自动化学报
ACTA AUTOMATICA SINICA

视觉工业缺陷检测: 从非基础模型到基础模型的演进与协同综述

杨天乐 常璐瑶 言嘉栋 李俊涛 张可 张民

Visual Industrial Defect Detection: A Survey on the Evolution and Synergy From Non-foundation Model to Foundation Model

YANG Tian-Le, CHANG Lu-Yao, YAN Jia-Dong, LI Jun-Tao, ZHANG Ke, ZHANG Min

在线阅读 View online: https://doi.org/10.16383/j.aas.c250594

视觉工业缺陷检测: 从非基础模型到基础模型的演进与协同综述

杨天乐[1] 常璐瑶[2] 言嘉栋[1] 李俊涛[1] 张 可[1] 张 民[3]

摘 要 随着工业产品日益丰富且精密复杂, 视觉工业缺陷检测技术备受关注. 近年来, 基础模型 (FM) 凭借其从海量数据中获取的广博先验知识, 在泛化能力及少样本与零样本场景中展现出强大的潜力. 然而, 梳理现有方法可以发现一个值得关注的现象: 当前许多先进的 FM 方法, 其性能的显著提升并非单纯依赖 FM 的应用, 而是通过将 FM 强大的通用表征能力与非基础模型 (NFM) 方法中成熟、高效的任务导向型原理 (例如对比学习、知识蒸馏和异常合成) 进行战略性融合. 为系统性地分析并揭示这一协同范式, 首先分别对 NFM 和 FM 方法进行系统性综述, 并从多维度比较两种方法, 分析各自的优势与局限. 在此基础上, 深入剖析 NFM 策略如何从其原始架构中解耦出来, 并被重新用于增强 FM 在工业应用中的性能, 同时构建协同机制适配性矩阵. 此外, 进一步探讨该范式在实际场景中的落地局限. 在对比数据的支持下, 分析结果表明, FM 的通用知识与 NFM 的任务特化优势之间存在巨大的协同潜力, 这为未来的研究指明了一条有效的借鉴思路.

关键词 工业缺陷检测; 少样本学习; 基础模型; 非基础模型

引用格式 杨天乐, 常璐瑶, 言嘉栋, 李俊涛, 张可, 张民. 视觉工业缺陷检测: 从非基础模型到基础模型的演进与协同综述. 自动化学报, 2026, **52**(8): 1593−1614

DOI 10.16383/j.aas.c250594 **CSTR** 32138.14.j.aas.c250594

Visual Industrial Defect Detection: A Survey on the Evolution and Synergy From Non-foundation Model to Foundation Model

YANG Tian-Le[1] CHANG Lu-Yao[2] YAN Jia-Dong[1] LI Jun-Tao[1] ZHANG Ke[1] ZHANG Min[3]

Abstract As industrial products become increasingly abundant and sophisticated, visual industrial defect detection technology has received much attention. In recent years, foundation model (FM) has demonstrated powerful potential, particularly in generalization as well as in few-shot and zero-shot scenes, by leveraging broad prior knowledge acquired from vast amounts of data. However, upon reviewing existing methods, a noteworthy phenomenon can be observed: The significant performance improvements of many current advanced FM methods do not rely solely on the application of FM, but rather through the strategic fusion of the FM′s powerful general-purpose representation capabilities with mature, efficient, task-specific principles from non-foundation model (NFM) methods, such as contrastive learning, knowledge distillation, and anomaly synthesis. To systematically analyze and reveal this synergistic paradigm, a systematic survey of both NFM and FM methods is first provided, and the two methods are compared from multiple dimensions to analyze their respective advantages and limitations. Building on this, an in-depth analysis is conducted on how NFM strategies can be decoupled from their original architectures and repurposed to enhance the performance of FM in industrial applications, while an adaptability matrix for collaborative mechanisms is simultaneously constructed. Furthermore, the practical deployment limitations of this paradigm in real-world scenes are further explored. Supported by comparative data, the analysis results indicate that there is significant synergistic potential between the general knowledge of FM and the task-specialization advantages of NFM, pointing to an effective direction for future research.

Keywords industrial defect detection; few-shot learning; foundation model; non-foundation model

Citation Yang Tian-Le, Chang Lu-Yao, Yan Jia-Dong, Li Jun-Tao, Zhang Ke, Zhang Min. Visual industrial defect detection: A survey on the evolution and synergy from non-foundation model to foundation model. *Acta Automatica Sinica*, 2026, **52**(8): 1593−1614

收稿日期 2025-11-03 录用日期 2026-03-04
Manuscript received November 3, 2025; accepted March 4, 2026
国家自然科学基金 (62506253) 资助
Supported by National Natural Science Foundation of China (62506253)
本文责任编委 陈胜勇
Recommended by Associate Editor CHEN Sheng-Yong

1. 苏州大学计算机科学与技术学院 苏州 215008 2. 华中师范大学计算机学院 武汉 430079 3. 哈尔滨工业大学 (深圳) 计算与智能研究院 深圳 518055
1. School of Computer Science and Technology, Soochow University, Suzhou 215008 2. School of Computer Science, Central China Normal University, Wuhan 430079 3. Faculty of Computing and Intelligence, Harbin Institute of Technology, Shenzhen 518055

视觉缺陷检测[1], 也称为视觉异常检测, 是人工智能算法的一个关键应用领域. 这项任务在确保工业产品质量方面扮演着至关重要的角色. 传统的工业异常检测算法 [2−5] 专注于对正常特征的统计分布进行建模, 并通过分析输入样本与这些学习到的模式的偏差来检测异常. 为增强模型识别异常模式的能力, 一些方法如文献 [6–7] 进一步探索正常与异常特征之间的对比学习机制[8]. 这些方法通常依赖大量高质量的训练数据来建立可靠的特征分布和对比关系. 然而, 在真实的工业场景中, 由于产品和缺陷的多样性与复杂性, 获取特定的高质量训练数据具有挑战性[9–10]. 例如, 在芯片缺陷检测中, 芯片种类繁多、缺陷类别众多, 包括结构缺陷和纹理缺陷, 这使得为各种产品和缺陷收集数据变得困难. 在这种情况下, 传统模型难以达到令人满意的检测效果. 近来, 随着视觉和语言基础模型 (foundation models, FM)(如 CLIP[11]、GPT[12–13] 和 SAM (segment anything model)[14]) 的发布, 基于这些模型的工业缺陷检测算法在 2D 和 3D 视觉环境中均取得显著进展, 尤其是在数据有限的少样本和零样本场景中[15]. 基础模型自身拥有强大的通用视觉和语言理解能力, 因此, 如何在没有额外训练样本和标注的情况下, 将其基础知识有效应用于工业检测问题, 是一个重要课题. 本文将不同基础模型在 2D 和 3D 工业缺陷检测中的应用分类如下:

1) SAM-2D: 视觉先验知识的应用. 作为一种强大的视觉分割基础模型, SAM 提供了通过在海量数据上进行广泛预训练而获得的语义先验信息, 显著增强了工业缺陷检测的准确性. 在基于 SAM 的 2D 工业缺陷检测任务中, 研究者们开发各种方法 [16−21], 以针对工业场景特定地提示 SAM. 此外, 基于 SAM 生成的掩码进行目标匹配, 用以识别缺陷区域.

2) CLIP-2D: 短文本与图像的语义匹配. 像 CLIP 这样的图文基础模型展示了细粒度的图文匹配能力. 这种能力有效地将微妙的视觉线索与描述性文本联系起来, 因此对缺陷检测尤其有益. 在基于 CLIP 的 2D 工业缺陷检测任务中[20, 22–33], 设计和学习合适的文本提示, 同时在细粒度级别上对齐图像信息, 对于进一步提升性能至关重要. 文本提示模板的设计已得到广泛研究.

3) GPT-2D: 长文本语义先验. 像 GPT 这样的大语言模型 (large language model, LLM) 可以生成长篇描述, 使其非常适用于需要详细解释和结构化描述的复杂场景. 因此, 在基于 GPT 的 2D 工业缺陷检测方法中[34–40], 一个关键挑战是设计提示以获得全面的文本描述, 并有效利用这些文本信息.

4) CLIP-3D: 应用于跨维度视觉任务的短文本图像语义匹配先验. 3D 缺陷检测由于其复杂的空间信息而面临更大挑战. 为解决这一问题, 像 CLIP 这样的图文基础模型通过跨模态信息互补提供了一个有前景的解决方案[41–43]. 这些模型有效地将视觉信息与文本描述相结合, 从而实现更精确的高维空间建模, 捕捉 3D 结构中的复杂缺陷.

尽管 FM 在工业缺陷检测中展现出广阔的应用前景, 但非基础模型 (non-foundation model, NFM) 方法由于其较小的参数量和较高的计算效率, 在特定应用场景中仍具有不可替代的优势. 基于此, 本文也对 NFM 方法进行回顾, 包括 2D 统计建模[44–48]、2D 异常数据合成[49–60]、2D/3D 跨模态知识蒸馏[43, 61–66] 以及基于 3D 生成模型的算法[41, 67–70]. 实际上, 这些 NFM 方法能够为 FM 的应用提供重要指导. 梳理现有研究可以发现, 许多性能优秀的 FM 方法, 其成功的关键恰恰在于没有摒弃 NFM 方法的范式. 为此, 本文系统性地分析成熟的 NFM 策略如何能够被调整, 以增强 FM 在工业缺陷检测领域的专业能力. 通过对实验数据的对比回顾, 分析结果揭示了这种有效融合的内在机制, 并为未来的研究者构建更强大、更专业的模型提供了清晰、可行的借鉴思路.

本综述的组织结构如下: 首先, 简述工业缺陷检测从 NFM 到 FM 的发展趋势, 阐明当前 FM 在应用中面临的挑战, 并提出本文关于两者“战略性融合”的核心研究视角; 在第 1 节中, 详细比较 FM 和 NFM 方法, 重点关注它们在训练目标、模型架构、算法框架和性能上的差异; 第 2 节和第 3 节分别回顾主流的 NFM 和 FM 方法, 展示各自领域内的技术演进; 在此基础上, 第 4 节深入分析一系列前沿的 FM 方法, 通过具体案例剖析它们如何借鉴和融合 NFM 策略以实现性能突破; 第 5 节从宏观视角探讨 FM 与 NFM 的适配性矩阵及当前协同范式的局限性; 在第 6 节中, 探讨大模型持续面临的挑战, 并指出未来值得进一步探索的潜在方向. 本文所调研方法的详细组织结构如图 1 所示.

1 FM 与 NFM 方法的比较

随着工业检测需求的多样化, 模型训练目标、结构、规模和性能的差异已成为影响方法选择的关键因素. 在深入对比前, 需首先明确工业缺陷检测领域的三大核心任务方向. 根据产出粒度与技术逻辑, 工业缺陷检测可划分为:

1) 异常检测与定位: 旨在识别偏离正常分布的

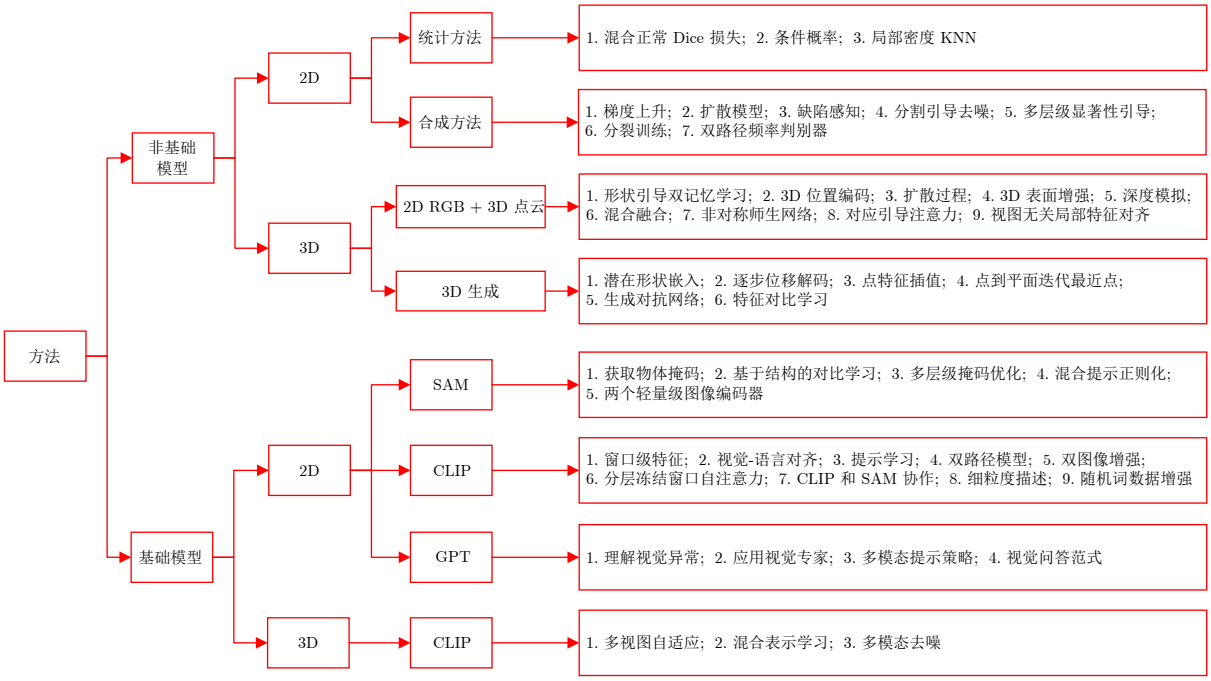


图 1　调研方法的组织结构

Fig. 1　Organization structure of surveyed methods

未知模式[1]. NFM 多采用统计分布建模[45] 或重建偏差分析[71], 而 FM 则侧重利用海量先验进行零样本识别[22].

2) 缺陷语义分割: 侧重于像素级缺陷边界的提取. NFM 常依赖有监督训练或数据合成[56], FM (如 SAM[14]) 则通过空间先验提供高保真的分割掩码.

3) 缺陷目标检测: 侧重于在复杂背景中框选特定类别的缺陷. NFM 依赖特定类别的标注数据, 而 FM 则能通过文本提示实现开放类别的灵活检测.

本节将从模型训练目标、预训练语料与知识来源、模型结构与规模、框架及模型性能五个维度, 对比 FM 与 NFM 在上述任务中的技术路径. 比较摘要如图 2 所示.

1.1　模型训练目标

FM 和 NFM 在数据需求、训练方法、计算资源以及特征学习的广度上表现出显著差异, 这导致了二者训练目标的差异. 2D NFM 方法依赖传统的网络架构, 利用生成对抗网络 (generative adversarial network, GAN)[72] 和扩散模型[73] 等技术生成多样化的异常样本, 以补偿数据不足; 同时, 此类方法高度强调特征选择和重建策略[74−77]. 3D NFM 方法专注于检测点云数据中的几何缺陷和缺失区域, 使用高效架构以减少计算开销; 此外, 相关研究还引入了创新技术, 如归一化流[78−79] 和基于扩散的重建机制, 以增强准确性和鲁棒性, 避免对设计文件或模型库的依赖.

相比之下, 2D FM 方法 (如 CLIP、SAM、GPT) 利用视觉−语言模型的跨模态能力[80−82] 来提高异常检测的准确性和效率. 其主要目标包括: 1) 通过无监督或零样本学习识别未知的异常类别, 减少对标注数据的依赖; 2) 生成可解释的检测结果, 用颜色、形状和类别等术语描述异常; 3) 通过将特定的异常观察模块与 FM 集成, 解决复杂异常, 从而提高准确性; 4) 增强模型的可扩展性和适应性, 使其能快速适应不同的工业场景; 5) 集成多模态信息以实现精确的异常定位和识别. 3D FM 方法专注于点云数据的几何特征[83−84] 和多视图融合[85], 解决数据不完整和噪声干扰问题[86−87]. 此类方法通过多视图渲染进行分类和分割, 同时也处理多模态数据间的不一致性.

总之, NFM 方法强调特征选择、计算效率和数据合成, 更适用于资源受限的环境; FM 方法则侧重于跨模态学习和生成能力, 在数据稀缺场景下表现出色, 适用于多任务和多域检测.

1.2　预训练语料与知识来源

FM 与 NFM 泛化能力的本质差异源于其预训练语料的规模与特性. NFM 的任务特化优势体现在对目标域分布的精准刻画. NFM 通常不依赖大规模预训练, 而是侧重于对工业场景内正常特征

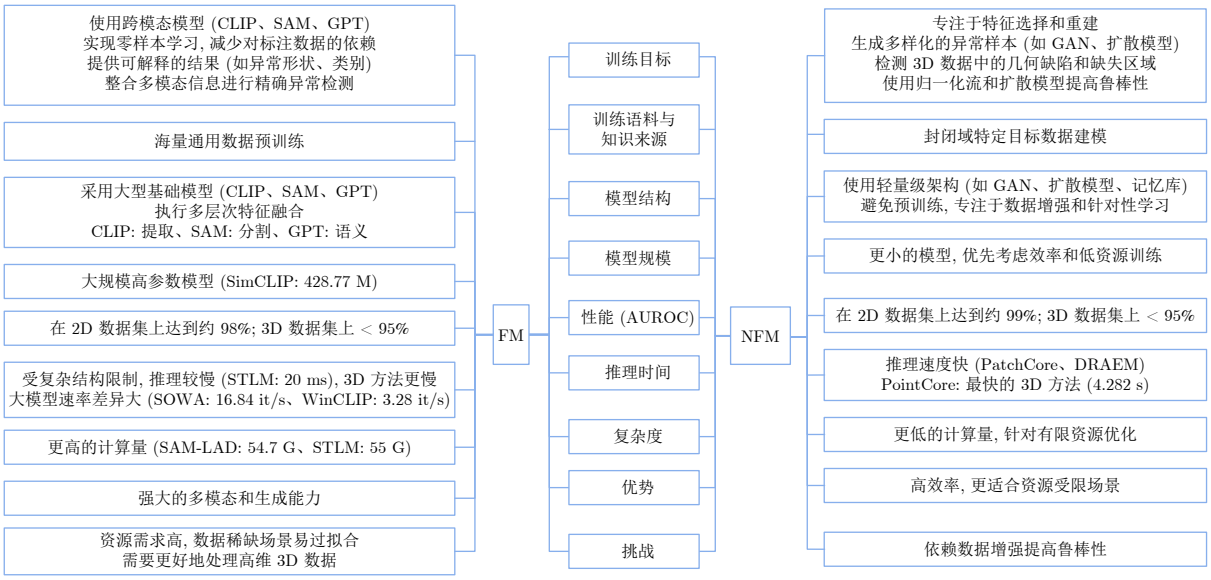


图 2 FM 与 NFM 方法的比较摘要

Fig. 2 A summary of the comparison between FM and NFM methods

分布的显式建模[1, 51]. 通过对特定产品数据的统计分析[45] 或基于领域知识的缺陷合成[49], NFM 能够在受控的生产环境下表现出更高的检测灵敏度与算力效率.

与 NFM 的封闭域训练不同, FM 的广博先验知识主要构建于海量通用语料之上: SAM[14] 基于包含 11 亿个分割掩码的 SA-1B 数据集, 其庞大的几何拓扑信息赋予模型极强的空间感知能力; CLIP[11] 依托 4 亿互联网图文对, 建立了自然纹理与语义描述的深层关联, 为零样本缺陷识别提供了文本指引; GPT 类基础模型[34] 则通过海量指令微调语料, 掌握了复杂的逻辑推理与成因分析能力, 能够根据工业需求生成结构化的诊断报告[37]. 这些通用语料虽未包含特定工业缺陷, 但其提供的通用表征具备极高的可迁移性. 这种从“领域专长”到“通用知识”的互补关系, 构成了后续讨论 FM 与 NFM 协同范式的理论基础.

1.3 模型结构与规模

1.3.1 模型结构

NFM 方法主要包括教师–学生架构[88−90]、分布图[91−92]、记忆库[93−94]、基于自动编码器[95−97]、基于 GAN、基于 Transformer 和基于扩散的框架. 这些方法不依赖大规模数据或预训练任务, 更专注于局部特征选择、样本生成和增强, 旨在优化有限数据下的特征学习和异常检测能力.

相比之下, FM 方法依赖强大的视觉–语言协同机制, 集成了如 CLIP、SAM 和 GPT 等大规模基础模型. 这些模型采用多级特征融合来建立协同工作流: CLIP 对图像和点云数据执行多模态特征提取和对齐; SAM 进行细粒度分割以分离潜在的异常区域; GPT 则提供对检测结果的语义理解和描述, 帮助用户快速获得分析结论. 为应对少样本和零样本学习的挑战[98], CLIP 的预训练知识使其能在未标注数据上进行有效推理, 从而增强了检测模型的泛化能力.

1.3.2 模型规模

NFM 方法的参数量通常较小, 主要通过高效的特征选择、对抗性训练和自监督学习来优化模型. 这些方法能在资源受限的环境中实现更高效的训练. 反观 FM 方法, 它们通常依赖庞大的参数量, 利用复杂的网络架构和跨模态学习来处理错综复杂的异常检测任务, 这导致了更高的训练时间和计算资源需求. 例如, SimCLIP[26] 的参数量为 428.77 M. 由于快速推理是必然趋势, 新发布的 FM 方法正试图探索加速推理的方法. 例如, 基于 SAM 的 STLM[16] 仅需 16.56 M 进行推理, 使其成为最高效的方法之一.

1.4 框架

FM 方法和 NFM 方法的框架如图 3 所示. NFM 方法侧重于设计特定任务的模型. 例如, 基于重建的异常检测方法通过学习重建过程来训练一个能准确重建正常数据的模型. 在数据预处理阶段, 由于异常样本稀少, 一些方法使用异常合成策略来扩充数据集. 通过使用目标数据训练模型, 模型逐渐改进, 最终生成专门用于检测或分割异常区域的

模型.

FM 方法主要利用嵌入于基础模型中的先验知识, 这些模型已在大型通用数据集上进行了预训练, 拥有强大的特征表示能力. 因此, 微调这些模型通常仅需少量样本. 不同类型的基础模型, 如 SAM 和 CLIP, 可以处理不同模态的数据, 包括图像和文本信息. 在训练过程中, FM 方法侧重于设计合适的损失函数, 以使基础模型更有效地适应工业应用中的异常检测任务. 最终, 训练好的模型能实现对工业图像中异常区域的精确分割或定位.

1.5　模型性能

1.5.1　主流数据集与评价指标

在对比两类方法的具体性能前, 需首先梳理工业缺陷检测的主流基准数据集. 如表 1 所示, 不同的数据集侧重于不同的检测挑战: MVTec AD[9] 主要考察纹理和物体表面的细微缺陷; VisA[10] 引入了复杂背景和多对象场景的干扰; MVTec 3D-AD[99] 专门针对空间几何形变异常; Real3D-AD[100] 则进一步提供了高分辨率的点云数据, 聚焦于高精度几何特征下的缺陷检测. 这些数据集的多样性决定了 FM 与 NFM 方法在不同场景下的适用性差异. 图 4 展示了这些数据集的部分样本.

评价指标的现状与局限性. 上述基准数据集涵盖了纹理、结构及空间几何等多种工业场景, 为算法验证提供了丰富的数据支持. 在现有的性能评估体系中, AUROC 是衡量模型判别能力的核心指标, 被广泛用于各类 SOTA 方法的比较. 然而, 这种单一的评价导向在工业实战中存在显著局限: 它倾向于被大量的背景像素主导, 导致对微小缺陷的漏检不敏感, 且无法反映模型在算力受限设备上的运行效率. 为全面评估基础模型与非基础模型在复杂工业场景下的真实性能, 本文构建了如表 2 所示的多维度评测体系, 建议从以下三个层面进行综合考量:

1) 缺陷感知维度: 针对细微缺陷, AUROC 往往掩盖了局部定位的失效, 因此对于分割任务, 应优先引入 PRO (per-region overlap) 指标以强调对不同大小缺陷区域的同等关注; 同时, 针对 3D 几何形变及逻辑异常等特殊任务, 建议补充 Chamfer distance 及多分类精度等专用指标.

2) 工业部署维度: 单纯的精度优势并不足以支撑落地, 针对产线实时性需求, 推理延迟、帧率 (frames per second, FPS) 与显存占用往往是决定模型是否可用的“一票否决”考量, 应将其纳入标准评估报告.

3) 场景适应维度: 场景适应的核心在于跨域泛化能力, 即从通用预训练域到特定工业域的迁移. 鉴于工业界对于新产线“冷启动”的迫切需求, 评估重

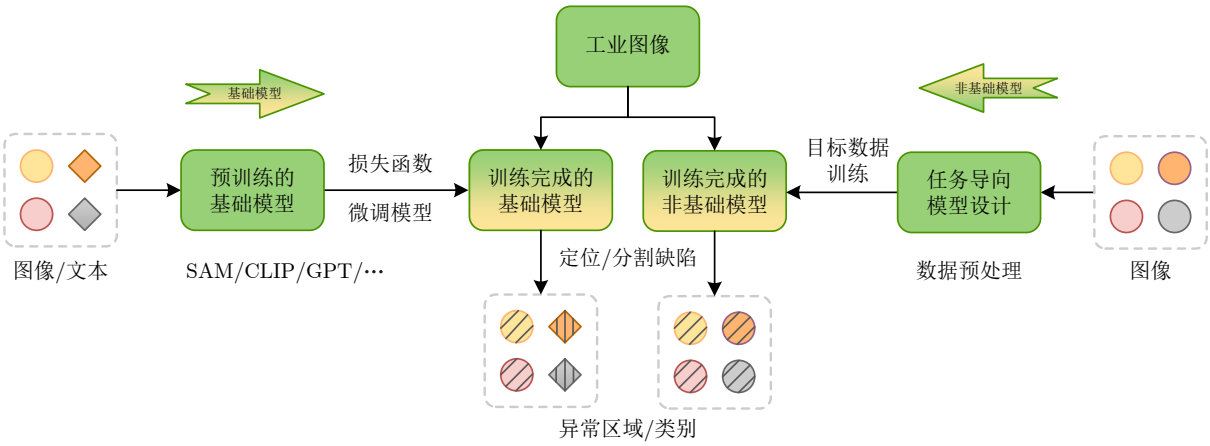


图 3　FM 与 NFM 方法的框架

Fig. 3　The framework of FM and NFM methods

表 1　主流工业缺陷检测数据集、核心挑战与典型应用场景汇总

Table 1　Summary of mainstream industrial defect detection datasets, core challenges and typical application scenes

主流数据集	场景领域	核心挑战	典型 FM 方法
MVTec AD[9]	通用元件/纹理	复杂纹理异常、微小缺陷分布	WinCLIP[22]
VisA[10]	复杂背景/多对象	多对象复杂背景、结构不规则	SAA+[17]
Real3D-AD[100]	精密工业品	高频几何形变、全视角无盲区	Group3AD[70]
MVTec 3D-AD[99]	3D 零部件	空间几何形变、高维空间建模	PointAD[42]

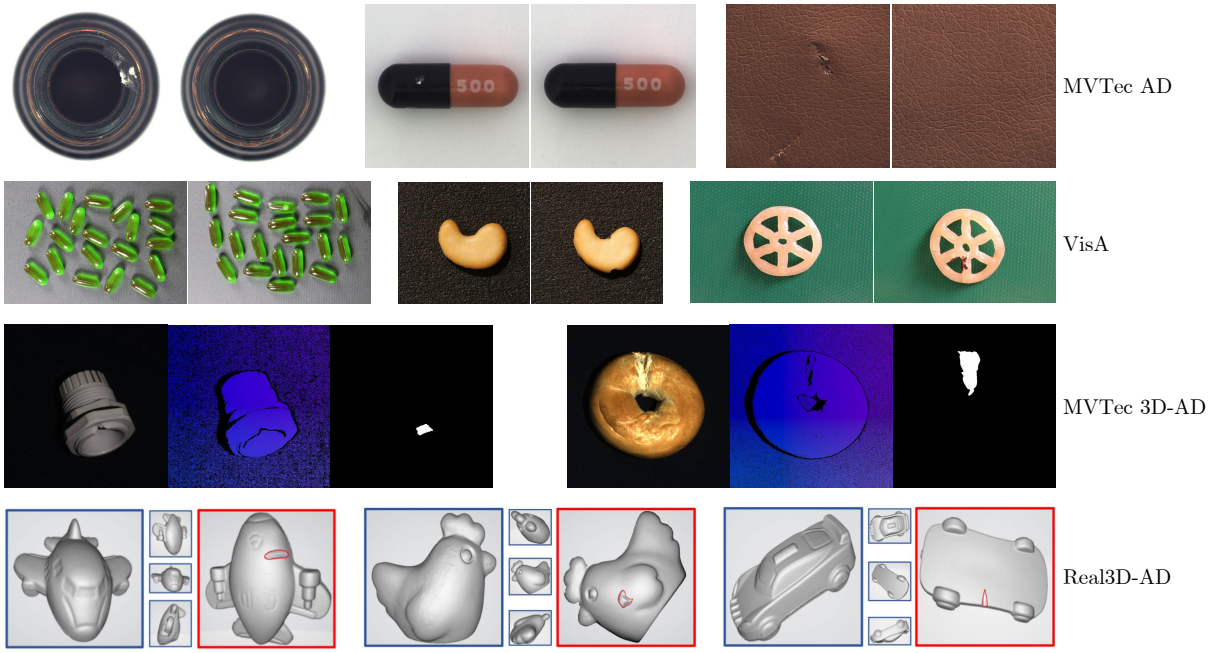


图 4 主流工业缺陷检测数据集与典型缺陷样本示例

Fig. 4 Mainstream industrial defect detection datasets and typical defect sample examples

表 2 多维度评测体系

Table 2 Multi-dimensional evaluation framework

评测维度	场景/任务	推荐指标
缺陷感知	2D 分类	I-AUROC、F1-max
	2D 分割	P-AUROC、PRO
	3D 检测	3D-PRO、Chamfer
	逻辑异常	Acc.、Conf. Matrix
工业部署	实时产线	FPS、Latency
	边缘设备	Params、Memory
场景适应	少样本	Performance Gain
	域迁移	Domain Adapt.、Acc.

点应在于模型能否利用极少样本迅速弥合域差异.

1.5.2 AUROC 性能

如表 3 所示, 以常用的 MVTec AD 数据集为例, 2D NFM 方法通常能达到接近 99% 的 AUROC 值, 表现最佳. 相比之下, 一些 2D FM 方法的 AUROC 值在 98% 左右. 在 3D 方法中, FM 和 NFM 的平均 AUROC 值均低于 95%. 这表明 2D 方法总体上优于 3D 方法, 且 NFM 目前在此背景下比 FM 更成熟.

1.5.3 推理时间

由于其庞大的参数量和复杂的模型结构, FM 方法通常需要更多的推理时间. 尽管 STLM 以 20 ms 的平均推理时间显著优化了推理效率, 但它仍落后于 DRAEM[71]、FastFlow[103] 和 PatchCore[104] 等具有更短推理时间的 NFM 方法. 3D 方法通常需要更长的推理时间; 然而, 一些方法, 如 SOWA[29], 仍表现出优异的推理速度, 速率为 16.84 it/s (相比之下, WinCLIP[22] 为 3.28 it/s, APRIL-GAN[30] 为 1.82 it/s). 与 BTF[63]、M3DM[101] 和 Reg 3D-AD[100] 相比, PointCore[68] 在 Real3D-AD[100] 上实现了最高的 AUROC 并且速度最快, 平均每个目标推理时间为 4.282 s (不包括 BTF[63] 的 2.19 s). Shape-Guided 方法[61] 的每个样本推理时间为 2.05 s, 优于 BTF[63].

1.5.4 计算复杂度

由于其庞大的模型规模, FM 方法的浮点运算次数 (floating-point operations, FLOPs) 通常高于 NFM 方法. 例如, STLM[16] 的 FLOPs 为 55 G. 采用 Transformer 和上采样特征图的 SAM-LAD[18] 的 FLOPs 为 54.7 G, 高于基于 CNN 的 NFM 方法, 如 AE[2] (5.0 G) 和 f-AnoGan[105] (7.7 G). 解决 FM 的计算复杂度已成为一个热门研究方向. 为推理效率而优化的 SimCLIP[26], 需要更少的 FLOPs (513.75 G), 低于 SOTA 的提示学习方法如 CoOp[106] (520.46 G) 和 Co-CoOp[107] (520.46 G), 同时保持相同的参数量. 然而, SimCLIP[26] 的 FLOPs 比 STLM[16] 和 SAM-LAD[18] 高出一个数量级, 因为 STLM[16] 使用了来自固定 SAM 教师的蒸馏, 而 SAM-LAD[18] 采用了 FeatUp[108] 的上采样因子; STLM[16] 和 SAM-LAD[18] 在推理过程中都不使用基础模型.

表 3 不同 NFM 与 FM 方法的简要总结与概览

Table 3 A brief summary and overview of different NFM and FM methods

类别	子类别	方法	描述	发表平台及年份	性能 (%)	数据集
非基础模型方法	2D 统计	SOFS[44]	引入异常先验图和混合正常 Dice 损失	IEEE TII 2025	93.3	MVTec AD
		PNI[45]	利用位置和邻域信息	ICCV 2023	99.6	
		REB[46]	减少领域和局部密度偏差	KBS 2024	99.5	
		BGAD[47]	通过拉近正常样本同时推开异常样本来强化决策边界	CVPR 2023	99.3	
		COAD[48]	通过受控的过拟合增强模型对异常的敏感性	ICLR 2025	99.9	
	2D 合成	GLASS[49]	基于高斯噪声和梯度上升的异常合成	ECCV 2024	99.9	
		AdaBLDM[50]	具有特征编辑功能的潜在扩散模型	arXiv 2024	—	
		RealNet[51]	强度可控的扩散异常合成	CVPR 2024	99.6	
		CAGEN[52]	文本引导的可控异常生成	ICASSP 2024	97.7	
		AnomalyXFusion[53]	用于增强样本保真度的多模态异常合成	arXiv 2024	99.2	
		AnomalyDiffusion[54]	空间异常嵌入, 自适应注意力重加权机制	AAAI 2024	99.2	
		DFMGAN[55]	在 StyleGAN2 中使用缺陷感知残差块	AAAI 2023	—	
		DeSTSeg[56]	去噪学生编码器–解码器, 自适应多级特征融合	CVPR 2023	98.6	
		CutSwap[57]	利用显著性指导以纳入语义线索	SIVP 2024	98.0	
		Split Training[58]	缓解过拟合问题的拆分训练策略	AAAI 2024	98.3	
		DFD[59]	具有双路径频率判别器的频域分析	KBS 2024	93.3	
		PBAS[60]	利用正常样本特征的紧凑分布指导特征级异常合成的方向	IEEE TCSVT 2025	99.8	
	2D RGB + 3D 点云	Shape-Guided[61]	用于颜色和形状异常定位的协同专家模型	ICML 2023	94.7	MVTec 3D-AD
		CPMF[62]	结合手工 PCD 描述符和预训练的 2D 神经网络	Pattern Recognition 2024	92.9	
		Back to the Feature[63]	具有 PatchCore 的手工 3D 表示	CVPR 2023	97.8	
		TransFusion[64]	利用基于透明度的扩散解决过度泛化和细节丢失问题	ECCV 2025	98.2	
		3DSR[65]	深度感知离散自动编码器和模拟深度生成过程	WACV 2024	97.8	
		M3DM[101]	一种混合融合方案, 以减少多模态特征间的干扰并鼓励特征交互	CVPR 2023	94.5	
		AST[66]	引入网络以补偿归一化流错误估计的似然性	WACV 2023	93.7	
	3D 生成	R3D-AD[67]	克服记忆库模块导致的低效和 MAE 不正确重建导致的低性能	ECCV 2024	73.4	Real3D-AD
		Reg3D-AD[100]	一种双特征表示方法, 以保留训练原型的局部和全局特征	NeurIPS 2023	70.4	Real3D-AD
		PointCore[68]	降低推理中的计算成本和错配干扰	arXiv 2024	82.9	Real3D-AD
		Uni-3DAD[69]	对无模型工业产品具有显著的适应性	Expert Syst. Appl. 2025	—	MVTec 3D-AD
		Group3AD[70]	通过组级别特征对比学习提高 3D 异常检测的分辨率和准确性	ACM MM 2024	75.1	Real3D-AD
基础模型方法	基于 2D SAM	ClipSAM[20]	具有多级提示的分层掩码细化	Neurocomputing 2025	92.3	MVTec AD
		UCAD[19]	使用 SAM 的基于结构的对比学习	AAAI 2024	93.0	
		SAM-LAD[18]	使用 SAM 获取查询和参考图像的对象掩码, 并提取对象特征进行匹配	KBS 2025	98.4	
		SAA+[17]	混合提示正则化	IEEE T-CYB 2025	—	
		STLM[16]	利用 SAM 作为教师指导学生网络	ACM TOMM 2025	98.3	
		SPT[21]	调整 SAM 以更好地理解图像中不同区域间关系	AAAI 2025	—	
	基于 2D CLIP	WinCLIP[22]	组合式提示集成, 参考关联方法	CVPR 2023	93.1	
		AnoCLIP[102]	局部感知视觉令牌, 领域感知提示, 测试时自适应方法	arXiv 2024	—	
		AnomalyCLIP[23]	对象无关的文本提示模板, 全局异常损失函数	ICLR 2024	—	
		AdaCLIP[24]	混合 (静态和动态) 可学习提示, 混合语义融合模块	ECCV 2024	—	
		VCP-CLIP[25]	视觉上下文提示模型	ECCV 2024	—	

表 3 不同 NFM 与 FM 方法的简要总结与概览 (续表)

Table 3 A brief summary and overview of different NFM and FM methods (continued table)

类别	子类别	方法	描述	发表平台及年份	性能 (%)	数据集
基础模型方法	基于 2D CLIP	SimCLIP[26]	多层次视觉适配器, 隐式提示学习, 先验感知优化算法	ACM MM 2024	95.3	MVTec AD
		CLIP-AD[27]	文本提示分布, 通过线性层促进对齐	IJCAI 2024 Workshop	—	
		CLIP-FSAC[28]	两阶段训练策略, 视觉驱动的文本特征, 融合−文本匹配任务	IJCAI 2024	95.5	
		ClipSAM[20]	CLIP 与 SAM 协作, 统一的多尺度跨模态交互, 多级掩码细化	Neurocomputing 2025	92.3	
		SOWA[29]	层次化冻结窗口自注意力, 双重可学习提示	ACM MM 2024	—	
		SAA+[17]	混合提示, 领域专家知识和目标图像上下文	IEEE T-CYB 2025	—	
		APRIL-GAN[30]	采用状态和模板集成组合, 基于记忆库的方法	arXiv 2023	92.0	
		PromptAD[31]	提示学习, 语义串联, 显式异常边界	CVPR 2024	94.6	
		FiLo[32]	细粒度描述, 可学习向量, 位置增强的高质量定位方法	ACM MM 2024	—	
		Dual-Image Enhanced CLIP[33]	双图像特征增强, 使用伪异常合成的测试时自适应	arXiv 2024	—	
	基于 2D GPT	AnomalyGPT[34]	轻量级且基于视觉−文本特征匹配的解码器, 提示嵌入	AAAI 2024	94.1	
		Myriad[37]	应用视觉专家, 视觉专家分词器	arXiv 2025	94.1	
		ALFA[38]	运行时提示自适应策略, 细粒度对齐器	ACM MM 2024	94.5	
		GPT-4V-AD[39]	视觉问答范式, 颗粒化区域划分, 提示设计, Text2Segmentation 方法	IJCAI 2024 Workshop	—	
		Customizable-VLM[36]	通过提示将专家知识作为外部记忆集成, 以增强基础模型	IEEE CSCWD 2025	82.9	
		LogiCode[40]	使用 LLM 提取图像逻辑并生成代码以进行逻辑异常检测	IEEE TASE 2025	—	
	基于 3D CLIP	CLIP3D-AD[41]	无需记忆库和大量训练样本, 解决少样本异常分类和分割	ACM MM 2024	—	MVTec 3D-AD
		PointAD[42]	混合表示学习框架	NeurIPS 2024	97.2	
		M3DM-NR[43]	使用疑似异常图实现去噪	IEEE TPAMI 2025	94.5	

基于以上分析, FM 方法凭借其强大的多模态能力和庞大的参数量, 在复杂的工业检测和跨领域应用中展现出强大潜力. 然而, 在数据稀缺场景下的过拟合以及推理效率等挑战仍然是瓶颈, 尤其是在处理 3D 高维数据时, 仍有充足的探索空间. 另一方面, NFM 方法依赖于有针对性的特征提取和高效计算, 使其在实时推理和计算资源受限的工业场景中更具优势.

2 非基础模型方法

在基础模型兴起之前, 轻量级模型长期主导着视觉工业缺陷检测领域. 这类模型通常不依赖海量预训练知识, 而是通过优化的架构、特征提取技术和计算效率来提高检测精度. 这些方法特别适用于需要快速推理的资源受限场景, 为在工业环境中进行实际部署提供了显著优势. 本节将讨论当前轻量级模型方法的四种主要类型: 统计学方法、异常合成策略、结合 2D RGB 与 3D 点云的方法以及 3D 生成方法. 表 3 展示了基于 NFM 的不同方法的总结和概述. 图 5 呈现了 NFM 发展的主要时间线.

2.1 统计学方法

统计学方法为提高异常检测模型性能提供了有效的理论基础. Zhang 等[44] 对传统的 Dice 损失进行了优化, 提出了一种混合正常 Dice 损失. 该损失函数在模型预测假阳性时施加较大惩罚, 从而优先防止此类错误预测. Bae 等[45] 提出了 PNI 算法, 以解决位置和邻域信息对正常特征分布的影响. 该算法采用基于邻域特征的条件概率, 使用多层感知器 (multilayer perceptron, MLP) 网络对正常特征的分布进行建模. 此外, 该方法通过在每个位置构建代表性特征的直方图来有效捕获位置信息. Lyu 等[46] 考虑了局部特征密度的变化, 并提出了局部密度 k 近邻 (local density k-nearest neighbors, LDKNN) 方法, 以减少 patch-level 特征中的密度偏差. COAD[48] 将过拟合视为一种可控机制, 通过受控的过拟合来增强对异常的敏感性. 它引入了畸变保留商 (aberrance retention quotient, ARQ) 度量, 以精确量化过拟合程度, 从而确定一个最佳的“黄金过拟合区间”(最佳 ARQ) 来优化异常检测性能. BGAD[47] 设计了一种边界引导的半推拉 (boundary-guided

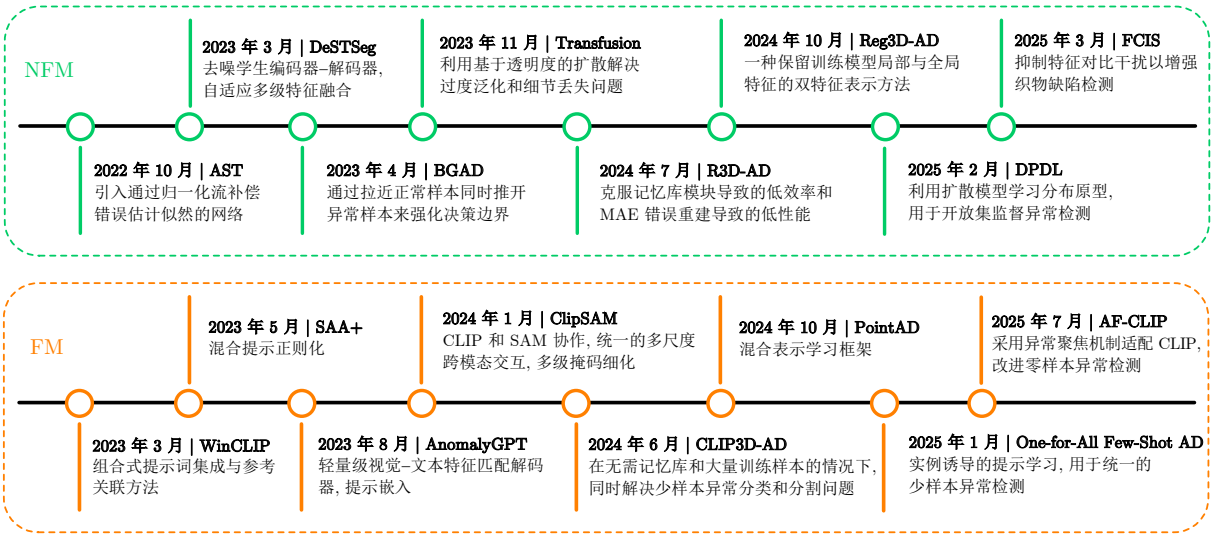


图 5　NFM 与 FM 模型发展中的代表性方法

Fig. 5　Representative methods along the development of NFM and FM model

semi-push-pull, BG-SPP) 损失. 首先, 它通过学习正常样本特征分布来生成一个显式边界. 在此基础上, 它拉近正常样本, 同时推远异常样本, 从而强化决策边界. BGAD 使模型仅用少量异常样本就能有效区分已见和未见的异常. 然而, 异常样本的稀缺性仍可能导致特征学习效率低下, BGAD 并未完全解决这一关键问题.

总结以上统计学方法可见, 其核心逻辑在于通过对正常特征分布的显式建模 (如 PNI[45] 的邻域信息采样) 或决策边界的精细调控 (如 BGAD[47] 的半推拉损失), 构建稳健的异常判别准则. 这类方法的主要优点在于计算开销低、模型解释性强, 且在数据分布相对稳定的受控工业环境下表现优异; 其局限性在于, 面对高维、复杂的背景纹理时, 单一的统计分布往往难以捕捉细微的模式偏移. 由此可见, 如何将传统的统计分布先验与自监督表征学习相结合 (如 COAD[48] 的受控过拟合机制), 以提升模型在复杂场景下的灵敏度, 是该方向的重要演进思路.

2.2　异常合成策略

异常合成策略旨在通过生成多样化且逼真的异常样本来增强异常检测模型的性能. 广义上, 异常合成策略可分为以下几类.

2.2.1　生成模型

基于去噪扩散概率模型 (denoising diffusion probabilistic model, DDPM), Zhang 等[51] 提出在逆向扩散过程中引入额外噪声, 以控制生成的异常样本的强度. 此外, Hu 等[53] 聚合了多种模态特征, 并将它们集成到一个统一的嵌入空间中, 优化了模态对齐, 进而通过基于扩散步骤自适应调整嵌入, 来促进受控生成. Jiang 等[52] 通过使用文本提示和二元掩码微调 ControlNet 模型, 增强了异常生成的可控性. Hu 等[54] 提出了 AnomalyDiffusion, 使用潜在扩散模型 (latent diffusion model, LDM) 来生成异常图像. 该模型将空间异常嵌入与自适应注意力机制相结合, 以改善生成的异常与其对应掩码之间的对齐. Li 等[50] 在混合潜在扩散模型 (blended latent diffusion model, BLDM)[109] 的基础上进行了多项创新: 首先, 设计了一种新颖的“缺陷三元图”(defect trimap) 来描绘生成图像中的目标掩码和缺陷区域; 其次, 在潜在空间和像素空间中引入了级联的“编辑”阶段, 以确保结构连贯性和细节保真度; 此外, 还提出了图像编码器的在线自适应, 以进一步提高图像质量. Duan 等[55] 训练了一个数据高效的 Style-GAN2 作为正常图像的骨干网络, 随后通过添加缺陷感知残差块来生成缺陷掩码, 并操纵掩码区域内的特征, 生成新的缺陷图像.

2.2.2　数据增强技术

基于正常特征, Chen 等[49] 通过梯度上升和截断投影引导高斯噪声, 在正常点周围合成弱异常. 此外, 该研究使用 Perlin 噪声创建二元掩码, 并将其与外部纹理结合, 以合成距离正常点更远的强异常. Zhang 等[56] 使用 Perlin 噪声生成异常图像, 并将其作为学生网络的输入. 通过训练学生网络去除合成的异常噪声, 有效增强了学生网络表示异常样本特征的能力, 从而提高了教师–学生框架在异常检测中的性能. Qin 等[57] 引入语义信息来生成异常样本. 该方法利用 LayerCAM 从图像中提取显著特

征, 并进行聚类以识别最重要的区域, 随后通过选择相似的 patch 对并交换其位置, 以这种方式生成的负样本更微妙但更逼真. Lin 等[58] 开发了一个全面的异常模拟框架, 结合了透明和不透明异常的重建策略. 通过使用选择性增强和基于分割的训练策略, 该框架成功解决了异常生成多样性、重建质量和过拟合的挑战. Bai 等[59] 发现小异常在频域中变得更明显. 通过将空间图像转换为多频率表示, 判别器学习正常图像和伪异常之间的联合表示, 从而提高了少样本异常检测的性能. PBAS[60] 首先学习一个带中心约束的正常样本特征的紧凑分布, 作为近似的决策边界, 用于指导特征级异常合成的方向; 随后, 在合成的异常和正常特征之间进行二元分类, 进一步优化决策边界, 以确保合成的异常不与正常样本重叠.

从早期的 Perlin 噪声增强到近期基于潜在扩散模型 (LDM) 的可控生成 (如 AnomalyDiffusion[54] 和 RealNet[51]), 异常合成策略显著缓解了工业领域“正负样本极度不均衡”的瓶颈. 这类方法的技术优势在于通过人为构造的伪异常“锐化”了模型对正常特征边界的认知, 使无监督任务转化为准监督任务. 然而, 其核心挑战在于合成样本的保真度与物理真实性之间的“虚实鸿沟”: 若合成缺陷与真实分布偏离过大, 可能导致模型在实际应用中出现严重的误报. 因此, 未来研究正趋向于利用文本引导 (如 CAGEN[52]) 或梯度引导 (如 GLASS[49]) 等策略来生成更具任务针对性的高质量缺陷样本.

2.3 结合 2D RGB 与 3D 点云的方法

结合 2D RGB 与 3D 点云的方法, 通过融合两种模态的特征——2D RGB 图像丰富的颜色和纹理特征以及 3D 点云提供的空间和几何信息, 提升了传统方法因单一模态数据缺失而受限的检测能力.

Chu 等[61] 提出了一个形状引导的基于专家的学习框架, 该框架采用两个专家模型分别检测 3D 结构和颜色外观中的异常, 并使用双记忆库和形状引导的推理方法来定位测试样本中的缺陷. 该模型利用神经隐式函数 (neural implicit functions, NIFs)[110] 来表示局部形状, 并通过符号距离场细化点云的复杂结构, 实现了点级异常预测. 这显著提高了异常定位的准确性, 同时降低了计算和内存成本. CPMF[62] 通过将点云投影到 2D 来生成伪 2D 表示, 并使用预训练的 2D 神经网络提取语义特征. 这些特征补充了从手工制作的点云描述符中提取的 3D 局部特征, 并通过特征对齐和融合模块统一为全局语义和局部几何的点云表示. Horwitz 等[63] 指出, 3D 方法目前被 2D 方法超越, 并提出了一个结合旋转不变的手工特征表示与基于深度学习的颜色特征的解决方案, 以提高 3D 异常检测性能. TransFusion[64] 通过迭代增加异常区域的透明度, 并逐渐用正常外观替换它们, 同时保留非异常区域的正常外观, 解决了过度泛化和细节丢失的问题. Zavrtanik 等[65] 引入了 3DSR, 其中 DADA 学习了一个通用的离散潜在空间, 联合建模 RGB 和深度数据. 3DSR 在 DADA 学习到的特征空间中执行判别式异常检测. M3DM[101] 为 RGB、3D 和融合特征构建了三个独立的记忆库, 并通过决策层融合 (decision layer fusion, DLF) 综合考虑来自这些记忆库的决策来进行异常检测. 为了更好地将 3D 点云特征与 2D RGB 特征对齐, 引入了点特征对齐 (point feature alignment, PFA). Rudolph 等[66] 提出了非对称学生–教师 (asymmetric student-teacher, AST) 网络, 采用归一化流进行密度估计作为教师网络, 并使用传统的前馈网络作为学生网络, 解决了先前方法中因学生和教师架构相似而导致的异常数据输出差异不足的问题.

现有工作显示, 跨模态融合方法通过引入 3D 几何信息 (如点云法向量、局部形状描述符), 有效弥补了传统 2D 方法在处理深度异常 (如凹坑、缺失) 时的短板. 无论是 M3DM[101] 的决策层融合, 还是 Shape-Guided[61] 的双专家协同, 其主要优点是显著提升了检测的鲁棒性, 能够识别单一模态下不可见的缺陷. 但其代价在于特征对齐过程引入了额外的计算复杂度, 且对多传感器标定的精度要求极高. 未来的演进逻辑在于开发更轻量化的跨模态对齐机制 (如 AST[66]), 在保证实时性的前提下, 实现颜色纹理与几何结构的深度表征对齐.

2.4 3D 生成方法

3D 生成方法使用生成模型来重建正常样本或缺失区域, 旨在减少计算开销并提高模型鲁棒性, 尤其有助于解决无模型产品 (model-free products) 和难以识别缺失区域的挑战.

Zhou 等[67] 采用基于扩散模型的数据分布转换, 在输入中完全掩盖异常几何形状, 在逆向扩散过程中学习逐步位移, 并显式控制异常形状的重建. 此外, 该研究还提出了一种称为 Patch-Gen 的 3D 异常模拟策略, 旨在生成逼真的缺陷形状, 并弥合训练和测试数据之间的差距. R3D-AD[67] 解决了 3D 异常检测中与计算存储开销和检测未掩盖区域异常相关的挑战. PointCore[68] 仅需一个记忆库来存储局部 (坐标) 和全局 (PointMAE) 表示, 为这些局部–全局特征分配不同优先级, 以降低计算成本并减轻

推理过程中的特征错配干扰. 使用基于排序的归一化方法来消除不同异常分数之间的分布差异, 同时应用迭代最近点 (iterative closest point, ICP) 算法来局部优化点云配准结果, 增强了决策的鲁棒性. Liu 等[69] 提出了一个双分支结构: 基于特征的分支和基于重建的分支分别检测表面缺陷和缺失区域. 其中, 重建分支首次引入了生成对抗网络反演 (GAN-Inversion) 来生成与输入最相似的正常样本, 从而减少误报. Zhu 等[70] 引入了簇间均匀性网络 (inter-cluster uniformity network, IUN) 和簇内对齐网络 (intra-cluster alignment network, IAN), 二者分别在特征空间中实现了簇间分散和簇内对齐, 增强了特征的均匀性和一致性. 此外, 自适应组中心选择设计专注于具有潜在问题的区域, 尤其是那些局部几何变化显著的位置, 从而提高了模型的敏感性.

3D 生成式方法代表了工业检测从“特征工程”向“分布重建”的转变. 通过学习正常样本的潜在分布并执行异常重建 (如 R3D-AD[67]), 这类方法在处理“无模型”工业产品时表现出极强的自适应性. 其显著优势在于能够直接定位复杂的几何畸变, 而无需依赖庞大的存储库. 然而, 原生 3D 重建 (如基于 GAN 或扩散模型) 往往面临计算效率低下的瓶颈, 难以直接应用于高速产线. 因此, 近期如 Point-Core[68] 等工作通过引入高效的局部–全局特征记忆方案, 试图在检测精度、重建质量与推理效率之间寻找新的平衡点.

3 基础模型方法

上述 NFM 方法在特定工业场景下表现优异, 但受限于数据规模和泛化能力, 在面对未知缺陷时往往力不从心. 近年来, 视觉–语言模型在异常检测方面显示出显著优势. 这些模型通过有效结合视觉信息和语言线索, 能够更好地理解和描述图像中的复杂特征. 与传统的异常检测方法相比, 视觉–语言模型能够利用丰富的上下文信息, 减少对人工标注和领域知识的依赖, 从而在零样本和少样本场景下实现更准确的检测与定位. 本部分将介绍基于视觉–语言模型及相关基础模型的四类主要方法: 基于 2D SAM、2D CLIP、2D GPT 和 3D CLIP 的方法, 并进一步讨论其他 3D 基础模型在工业缺陷检测中的应用与挑战. 表 3 对基于 FM 的不同方法也进行了总结和概述. 图 5 展示了 FM 发展过程中最重要和最流行工作的时间线.

3.1 基于 2D SAM 的方法

作为一种基础模型, SAM[14, 111–112] 具有提取高质量分割掩码的强大能力. 通过利用大规模预训练数据, 它可以在各种场景下对任何目标执行实例分割, 而无需特定任务的训练. 因此, 基于 SAM 的方法在零样本异常检测任务中表现出良好性能. Cao 等[17] 利用 SAM 和级联提示引导的目标检测模型[113] 构建了一个基础框架, 即 SAA (segment any anomaly). SAA 通过简单的语言提示 (如“缺陷”或“异常”) 生成初步的异常区域, 随后进行细化处理. 他们进一步引入了一种混合提示正则化技术, 将框架增强为 SAA+ (segment any anomaly+). 为了更好地处理 SAM 生成的掩码, Hou 等[20] 提出了 ClipSAM, 该方法结合了 CLIP 和 SAM 的优势. ClipSAM 使用 CLIP 的语义理解能力进行异常定位和粗略分割, 然后将结果作为 SAM 的提示约束来细化分割结果. 在 SAM-LAD 中, Peng 等[18] 引入 SAM 来获取查询图像和参考图像的目标掩码. 每个目标掩码与整个图像的特征图相乘, 以获得目标特征图, 然后用于目标匹配. 在此基础上, 提出了一种异常测量模型 (anomaly measurement model, AMM) 来检测逻辑和结构异常. 在 UCAD 中, Liu 等[19] 使用 SAM 通过结构化对比学习 (structured contrastive learning, SCL) 来增强异常检测. 通过将 SAM 生成的掩码视为结构, 同一掩码内的特征被拉近, 而其他特征被推远, 这改善了用于异常检测的特征表示. Li 等[16] 提出了一个 SAM 引导的双流轻量级模型 (SAM-guided two-stream lightweight model, STLM), 优先考虑效率和移动兼容性. 一个流提取特征以区分正常与异常区域, 另一个流重建无异常图像以增强区分度. 通过共享的掩码解码器和特征聚合, STLM 提供了精确的异常图. 在 SPT 中, Yang 等[21] 引入了一种视觉关系感知适配器 (VRA-Adapter), 以帮助 SAM 更好地理解图像中不同区域之间的关系, 增强 SAM 对异常模式的细粒度理解.

综上所述, 基于 SAM 的方法核心在于利用其“万物皆可分割”的空间先验. 从早期直接利用语言提示触发分割的 SAA+[17], 到近期通过结构化对比学习[19] 或轻量级双流架构[16] 实现任务特化, 研究重点已从简单的零样本应用演进为如何赋予 SAM 工业级的语义辨识力. 虽然 SAM 能够提供高保真的异常边界, 这对量化缺陷几何参数至关重要, 但其原生权重缺乏对特定缺陷类别的感知. 由此可见, 未来的技术突破点在于开发更高效的视觉关系适配器 (如 SPT[21]), 以在保持分割精度的同时, 实现语义特征与空间结构的深度解耦.

3.2 基于 2D CLIP 的方法

基于 CLIP 的方法使用大规模预训练的视觉–

语言模型, 将图像编码与文本线索相结合, 从而加强视觉特征与语言信息之间的关系, 并在零样本和少样本场景中表现尤为出色.

3.2.1 文本提示

Jeong 等[22] 提出了一种称为 WinCLIP 的窗口级异常检测方法, 通过组合式提示集成和多尺度特征提取实现零样本异常分割[11]. Zhou 等[23] 设计了一种目标无关的文本提示模板, 通过学习通用的正常与异常提示[107], 并结合全局和局部上下文信息, 来捕获图像中的异常区域. APRIL-GAN[30] 将文本提示集成策略与线性层[114–115] 相结合, 以提高零样本和少样本异常检测的性能. SimCLIP[26] 通过引入具有隐式提示调整的多级视觉适配器, 减少了对手工设计提示的依赖. Qu 等[25] 使用视觉上下文提示来激活 CLIP 的异常语义能力, 无需针对特定产品的提示. Cao 等[24] 提出通过结合静态和动态的可学习提示来优化零样本异常检测性能. Li 等[31] 提出通过语义连接将正常提示转换为异常提示, 为提示学习构建大量负样本.

3.2.2 细粒度对齐

Gu 等[32] 通过细粒度的局部描述和优化的视觉编码器提高了异常定位的准确性. Zhang 等[33] 提出使用双图像作为视觉参考来提高异常检测的准确性. Zuo 等[28] 提出通过两阶段训练策略和图文跨注意力模块, 可以有效提高少样本异常分类的性能. Chen 等[27] 提出通过代表性向量选择和分阶段的双路径模型来实现细粒度对齐. SAA+[17] 通过提示引导的目标检测和细化技术实现更精确的异常定位. ClipSAM[20] 结合了 CLIP 和 SAM[116] 的优势, 改善了零样本异常分割, 并利用 CLIP 的语义理解能力进行异常定位和细粒度分割[117]. Hu 等[29] 提出了一种层次化冻结窗口自注意力机制[118], 通过结合多级适配器来捕获不同层次的特征, 以实现细粒度定位[119].

由此可见, 基于 CLIP 的工业检测方法正经历从“手工提示工程”向“域自适应对齐”的范式转变. 虽然 WinCLIP[22] 奠定了零样本检测的基础, 但近期如 SimCLIP[26] 和 SOWA[29] 等工作表明, 引入多级视觉适配器与层次化自注意力机制, 对于弥合自然图像语义与微观工业纹理之间的域分布鸿沟至关重要. 分析发现, 文本驱动类方法虽然泛化能力强, 但在处理定位精度要求极高的像素级任务时, 仍需借助异常合成或测试时自适应 (test-time adaptation, TTA) 策略来强化其判别力. 未来的研究重心将趋向于在不破坏 CLIP 原始语义空间的前提下, 实现全局语义与局部细粒度特征的精细化对齐.

3.3 基于 2D GPT 的方法

基于 GPT 的方法利用大语言模型在自然语言理解和自适应学习方面的优势, 通过生成具体的文本描述来支持异常检测任务, 同时能够根据实时输入调整其处理策略, 以适应不断变化的检测环境和异常类型.

AnomalyGPT[34] 率先将大型视觉–语言模型 (large vision-language models, LVLMs)[120–121] 应用于异常检测任务, 并支持多轮对话, 展示了出色的少样本学习能力. 随后, Cao 等[35] 和 Xu 等[36] 探索了如何使用 LVLMs 进行跨各种领域的一般异常检测任务. 同时, 他们将来自不同模态的信息 (如领域知识、类别上下文和参考图像) 作为提示, 以提高 LVLMs 的检测性能. Myriad[37] 通过将视觉专家与大规模多模态模型相结合, 减少了对标记数据的依赖. Zhu 等[38] 提出了一种运行时提示自适应策略来生成信息丰富的异常提示, 结合细粒度对齐器, 可以实现精确的异常定位, 并增强模型的动态适应性, 使其在多样化的工业场景中更加有用. Zhang 等[39] 探索了面向视觉问答 (visual question answering, VQA) 的 GPT-4 with Vision 在异常检测中的潜力, 引入了一个 GPT-4V-AD 框架, 该框架集成了颗粒化区域划分、提示设计和 Text2Segmentation (文本到分割). LogiCode[40] 充分利用了大语言模型的推理能力. 该方法从正常图像中提取逻辑关系, 并生成可执行的 Python 代码, 以自动检测测试图像中的逻辑异常. 该系统还提供异常的具体位置和详细解释. LogiCode 的创新框架突破了传统的异常检测方法, 提供了一种更智能的解决方案.

从 AnomalyGPT[34] 的多轮对话能力到 LogiCode[40] 的逻辑代码生成, 基于 GPT 的方法将缺陷检测从单纯的视觉感知提升到了逻辑推理的高度. 这类方法的独特优势在于其强大的上下文学习能力, 能通过专家提示 (如 Customizable-VLM[36]) 快速适配新场景. 然而, 如何克服大语言模型的“幻觉”现象并降低其在边缘设备上的推理开销, 仍是目前的瓶颈. 未来的演进趋势可能在于构建“视觉专家–逻辑推理”的协同范式, 即利用轻量化检测器确保感知精度, 而利用大模型实现缺陷成因分析与质量追溯的闭环.

3.4 基于 3D CLIP 的方法

与主要处理 RGB 图像的传统 2D CLIP 模型相比, 3D CLIP 处理的是三维数据——点云, 其包含更复杂的空间结构和几何信息. 目前, 3D 异常检测的研究相比其 2D 对应领域尚不发达, 这很大程

度上是因为 CLIP 最初是在 2D RGB 图像与文本对上训练的. 因此, 3D CLIP 在将点云数据与图像集成到多模态框架中时面临挑战. 近期的研究在克服 3D 数据处理中的模态鸿沟方面取得了显著进展, 采用了多模态噪声降低、3D 数据的多视图处理与融合, 以及集成零样本学习等技术来提高性能.

Zuo 等[41] 提出了一种多视图融合模块, 集成了来自不同视角的 2D 图像特征, 从而增强了点云数据的表示能力, 并克服了直接处理 3D 点云时由模态差异带来的挑战. PointAD[42] 通过将 3D 点云从多个视图渲染为 2D 图像[122], 然后通过混合表示学习联合优化 2D 和 3D 特征, 实现了零样本 3D 异常检测. M3DM-NR[43] 通过三阶段多模态噪声去除方法显著提高了数据质量并减少了噪声干扰; 具体而言, 其利用预训练的 CLIP 和 Point-Bind 模型, 并采用多尺度特征比较和加权来提高训练样本的质量, 增加整体数据纯度.

针对 3D 数据的处理, 目前的文献呈现出从“多视图投影”向“原生几何表征”过渡的特征. 虽然 PointAD[42] 利用 CLIP 处理渲染后的 2D 图像, 能有效迁移成熟的视觉知识, 但投影过程不可避免地造成了空间邻域信息的丢失. 相比之下, M3DM-NR[43] 等工作展示了多模态特征对齐在提升数据纯度方面的潜力. 目前的挑战在于, 基础模型严重缺乏针对精密零件 CAD 模型或微米级几何畸变的预训练. 因此, 探索原生 3D 基础模型与工业特定几何特征的对齐机制, 是实现高鲁棒性 3D 缺陷检测的关键路径.

3.5 其他 3D 基础模型应用与挑战

目前基于 3D CLIP 的跨模态对齐方法在零样本和少样本 3D 异常检测中占据主流地位, 但学术界近年来也开始积极探索将 SAM 和 LLM 扩展至 3D 感知领域. 例如, Point-SAM[123] 和 SA3D[124] 尝试将 SAM 强大的 2D 分割能力迁移至 3D 点云; Point-Bind[125]、Point-LLM[126] 及 3D-LLM[127] 则致力于通过对齐点云编码器与大语言模型, 实现对 3D 场景的语义理解和问答.

然而, 在工业缺陷检测领域, 上述模型尚未得到广泛的实际应用. 这主要是因为工业场景对检测精度、推理速度和数据分布有着特殊要求, 直接迁移通用 3D 基础模型面临以下严峻挑战:

1) 细粒度感知的缺失与分辨率冲突. 现有的 3D-LLM 多基于室内场景 (如 ScanNet[86]) 或通用物体数据集预训练, 其训练目标侧重于生成高层级的语义描述 (如“这是一把椅子”). 相比之下, 工业缺陷检测通常需要识别微米级的细微特征 (如“表面划痕”或“微小凹坑”)[99–100]. 通用 3D 大模型往往缺乏这种细粒度的空间几何感知能力, 难以捕捉工业级的微小形变异常.

2) 2D 到 3D 的投影一致性难题 (针对 SAM 类方法). 将 SAM 应用于 3D 点云[123–124] 通常依赖多视图投影策略, 即先将 3D 数据投影为 2D 图像进行分割, 再反投影回 3D 空间. 在工业产品复杂的拓扑结构和自遮挡情况下, 不同视角的分割掩码往往存在冲突, 导致 3D 边界模糊或不一致. 此外, 这种“渲染−分割−融合”的流水线引入了巨大的计算开销, 难以满足工业产线毫秒级的实时性需求.

3) 领域数据的模态鸿沟. 目前的 3D 基础模型极度缺乏工业 CAD 模型或点云数据的预训练[42]. 工业点云具有高密度、高精度的特点, 与激光雷达或 RGB-D 相机采集的稀疏、含噪点云[86] 在数据分布上存在显著差异. 这种严重的域偏移导致通用 3D 模型在直接应用于工业样本时, 泛化性能大幅下降[42–43].

综上所述, 虽然 SAM 和 LLM 在通用 3D 视觉中展现了潜力, 但要在工业缺陷检测中落地, 仍需在轻量化 3D 适配器设计、细粒度几何特征对齐以及工业 3D 预训练微调等方面进行深入探索.

4 借鉴成熟 NFM 方法增强 FM 在工业缺陷检测中的应用

FM 方法和 NFM 方法并非两种完全独立的范式, 实际上, 当前许多先进的 FM 方法都有 NFM 的影子, 是 FM 与 NFM 方法原理的战略性融合. 这种融合的本质, 是将 NFM 方法的核心机制 (如对比学习、知识蒸馏等) 从其原始的、通常较为轻量的网络架构中解耦, 并将其作为一种优化策略重塑并应用于更强大的 FM 特征提取器之上. 在此范式中, FM 提供高质量、高泛化性的特征表示, 而 NFM 的策略则精确地引导这些特征服务于特定的异常检测目标. 这种协同作用不仅显著提升了检测的精度与鲁棒性, 更实现了 FM 的通用知识与 NFM 的任务特异性优势互补, 从而在性能与效率之间取得了更优的平衡. 如图 6 所示, 本文系统总结了 SAM、CLIP 和 GPT 三类基础模型与 NFM 方法的协同增强嵌入机制; 表 4 则进一步讨论了几种代表性的增强方法及其性能.

4.1 SAM: 利用空间先验优化学习目标

SAM 强大的零样本分割能力使其成为一个理想的空间先验信息来源.

在 UCAD 中, Liu 等[19] 利用 SAM 为经典的对

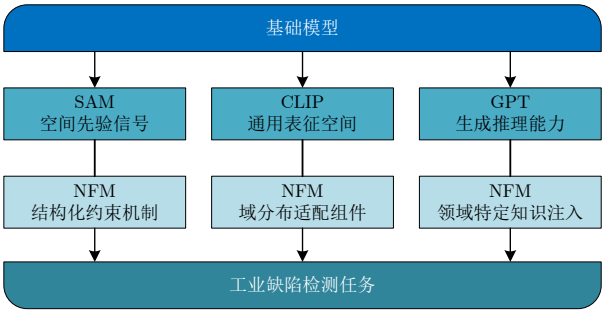


图 6 FM 与 NFM 策略的协同增强范式

Fig. 6 Synergy enhancement paradigm of FM and NFM strategies

比学习框架赋能. 其核心逻辑并非简单地套用 SAM 生成的分割掩码, 而是将其作为一种结构化先验引入特征学习过程. 具体而言, 该方法将 SAM 编码器的多级特征投影到统一的度量空间, 并利用高保真掩码为像素特征的正负样本定义提供结构化依据: 属于同一掩码区域的特征定义为正样本对, 不同区域的特征则互为负样本对. 通过这种方式, 传统的对比学习目标被成功转化为 FM 特征空间内的结构化对齐约束, 其对比损失定义为:

$$L_{cont} = -\log \frac{\exp\left(\frac{f_i \cdot f^+}{\tau}\right)}{\sum_j \exp\left(\frac{f_i \cdot f_j}{\tau}\right)} \tag{1}$$

其中, f_i 表示当前的锚点像素特征; f^+ 表示与 f_i 处于同一掩码区域内的正样本特征; f_j 表示归一化项中参与对比的所有样本特征; τ 是用于缩放特征点积相似度的温度超参数. 这种机制显著增强了模型对微小纹理偏移的辨识力, 使其得以在更具语义一致性的空间结构上进行优化. 该策略的有效性在性能指标上得到充分验证: 像素级 AUROC (P-AUROC) 和图像级 AUROC (I-AUROC) 分别从基线的 69.30% 和 18.30% 大幅提升至 93.00% 和 45.60%, 实现了显著的性能跃升.

SAA+[17] 则直接利用 SAM 强大的提示驱动分割能力, 并通过注入领域知识对其进行优化. 该方法引入了一种知识驱动策略, 将领域专家知识 (例如“划痕”、“污渍”) 编码为丰富的文本提示. 然后将这些提示与从图像中提取的视觉上下文相结合, 形

表 4 基础模型增强方法及其在 MVTec AD 上的性能比较 (%)

Table 4 Comparison of foundation model enhancement methods and their performance on MVTec AD (%)

FM 类型	FM 方法	增强方法	与 NFM 方法的联系	性能 (MVTec)			
				指标	基线	增强后	提升
SAM	UCAD	由高保真 SAM 掩码引导的对比学习	对比学习 (BGAD、Group3AD): 利用对比学习框架, 使用 SAM 掩码定义正负样本	P-AUROC I-AUROC	69.30 18.30	93.00 45.60	↑23.70 ↑27.30
	SAA+	使用专家知识和图像上下文的混合提示正则化	知识驱动: 将领域专家先验知识编码到混合提示中以指导基础模型	F_p F_r	30.95 29.17	34.85 34.07	↑3.90 ↑4.90
	STLM	SAM 引导的双流轻量级架构	学生–教师架构 (AST、DeSTSeg)/知识重建 (R3D-AD): 一个流基于学生–教师架构生成判别性特征; 另一个流重建无异常图像	P-AUROC I-AUROC	—	98.26 99.05	—
CLIP	AnoCLIP	使用合成噪声扰动的测试时自适应	异常合成 (GLASS、CAGEN): 合成数据用于轻量级适配器的在线优化	P-AUROC I-AUROC	88.90 —	90.60 —	↑1.70 —
	Dual-Image Enhanced CLIP	使用合成异常的测试时自适应	异常合成: 类似于 AnoCLIP	P-AUROC I-AUROC	85.30 91.60	92.80 93.20	↑7.50 ↑1.60
	SimCLIP	多级视觉适配器与隐式提示学习	适配器/提示学习 (SimCLIP): 插入轻量级模块对齐工业纹理与预训练分布的域差异	P-AUROC I-AUROC	—	95.60 95.30	—
	WinCLIP	组合式提示集成与参考关联方法	记忆库机制 (PatchCore): 引入外部记忆库存储正常特征以辅助少样本决策	P-AUROC I-AUROC	85.10 91.80	95.20 93.10	↑10.10 ↑1.30
	APRIL-GAN	线性适配层与记忆库结合	记忆库机制/对比学习: 结合分布存储与线性层以优化少样本下的稳定性	P-AUROC I-AUROC	—	95.10 92.00	—
GPT	AnomalyGPT	在模拟数据上训练, 结合特征匹配解码器和提示学习器	异常合成 (AnomalyXFusion): 通过生成大量“异常图像–文本描述”数据对来训练 LVLMs, 实现视觉–语言对齐	P-AUROC I-AUROC	—	93.10 97.40	—
	Myriad	由专业“视觉专家”引导的 LoRA 自适应	学生–教师知识蒸馏 (AST): 专业的检测器作为“教师”, 生成异常图以指导 LMM (学生) 的注意力	Accuracy	92.00	94.20	↑2.20
	Customizable-VLM	多模态提示策略	知识驱动: 将专家知识编码到提示中	I-AUROC	—	82.90	—

成“混合提示”. 通过正则化技术, 引导 SAM 同时关注专家指定的语义概念和图像的内在视觉特征. 提示成为领域专家与 SAM 模型之间交互的界面, 使专家能够使用自然语言更精确地“指导”SAM 的分割行为. 这种专家知识注入的有效性, 也体现在了性能指标的提升上: 像素级 F-score (F_p) 和区域级 F-score (F_r) 分别从 30.95% 和 29.17% 提升到 34.85% 和 34.07%.

STLM[16] 通过整合师生架构与知识重建原理, 构建了一个高效的协同双流架构. 在知识蒸馏流中, 该模型采用了多层级的特征映射策略, 提取固定 SAM 教师网络的中间层特征图作为监督信号, 引导轻量级学生网络进行知识迁移, 从而实现 FM 通用知识向工业任务的有效“下沉”. 与此同时, 架构中的重建流专注于学习输入图像的无异常表征, 并通过引入重建损失以精确捕捉缺陷产生的残差信息:

$$L_{rec} = \|I - \hat{I}\|_2^2 \tag{2}$$

其中, I 为原始输入图像; $\hat{I}$ 为重建后的无异常图像. 通过聚合这两条流的特征, STLM 不仅利用 NFM 的重建原理锐化了异常区域的边缘定位精度, 还保留了 FM 赋予的强判别表征能力. 这种深度融合范式使其在性能与效率之间取得了极佳平衡, 其像素级与图像级 AUROC 分别达到了 98.26% 和 99.05%, 且保持了极高的推理效率.

4.2 CLIP: 增强视觉–语言对齐

CLIP 能够理解图像与文本间的语义关联, 但在处理工业高分辨率纹理时, 其预训练特征往往面临严重的域分布差异与定位精度瓶颈, 且在少样本场景下表现出性能波动.

AnoCLIP[102] 和 Dual-Image Enhanced CLIP[33] 均采用了异常合成的思想, 通过在模型中引入人为构造的伪异常干扰, 引导轻量级适配器进行针对性的在线优化或微调, 以锐化模型对特定工业缺陷的判别力. 具体而言, AnoCLIP 在推理过程中通过向测试图像添加“合成噪声扰动”来创建伪异常, 强制适配器在像素级别上学习异常边界; Dual-Image Enhanced CLIP 也采用类似的在线合成策略, 以强化模型对工业异常分布的敏感性. 这种策略有效强化了模型对“正常”特征分布的理解, 通过性能指标得以证明: AnoCLIP 的像素级 AUROC 从 88.90% 提高到 90.60%, 而 Dual-Image Enhanced CLIP 的像素级和图像级 AUROC 分别从 85.30% 和 91.60% 提高到 92.80% 和 93.20%.

SimCLIP[26] 通过引入多级视觉适配器并结合隐式提示调整, 缓解了自然图像与工业纹理间的域分布差异. 该方法允许模型在不破坏 CLIP 原始语义空间的前提下, 利用轻量级适配层实现特征空间的精细对齐. AdaCLIP[24] 则探索了提示学习的潜力, 通过优化文本或视觉上下文提示来激活 CLIP 的异常语义感知, 使其更契合特定的工业分布.

WinCLIP[22] 与 APRIL-GAN[30] 则借鉴了 NFM 中成熟的记忆库机制. WinCLIP 通过组合式提示集成与窗口级特征关联, 构建了基于正常样本原型的参考关联方法, 有效弥补了零样本能力在应对复杂缺陷时的偏差. APRIL-GAN 将记忆库与线性适配层相结合, 利用对比学习强化了模型在极少样本下的稳定性. 这种将基础模型的通用表示与记忆库的分布存储相结合的策略, 使得模型在面对未见异常时具备更稳健的决策边界.

4.3 GPT (LVLMs): 将生成能力与领域特定知识对齐

通用的 LVLMs 或许能够描述一张图片里的内容, 但其本身并不知道如何以一种结构化的、对工业生产有用的方式来描述一个缺陷. 所以基于 GPT 这类大型视觉语言模型的方法, 其核心挑战就在于如何将其强大的通用图像理解和文本生成能力, 与工业检测所需的领域特有知识进行对齐.

AnomalyGPT[34] 利用异常合成方法, 通过生成一个大规模、由“异常图像–文本描述”对组成的模拟数据集来微调模型. 在这个数据集中, 每张合成的异常图像都配有一段符合工业标准的专业文本描述. 通过在此数据集上的学习, AnomalyGPT 掌握了从特定视觉缺陷模式到专业术语描述的精确映射.

Myriad[37] 采用了教师–学生知识蒸馏范式. 在此框架中, 一个专门的缺陷检测器充当“视觉专家”(教师), 负责生成高质量的异常热图, 以识别图像中的潜在缺陷区域. 随后, 作为“学生”的大型多模态模型使用 LoRA 技术进行微调, 使其注意力机制与教师生成的异常热图对齐. 这一知识蒸馏过程将检测准确率从 92.00% 提高到了 94.20%.

Customizable-VLM[36] 采用了多模态知识驱动策略. 该方法允许专家在提示中不仅使用文本指令, 还可以附上“正常品”和“已知缺陷”的示例图片. 这种多模态提示将专家知识直接编码到推理查询中, 使得模型能通过强大的上下文学习能力, 在推理时被即时、动态地指导, 从而快速适应新任务, 无需重新训练.

5 协同范式的深度分析与讨论

在第 4 节中, 通过具体案例展示了 FM 与 NFM

方法融合在提升检测精度方面的有效性. 然而, 面对日益复杂的工业场景, 如何选择最适合的基础模型与 NFM 策略进行组合, 仍缺乏系统的理论指导. 为了更深入地指导工业实践, 本节将从宏观视角出发, 首先探讨不同类型基础模型与 NFM 机制之间的最佳适配策略, 并构建适配性矩阵; 随后将客观剖析当前融合范式在实际落地中仍面临的局限性.

5.1 FM 与 NFM 方法的场景导向差异化适配分析

虽然 FM 与 NFM 方法的结合已展现出巨大潜力, 但不同的工业场景与缺陷类型对模型的感知粒度、语义理解及推理能力有着截然不同的需求. 盲目的"模型叠加"不仅可能导致计算冗余, 甚至可能在特定缺陷上引入噪声. 基于对现有文献的梳理, 本文提出"场景–缺陷–模型"的三维适配原则, 并构建了适配性矩阵, 如表 5 所示, 旨在明确不同工况下优先考虑的协同路径. 深入的分析揭示了以下核心适配规律.

5.1.1 SAM: 几何先验与语义增强

SAM 拥有强大的几何先验, 但缺乏对"缺陷"概念的内生理解. 针对不同缺陷类型, 需采取差异化的 NFM 协同策略:

● 针对具有明确语义的对象级缺陷 (如异物入侵、组件缺失), 适配重点在于"语义注入". 此时应优先采用知识驱动与提示工程策略 (如 SAA+[17]), 将专家知识转化为文本或点提示, 引导 SAM 在零样本下识别特定类别的异常.

● 针对缺乏语义的微小纹理缺陷 (如金属表面的微米级划痕、织物同色异谱疵点), 单纯的提示工程往往失效. 此时需结合对比学习机制 (如 UCAD[19]), 利用 SAM 的高保真掩码构建结构化正负样本对, 强制模型在特征空间拉开正常纹理与微小瑕疵的距离.

5.1.2 CLIP: 域适配与细粒度定位

CLIP 虽然具备强大的泛化语义, 但在面对高频工业纹理时常面临域漂移与定位模糊的问题. 适配策略需根据数据分布与任务目标进行调整:

● 针对自然图像与工业纹理的强域分布差异 (如复杂的皮革纹理、PCB 线路), 直接应用 CLIP 效果不佳. 此时适配器 (adapter) 或提示学习 (如 SimCLIP[26]) 是最佳选择, 通过插入轻量级模块实现特征空间的低成本对齐.

● 针对像素级定位精度要求极高的场景 (如半导体晶圆检测), CLIP 的 patch-level 特征分辨率不足. 此时需引入异常合成策略 (如 AnoCLIP[102]), 通过生成伪异常强制模型学习精细的像素级决策边界.

● 针对多品种小批量 (few-shot) 场景, 记忆库机制 (如 WinCLIP[22]) 不可或缺, 它通过显式存储正常样本分布, 有效弥补了 CLIP 在极少样本下零样本推理的不稳定性.

5.1.3 GPT/LVLMs: 逻辑推理与知识驱动

对于 GPT 类模型, 其核心优势在于逻辑推理而非像素级感知.

● 针对逻辑异常与因果诊断 (如装配顺序错误、零部件兼容性问题), 这是传统 CV 模型难以处理的盲区. 此时应采用上下文学习与知识驱动策略 (如 Customizable-VLM[36]), 通过注入多模态专家示例, 激活大模型的推理能力.

● 针对实时性要求高的在线检测, 直接部署 LVLMs 并不可行. 此时应采用知识蒸馏策略, 将大模型的理解能力转移至轻量级学生网络, 实现"大模型训练, 小模型推理".

5.2 协同范式的局限性

尽管"FM + NFM"的协同范式在 MVTec AD 数据集上实现了显著的性能跃升, 如表 4 所示, 但审视其在真实工业环境中的应用潜力, 当前的融合策略仍存在以下局限性, 这也是未来研究亟须解决的关键问题:

1) "松耦合"带来的计算与工程冗余: 当前大多数方法采用的是"串联式"或"外挂式"融合. 例如, ClipSAM[20] 或 SAA+[17] 往往需要先运行一个庞大的基础模型提取特征或掩码, 再运行后续的判别模块. 这种松耦合导致推理流程割裂, 计算开销往往是 FM 与 NFM 之和, 而非优化后的整体. 与 PatchCore[104] 等传统高效 NFM 方法相比, 部分未经过蒸馏优化的 FM 融合方法难以满足高速产线 (通常要求 $>$ 20 FPS) 的实时性要求.

2) 特征空间的强制对齐损失: FM 的特征空间通常基于海量自然图像 (如 WIT-400M) 训练获得[11], 而 NFM 通常针对特定的工业纹理进行优化[9]. 两者之间存在本质的域分布差异. 当前许多方法 (如 SimCLIP[26], APRIL-GAN[30] 和 CLIP-AD[27]) 试图通过简单的线性层或适配器强行对齐两者, 这往往会导致细粒度信息的丢失. FM 的高层语义特征可能掩盖了 NFM 本应捕捉到的微小纹理缺陷, 导致在极微小缺陷 (如微米级裂纹) 检测上的表现甚至不如纯粹的 NFM.

3) 跨模态迁移的几何失真: 在 3D 缺陷检测领域, 当前主流的"FM + NFM"范式 (如 PointAD[42] 和 CLIP3D-AD[41]) 多采用将 3D 点云投影为 2D 图

表 5　FM 面临的挑战与 NFM 协同机制适配性矩阵

Table 5　Adaptability matrix of FM facing challenges and NFM collaborative mechanisms

FM 类型	主要局限/挑战	适用场景/缺陷类型	推荐 NFM 协同机制	协同原理简述	代表性方法
SAM	缺乏语义感知 (不识“缺陷”类别)	语义/对象级缺陷 (如异物、组件缺失)	知识驱动/提示工程	将专家知识 (如“划痕”) 编码为提示, 引导 SAM 聚焦特定异常语义	SAA+[17] ClipSAM[20]
	特征判别性不足 (仅基于边缘分割)	微小纹理/隐蔽缺陷 (如微划痕、同色疵点)	对比学习	利用掩码构造结构化正负样本, 强化特征空间对微小差异的敏感度	UCAD[19]
CLIP	定位精度低 (分割分辨率差)	像素级表面缺陷 (需高精度定位)	异常合成	生成伪异常微调适配器, 强制模型学习像素级边界	AnoCLIP[102] Dual-Image[33]
	域分布差异 (自然与工业)	复杂工业纹理 (强背景干扰)	适配器/提示学习	插入轻量级模块对齐特征空间, 弥合自然图像与工业纹理鸿沟	SimCLIP[26] AdaCLIP[24]
	样本极少 (需分布参考)	多品种小批量 (少样本/零样本)	记忆库机制	引入外部记忆库存储正常分布, 辅助少样本下的决策稳定性	WinCLIP[22] APRIL-GAN[30]
GPT/ LVLMs	缺乏领域知识 (描述不专业)	逻辑/因果异常 (需推理诊断)	知识驱动/上下文学习	注入专家示例与多模态提示, 激活模型的逻辑推理与泛化能力	Customizable-VLM[36] AnomalyGPT[34]
	推理开销大 (难以实时)	在线实时检测 (高 FPS 需求)	知识蒸馏	大模型作为“教师”生成伪标签, 指导轻量级“学生”网络学习	STLM[16] Myriad[37]

像再利用 CLIP 处理的策略. 这种降维操作不可避免地破坏了点云原始的拓扑结构和空间邻域信息, 导致深度方向的微小形变 (如凹坑、凸起) 在投影视图中被弱化甚至消失, 限制了基础模型在复杂 3D 几何缺陷检测中的上限.

4) 数据隐私与云端依赖的矛盾: 以 GPT-4V[35, 39] 为代表的视觉语言模型展示了强大的逻辑推理能力, 但这类模型多依赖云端 API. 出于数据保密性的严格要求, 高端制造业极少允许生产数据上云. 目前的融合范式尚未很好地解决如何在本地边缘设备上部署大模型, 或如何在保证隐私的前提下有效利用云端 FM 的推理能力.

5) 工业可靠性与决策可追溯性的张力: 目前的融合机制多侧重于提升统计意义上的检测精度, 却忽视了工业现场对决策逻辑可追溯性的严苛要求. 基础模型的“黑盒”属性导致其在面对罕见样本或极端工况时, 可能产生缺乏物理一致性的非确定性输出. 这种“感知能力”与“决策可信度”之间的断层, 仍是限制基础模型在高端精密制造等安全性敏感领域深度落地的重大隐患.

6　结束语

本文对视觉工业缺陷检测技术进行了全面回顾, 系统梳理了从传统非基础模型到新兴基础模型的演进历程, 并深入探讨了两者之间的协同范式. 研究表明, 虽然基础模型凭借海量数据积累的先验知识在泛化能力和少样本场景中展现出显著优势, 但工业缺陷检测的未来并非简单的模型替代, 而是非基础模型任务导向型的成熟原理与基础模型通用表征能力的深度战略协同. 这种协同演进不仅是技术架构的叠加, 更是工业视觉领域迈向更高效、更通用智能的关键路径.

然而, 基础模型在工业落地过程中仍面临着显著的瓶颈. 首先是庞大的计算开销与产线实时性要求之间的矛盾, 其原因在于目前的协同策略多采用松耦合的插件式设计, 导致推理过程资源冗余, 难以满足高速产线的响应需求; 其次是基础模型的通用特征空间与工业微观缺陷之间的域分布鸿沟, 原因在于基础模型多基于自然图像预训练, 其表征空间对细微纹理偏差的捕捉能力受限, 导致细粒度定位精度不尽如人意; 此外, 在三维检测维度, 过度依赖降维投影的策略造成了严重的几何拓扑失真, 原因在于视图转换不可避免地破坏了点云原始的空间邻域关系; 最后, 大模型在复杂逻辑判断上的不确定性与工业严苛标准之间仍存在冲突, 成因在于通用知识缺乏与工业专家经验及物理约束的深度对齐.

针对上述挑战, 未来研究应致力于构建更深层次的内生式演进机制. 一方面, 应重点推动知识转移技术的研究, 通过高效的蒸馏手段将大模型的泛化判别力注入轻量化检测器中, 实现感知能力与实时性能的解耦, 从而解决落地效率问题. 另一方面, 应加强针对工业特定分布的适配层与微调算法开发, 在不破坏基础模型原始语义空间的前提下, 锐化其对微小形变与异常模式的敏感度. 在三维感知层面, 发展原生空间内的几何特征表征技术, 是突破投影损耗、实现高保真感知的重要方向. 同时, 通过引入领域专家知识与因果推理机制, 将缺陷检测从视觉识别提升至认知决策层面, 实现从缺陷发现到根因诊断的智能闭环. 总之, 基础模型与非基础模型的深度协同将持续驱动工业检测技术的边界扩张, 未来的突破关键在于如何在保留通用知识广度的同时, 精准触达工业场景所需的专业深度与效率边界.

参考文献

1 Pang G S, Shen C H, Cao L B, Hengel A V D. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 2022, **54**(2): Article No. 38

2 Bergmann P, Löwe S, Fauser M, Sattlegger D, Steger C. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. arXiv preprint arXiv: 1807.02011, 2019.

3 Gong D, Liu L Q, Le V, Saha B, Mansour M R, Venkatesh S, et al. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Republic of Korea: IEEE, 2019. 1705−1714

4 Liu Y F, Zhuang C Q, Lu F. Unsupervised two-stage anomaly detection. arXiv preprint arXiv: 2103.11671, 2021.

5 Deng H Q, Li X Y. Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 9727−9736

6 Liu Z K, Zhou Y M, Xu Y S, Wang Z L. SimpleNet: A simple network for image anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 20402−20411

7 Liang Y, Hu Z G, Huang J J, Di D L, Su A Y, Fan L. ToCoAD: Two-stage contrastive learning for industrial anomaly detection. *IEEE Transactions on Instrumentation and Measurement*, 2025, **74**: Article No. 5012009

8 Hu H G, Wang X Y, Zhang Y, Chen Q, Guan Q. A comprehensive survey on contrastive learning. *Neurocomputing*, 2024, **610**: Article No. 128645

9 Bergmann P, Fauser M, Sattlegger D, Steger C. MVTec AD——A comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 9584−9592

10 Zou Y, Jeong J, Pemula L, Zhang D Q, Dabeer O. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: Proceedings of the 17th European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer, 2022. 392−408

11 Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). Virtual Event: PMLR, 2021. 8748−8763

12 Zhu D Y, Chen J, Shen X Q, Li X, Elhoseiny M. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024.

13 Yang Z Y, Li L J, Lin K, Wang J F, Lin C C, Liu Z C, et al. The dawn of LMMs: Preliminary explorations with GPT-4V(ision). arXiv preprint arXiv: 2309.17421, 2023.

14 Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, et al. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 4015−4026

15 Rani A, Ortiz-Arroyo D, Durdevic P. Advancements in point cloud-based 3D defect detection and classification for industrial systems: A comprehensive survey. arXiv preprint arXiv: 2402.12923, 2024.

16 Li C H, Qi L, Geng X. A SAM-guided two-stream lightweight model for anomaly detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025, **21**(2): Article No. 67

17 Cao Y K, Xu X H, Cheng Y Q, Sun C, Du Z W, Gao L, et al. Personalizing vision-language models with hybrid prompts for zero-shot anomaly detection. *IEEE Transactions on Cybernetics*, 2025, **55**(4): 1917−1929

18 Peng Y, Lin X, Ma N C, Du J Y, Liu C W, Liu C J, et al. SAM-LAD: Segment anything model meets zero-shot logic anomaly detection. *Knowledge-Based Systems*, 2025, **314**: Article No. 113176

19 Liu J Q, Wu K, Nie Q, Chen Y, Gao B B, Liu Y, et al. Unsupervised continual anomaly detection with contrastively-learned prompt. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 3639−3647

20 Li S Z, Cao J J, Ye P, Ding Y H, Tu C J, Chen T. ClipSAM: CLIP and SAM collaboration for zero-shot anomaly segmentation. *Neurocomputing*, 2025, **618**: Article No. 129122

21 Yang H Y, Chen H, Wang A, Chen K, Lin Z J, Tang Y L, et al. Promptable anomaly segmentation with SAM through self-perception tuning. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press, 2025. 13017−13025

22 Jeong J, Zou Y, Kim T, Zhang D Q, Ravichandran A, Dabeer O. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 19606−19616

23 Zhou Q H, Pang G S, Tian Y, He S B, Chen J M. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024.

24 Cao Y K, Zhang J N, Frittoli L, Cheng Y Q, Shen W M, Boracchi G. AdaCLIP: Adapting CLIP with hybrid learnable prompts for zero-shot anomaly detection. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 55−72

25 Qu Z, Tao X, Prasad M, Shen F, Zhang Z T, Gong X Y, et al. VCP-CLIP: A visual context prompting model for zero-shot anomaly segmentation. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 301−317

26 Deng C H, Xu H T, Chen X L, Xu H D, Tu X T, Ding X H, et al. SimCLIP: Refining image-text alignment with simple prompts for zero-/few-shot anomaly detection. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024. 1761−1770

27 Chen X H, Zhang J N, Tian G Z, He H Y, Zhang W H, Wang Y B, et al. CLIP-AD: A language-guided staged dual-path model for zero-shot anomaly detection. In: Proceedings of the 4th International Workshop and 1st International Workshop on Human Activity Recognition and Anomaly Detection. Jeju, Republic of Korea: Springer, 2024. 17−33

28 Zuo Z, Wu Y, Li B Q, Dong J H, Zhou Y, Zhou L, et al. CLIP-FSAC: Boosting CLIP for few-shot anomaly classification with synthetic anomalies. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence. Jeju, Republic of Korea: IJCAI, 2024. Article No. 203

29 Hu Z, Zhang Z. SOWA: Adapting hierarchical frozen window self-attention to visual-language models for better anomaly detection. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024.

30 Chen X H, Han Y, Zhang J N. APRIL-GAN: A zero-/few-shot anomaly classification and segmentation method for CVPR 2023 VAND workshop challenge tracks 1&2: 1st place on zero-

shot AD and 4th place on few-shot AD. arXiv preprint arXiv: 2305.17382, 2023.

31 Li X F, Zhang Z Z, Tan X, Chen C W, Qu Y Y, Xie Y, et al. PromptAD: Learning prompts with only normal samples for few-shot anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 16838−16848

32 Gu Z P, Zhu B K, Zhu G B, Chen Y Y, Li H, Tang M, et al. FiLo: Zero-shot anomaly detection by fine-grained description and high-quality localization. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024. 2041−2049

33 Zhang Z X, Deng H Q, Bao J N, Li X Y. Dual-image enhanced CLIP for zero-shot anomaly detection. arXiv preprint arXiv: 2405.04782, 2024.

34 Gu Z P, Zhu B K, Zhu G B, Chen Y Y, Tang M, Wang J Q. AnomalyGPT: Detecting industrial anomalies using large vision-language models. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. Article No. 215

35 Cao Y K, Xu X H, Sun C, Huang X N, Shen W M. Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (GPT-4V) takes the lead. arXiv preprint arXiv: 2311.02782, 2023.

36 Xu X H, Cao Y K, Zhang H X, Sang N, Huang X N. Customizing visual-language foundation models for multi-modal anomaly detection and reasoning. In: Proceedings of the 28th International Conference on Computer Supported Cooperative Work in Design (CSCWD). Compiegne, France: IEEE, 2025. 1443−1448

37 Li Y Z, Wang H L, Yuan S H, Liu M, Zhao D B, Guo Y W, et al. Myriad: Large multimodal model by applying vision experts for industrial anomaly detection. arXiv preprint arXiv: 2310.19070, 2025.

38 Zhu J Q, Cai S F, Deng F, Ooi B C, Wu J R. Do LLMs understand visual anomalies? Uncovering LLM′s capabilities in zero-shot anomaly detection. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024. 48−57

39 Zhang J N, He H Y, Chen X H, Xue Z C, Wang Y B, Wang C J, et al. GPT-4V-AD: Exploring grounding potential of VQA-oriented GPT-4V for zero-shot anomaly detection. In: Proceedings of the 4th International Workshop and 1st International Workshop on Human Activity Recognition and Anomaly Detection. Jeju, Republic of Korea: Springer, 2024. 3−16

40 Zhang Y H, Cao Y K, Xu X H, Shen W M. LogiCode: An LLM-driven framework for logical anomaly detection. *IEEE Transactions on Automation Science and Engineering*, 2025, **22**: 7712−7723

41 Zuo Z, Dong J, Wu Y, Qu Y, Wu Z. CLIP3D-AD: Extending CLIP for 3D few-shot anomaly detection with multi-view images generation. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024.

42 Zhou Q H, Yan J T, He S B, Meng W C, Chen J M. PointAD: Comprehending 3D anomalies from points and pixels for zero-shot 3D anomaly detection. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc., 2024. Article No. 2695

43 Wang C J, Zhu H K, Peng J L, Wang Y, Yi R, Wu Y S, et al. M3DM-NR: RGB-3D noisy-resistant industrial anomaly detection via multimodal denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, **47**(11): 9981−9993

44 Zhang Z L, Niu C, Zhao Z B, Zhang X W, Chen X F. Small object few-shot segmentation for vision-based industrial inspection. *IEEE Transactions on Industrial Informatics*, 2025, **21**(6): 4388−4399

45 Bae J, Lee J H, Kim S. PNI: Industrial anomaly detection using position and neighborhood information. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 6350−6360

46 Lyu S, Mo D M, Wong W K. REB: Reducing biases in representation for industrial anomaly detection. *Knowledge-Based Systems*, 2024, **290**: Article No. 111563

47 Yao X C, Li R Q, Zhang J, Sun J, Zhang C Y. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 24490−24499

48 Qian L, Zhu B, Chen Y, Tang M, Wang J. Friend or foe? Harnessing controllable overfitting for anomaly detection. In: Proceedings of the International Conference on Learning Representations (ICLR). Singapore: OpenReview, 2025.

49 Chen Q Y, Luo H Y, Lv C K, Zhang Z T. A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 37−54

50 Li H X, Zhang Z X, Chen H, Wu L, Li B, Liu D Y, et al. A novel approach to industrial defect generation through blended latent diffusion model with online adaptation. arXiv preprint arXiv: 2402.19330, 2024.

51 Zhang X M, Xu M, Zhou X Z. RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 16699−16708

52 Jiang B L, Xie Y Q, Li J W, Li N Q, Jiang Y, Xia S T. CA-GEN: Controllable anomaly generator using diffusion model. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Republic of Korea: IEEE, 2024. 3110−3114

53 Hu J, Huang Y W, Lu Y L, Xie G Y, Jiang G N, Zheng Y F, et al. AnomalyXFusion: Multi-modal anomaly synthesis with diffusion. arXiv preprint arXiv: 2404.19444, 2024.

54 Hu T, Zhang J N, Yi R, Du Y Z, Chen X, Liu L, et al. AnomalyDiffusion: Few-shot anomaly image generation with diffusion model. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 8526−8534

55 Duan Y X, Hong Y, Niu L, Zhang L Q. Few-shot defect image generation via defect-aware feature manipulation. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, DC, USA: AAAI Press, 2023. 571−578

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

57 Qin J J, Gu C Z, Yu J, Zhang C. Multilevel saliency-guided self-supervised learning for image anomaly detection. *Signal, Image and Video Processing*, 2024, **18**(8): 6339−6351

58 Lin J, Yan Y P. A comprehensive augmentation framework for anomaly detection. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 8742−8749

59 Bai Y H, Zhang J N, Chen Z F, Dong Y H, Cao Y K, Tian G Z. Dual-path frequency discriminators for few-shot anomaly detection. *Knowledge-Based Systems*, 2024, **302**: Article No. 112397

60 Chen Q Y, Luo H Y, Gao H, Lv C K, Zhang Z T. Progressive boundary guided anomaly synthesis for industrial anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025, **35**(2): 1193−1208

61 Chu Y M, Liu C, Hsieh T I, Chen H T, Liu T L. Shape-Guided dual-memory learning for 3D anomaly detection. In: Proceedings of the 40th International Conference on Machine Learning (ICML). Honolulu, USA: PMLR, 2023. Article No. 245

62 Cao Y K, Xu X H, Shen W M. Complementary pseudo multimodal feature for point cloud anomaly detection. *Pattern Recognition*, 2024, **156**: Article No. 110761

63 Horwitz E, Hoshen Y. Back to the feature: Classical 3D features are (almost) all you need for 3D anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 2968−2977

64 Fučka M, Zavrtanik V, Skočaj D. TransFusion——A transparency-based diffusion model for anomaly detection. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2025. 91−108

65 Zavrtanik V, Kristan M, Skočaj D. Cheating depth: Enhancing 3D surface anomaly detection via depth simulation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2024. 2153−2161

66 Rudolph M, Wehrbein T, Rosenhahn B, Wandt B. Asymmetric student-teacher networks for industrial anomaly detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2023. 2591−2601

67 Zhou Z Y, Wang L, Fang N Y, Wang Z L, Qiu L M, Zhang S Y. R3D-AD: Reconstruction via diffusion for 3D anomaly detection. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 91−107

68 Zhao B Z, Xiong Q W, Zhang X H, Guo J F, Liu Q, Xing X F, et al. PointCore: Efficient unsupervised point cloud anomaly detector using local-global features. arXiv preprint arXiv: 2403.01804, 2024.

69 Liu J Y, Mou S C, Gaw N, Wang Y N. Uni-3DAD: GAN-inversion aided universal 3D anomaly detection on model-free products. *Expert Systems With Applications*, 2025, **272**: Article No. 126665

70 Zhu H Z, Xie G Y, Hou C B, Dai T, Gao C, Wang J B, et al. Towards high-resolution 3D anomaly detection via group-level feature contrastive learning. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024. 4680−4689

71 Zavrtanik V, Kristan M, Skočaj D. DRÆM——A discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 8310−8319

72 Al-Fakih A, Koeshidayatullah A, Mukerji T, Kaka S I. Enhanced anomaly detection in well log data through the application of ensemble GANs. arXiv preprint arXiv: 2411.19875, 2024.

73 Bhosale A, Mukherjee S, Banerjee B, Cuzzolin F. Anomaly detection using diffusion-based methods. arXiv preprint arXiv: 2412.07539, 2024.

74 Kim S, Lee S Y, Bu F C, Kang S, Kim K, Yoo J, et al. Rethinking reconstruction-based graph-level anomaly detection: Limitations and a simple remedy. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 3040

75 Yao H, Liu M, Yin Z C, Yan Z F, Hong X P, Zuo W M. GLAD: Towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 1−17

76 Rafieizadeh H, Zare H, Parsa M G, Davardoust H, Bagheri M S. DCOR: Anomaly detection in attributed networks via dual contrastive learning reconstruction. arXiv preprint arXiv: 2412.16788, 2024.

77 Patra S, Taieb S B. Revisiting deep feature reconstruction for logical and structural industrial anomaly detection. arXiv preprint arXiv: 2410.16255, 2024.

78 Lee K, Kim M, Jun Y, Woo S S. GDFlow: Anomaly detection with NCDE-based normalizing flow for advanced driver assistance system. arXiv preprint arXiv: 2409.05346, 2024.

79 Zhou Y X, Xu X, Sun Z, Song J K, Cichocki A, Shen H T. VQ-Flow: Taming normalizing flows for multi-class anomaly detection via hierarchical vector quantization. *IEEE Transactions on Multimedia*, 2025, **27**: 6884−6895

80 Liu X F, Xing F X, Zhuo J C, Stone M, Prince J L, El Fakhri G, et al. Speech motion anomaly detection via cross-modal translation of 4D motion fields from tagged MRI. In: Proceedings of the SPIE 12926, Medical Imaging 2024: Image Processing. San Diego, USA: SPIE, 2024.

81 Tu Y P, Zhang B S, Liu L, Li Y X, Zhang J N, Wang Y B, et al. Self-supervised feature adaptation for 3D industrial anomaly detection. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 75−91

82 Li J T, Wang X Y, Zhao H W, Zhong Y F. Learning a cross-modality anomaly detector for remote sensing imagery. *IEEE Transactions on Image Processing*, 2024, **33**: 6607−6621

83 Arav R, Wittich D, Rottensteiner F. Evaluating saliency scores in point clouds of natural environments by learning surface anomalies. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2025, **224**: 235−250

84 Ye J N, Zhao W G, Yang X, Cheng G L, Huang K Z. PO3AD: Predicting point offsets toward better 3D point cloud anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 1353−1362

85 Hao S, Fu W T, Chen X Z, Jin C X, Zhou J J, Yu S Q, et al. Network anomaly traffic detection via multi-view feature fusion. arXiv preprint arXiv: 2409.08020, 2025.

86 Dai A, Chang A X, Savva M, Halber M, Funkhouser T, Nießner M. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 2432−2443

87 Uy M A, Pham Q H, Hua B S, Nguyen T, Yeung S K. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Republic of Korea: IEEE, 2019. 1588−1597

88 Zhou Q H, He S B, Liu H Y, Chen T, Chen J M. Pull & push: Leveraging differential knowledge distillation for efficient unsupervised anomaly detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, **33**(5): 2176−2189

89 Chen Z N, Luo X S, Wang W Q, Zhao Z C, Su F, Men A. Filter or compensate: Towards invariant representation from distribution shift for anomaly detection. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press, 2025. 2420−2428

90 Liu X Y, Wang J Y, Leng B, Zhang S. Unlocking the potential

of reverse distillation for anomaly detection. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press, 2025. 5640−5648

91 Liu H W, Xu X D, Li E H, Zhang S C, Li X L. Anomaly detection with representative neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(6): 2831−2841

92 Zhou J X, Wu Y. Outlier-probability-based feature adaptation for robust unsupervised anomaly detection on contaminated training data. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, **34**(10): 10023−10035

93 Xing P, Li Z C. Visual anomaly detection via partition memory bank module and error estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, **33**(8): 3596−3607

94 Zhang F H, Zhu H Y, Cen Y G, Kan S C, Zhang L N, Vadakkepat P, et al. Low-shot unsupervised visual anomaly detection via sparse feature representation. *IEEE Transactions on Neural Networks and Learning Systems*, 2025, **36**(5): 7903−7917

95 Zhou Y J, Song X C, Zhang Y R, Liu F X, Zhu C, Liu L Q. Feature encoding with autoencoders for weakly supervised anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(6): 2454−2465

96 Ramírez R A, Khan A, Bekkouch I E I, Sheikh T S. Anomaly detection based on zero-shot outlier synthesis and hierarchical feature distillation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(1): 281−291

97 Liang Y X, Li X S, Huang X, Zhang Z Q, Yao Y. An automated data mining framework using autoencoders for feature extraction and dimensionality reduction. arXiv preprint arXiv: 2412.02211, 2024.

98 Chen J Y, Wang C B, Hong Y F, Mi R, Zhang L J, Wu Y R, et al. A survey on anomaly detection with few-shot learning. In: Proceedings of the 8th International Conference on Cognitive Computing (ICCC). Bangkok, Thailand: Springer, 2024. 34−50

99 Bergmann P, Jin X, Sattlegger D, Steger C. The MVTec 3D-AD dataset for unsupervised 3D anomaly detection and localization. In: Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP). Virtual Event: SCITEPRESS, 2022. 202−213

100 Liu J Q, Xie G Y, Chen R T, Li X P, Wang J B, Liu Y, et al. Real3D-AD: A dataset of point cloud anomaly detection. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 1324

101 Wang Y, Peng J L, Zhang J N, Yi R, Wang Y B, Wang C J. Multimodal industrial anomaly detection via hybrid fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 8032−8041

102 Deng H Q, Zhang Z X, Bao J N, Li X Y. Bootstrap fine-grained vision-language alignment for unified zero-shot anomaly localization. arXiv preprint arXiv: 2308.15939, 2024.

103 Yu J W, Zheng Y, Wang X, Li W, Wu Y S, Zhao R, et al. FastFlow: Unsupervised anomaly detection and localization via 2D normalizing flows. arXiv preprint arXiv: 2111.07677, 2021.

104 Roth K, Pemula L, Zepeda J, Schölkopf B, Brox T, Gehler P. Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 14298−14308

105 Schlegl T, Seeböck P, Waldstein S M, Langs G, Schmidt-Erfurth U. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 2019, **54**: 30−44

106 Zhou K Y, Yang J K, Loy C C, Liu Z W. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022, **130**(9): 2337−2348

107 Zhou K Y, Yang J K, Loy C C, Liu Z W. Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 16816−16825

108 Fu S, Hamilton M, Brandt L E, Feldman A, Zhang Z T, Freeman W T. FeatUp: A model-agnostic framework for features at any resolution. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2024.

109 Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 10674−10685

110 Ma B R, Liu Y S, Zwicker M, Han Z Z. Surface reconstruction from point clouds by learning predictive context priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 6316−6327

111 Cao Y K, Xu X H, Liu Z G, Shen W M. Collaborative discrepancy optimization for reliable image anomaly localization. *IEEE Transactions on Industrial Informatics*, 2023, **19**(11): 10674−10683

112 Wan Q, Gao L, Li X Y, Wen L. Industrial image anomaly localization based on Gaussian clustering of pretrained feature. *IEEE Transactions on Industrial Electronics*, 2022, **69**(6): 6182−6192

113 Liu S L, Zeng Z Y, Ren T H, Li F, Zhang H, Yang J, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 38−55

114 Lin T Y, Goyal P, Girshick R, He K M, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2999−3007

115 Milletari F, Navab N, Ahmadi S A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the 4th International Conference on 3D Vision (3DV). Stanford, USA: IEEE, 2016. 565−571

116 Wang H X, Vasu P K A, Faghri F, Vemulapalli R, Farajtabar M, Mehta S, et al. SAM-CLIP: Merging vision foundation models towards semantic and spatial understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 3635−3647

117 Wang Z Q, Lu Y, Li Q, Tao X Q, Guo Y D, Gong M M, et al. CRIS: CLIP-driven referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 11686−11695

118 Xing Y J, Wang X, Li Y B, Huang H, Shi C. Less is more: On the over-globalizing problem in graph Transformers. In: Proceedings of the 41st International Conference on Machine Learning (ICML). Vienna, Austria: PMLR, 2024. 54656−54672

119 Sun X M, Hu P, Saenko K. DualCoOp: Fast adaptation to multi-label recognition with limited annotations. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 2216

120 Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C L, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Proceedings of the 36th International Conference on Neural Information Processing Systems.

New Orleans, USA: Curran Associates Inc., 2022. Article No. 2011

121 Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, et al. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv: 2302.13971, 2023.

122 Zhang R R, Guo Z Y, Zhang W, Li K C, Miao X P, Cui B, et al. PointCLIP: Point cloud understanding by CLIP. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 8552–8562

123 Zhou Y C, Gu J Y, Chiang T Y, Xiang F B, Su H. Point-SAM: Promptable 3D segmentation model for point clouds. In: Proceedings of the 13th International Conference on Learning Representations (ICLR). Singapore: ICLR, 2025.

124 Cen J Z, Fang J M, Zhou Z W, Yang C, Xie L X, Zhang X P, et al. Segment anything in 3D with radiance fields. *International Journal of Computer Vision*, 2025, **133**(8): 5138–5160

125 Guo Z Y, Zhang R R, Zhu X Y, Tang Y W, Ma X Z, Han J M, et al. Point-Bind & Point-LLM: Aligning point cloud with multi-modality for 3D understanding, generation, and instruction following. arXiv preprint arXiv: 2309.00615, 2023.

126 Xu R S, Wang X L, Wang T, Chen Y L, Pang J M, Lin D H. PointLLM: Empowering large language models to understand point clouds. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 131–147

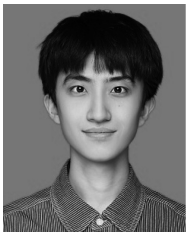
127 Hong Y N, Zhen H Y, Chen P H, Zheng S H, Du Y L, Chen Z F, et al. 3D-LLM: Injecting the 3D world into large language models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates, Inc., 2023. Article No. 900

杨天乐 苏州大学计算机科学与技术学院硕士研究生. 主要研究方向为视觉与多模态大语言模型的应用.
E-mail: tlyang@stu.suda.edu.cn
(**YANG Tian-Le** Master student at the School of Computer Science and Technology, Soochow University. His main research interest is the application of vision and multimodal large language models.)

常璐瑶 华中师范大学计算机学院硕士研究生. 主要研究方向为计算机视觉. E-mail: changluyao001@163.com
(**CHANG Lu-Yao** Master student at the School of Computer Science, Central China Normal University. Her main research interest is computer vision.)

言嘉栋 苏州大学计算机科学与技术学院硕士研究生. 主要研究方向为多模态大语言模型的应用.
E-mail: jdyan24@stu.suda.edu.cn
(**YAN Jia-Dong** Master student at the School of Computer Science and Technology, Soochow University. His main research interest is the application of multimodal large language models.)

李俊涛 苏州大学计算机科学与技术学院副教授. 主要研究方向为预训练语言模型, 文本生成与对话系统.
E-mail: ljt@suda.edu.cn
(**LI Jun-Tao** Associate professor at the School of Computer Science and Technology, Soochow University. His research interests include pre-trained language models, text generation, and dialogue systems.)



张 可 苏州大学计算机科学与技术学院副教授. 主要研究方向为视觉大模型与多模态大语言模型的算法基础与应用. 本文通信作者.
E-mail: kzhang19@suda.edu.cn
(**ZHANG Ke** Associate professor at the School of Computer Science and Technology, Soochow University. His research interests include the algorithmic foundations and applications of large-scale vision models and multimodal large language models. Corresponding author of this paper.)

张 民 哈尔滨工业大学 (深圳) 计算与智能研究院教授. 主要研究方向为自然语言处理, 人工智能与大模型.
E-mail: zhangminmt@hotmail.com
(**ZHANG Min** Professor at the Faculty of Computing and Intelligence, Harbin Institute of Technology, Shenzhen. His research interests include natural language processing, artificial intelligence, and large models.)