训练、按类调参”的原则,并未缓解跨品类切换的部署成本。

针对上述问题,本文以跨品类、零样本迁移、免下游训练为目标,提出一种多尺度自监督特征融合的工业视觉缺陷检测方法。该方法以冻结的自监督视觉主干网络 (DINOv2-ViT-B/14) 为特征提取器,通过多 Hook 并行抓取浅、中、深 3 层 patch token 特征;采用等权得分级融合机制,使同一套超参数可在多组工业数据集上直接复用;各层核心子集采样均分配独立的确定性种子,以保障跨进程、跨设备的数值一致性。

## 1 相关工作

### 1.1 基于预训练特征的视觉缺陷检测

PatchCore 提出“块特征-核心子集-最近邻打分”

的范式,其后续工作沿不同方向展开:PaDiM<sup>[6]</sup>以马氏距离建模块特征分布;DRAEM<sup>[7]</sup>通过合成异常和重建网络进行自监督训练;SimpleNet<sup>[8]</sup>在特征上增加判别器与噪声扰动;CFA<sup>[9]</sup>利用耦合超球进行目标自适应;GraphCore<sup>[10]</sup>在少样本下引入图卷积与旋转不变特征。上述方法均依赖正常样本建立参考分布或记忆库。本文沿用 PatchCore 的核心子集与最近邻框架,并将其主干网络从 ResNet 转换为 DINOv2-ViT-B/14,从而将多尺度特征融合方式由特征拼接调整为得分级融合,同时每层独立维护核心子集记忆库并分配确定性采样种子。

### 1.2 跨模态、自监督与多尺度特征聚合

跨模态、零样本或少样本检测方法将 CLIP 等模型的语义引入视觉缺陷检测领域:WinCLIP 通过滑窗与提示集成实现零样本检测;AnomalyCLIP 学习对象无关的可调文本提示并进行对比训练。MVFA 在冻结的 CLIP 视觉编码器多个层级上分别插入轻量级特征适配器,对浅层至深层特征进行逐级适配,在少样本条件下实现跨域缺陷检测。该类方法支持开放词表与少样本检测,但检测精度对提示模板、文本词集敏感。本文不依赖 CLIP 视觉特征,仅利用 CLIP 文本通路进行缺陷类型归因,将文本路径解耦为后处理模块,避免提示工程对检测精度的影响。

在视觉主干网络中,本文以冻结的 DINOv2-ViT-B/14<sup>[5]</sup>为唯一特征提取源。其中,DINOv2 各层均保留 patch 级信息,便于多 Hook 抓取;其自蒸馏目标使特征对局部异质性敏感,与视觉缺陷检测需求契合,且公开权重稳定,便于复现;与 MAE<sup>[11]</sup>、DINO<sup>[12]</sup>等其他自监督视觉模型相比,DINOv2 在 ViT 上的 patch token 多层迁移性更优。

## 2 MultiHook-AD 缺陷检测框架

### 2.1 整体架构

本文提出的多尺度自监督特征融合的工业视觉缺陷检测方法实现框架为 MultiHook-AD (multi-hook anomaly detection)。该框架由多层视觉特征提取、核心子集记忆库、多尺度最近邻打分与融合、后处理、

文本探针归因(与检测通路解耦)5个模块串联组成, 其整体架构及推理流程如图1所示。

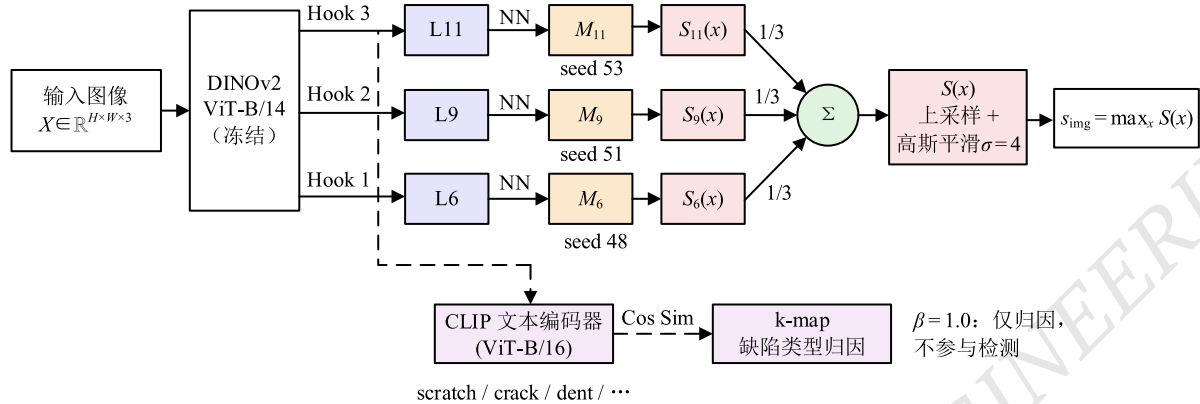


图1 MultiHook-AD 缺陷检测框架整体架构及推理流程

1) 多层视觉特征提取模块。输入图像 $X \in \mathbb{R}^{H \times W \times 3}$ 经冻结的 DINOv2-ViT-B/14 视觉主干网络完成单次前向计算, 并通过多 Hook 机制同时抓取 L6、L9、L11 层的 patch token 特征, 每层输出  $32 \times 32$  个 768 维特征  $f_l(x)$ 。

2) 核心子集记忆库模块。利用贪心核心子集对每层正常样本特征进行采样(采样比例固定为 0.01), 构建 3 个独立的记忆库  $M_6$ 、 $M_9$ 、 $M_{11}$ , 每个记忆库约含 500 个块特征。

3) 多尺度最近邻打分与融合模块。通过每层特征  $f_l(x)$  与对应记忆库  $M_l$  计算最近邻距离, 作为单层得分图:

$$s_l(x) = \min_{m \in M_l} \|f_l(x) - m\|_2, \quad l \in \{6, 9, 11\}$$

各层得分图按等权( $w_l = 1/3$ )进行得分级融合:

$$s(x) = \sum_{l \in \{6, 9, 11\}} w_l \cdot s_l(x), \quad w_l = 1/3$$

4) 后处理模块。融合后的得分图  $S(x)$  经双线性上采样至原图尺寸, 并经高斯平滑 ( $\sigma=4$ ) 后, 可得到像素级异常图像。图像级分数取得分图在所有空间位置上的最大值:

$$s_{\text{img}} = \max_x s(x)$$

5) 文本探针归因模块(与检测通路解耦)。CLIP

文本编码器 (ViT-B/16) 对用户指定的缺陷词集  $W = \{\text{scratch}, \text{crack}, \text{dent}, \dots\}$  进行编码, 并与 L11 层的块特征计算相似度, 仅用于生成 k-map 缺陷类型可视化, 不参与检测。

## 2.2 多尺度选层

本文通过层组合消融实验讨论选层的合理性。在 MVTec-AD 数据集全部 15 类、3 个随机种子 (SEED  $\in \{42, 7, 123\}$ ) 上比较 7 种层组合, 各层组合均采用等权得分级融合机制, 层组合消融实验结果如表 1 所示。其中, P-AUROC 为像素级 AUROC; PRO 为按缺陷连通域加权的区域召回率, 该指标由 MVTec-AD 数据集原作者定义<sup>[13-14]</sup>。

表1 层组合消融实验结果(均值±标准差)

| 层组合       | I-AUROC                | P-AUROC                | PRO                    |
|-----------|------------------------|------------------------|------------------------|
| L6        | 0.978 0±0.004 0        | 0.958 2±0.000 9        | 0.796 6±0.010 5        |
| L9        | 0.946 6±0.003 1        | 0.963 3±0.000 5        | 0.776 2±0.015 3        |
| L11       | 0.985 1±0.001 4        | 0.960 0±0.000 5        | 0.797 5±0.005 4        |
| L6+L9     | 0.957 2±0.002 9        | <b>0.968 8±0.000 2</b> | 0.782 4±0.033 0        |
| L6+L11    | 0.986 7±0.001 0        | 0.962 7±0.000 4        | 0.805 0±0.016 0        |
| L9+L11    | 0.986 5±0.002 0        | 0.966 7±0.000 4        | <b>0.816 7±0.009 1</b> |
| L6+L9+L11 | <b>0.987 4±0.001 7</b> | 0.968 4±0.000 4        | 0.816 3±0.012 1        |

由表 1 可知, L6+L9+L11 三层组合的 I-AUROC 最高 (0.987 4±0.001 7), 综合考虑图像级与区域级检

测精度，本文采用 L6+L9+L11 三层组合。

从跨种子稳定性方面看：单层 I-AUROC 标准差由小到大依次为 L11 (0.0014)、L9 (0.0031)、L6 (0.0040)，说明 L11 最稳定、L6 标准差最大；多层组合 (L6+L11、L9+L11、L6+L9+L11) 的标准差均明显低于单层标准差，验证了“多层的等权得分级融合通过跨层平均降低跨种子标准差”机制的有效性。

从直观语义方面看：L6 主要提供局部几何与纹理细节；L9 主要提供中层语义部件信息；L11 主要提供整体语义上下文，三者互补可分别支撑划痕（局部几何）、裂纹（部件破坏）与错位（语义不合理）等不同尺度的缺陷类型。

### 2.3 文本通路的解耦设计

一般形式的检测分数为

$$s_{\text{det}}(x) = (1 - \beta) \cdot s_{\text{vis}}(x) + \beta \cdot s_{\text{txt}}(x)$$

式中： $S_{\text{vis}}(x)$  为视觉路径得分； $S_{\text{txt}}(x)$  为文本路径得分，由 patch、token 特征与缺陷词嵌入的余弦相似度确定； $\beta$  为融合系数。

本文将文本路径融合系数置零 ( $\beta=0$ )，检测分数完全由视觉路径给出，文本路径仅在后处理阶段用于生成 k-map（判断属于哪类缺陷）。

## 3 缺陷类型归因与可视化

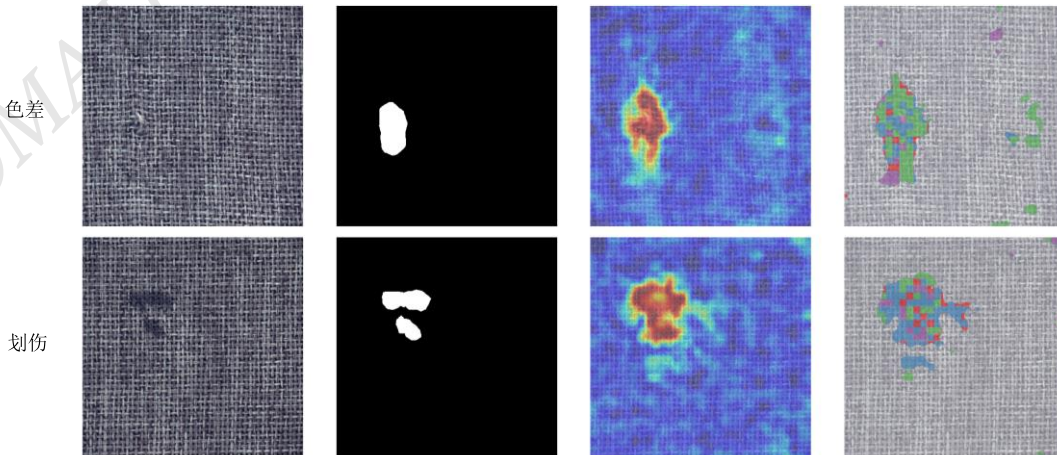
在工业生产线中，视觉缺陷检测除了判断是否存在

在缺陷外，还需确定缺陷类型、置信度等可解释信息，以支持阈值设定、报废分级等后续决策。本文借助 CLIP 文本探针实现 MultiHook-AD 缺陷检测框架的开放词表缺陷类型归因与可视化。其中，开放词表是指用户可在测试阶段任意修改词集而无需重新训练记忆库或重新评估检测精度，归因的精细程度取决于所选词集与 CLIP 嵌入语义的匹配程度。

借助 CLIP 文本探针，本文以 k-map 形式输出每个像素对各缺陷词（如污渍 stain、划痕 scratch、裂纹 crack、凹痕 dent、孔洞 hole）的相似度分布。记第  $k$  个缺陷词的文本嵌入为  $t_k$ 、像素  $x$  处的特征为  $f_{11}(x)$ （取 L11，深层语义更适合类型判别），则像素  $x$  的类型预测为

$$k^*(x) = \underset{k}{\operatorname{argmax}} \cos(f_{11}(x), t_k)$$

MVTec-AD 数据集 carpet 类上的 k-map 输出如图 2 所示。测试时，输入的候选缺陷词集 {stain, scratch, hole, knot} 无需与实际缺陷类型一一对应，以模拟工程师面对未知缺陷的开放词表场景。为突出高置信区域，图 2 的第 3、4 列仅显示检测分数前 10% 的像素。其中，k-map 不参与检测分数计算，仅在视觉最近邻路径已判定某像素异常后，给出缺陷类型的辅助标签。因此本文开放词表是指归因层的能力，而非模型对未知缺陷的检测能力。



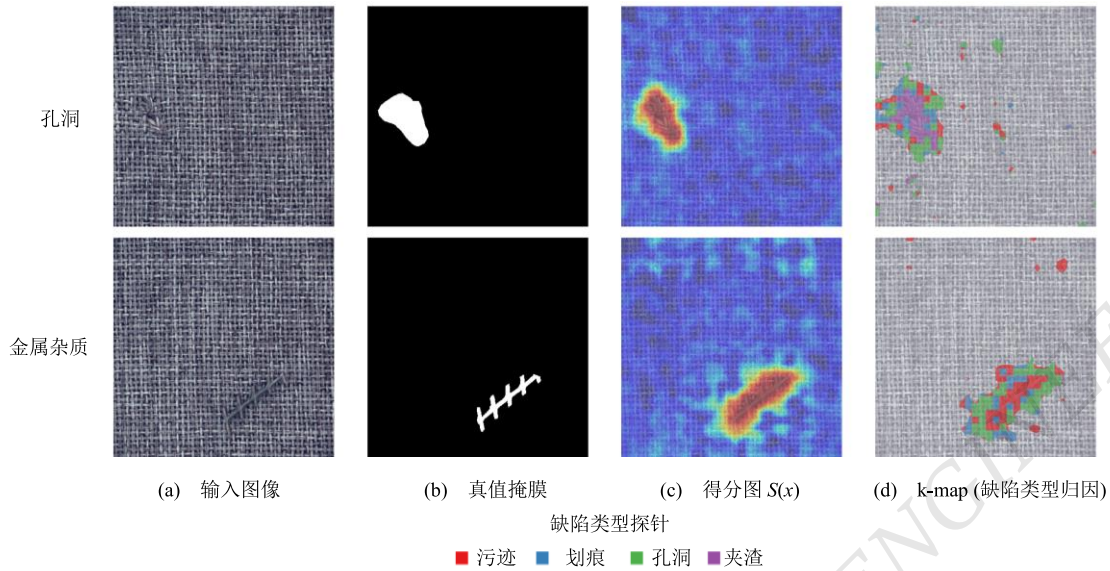


图2 MVtec-AD 数据集 carpet 类上的 k-map 缺陷类型归因可视化

由图2可知：第3列显示的视觉最近邻检测分数能准确定位缺陷区域，与第2列真值掩膜吻合；第4列在高分像素处叠加相似度最大的候选缺陷词，同一模型对不同形态的缺陷给出具有区分度的类型标签，如第1行颜色斑块状缺陷多被归因为污迹，第4行线状磨损出现划痕与污迹的混合。

## 4 实验

### 4.1 实验设置

本实验以 MVtec-AD 数据集全部 15 类为主表；VisA 数据集<sup>[15]</sup>全部 12 类与 Real-IAD 数据集前 6 类用于跨数据集泛化性能测试<sup>[16]</sup>；自建的发动机缸套缺陷图像数据集用于实际工业场景验证。

MultiHook-AD 缺陷检测框架的视觉主干网络为冻结的 DINOv2-ViT-B/14，输入分辨率为  $448 \times 448$ ，

patch 尺寸为 14。实验基于 PyTorch 2.x 实现，批大小为 32，使用单张 RTX 3090 24 GB 显卡。

对比基线包括 6 种方法：PatchCore-DINOv2、RawDINOv2-NN、WinCLIP-stub、AnomalyCLIP-stub、GraphCore-stub 与 MVFA-stub，其中后缀-stub 表示仅保留方法核心模块、不复现完整训练流程，这有利于消除主干网络、分辨率与数据预处理等无关变量带来的影响，使复现结果反映方法本身的差异。本文将所有对比基线统一在 DINOv2-ViT-B/14 主干网络、输入分辨率  $448 \times 448$  与同一核心子集采样流程下复现，结果如表 2 所示。

为消除单次运行的随机性，本文在  $SEED \in \{42, 7, 123\}$  3 个种子下重复完整流程，并以均值  $\pm$  标准差 (mean  $\pm$  std) 形式呈现；其中 GraphCore-stub 与 MVFA-stub 方法仅显示  $SEED = 42$  的单个结果。

表2 MVtec-AD 数据集 15 类 3 个种子 ( $SEED \in \{42, 7, 123\}$ ) 对比实验结果 (均值 $\pm$ 标准差)

| 方法                  | I-AUROC                                 | P-AUROC                                 | PRO                                     |
|---------------------|-----------------------------------------|-----------------------------------------|-----------------------------------------|
| 本文方法 (MultiHook-AD) | <b>0.987 4 <math>\pm</math> 0.001 7</b> | <b>0.968 4 <math>\pm</math> 0.000 4</b> | <b>0.816 3 <math>\pm</math> 0.012 1</b> |
| PatchCore-DINOv2    | 0.944 2 $\pm$ 0.009 8                   | 0.963 2 $\pm$ 0.000 5                   | 0.769 2 $\pm$ 0.014 5                   |
| RawDINOv2-NN        | 0.942 0 $\pm$ 0.015 4                   | 0.959 7 $\pm$ 0.004 7                   | 0.747 0 $\pm$ 0.016 6                   |
| WinCLIP-stub        | 0.947 4 $\pm$ 0.012 2                   | 0.964 0 $\pm$ 0.000 9                   | 0.763 2 $\pm$ 0.033 9                   |
| AnomalyCLIP-stub    | 0.939 6 $\pm$ 0.010 7                   | 0.964 5 $\pm$ 0.000 4                   | 0.785 5 $\pm$ 0.009 7                   |
| GraphCore-stub      | 0.933 6                                 | 0.960 1                                 | 0.794 5                                 |
| MVFA-stub           | 0.986 1                                 | 0.962 0                                 | 0.781 3                                 |

由表 2 可知:

1) 在图像级精度方面, 本文方法的 I-AUROC 均值为 0.987 4, 较 MVFA-stub 方法的单点值 0.986 1 提高了 0.13 个百分点, 该差距落入种子噪声范围 ( $\text{std} = 0.0017$ ); 考虑到 MVFA-stub 方法仅显示 SEED = 42 的单点结果, 而本文方法显示了 3 个种子的均值, 两者比较口径不统一, 故不能据此认为该差距具有统计显著性; PatchCore-DINOv2 方法的 I-AUROC 为 0.944 2, 低于 Anomalib 报告值 0.987 0 约 4 个百分点, 这是本文统一主干网络与单层配置所致, 非实现误差, 升级为多层配置后回升至 0.979 2 (见 4.2 节表 3);

2) 在跨种子标准差方面, 本文方法的 I-AUROC 标准差为 0.001 7, 仅为其他 4 种方法 3 个种子同主干基线标准差 (0.009 8~0.015 4) 的 11%~20%, 其原因

详见 4.2 节;

3) 在区域级精度方面, 本文方法的 PRO 为 0.816 3, 较 MVFA-stub 方法的单点值 0.781 3 提高了 3.5 个百分点, 表明多层 Hook 的等权得分级融合机制在缺陷连通域尺度上具有更完整的覆盖能力。

#### 4.2 跨种子标准差与基线增益的归因消融实验

为验证表 2 中本文方法标准差较低的原因, 针对 PatchCore-DINOv2 方法进行 2×2 对照消融实验: {单层 L9, 多层 L6+L9+L11 等权融合} × {全局随机种子, 确定性种子  $\text{seed}_l = \text{SEED} + l$ }; 运行设置为 MVTEC-AD 数据集的 5 类 (bottle、carpet、cable、screw、transistor) × 3 个种子 ( $\text{SEED} \in \{42, 7, 123\}$ ), 共 4 种配置 × 5 类 × 3 个种子 = 60 次, 实验结果如表 3 所示。

表 3 PatchCore-DINOv2 方法 2×2 归因消融 (5 类×3 个种子) 实验结果 (均值±标准差)

| 配置                             | I-AUROC                  | P-AUROC                  | PRO                      |
|--------------------------------|--------------------------|--------------------------|--------------------------|
| 单层 L9 + 全局随机种子 (PatchCore 原配置) | 0.920 6 ± 0.012 4        | 0.952 7 ± 0.002 3        | 0.709 5 ± 0.046 6        |
| 单层 L9 + 确定性种子                  | 0.925 3 ± 0.010 8        | 0.950 4 ± 0.001 1        | 0.724 5 ± 0.044 3        |
| 多层 L6+L9+L11 + 全局随机种子          | 0.979 2 ± 0.000 8        | 0.955 3 ± 0.000 4        | <b>0.754 1 ± 0.010 4</b> |
| 多层 L6+L9+L11 + 确定性种子           | <b>0.978 0 ± 0.002 6</b> | <b>0.957 0 ± 0.000 6</b> | 0.753 9 ± 0.005 7        |

由表 3 可知:

1) 在标准差上, 4 种配置的 P-AUROC 差异不超过 0.005 且标准差均很小, 说明像素级指标对单层/多层与种子设置均不敏感, 故归因分析聚焦 I-AUROC 与 PRO; 单独引入确定性种子时, 标准差仅小幅下降 (I-AUROC 由 0.012 4 降至 0.010 8, 降幅为 13%; PRO 由 0.046 6 降至 0.044 3, 降幅为 5%); 单独引入多层 Hook 时, 标准差大幅下降 (I-AUROC 由 0.012 4 降至 0.000 8, 降幅为 93%; PRO 由 0.046 6 降至 0.010 4, 降幅为 78%); 两者叠加在 PRO 上呈互补效应, 标准差进一步降至 0.005 7 (综合降幅为 88%), 表明跨种子标准差减小的主要原因是多层 Hook 的等权得分级融合机制, 确定性种子规则的独立贡献较小;

2) 在均值上, 从单层升级到多层, 5 类缺陷的均值 I-AUROC 从 0.920 6 升至 0.979 2 (增加 5.86 个百分点), 已覆盖本文方法主表 (0.987 4) 相对 PatchCore 单层 (5 类对应 0.920 6) 的大部分差距; 其余方法相对本文方法 0.987 4 的差距仅为 0.82 个百分点, 即为本文方法在多层 Hook 之外的其他设计 (文本通路解耦、归一化处理等) 所贡献的净增益。

#### 4.3 跨数据集泛化

直接复用 MVTEC-AD 数据集上的全部超参数, 对 VisA、Real-IAD 2 个外部数据集进行零样本迁移测试:

1) VisA 数据集全部 12 类 (均值), I-AUROC 为 0.826 6, P-AUROC 为 0.992 9, PRO 为 0.802 3; P-AUROC 高于主表, 表明多层 DINOv2 特征对局部

异质性的表征具有跨数据集稳定性；

2) Real-IAD 数据集前 6 类跨数据集零样本迁移结果如表 4 所示。

表 4 Real-IAD 数据集前 6 类跨数据集零样本迁移结果 (SEED=42)

| 类别             | I-AUROC                  | P-AUROC                  | PRO                      |
|----------------|--------------------------|--------------------------|--------------------------|
| audiojack      | 0.852 8                  | 0.982 8                  | 0.877 3                  |
| bottle_cap     | 0.958 5                  | 0.992 9                  | 0.861 5                  |
| button_battery | 0.853 7                  | 0.979 4                  | 0.651 5                  |
| end_cap        | 0.821 1                  | 0.974 8                  | 0.686 8                  |
| eraser         | 0.887 3                  | 0.984 7                  | 0.723 5                  |
| fire_hood      | 0.875 3                  | 0.993 5                  | 0.819 9                  |
| 均值±标准差         | <b>0.874 8 ± 0.046 9</b> | <b>0.984 7 ± 0.007 4</b> | <b>0.770 1 ± 0.095 4</b> |

由表 4 可知: Real-IAD 数据集前 6 类的 P-AUROC、PRO 与主表处于同一量级, 表明跨数据集零样本迁移在该 6 类上成立; I-AUROC 为 0.874 8, 数值偏低, 主要源于其类间外观差异较大, 单一图像级阈值难以同时适配, 需以少量正常样本进行阈值校准。

关于测试子集的选择依据: Real-IAD 数据集共包含 30 类工业产品, 本文按官方类别列表字母序选取前 6 类, 属便利采样, 不涉及依据方法表现的主观挑类; 所选 6 类工业产品涵盖音频接口、瓶盖、纽扣电池、端盖、橡皮与消防罩等结构与材质差异较大的类别, 其 I-AUROC 标准差为 0.046 9, 也表明未刻意选取易检类别。受计算预算与篇幅限制, 其余 24 类工业产品 (含印制电路板、纺织品等) 尚未覆盖, 因此跨数据集零样本迁移的结论仅在该 6 类子集层面成立, 完整 30 类的验证留作后续工作。

#### 4.4 发动机缸套场景

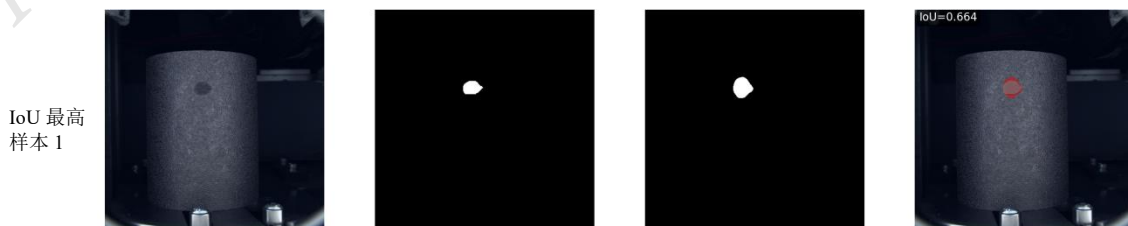
本文提出图内 patch 适配方案: 将每幅图像中缺陷掩膜取反后的非缺陷区域视为正常块参考, 并由此构建图内记忆库, 从而无需外部正常样本。本文采用自建的发动机缸套缺陷图像数据集进行实验验证。该数据集采集自某工厂的发动机缸套外圆在线检测工位 (300 万像素相机、8 mm 镜头、隧道光源), 共 569 幅图像, 全部为含缺陷样本, 缺陷形态覆盖划痕、磕碰、锈蚀、涂料眼等典型类型。每幅图像均配有人工标注的像素级缺陷掩膜: 标注通过 HALCON 工具完成; 因原始类型标签一致性不足, 统一进行二值化处理, 掩膜值  $\geq 1$  的像素视为缺陷。数据集按 7:3 划分为参考集与测试集 (划分种子固定为 42); 单图非缺陷区域面积占比的 5%、50%、95% 分位数分别为 78%、92%、99%, 均满足图内记忆库的构建条件。实验结果如表 5 所示。

表 5 发动机缸套场景实验结果 (3 个种子 SEED  $\in \{42, 7, 123\}$ ) (均值±标准差)

| 方法               | P-AUROC                  | PRO                      |
|------------------|--------------------------|--------------------------|
| PatchCore-DINOV2 | 0.960 0 ± 0.004 2        | 0.691 8 ± 0.054 3        |
| AnomalyCLIP-stub | 0.963 5 ± 0.005 0        | 0.654 7 ± 0.047 2        |
| MultiHook-AD     | <b>0.967 3 ± 0.001 5</b> | <b>0.790 3 ± 0.029 7</b> |

由表 5 可知, 本文方法的 P-AUROC、PRO 均优于同主干网络的 PatchCore-DINOV2、AnomalyCLIP-stub 方法, 表明多层 Hook 的等权得分级融合机制在无外部正常样本的图内适配场景下, 能提升检测精度与跨种子稳定性。

本文方法在发动机缸套缺陷图像数据集上的缺陷检测可视化结果如图 3 所示。



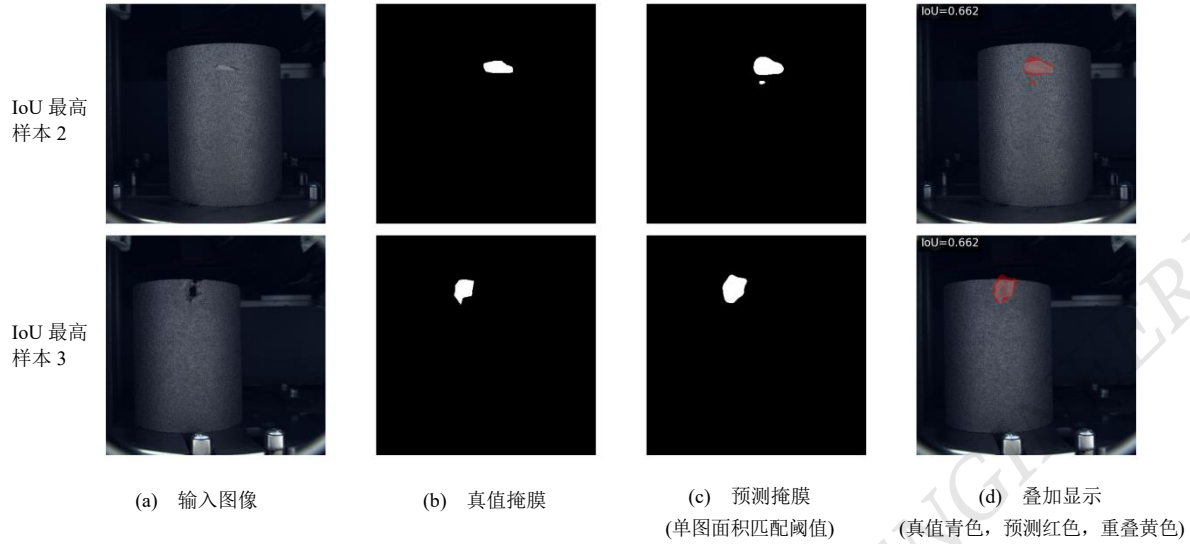


图3 发动机缸套缺陷图像数据集上的缺陷检测可视化结果 (IoU 最高的 3 个样本)

由图3可知,检测的异常区域与真值掩膜形态吻合(3个样本的IoU均约为0.66),表明图内patch适配方案在无外部正常样本的条件下仍能准确定位缺陷区域。

#### 4.5 GPU加速部署评测

MultiHook-AD缺陷检测框架在MVTec-AD数据集3类(bottle、carpet、hazelnut)×3个种子上测得:工作站级RTX 3090单图FP32推理耗时为 $28.35 \pm 1.68$  ms (FPS  $35.4 \pm 2.1$ ); FP16推理耗时为 $25.37 \pm 6.01$  ms (FPS  $43.1 \pm 5.2$ ); I-AUROC均接近饱和(>0.997), FP16相对单图FP32的I-AUROC差异不到0.001,表明本文方法可适配多数生产线节拍。

## 5 结论

面向柔性生产过程中,产品频繁切换、跨品类快速复用等实际需求,本文提出一种多尺度自监督特征融合的工业视觉缺陷检测方法。该方法通过多Hook抓取浅、中、深3层patch token特征,按等权得分级融合机制形成统一的异常图像,支持同一组超参数跨数据集直接复用。实验表明,本文方法在MVTec-AD数据集全部15类、VisA数据集全部12类及Real-IAD数据集前6类上均保持稳定的图像级与像素级精度,跨种子标准差仅为同主干基线的11%~20%。

未来工作将继续研究跨品类阈值校准与漏检、误判分布的精细分析,补充与AnomalyDINO<sup>[17]</sup>、Dinomaly<sup>[18]</sup>、MuSc<sup>[19]</sup>等同主干网络方法的端到端公平对照,并完成Real-IAD数据集完整30类工业零件的零样本迁移验证。

©The author(s) 2024. This is an open access article under the CC BY-NC-ND 4.0 License (<https://creativecommons.org/licenses/by-nc-nd/4.0/>)

## 参考文献

- [1] ROTH K, PEMULA L, ZEPEDA J, et al. Towards total recall in industrial anomaly detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 14318-14328.
- [2] JEONG J, ZOU Y, KIM T, et al. WinCLIP: Zero-/Few-Shot anomaly classification and segmentation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 19606-19616.
- [3] ZHOU Q, PANG G, TIAN Y, et al. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection[C]// International Conference on Learning Representations (ICLR). Vienna: OpenReview, 2024.
- [4] HUANG C, JIANG A, FENG J, et al. Adapting visual-language models for generalizable anomaly detection in medical images [C]//Proceedings of the IEEE/CVF Conference on Computer

- Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 11375-11385.
- [5] OQUAB M, DARCET T, MOUTAKANNI T, et al. DINOv2: Learning robust visual features without supervision[J]. arXiv Preprint arXiv:2304.07193, 2023.
- [6] DEFARD T, SETKOV A, LOESCH A, et al. PaDiM: A patch distribution modeling framework for anomaly detection and localization[C]//Pattern Recognition. ICPR International Workshops and Challenges, Lecture Notes in Computer Science, vol 12664. Cham: Springer, 2021: 475-489.
- [7] ZAVRTANIK V, KRISTAN M, SKOČAJ D. DRAEM: A discriminatively trained reconstruction embedding for surface anomaly detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 8310-8319.
- [8] LIU Z, ZHOU Y, XU Y, et al. SimpleNet: A simple network for image anomaly detection and localization[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 20402-20411.
- [9] LEE S, LEE S, SONG B C. CFA: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization[J]. IEEE Access, 2022, 10: 78446-78454.
- [10] XIE G, WANG J, LIU J, et al. Pushing the limits of fewshot anomaly detection in industry vision: Graphcore[J]. arXiv Preprint arXiv:2301.12082, 2023.
- [11] HE K, CHEN X, XIE S, et al. Masked autoencoders are scalable vision learners[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 15979-15988.
- [12] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 9630-9640.
- [13] BERGMANN P, FAUSER M, SATTLEGGER D, et al. MVTec AD-A comprehensive real-world dataset for unsupervised anomaly detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 9584-9592.
- [14] BERGMANN P, BATZNER K, FAUSER M, et al. The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection[J]. International Journal of Computer Vision, 2021, 129(4): 1038-1059.
- [15] ZOU Y, JEONG J, PEMULA L, et al. SPot-the-difference self-supervised pre-training for anomaly detection and segmentation[C/OL]. European Conference on Computer Vision (ECCV). Cham: Springer, 2022:392-408 [2026-08-16]. <https://arxiv.org/abs/2207.14315>.
- [16] WANG C, ZHU W, GAO B B, et al. Real-IAD: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2024: 22883-22892.
- [17] DAMM S, LASZKIEWICZ M, LEDERER J, et al. Anomaly-DINO: Boosting patch-based few-shot anomaly detection with DINOv2[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025:1319-1329.
- [18] GUO J, LU S, ZHANG W, et al. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection [C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2025: 20405-20415.
- [19] LI X, HUANG Z, XUE F, et al. MuSc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images[C]//International Conference on Learning Representations, 2024, 2024: 2172-2197.

#### 作者简介:

陈永彬, 男, 1990 年生, 博士研究生, 副教授, 主要研究方向: 智能测控、AI 大模型。E-mail: chenyb@gdmec.edu.cn

黄嘉思, 女, 1984 年生, 硕士研究生, 副教授, 主要研究方向: AI 智能体。E-mail: 245504588@qq.com

黄国健, 男, 1981 年生, 博士研究生, 教授级高工, 主要研究方向: 力学研究。E-mail: guojian.huang@gmail.com