自动化学报
ACTA AUTOMATICA SINICA

基于文本解耦的零样本异常检测

赵梦阳 郭强

Zero-shot Anomaly Detection via Textual Decoupling

ZHAO Meng-Yang, GUO Qiang

在线阅读 View online:
http://www.aas.net.cn/article/current/250f47d68532c630847ae92a35a5d59d4f7c0382bf0323a355db178ec7fb8c78

您可能感兴趣的其他文章

基于多尺度流模型的视觉异常检测研究
Research on Visual Anomaly Detection Based on Multi-scale Normalizing Flow
自动化学报. 2024, 50(3): 640-648 https://doi.org/10.16383/j.aas.c230476

图像异常检测研究现状综述
Review of Image Anomaly Detection
自动化学报. 2022, 48(6): 1402-1428 https://doi.org/10.16383/j.aas.c200956

融合包注意力机制的监控视频异常行为检测
Abnormal Behavior Detection Algorithm With Video-bag Attention Mechanism in Surveillance Video
自动化学报. 2022, 48(12): 2951-2959 https://doi.org/10.16383/j.aas.c190805

融合知识的多视图属性网络异常检测模型
Multi-view Outlier Detection for Attributed Network Based on Knowledge Fusion
自动化学报. 2023, 49(8): 1732-1744 https://doi.org/10.16383/j.aas.c220629

基于语言视觉对比学习的多模态视频行为识别方法
Multi-modal Video Action Recognition Method Based on Language-visual Contrastive Learning
自动化学报. 2024, 50(2): 417-430 https://doi.org/10.16383/j.aas.c230159

一种基于深度迁移学习的滚动轴承早期故障在线检测方法
A New Deep Transfer Learning-based Online Detection Method of Rolling Bearing Early Fault
自动化学报. 2022, 48(1): 302-314 https://doi.org/10.16383/j.aas.c190593

基于文本解耦的零样本异常检测

赵梦阳 [1] 郭 强 [1, 2]

摘 要 随着大规模视觉−语言预训练模型的不断发展, 零样本异常检测逐渐成为一个重要的研究方向. 该任务要求模型能够直接对类别未知的异常样本进行有效检测与定位, 而无需依赖目标领域的训练数据. 由于异常检测与异常定位任务分别需要全局和局部不同粒度的语义信息, 现有方法共享文本提示的设计使模型无法同时满足二者的需求, 导致性能难以兼顾. 为此, 提出一种基于文本解耦的零样本异常检测方法, 其核心是为每个任务分别设计独立的提示并进行优化. 同时, 针对模型在异常定位任务中跨数据集泛化能力较弱的问题, 提出原型对齐模块. 该模块通过优化图像块特征与原型之间的距离, 提升模型的异常定位能力. 此外, 考虑到仅依赖图像的全局特征难以充分识别细微异常, 进一步设计了异常特征增强策略, 通过聚焦于潜在的异常区域以提升异常检测的性能. 实验结果表明, 所提出方法在 MVTec AD、ISIC、BrainMRI 等公开数据集上均取得优良性能, 验证了其有效性与泛化能力.

关键词 异常检测; 对比语言−图像预训练; 零样本; 文本解耦

引用格式 赵梦阳, 郭强. 基于文本解耦的零样本异常检测. 自动化学报, xxxx, **xx**(x): x−xx

DOI 10.16383/j.aas.c260031 **CSTR** 32138.14.j.aas.c260031

Zero-shot Anomaly Detection via Textual Decoupling

ZHAO Meng-Yang[1] GUO Qiang[1, 2]

Abstract With the continuous advancement of large pre-trained vision-language, zero-shot anomaly detection has gradually emerged as a significant research direction. This task requires the models to effectively detect and localize anomalies of unknown categories directly, without relying on training data from the target domain. Since anomaly detection and anomaly localization tasks require semantic information of different granularity globally and locally, the design of shared textual prompts in the existing methods makes it impossible for the model to satisfy the needs of both at the same time, which makes it difficult to balance performance. To address this issue, this paper proposes a text-decoupled zero-shot anomaly detection method. Its core lies in designing and optimizing independent prompts for each task. Furthermore, to tackle the model′s weak cross-dataset generalization in anomaly localization, a prototype alignment module is introduced. This module enhances anomaly localization capabilities by optimizing the distance between image patch features and prototypes. Finally, considering that relying solely on global image features is insufficient for identifying subtle anomalies, an anomaly feature enhancement strategy is designed to improve anomaly detection performance by focusing on potential anomalous regions. Experimental results demonstrate that the proposed method achieves excellent performance on public datasets such as MVTec AD, ISIC, and BrainMRI, validating its effectiveness and generalization capability.

Keywords anomaly detection; contrastive language-lmage pre-training; zero-shot; text decoupling

Citation Zhao Meng-Yang, Guo Qiang. Zero-shot anomaly detection via textual decoupling. *Acta Automatica Sinica*, xxxx, **xx**(x): x−xx

异常检测是计算机视觉中的一项关键任务, 其目标是从图像中识别与正常模式显著偏离的样本或区域. 该任务在工业异常检测[1−6]、医疗诊断[7−9] 和视频监控[10−11] 等多个实际场景中具有应用价值. 在实际应用中, 高质量异常样本的标注成本昂贵且异常类型难以预知, 促使无监督异常检测方法成为研究主流. 这类方法通常仅利用正常样本进行训练, 其核心是学习正常数据的分布. 在测试阶段, 与所学分布存在显著差异的样本被识别为异常. 然而, 无监督方法不仅严重依赖于大量未标注的正常样本, 还要求训练数据与测试环境在分布上具有一致性. 这两个前提在许多现实场景中难以同时满足: 医学影像因隐私保护无法共享, 工业制造中产品型号频

收稿日期 2026-01-13 录用日期 2026-05-26
Manuscript received January 13, 2026; accepted May 26, 2026
国家自然科学基金 (61873145), 山东省重点研发计划 (2024TSGC0109) 资助
Supported by National Natural Science Foundation of China (61873145) and Key Research and Development Program of Shandong Province (2024TSGC0109)
本文责任编委 刘青山
Recommended by Associate Editor LIU Qing-Shan
1. 山东财经大学计算机与人工智能学院 济南 250014 2. 山东省数字经济轻量智算与可视化重点实验室 济南 250014
1. School of Computing and Artificial Intelligence, Shandong University of Finance and Economics, Jinan 250014 2. Shandong Key Laboratory of Lightweight Intelligent Computing and Visualization for Digital Economy, Jinan 250014

繁更迭导致历史数据无法代表新生产样本, 或是某些极端情况下根本无法获取任何训练数据. 这些限制促使零样本异常检测逐渐受到关注, 其目标是在不依赖任何特定任务的训练样本的前提下, 同时完成图像级的异常检测与像素级的异常定位.

近年来, 大规模视觉–语言预训练模型在各种下游任务中展现出强大的零样本识别能力, 其中, 对比语言–图像预训练(contrastive language-image pre-training, CLIP)[12] 已成为零样本异常检测的重要基础[13–15]. 该类方法通常将图像与描述正常和异常的文本提示映射到同一特征空间, 并通过计算余弦相似度进行异常检测与定位, 其判别依据是: 与异常文本更相似的图像或区域被识别为异常. Jeong 等[13] 提出的 WinCLIP 采用多尺度窗口策略来获取局部特征, 并将其与人工文本提示进行匹配. 然而, 由于 CLIP 原本设计用于识别图像中的前景对象类别, 而非区分图像中正常与异常区域, 因此 WinCLIP 的异常检测性能存在一定局限.

为缓解 CLIP 预训练目标与异常检测任务间的语义差异, 后续研究提出多种自适应方法. April-GAN[14] 和 MVFA-AD[15] 通过引入额外的线性层增强模型的域适应能力; 另一类工作则采用提示学习策略, 通过优化可学习的文本嵌入以更精准地表示异常特征. Zhou 等[16] 提出的 AnomalyCLIP 通过引入对象无关的可训练文本提示, 使模型能够更专注于异常特征的检测与定位.

然而, 上述方法中的同一文本表示难以满足异常检测与异常定位不同粒度的语义需求. 为此, 本文提出一种基于文本解耦的零样本异常检测. 具体而言, 首先为异常检测与异常定位任务分别设计不同的初始化文本, 形成两组独立的文本嵌入. 然后, 在训练过程中对这两组嵌入进行独立优化. 同时, 引入原型对齐模块, 其通过优化图像块特征与原型之间的距离来解决当前模型在异常定位中跨数据集泛化能力较弱的问题. 最后, 考虑到仅依赖图像的全局特征难以充分识别细微异常, 本文进一步提出异常特征增强策略, 该模块通过聚焦于图像中潜在的异常区域, 显著提升了模型的异常检测能力. 本文的主要贡献概括如下:

1) 提出文本解耦策略, 其核心是将统一的文本表示解耦为两组独立的可学习嵌入, 分别服务于异常检测与定位任务, 并进行独立优化;

2) 引入原型对齐模块, 通过自监督的方式优化图像块特征, 提升跨数据集异常定位的泛化能力;

3) 提出异常特征增强策略, 通过聚焦潜在异常区域, 强化模型对细微异常特征的检测能力, 弥补全局特征表示的不足.

4) 在 Mvtec AD, BrainMRI 等 12 个工业和医学领域的数据集上进行大量实验, 结果表明, 本文方法在精度和泛化能力方面优于现有的零样本异常检测模型.

1 相关工作

1.1 无监督异常检测方法

基于深度学习的无监督异常检测方法主要分为三类, 分别是基于图像重建的方法[17–20]、基于特征嵌入的方法[21–22] 以及基于知识蒸馏的方法[23–26].

基于图像重建的方法[17–20] 在训练阶段使用正常图像训练模型, 使其足以重建正常区域. 在推理时, 模型难以准确重建异常区域, 导致这些区域产生较高的重建误差, 从而被识别为异常. 基于特征嵌入的方法[21–22] 则借助预训练的深度神经网络提取正常样本特征, 并以正常样本分布为核心构建特征嵌入空间, 通过度量测试样本与该空间的距离以识别异常. 基于知识蒸馏的方法[23–26] 通常包含一个预训练的教师网络和一个可训练的学生网络. 训练阶段, 学生网络模仿教师网络对正常样本的特征表示; 推理阶段, 异常样本的出现会导致两个网络输出间产生显著差异, 从而实现异常检测.

无监督异常检测方法的核心思想在于通过对大量正常样本进行建模, 学习其分布或特征表示, 从而在测试阶段识别出与正常模式显著偏离的异常样本和区域. 然而, 此类方法高度依赖于充足且具有代表性的正常样本, 一旦正常样本数量不足或覆盖范围有限, 其检测准确性将显著下降.

1.2 大规模视觉–语言预训练模型

以 CLIP[12] 为代表的大规模视觉–语言预训练模型, 通过在大规模图像–文本对上进行对比学习, 实现视觉与语言模态之间的高度对齐. 该类模型旨在将大量成对的图像和文本映射到嵌入空间中, 并使得相似的图像和文本距离更近, 而不相关的图像和文本距离较远. 这一训练机制以及使用大规模的训练数据使模型学习到高度泛化的视觉语言表征, 具备了出色的零样本识别与迁移能力, 能够直接应用于未见过任务场景.

近年来, CLIP 已广泛引入到多个下游视觉任务中, 在语义分割[27]、目标检测[28]、图像生成[29] 以及异常检测[16] 等领域取得显著成果. 因此, 本文选择以 CLIP 作为基础模型, 开展零样本异常检测研究.

1.3 提示学习

提示学习 (prompt learning) 最初在自然语言

处理领域中得以提出并广泛应用, 随后逐渐扩展到多模态环境中, 尤其在大规模视觉–语言预训练模型中发挥重要作用. 其核心思想是通过设计或学习合适的提示, 将下游任务适配到预训练模型的能力范围内, 从而在不微调模型参数的情况下提升性能.

CoOp[30] 首次将提示学习引入 CLIP, 将一组可训练的上下文向量嵌入到文本输入空间中, 以替代原本手工设计的提示模板, 显著提高了模型在图像分类任务中的表现. 随后, DenseCLIP[31] 和 CoCoOp[32] 通过视觉上下文提示引导预训练的模型适应目标域. 本文针对异常检测与异常定位两个任务, 分别设计不同的文本初始化向量, 并通过不同的损失函数对它们进行独立优化. 本文方法使模型能够更精准地捕获与任务相关的语义特征, 从而在零样本异常检测场景中实现更精确的检测与定位性能.

1.4 零样本异常检测方法

零样本异常检测方法借助大规模视觉–语言预训练模型强大的泛化能力, 实现对未知类别异常的检测与定位. 现有方法根据是否依赖辅助训练数据可划分为两类: 无需训练的方法和基于辅助数据进行训练的方法.

第一类方法无需任何训练过程, 直接利用预训练模型进行零样本推理. 典型代表如 WinCLIP[13], 该方法通过手工构建描述正常与异常的文本提示, 将图像区域特征与文本特征在嵌入空间中进行余弦相似度比较, 进而通过多尺度特征插值与融合生成像素级异常热力图. 这类方法虽然简单高效, 但其性能在很大程度上依赖于人工设计提示的质量, 且难以适应不同领域的语义差异. 第二类方法则借助辅助数据进行训练, 以提升在目标域上的判别能力. April-GAN[14] 使用辅助数据来训练额外加入的投影层, 以弥合自然图像域与工业异常检测域之间的特征分布差异. AnomalyCLIP[16] 进一步引入可学习的文本嵌入向量, 替代传统手工设计的提示模板, 使模型能够自动学习与异常更相关的语义表示, 提高零样本条件下的检测与定位精度. AdaCLIP[33] 通过结合静态与动态可学习提示的混合提示策略, 有效提升模型在工业与医学领域的零样本异常检测性能. FiLo[34] 则采用细粒度描述与高质量定位相结合的方式, 利用大语言模型为每个类别生成详细的异常类型描述, 显著增强了异常检测的可解释性.

2 问题定义

在零样本异常检测任务中, 给定一个测试样本图像 $I \in \mathbf{R}^{H\times W}$, 模型的目标是同时生成一个图像级别的异常得分 S_{img} 和一个像素级别的异常得分图 $S_{\text{map}} \in \mathbf{R}^{H\times W}$, 这些评分被归一化到区间 $[0,\ 1]$, 其中值越接近 1 表示该图像或像素属于异常的可能性越高. 为训练模型, 使用一个辅助训练数据集 $I_{\text{train}} = \{(I_i,\ G_i)\}_{i=1}^{N_{\text{train}}}$, 其中每对数据包括一幅图像 I, 以及对应的掩码 $G \in \mathbf{R}^{H\times W}$. 在该掩码中, 若某一像素或整幅图像属于正常样本, 则其标注值为 0, 若属于异常, 则标注值为 1. 训练数据中同时包含正常与异常样本, 以便模型学习正常模式与异常模式. 在模型训练完成后, 使用全新的数据集 $I_{\text{test}} = \{I_i\}_{i=1}^{N_{\text{test}}}$ 来对其性能进行评估. 需要注意的是, 训练集与测试集之间满足 $I_{\text{train}} \cap I_{\text{test}} = \emptyset$, 即模型在测试阶段处理的是完全未在训练中出现过的样本.

3 本文方法

3.1 整体架构

本文采用 CLIP 作为主干网络, 利用其图像编码器提取输入图像的多个局部图像块特征和全局图像特征, 同时利用文本编码器处理描述正常与异常状态的语义提示, 并将二者映射到统一的嵌入空间. 通过计算视觉特征与文本特征之间的余弦相似度, 来实现图像级别的异常检测与像素级别的异常定位.

在文本分支, 为区分异常检测与定位任务的不同需求, 本文采用文本解耦策略 (如图 1 所示), 分别为检测与定位任务设计描述正常与异常的提示模板, 经文本编码器处理后, 取各自多个特征向量的均值作为可学习文本嵌入的初始值. 训练过程中, 两组文本嵌入分别通过图像级检测损失与像素级定位损失进行独立优化.

在图像分支, 输入图像经由图像编码器处理后, 同时提取多个局部图像块特征与全局图像特征. 其中, 全局图像特征经过异常特征增强策略优化后, 直接参与检测损失的计算. 同时, 为增强模型对异常区域的跨域敏感性, 在视觉编码器的首个 Transformer 层之后引入一个轻量化投影层, 来优化局部图像块特征与原型的距离. 最终, 经过对齐之后的局部图像块特征被聚合, 用于构建像素级定位损失. 整个模型在保持 CLIP 原始参数固定的前提下, 仅对少量参数进行训练, 实现高效的零样本异常检测与定位.

3.2 文本解耦

在零样本异常检测中, 如何有效设计文本提示

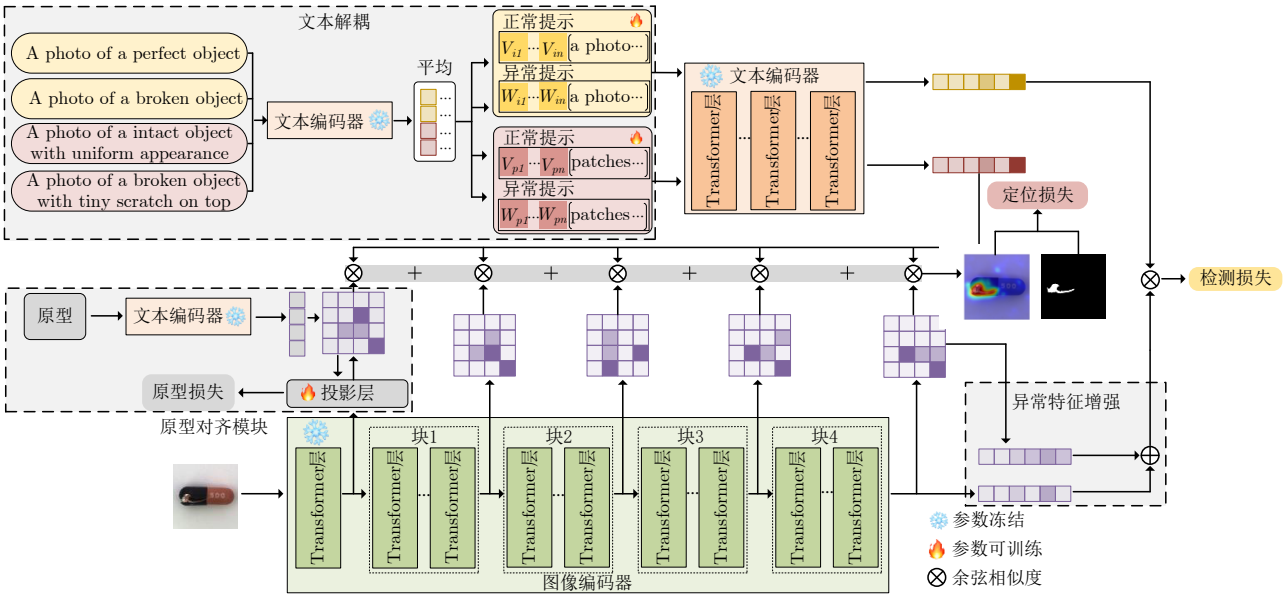


图 1 所提模型结构示意图

Fig. 1 The architecture of the proposed model

是一项关键挑战. 由于异常检测任务通常依赖于高层次的语义特征, 而异常定位任务则更需要捕捉中低层的局部细节特征, 若两者共用同一组文本提示, 可能导致模型在全局分类与局部定位目标之间产生优化冲突, 从而限制整体性能. 为此, 本文提出一种文本解耦策略, 分别为异常检测与异常定位任务设计独立的文本初始化向量, 并对两组可学习的文本嵌入进行分别优化.

针对检测与定位任务的不同需求, 首先设计不同的文本提示来进行初始化, 检测分支的文本提示侧重于描述整体图像的正常或异常状态, 而定位分支的提示则进一步融入局部异常的细节特征 (异常的种类、大小以及位置). 这种设计可以增强模型对像素级特征的理解. 受 WinCLIP[13] 的启发, 本文使用组合提示来设计初始文本提示. 对于检测任务, 提示由模板和状态构成. 假设模板是 A photo of a [] object, 状态可以是 perfect 或是 broken. 从而形成如 A photo of a perfect object 和 A photo of a broken object 的完整提示, 它们分别对应正常与异常样本的语义描述. 在定位任务中, 除模板和状态外, 还引入异常细节描述, 具体包括 with tiny scratch on top 等. 将模板、状态、异常细节描述三者组合后可生成 A photo of a broken object with tiny scratch on top 这类的细粒度提示, 使其更适配于局部异常区域的特征对齐需求. 在得到所有可能的模板、状态词及异常描述组合后, 分别计算每类任务中各组合文本嵌入的均值, 作为检测分支和定位分支可学习文本向量的初始值.

在完成文本初始化后, 本文在 AnomalyCLIP[16] 类别无关的可学习提示的基础上, 提出任务导向的可学习提示机制, 以更好地区分异常检测与定位任务的不同需求. 具体地, 首先将类别无关的可学习提示解耦如下:

$$\begin{cases} g_{in} = [V_{i1}]\,[V_{i2}]\cdots[V_{in}]\,[\text{object}] \\ g_{ia} = [W_{i1}]\,[W_{i2}]\cdots[W_{in}]\,[\text{damaged}]\,[\text{object}] \\ g_{pn} = [V_{p1}]\,[V_{p2}]\cdots[V_{pn}]\,[\text{object}] \\ g_{pa} = [W_{p1}]\,[W_{p2}]\cdots[W_{pn}]\,[\text{damaged}]\,[\text{object}] \end{cases} \tag{1}$$

其中, g_{in} 和 g_{ia} 分别表示检测任务中的正常与异常提示; g_{pn} 和 g_{pa} 分别表示定位任务中的正常与异常提示. $[V]$ 和 $[W]$ 分别代表正常和异常提示模板中的可学习词嵌入, 其参数在训练过程中通过梯度下降进行优化.

然后, 为进一步增强模型的任务区分能力, 在上述提示的基础上提出任务导向的可学习提示, 其定义如下:

$$\begin{cases} g_{in} = \mathcal{V}_i\,[\text{a photo of an object}] \\ g_{ia} = \mathcal{W}_i\,[\text{a photo of a damaged object}] \\ g_{pn} = \mathcal{V}_p\,[\text{patches of an object}] \\ g_{pa} = \mathcal{W}_p\,[\text{patches of a damaged object}] \end{cases} \tag{2}$$

其中,

$$\begin{cases} \mathcal{V}_i = [V_{i1}]\,[V_{i2}]\cdots[V_{in}] \\ \mathcal{W}_i = [W_{i1}]\,[W_{i2}]\cdots[W_{in}] \\ \mathcal{V}_p = [V_{p1}]\,[V_{p2}]\cdots[V_{pn}] \\ \mathcal{W}_p = [W_{p1}]\,[W_{p2}]\cdots[W_{pn}] \end{cases} \tag{3}$$

基于任务导向的可学习提示设计, 将可学习嵌入 $\mathcal{V}_i$ 和 $\mathcal{W}_i$ 分别与固定文本 a photo of an object 和 a photo of a damaged object 进行结合, 构建图像级检测提示 g_{in} 与 g_{ia}. 同时将 $\mathcal{V}_p$ 和 $\mathcal{W}_p$ 分别与 patches of an object 和 patches of a damaged object 结合, 形成像素级定位提示 g_{pn} 与 g_{pa}, 从而实现检测与定位任务在文本表示层面的有效区分与适配.

3.3 原型对齐模块

由于辅助训练数据规模有限, 且难以覆盖所有潜在应用领域和异常类型, 模型在训练阶段接触到的异常模式具有不可避免的局限性. 因此, 如何通过有限的训练样本使模型学习到更具判别力和泛化性的异常特征表示, 成为提升零样本异常定位性能的关键. 为此, 本文提出原型对齐模块, 该模块旨在通过原型学习使用自监督的方式来优化图像块特征.

为增强模型对局部异常的定位能力, 本文选取视觉编码器中局部特征最为丰富的第一个 Transformer 层的输出, 并在该层之后引入一个轻量化的可学习投影层, 用于将图像块特征对齐至预先定义的原型特征空间. 为保证计算效率, 本文采用一种轻量化结构, 该结构由两个 1×1 卷积层构成. 具体而言, 第一个 Transformer 层输出的特征 $\boldsymbol{X}$ 首先经过一次 1×1 卷积进行降维, 通道数缩减为原始维度的一半, 随后通过第二次卷积恢复至原始通道数. 最终, 通过残差连接将卷积后的结果与原始输入 X 相加, 形成投影层的输出 $\boldsymbol{Y}\in\mathbf{R}^{B\times N\times d}$ (B 为批次大小, N 为图像块的个数, d 为嵌入维度).

基于跨域语义对齐需求, 将简单的领域先验知识通过 CLIP 的文本编码器编码为固定的原型特征向量 $\boldsymbol{W}\in\mathbf{R}^{C\times d}$ (C 为原型类别数), 作为跨域语义对齐的锚点. 这些领域先验知识包括 A normal medical photo without damage, A damaged photo with industrial products 等语义概念. 在测试阶段, 对于未见类别的样本, 直接复用训练阶段构建的同一组原型, 无需针对新类别重新构造或调整. 这是因为原型描述聚焦于正常和异常的通用语义, 而非特定物体的细粒度概念, 因此能够为零样本场景下的跨域泛化提供稳定的语义锚点.

基于投影后的图像块特征 $\boldsymbol{Y}$ 和预设的原型向量 $\boldsymbol{W}$, 可以通过计算它们的相似度来为每个图像块生成伪标签 y. 具体实现中, 特征 $\boldsymbol{Y}$ 与原型向量 $\boldsymbol{W}$ 分别进行归一化处理以消除尺度影响, 进而计算余弦相似度矩阵 S, 并引入温度系数 τ 以调节相似度分布:

$$S[b,\ n,\ c]=\frac{\langle \mathrm{Norm}(\boldsymbol{Y}),\ [\mathrm{Norm}(\boldsymbol{W})]^{\mathrm{T}}\rangle}{\tau} \tag{4}$$

式中, $S\in\mathbf{R}^{B\times N\times C}$, $\mathrm{Norm}(\cdot)$ 表示归一化操作. 接着, 通过取每个图像块在所有原型类别上的最大相似度所对应的类别索引, 生成对应的伪标签, 即

$$y[b,\ n]=\underset{c\in[1,\ C]}{\arg\max}\, S[b,\ n,\ c] \tag{5}$$

此处, $y\in\mathbf{R}^{B\times N}$ 表示每个图像块被分配的原型类别. 在此基础上, 可以构建原型对齐损失函数 $\mathcal{L}_{pr}$, 从而推动图像块特征向其最接近的原型向量对齐:

$$\mathcal{L}_{pr}=CE\left(\mathcal{F}(S),\ \mathcal{F}(y)\right) \tag{6}$$

其中, CE 表示交叉熵损失函数, $\mathcal{F}(\cdot)$ 代表将张量展平的操作.

原型对齐损失通过最大化图像块特征与对应原型在语义空间中的一致性, 促使模型学习更具判别性的局部表示. 尤其在不同数据域中, 该模块能够强化模型对正常与异常模式的语义理解能力, 进而提升零样本场景下的泛化性能.

3.4 异常特征增强策略

CLIP 的视觉编码器能够通过自注意力机制有效捕捉不同图像块之间的语义关联, 从而生成信息丰富的全局特征表示, 所以现有方法大多直接基于视觉编码器最后一层的全局特征进行分类. 然而, 在异常检测任务中, 由于正常区域通常在图像中占据主导地位, 全局图像特征容易受到大量正常信息的干扰, 导致局部或细微的异常特征在全局表示中被稀释, 进而降低模型对异常特征的敏感度.

针对上述问题, 本文提出一种异常特征增强策略, 核心思想是利用视觉编码器最后一层输出特征的注意力权重, 定位模型在推理过程中关注的重点区域, 进而对全局特征进行调整, 增强潜在异常区域在全局特征中的贡献度, 提升模型对异常特征的感知与判别能力.

具体而言, 假设视觉编码器最后一层输出的注意力权重矩阵为 $A\in\mathbf{R}^{B\times(N+1)\times(N+1)}$, 通过提取类别标记 (class token) 对所有图像块的注意力权重向量 $\boldsymbol{a}=A[:,\ 0,\ 1:]\in\mathbf{R}^{B\times N}$ 可得到每个图像块对分类决策的重要性. 为使用注意力权重识别潜在的异常区域, 需将其转换为二值掩码 $M\in\{0,\ 1\}^{B\times N}$, 其定义如下:

$$M=\begin{cases}1, & \text{若 } \boldsymbol{a}\geq\mu(\boldsymbol{a})+\alpha*\sigma(\boldsymbol{a})\\ 0, & \text{否则}\end{cases} \tag{7}$$

式中, $\mu(\boldsymbol{a})$ 和 $\sigma(\boldsymbol{a})$ 分别表示注意力权重 $\boldsymbol{a}$ 的平均值和标准差, 超参数 α 用于控制异常区域筛选阈值

的敏感度, 在本文实验中将其设置为 $\alpha = 1.5$. 该掩码用于标识显著高于平均关注程度的图像块, 即最有可能包含异常的区域.

基于掩码 $\boldsymbol{M}$, 可以计算异常区域特征 $\boldsymbol{F}_{\mathrm{ar}} \in \mathbf{R}^{B\times d}$, 其计算方式是对掩码值为 1 的图像块特征取平均:

$$\boldsymbol{F}_{\mathrm{ar}} = \frac{\sum_{j=1}^{N} \boldsymbol{M}[b,\ j] \cdot \boldsymbol{F}_{\mathrm{p}}\,[b,\ j]}{\sum_{j=1}^{N} \boldsymbol{M}[b,\ j]} \tag{8}$$

其中, $\boldsymbol{F}_{\mathrm{p}}$ 为图像块特征. 最终, 通过将原始全局特征 $\boldsymbol{F}_{\mathrm{g}}$ 与异常区域特征 $\boldsymbol{F}_{\mathrm{ar}}$ 进行自适应融合, 得到增强后的全局特征 $\boldsymbol{F}_{\mathrm{eg}}$:

$$\boldsymbol{F}_{\mathrm{eg}} = \mathrm{Norm}\left[(1-\beta)\cdot \boldsymbol{F}_{\mathrm{g}} + \beta \cdot \boldsymbol{F}_{\mathrm{ar}}\right] \tag{9}$$

其中, $\beta \in [0,\ 1]$ 为融合权重.

在图像级异常检测损失的计算中, 本文采用增强后的全局特征 $\boldsymbol{F}_{\mathrm{eg}}$, 以利用其包含的异常区域信息, 提升模型对异常样本的判别能力.

3.5 训练和推理流程

3.5.1 CLIP 的训练推理流程

CLIP 模型由一个视觉编码器 $f(\cdot)$ 和一个文本编码器 $g(\cdot)$ 构成, 两者通常采用主流骨干网络 (如 ViT[35]) 实现. 该模型在大规模图像–文本对数据集 $\{(x_t,\ s_t)\}_{t=1}^{T}$ 上进行训练, 通过对比学习拉近匹配的图像–文本对在嵌入空间中的距离, 同时推远非匹配对的距离. 其训练过程旨在最大化配对样本 $f(x_t)$ 与 $g(s_t)$ 之间的余弦相似度 $\langle f(x_t),\ g(s_t)\rangle$, 从而实现跨模态语义对齐.

在推理阶段, 给定类别集合 $C = \{c_1,\ \cdots,\ c_n\}$ 和一个测试样本图像 I, CLIP 可通过下式计算 I 属于类别 c 的概率, 即

$$p\left(y = c \mid I\right) = \frac{\exp\left(\frac{\langle f(I),\ g(c)\rangle}{\tau}\right)}{\sum\limits_{c\in C}\exp\left(\frac{\langle f(I),\ g(c)\rangle}{\tau}\right)} \tag{10}$$

基于上述概率公式, 计算测试样本图像属于正常或异常类别的概率, 以此来实现图像级别的异常检测. 在异常定位任务中, 由于需要获取更细粒度的信息, 通常提取视觉编码器中间层的特征输出, 利用其丰富的局部特征实现精准的异常定位.

3.5.2 损失函数

在异常检测任务中, 通过计算增强后的全局特征 $\boldsymbol{F}_{\mathrm{eg}}$ 与检测任务的可学习提示之间的语义相似度, 来检测图像是否异常, 并使用交叉熵损失进行优化. 具体而言, 在获得增强全局特征 $\boldsymbol{F}_{\mathrm{eg}}$ 以及可学习的正常提示 g_{in} 和异常提示 g_{ia} 后, 将二者拼接为联合文本表示 $g_i = [g_{in},\ g_{ia}]$, 并通过如下损失函数进行优化:

$$\mathcal{L}_{\mathrm{det}} = \mathrm{CE}\left(\frac{\mathrm{Norm}\left(\boldsymbol{F}_{\mathrm{eg}}\right)\cdot[\mathrm{Norm}(g_i)]^{\mathrm{T}}}{\tau},\ t\right) \tag{11}$$

其中, t 为图像级别的真实标签 (0 表示正常, 1 表示异常). $\mathcal{L}_{\mathrm{det}}$ 通过最大化增强特征与对应类别文本提示之间的相似度, 促使模型准确识别全局图像中的异常情况.

在异常定位任务中, 提取视觉编码器中多个 ($k \in K$) 中间层的输出特征 $\boldsymbol{Y}^k \in \mathbf{R}^{B\times N\times d}$. 针对每一层特征, 计算其与定位任务相关文本提示的相似度, 得到相应的异常得分图 $\boldsymbol{M}^k \in \mathbf{R}^{B\times C\times H\times W}$ (H 和 W 分别为得分图的空间高度与宽度). 最后, 通过插值操作将 $\boldsymbol{M}^k$ 上采样至与真实像素级标签相同的空间尺寸. 定位损失的表达式如下:

$$\begin{aligned}\mathcal{L}_{\mathrm{loc}} = \sum_{k=1}^{K}\Big[&\mathcal{L}_{\mathrm{focal}}\left(M^k,\ \boldsymbol{Y}\right) + \mathcal{L}_{\mathrm{dice,\,1}}\left(M^k,\ \boldsymbol{Y}\right) + \\ &\mathcal{L}_{\mathrm{dice,\,0}}\left(M^k,\ \boldsymbol{Y}\right)\Big]\end{aligned} \tag{12}$$

其中, $\boldsymbol{Y} \in \{0,\ 1\}^{B\times H\times W}$ 为像素级真实标签, $\mathcal{L}_{\mathrm{focal}}$ 表示焦点损失[36], $\mathcal{L}_{\mathrm{dice,\,1}}$ 和 $\mathcal{L}_{\mathrm{dice,\,0}}$ 分别为异常区域和正常区域上的骰子损失[37].

综上所述, 本文方法的总体目标函数由原型对齐损失、检测损失与定位损失三部分构成, 其表达式如下:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{pr}} + \mathcal{L}_{\mathrm{det}} + \mathcal{L}_{\mathrm{loc}} \tag{13}$$

3.5.3 模型推理过程

在测试阶段, 给定输入图像 I, 使用式 (10) 来计算其被判定为异常的概率, 并将该概率值作为图像级别的异常得分 S_{img}. 当图像特征与异常文本嵌入在语义空间中高度对齐时, 得分趋近于 1, 表明该图像存在异常.

对于像素级异常定位, 提取视觉编码器中多层的中间特征, 并计算其与异常提示文本的相似度集合 $\boldsymbol{M}^k \in \mathbf{R}^{B\times C\times H\times W}$, 使用下式获得最终的像素级异常得分图 S_{map}:

$$S_{\mathrm{map}} = G_\sigma\left(\sum_{k\in K}\left(\frac{1}{2}\left(\boldsymbol{E} - M^k\right) + \frac{1}{2}M^k\right)\right) \tag{14}$$

其中, G_σ 表示标准差为 σ 的高斯平滑滤波器, $\boldsymbol{E}$ 为全 1 矩阵.

4 实验结果与分析

4.1 实验设置

本文在 12 个公开数据集上进行广泛的实验,

以全面评估所提出方法的性能. 这些数据集涵盖工业和医学领域. 在工业异常检测方面, 选用 MVTec AD[5]、VisA[38]、MPDD[39]、BTAD[40]、DAGM[41] 和 DTD-Synthetic[42] 六个基准数据集. 在医学领域, 本文使用皮肤癌检测数据集 ISIC[43]、结肠息肉检测数据集 CVC-ClinicDB[44] 和 CVC-ColonDB[45], 以及脑肿瘤检测数据集 Head_CT[46]、BrainMRI[47] 和 Br35H[48]. 这些数据集的详细统计信息如表 1 所示.

在异常检测任务的评估中, 本文采用图像级别的受试者工作特性曲线下面积 (area under the receiver operating characteristic, AUROC) 和 平均精度 (average precision, AP) 作为主要评价指标, 以全面衡量模型在区分正常与异常图像方面的性能. 对于异常定位任务, 本文使用像素级别的 AUROC 以及异常区域重叠度曲线下面积 (area under the PRO curve, AUPRO)[23] 作为评价指标.

为评估所提方法的性能, 本文在多个基准数据集上与当前先进的零样本异常检测方法进行比较, 包括基于大规模视觉–语言预训练的模型 CLIP[12]、基于可学习提示的方法 CoOp[30]、采用多尺度窗口匹配的 WinCLIP[13]、使用特征适配器的 April-GAN[14] 以及引入可学习提示的 AnomalyCLIP[16].

本文采用公开的预训练 CLIP 模型 (VIT-L/14@336px) 作为基础骨干网络, 并保持其所有权重冻结. 在特征提取方面, 从视觉编码器的第 1、6、12、18 和 24 层提取局部特征, 以获取多层次的语义信息. 所有输入图像均统一缩放至 518×518 分辨率. 在模型配置方面, 可学习提示的长度设为 12, 批次大小设置为 8. 优化器选用 Adam, 其中可学习提示的学习率为 0.001, 投影层的学习率为 0.005, 特征融合权重 β 设为 0.1.

考虑到零样本检测设定, 本文使用 MVTec AD 数据集对模型进行训练, 并在其余 11 个数据集上进行测试. 需特别说明的是, 在 MVTec AD 数据集上报告的测试结果, 其模型是在 VisA 数据集上训练得到的, 以体现严格的零样本评估设定. 训练共进行 15 个轮次, 自第 5 轮起冻结异常定位分支的可学习向量及投影层参数, 以稳定训练过程并避免过拟合. 在推理阶段, 使用标准差 $\sigma = 4$ 的高斯滤波器对生成的异常分数图进行平滑处理.

4.2　工业数据集零样本异常检测结果

表 2 详细展示了所提方法在工业数据集上的零样本异常检测结果. 从结果可以看出, 基线方法 CLIP[12] 及其提示优化版本 CoOp[30] 性能相对有限, 这主要是由于 CLIP 模型在预训练阶段主要聚焦于对象级语义对齐, 而非细粒度的异常特征判别. 通过引入手工设计的相关文本提示, WinCLIP[13] 能够在一定程度上提升异常检测的准确性. 进一步地, 通过利用辅助数据集进行轻量微调的方法 (如 April-GAN[14] 和 AnomalyCLIP[16]) 取得更为显著的性能提升. 这表明在零样本设定下, 辅助数据能够为模型提供有效的异常语义先验知识, 从而提升模型异常检测的能力.

在工业数据集上的实验结果表明, 本文方法在零样本异常检测任务中展现出卓越的综合性能. 在图像级异常检测方面, 本文方法在六个工业数据集上的平均 AUROC 和 AP 分别达到 89.0% 和 91.2%, 其中在四个数据集上取得最优性能. 特别是在 BTAD 数据集上, 本文方法以 91.8% 的 AUROC 和 94.8% 的 AP 显著超越所有对比方法, 显示出优异的异常检测能力. 在像素级异常定位任务中, 本

表 1　数据集的关键统计数据

Table 1　Key statistics the utilized datasets

数据集	领域	类别	模态	类别数	测试集中正常与异常样本数	任务
MVTec AD	工业	物体和纹理	摄影	15	(467, 1 258)	异常检测与定位
VisA		物体类		12	(962, 1 200)	
MPDD		物体类		6	(176, 282)	
BTAD		物体类		3	(451, 290)	
DAGM		纹理类		10	(6 996, 1 054)	
DTD-Synthetic		纹理类		12	(357, 947)	
ISIC	医学	皮肤	摄影	1	(0, 379)	异常定位
CVC-ClinicDB		结肠	内窥镜		(0, 612)	
CVC-ColonDB		结肠	内窥镜		(0, 380)	
Head_CT	医学	脑部	放射学 (CT)	1	(100, 100)	异常检测
BrainMRI			放射学 (MRI)		(98, 155)	
Br35H			放射学 (MRI)		(1 500, 1 500)	

表 2 不同方法在工业领域的性能比较（%）

Table 2 Comparison of the performance of different methods in the industrial domain (%)

任务	数据集	CLIP[12]	CoOp[30]	WinCLIP[13]	April-GAN[14]	AnomalyCLIP[16]	本文方法
图像级 (AUROC, AP)	MVTec AD	(74.1, 87.6)	(88.8, 94.8)	(91.8, 96.5)	(86.1, 93.5)	(91.5, 96.2)	(**92.0**, **96.8**)
	VisA	(66.4, 71.5)	(62.8, 68.1)	(78.1, 81.2)	(78.0, 81.4)	(82.1, 85.4)	(**83.7**, **85.8**)
	MPDD	(54.3, 65.4)	(55.1, 64.2)	(63.6, 69.9)	(73.0, 80.2)	(**77.0**, **82.0**)	(76.0, 80.6)
	BTAD	(34.5, 52.5)	(66.8, 77.4)	(68.2, 70.9)	(73.6, 68.6)	(88.3, 87.3)	(**91.8**, **94.8**)
	DAGM	(79.6, 59.0)	(87.5, 74.6)	(91.8, 79.5)	(94.4, 83.8)	(97.5, 92.3)	(**97.9**, **92.5**)
	DTD-Synthetic	(71.6, 85.7)	(-, -)	(93.2, 92.6)	(86.4, 95.0)	(**93.5**, 97.0)	(93.1, **97.1**)
	平均值	(63.4, 70.3)	(72.2, 75.8)	(81.1, 81.8)	(82.6, 83.8)	(88.3, 90.0)	(**89.0**, **91.2**)
像素级 (AUROC, AUPRO)	MVTec AD	(38.4, 11.3)	(33.3, 6.7)	(85.1, 64.6)	(87.6, 44.0)	(**91.1**, 81.4)	(90.8, **83.2**)
	VisA	(46.6, 14.8)	(24.2, 3.8)	(79.6, 56.8)	(94.2, 86.8)	(**95.5**, 87.0)	(95.3, **90.9**)
	MPDD	(62.1, 33.0)	(15.4, 2.3)	(76.4, 48.9)	(94.1, 83.2)	(96.5, 88.7)	(**97.4**, **91.9**)
	BTAD	(30.6, 4.4)	(28.6, 3.8)	(72.7, 27.3)	(60.8, 25.0)	(94.2, 74.8)	(**94.5**, **80.0**)
	DAGM	(28.2, 2.9)	(17.5, 2.1)	(87.6, 65.7)	(82.4, 66.2)	(**95.6**, **91.0**)	(**95.6**, 90.0)
	DTD-Synthetic	(33.9, 12.5)	(-, -)	(83.9, 57.8)	(95.3, 86.9)	(97.9, 92.3)	(**98.4**, **95.1**)
	平均值	(40.0, 13.2)	(23.8, 3.7)	(80.9, 53.5)	(85.7, 65.4)	(95.1, 85.9)	(**95.3**, **88.5**)

注: 加粗字体表示该方法在该数据集上取得最佳性能, 下划线表示次优性能.

文方法表现更为突出, 平均 AUROC 和 AUPRO 分别达到 95.3% 和 88.5%, 体现了优异的局部异常区域定位精度. 与之前的最佳方法 AnomalyCLIP 相比, 本文方法在图像级检测的 AP 指标上提升 1.3%, 在像素级定位的 AUPRO 指标上提升 3.0%, 这一显著的性能提升充分证明了本文方法的有效性和优越性.

4.3 医学数据集零样本异常检测结果

为验证模型在不同领域的泛化能力, 表 3 展示了本文方法在医学影像异常检测任务中的实验结果. 在图像级异常检测方面, 本文方法在 Head_CT、BrainMRI 和 Br35H 三个脑部医学数据集上均取得最优性能, 平均 AUROC 和 AP 分别达到 95.2% 和 95.7%, 较次优方法 AnomalyCLIP 分别提升 2.6% 和 3.1%, 展现出优异的跨模态适应能力. 在像素级异常定位任务中, 本文方法在 ISIC 皮肤癌数据集和 CVC-ClinicDB 结肠息肉数据集上均获得最佳性能, 平均 AUROC 和 AUPRO 分别达到 85.1% 和 74.1%, 显著优于其他对比方法. 这些结果充分证明了本文方法在复杂医学影像场景下具有良好的泛化性和鲁棒性.

4.4 消融实验

为深入评估模型中各组件的有效性, 本文在多个数据集上进行一系列的消融实验, 所选数据集涵盖工业与医疗领域, 具体包括: VisA[38]、BTAD[40]、CVC-ClinicDB[44]、ISIC[43]、Head_CT[46] 以及 BrainMRI[47].

4.4.1 文本解耦的有效性

表 4 展示了文本解耦对模型性能的消融实验结

表 3 不同方法在医学领域的性能比较（%）

Table 3 Comparison of the performance of different methods in the medical domain (%)

任务	数据集	CLIP[12]	CoOp[30]	WinCLIP[13]	April-GAN[14]	AnomalyCLIP[16]	本文方法
图像级 (AUROC, AP)	Head_CT	(56.5, 58.4)	(78.4, 78.8)	(81.8, 80.2)	(89.1, 89.4)	(93.4, 91.6)	(**94.1**, **94.7**)
	BrainMRI	(73.9, 81.7)	(61.3, 44.9)	(86.6, 91.5)	(89.3, 90.9)	(90.3, 92.2)	(**95.0**, **95.9**)
	Br35H	(78.4, 78.8)	(86.0, 87.5)	(80.5, 82.2)	(93.1, 92.9)	(94.6, 94.7)	(**96.5**, **96.5**)
	平均值	(69.6, 73.0)	(75.2, 70.4)	(83.0, 84.6)	(90.5, 91.1)	(92.8, 92.8)	(**95.2**, **95.7**)
像素级 (AUROC, AUPRO)	ISIC	(33.1, 5.8)	(51.7, 15.9)	(83.3, 55.1)	(89.4, 77.2)	(89.7, 78.4)	(**90.5**, **82.2**)
	CVC-ColonDB	(49.5, 15.8)	(40.5, 2.6)	(70.3, 32.5)	(78.4, 64.6)	(**81.9**, **71.3**)	(80.0, 70.6)
	CVC-ClinicDB	(47.5, 18.9)	(34.8, 2.4)	(51.2, 13.8)	(80.5, 60.7)	(82.9, 67.8)	(**84.9**, **69.4**)
	平均值	(43.4, 13.5)	(42.3, 7.0)	(68.3, 33.8)	(82.8, 67.5)	(84.8, 72.5)	(**85.1**, **74.1**)

注: 加粗字体表示该方法在该数据集上取得最佳性能, 下划线表示次优性能.

果. 在消融实验中, 基线模型使用随机初始化的单一文本提示同时处理异常检测与定位任务. 此外, 使用如下的损失函数进行训练:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda \cdot \mathcal{L}_{\text{loc}} \tag{15}$$

通过调节权重系数 λ 改变定位损失与检测损失在总损失中的比例, 进而分析不同比例对模型性能的影响.

由表 4 可知, 文本解耦策略的引入显著提升了模型性能的稳定性, 本文模型在不同 λ 参数设置下的结果保持不变, 而基线模型在不同 λ 取值下性能波动显著. 在 VisA 和 BTAD 数据集中, 随着 λ 值的减小, 图像级检测指标明显提升. 然而当 λ 从 0.5 降至 0.25 时, 像素级定位性能则出现明显下降. 这一现象表明, 使用单一文本提示难以有效平衡异常检测与异常定位任务之间的优化目标, 导致二者无法同时达到最优性能.

更重要的是, 本文方法在提升稳定性的同时, 也显著增强了模型的异常检测能力. 在 BTAD 数据集上, 图像级 AUROC 从基线最优的 63.8% 提升至 91.6%, AP 由 67.5% 提升至 94.1%. 像素级 AUROC 从 93.3% 改善至 93.6%, AUPRO 由 78.2% 提升至 79.3%. 同时, 在 VisA 数据集上, 图像级与像素级指标也都观察到提升. 消融实验结果表明, 文本解耦策略不仅使模型保持稳定性, 还实现了整体性能的全面提升.

此外, 本文还设计可视化实验, 直观验证了共享文本提示导致的优化冲突问题, 以及本文提出的文本解耦策略的有效性. 与上述实验设置相同, 通过逐步增大定位损失在总损失函数中的权重系数 λ 来观察模型输出, 分别可视化定位任务和检测任务中文本特征与视觉特征之间的余弦相似度. 图 2 展示了在定位任务中文本与每个图像块的相似度, 颜色越深代表相似度越高. λ 从左到右分别为 0.25、0.5、1、2、4. 表 5 展示了检测任务中文本与全局特征计算相似度后的异常得分.

表 4　文本解耦策略消融实验结果

Table 4　Ablation study results of text decoupling strategy

方法	数据集	不同 λ 值的性能对比				
		$\lambda = 4$	$\lambda = 2$	$\lambda = 1$	$\lambda = 0.5$	$\lambda = 0.25$
基线模型	VisA	(46.4, 57.0)	(58.0, 65.1)	(68.2, 72.6)	(76.3, 79.6)	(79.8, 82.3)
		(94.6, 90.1)	(95.0, 90.1)	(95.1, 89.8)	(95.0, 88.9)	(94.2, 85.0)
	BTAD	(27.7, 49.2)	(32.3, 50.6)	(42.5, 55.1)	(52.0, 59.8)	(63.8, 67.5)
		(93.2, 78.2)	(93.3, 78.1)	(93.0, 77.3)	(92.2, 74.8)	(90.2, 69.3)
基线模型 + 文本解耦	VisA	(82.5, 84.5)	(82.5, 84.5)	(82.5, 84.5)	(82.5, 84.5)	(82.5, 84.5)
		(95.2, 90.6)	(95.2, 90.6)	(95.2, 90.6)	(95.2, 90.6)	(95.2, 90.6)
	BTAD	(91.6, 94.1)	(91.6, 94.1)	(91.6, 94.1)	(91.6, 94.1)	(91.6, 94.1)
		(93.6, 79.3)	(93.6, 79.3)	(93.6, 79.3)	(93.6, 79.3)	(93.6, 79.3)

注: 对于每一个数据集, 单元格中的第 1 行数据是图像级指标 (AUROC, AP), 第 2 行数据是像素级指标 (AUROC, AUPRO).

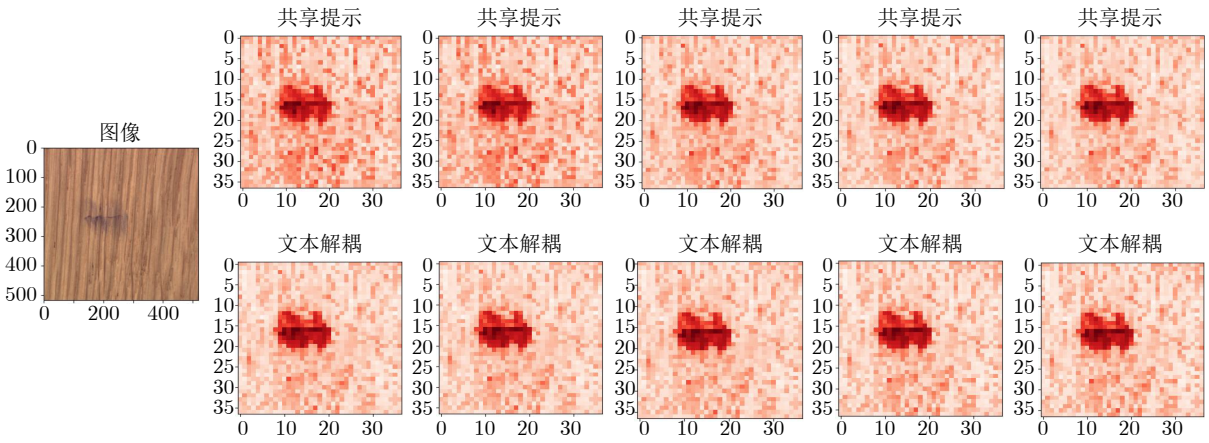


图 2　定位相似度可视化

Fig. 2　Visualization of localization similarity

表 5 检测任务的异常得分
Table 5 Abnormal scores in the detection task

	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$	$\lambda = 4$
共享提示	0.0795	0.0447	0.0404	0.0177	0.0159
文本解耦	0.1308	0.1543	0.1289	0.1273	0.1273

如图 2 所示, 当使用共享提示时, 随着定位损失权重 λ 的增大, 模型对正常区域的响应逐渐降低, 表明其异常定位能力有所增强. 然而, 从表 5 的检测异常得分可以看出, 检测分数持续下降, 意味着模型的异常检测能力在不断削弱. 这一结果清晰地揭示了优化冲突的存在: 共享文本提示被迫同时服务于不同粒度的语义需求, 当模型过度关注细粒度的定位特征时, 检测性能随之受损. 相比之下, 采用本文提出的文本解耦策略后, 检测与定位任务的相似度在不同 λ 取值下均保持稳定, 这表明, 独立的文本提示能够有效缓解优化冲突, 使两个任务各自维持最佳的语义对齐, 互不干扰.

4.4.2 原型对齐模块的有效性

根据表 6 所示的原型对齐模块消融实验结果, 所有参与评估的数据集在像素级指标上均取得提升, 在医学领域, 该模块使 CVC-ClinicDB 数据集的 AUPRO 指标提升约 0.6%. 在 ISIC 数据集上, AUPRO 指标提升约 1.9%. 在工业领域, VisA 数据集的 AUROC 和 AUPRO 分别提升约 0.1% 和 0.3%. BTAD 数据集的 AUROC 和 AUPRO 则分别提升约 1.0% 和 0.9%. 结果验证了该模块在增强模型语义理解能力方面的有效性. 即使不引入额外的医学数据集进行训练, 模型在跨域定位任务中仍能取得优于基线方法的性能. 这主要得益于模型可以从预先定义好的原型中学习不同领域的知识, 从而增强了模型对不同领域中正常与异常模式的判别能力, 最终提升了零样本设定下的泛化性能.

此外, 本文还针对原型对齐模块从原型数量和原型语义设计开展实验来验证该模块的稳定性和合理性. 为验证原型数量对模型性能的影响, 本文设置原型数量 $C \in \{4, 8, 12, 16\}$ 进行实验. 实验结果如表 7 所示.

表 6 原型对齐模块消融实验结果
Table 6 Ablation study results of prototype alignment module

原型对齐	医学领域数据集		工业领域数据集	
	CVC-ClinicDB	ISIC	VisA	BTAD
×	(84.8, 69.0)	(**90.5**, 80.7)	(95.2, 90.6)	(93.6, 79.3)
✓	(**84.9**, **69.4**)	(**90.5**, **82.2**)	(**95.3**, **90.9**)	(**94.5**, **80.0**)

注: 单元格中的数据是像素级指标 (AUROC, AUPRO).

表 7 的实验结果表明, 不同原型数量下的性能差异极小, 表明原型对齐模块对原型数量不敏感, 具有良好的稳定性. 即使仅使用 4 个原型, 也能取得与 16 个原型相近的性能.

在原型语义设计方面, 本文对比三种原型构造策略: 随机初始化原型向量、为数据集定制描述 (如“a normal colonoscopy image without polyps”)、使用通用描述 (如“a normal medical photo”). 相应的实验结果见表 8.

根据表 8 结果可知, 随机初始化的原型缺乏语义信息, 无法为图像块特征提供有意义的对齐锚点. 其性能甚至低于未使用原型对齐模块的基线, 表明无意义的原型可能引入噪声, 干扰模型学习. 由此验证了原型需要具备语义信息, 而非任意向量. 为数据集定制描述尽管能够取得较好的性能, 然而其效果会略低于通用描述. 其原因是特定描述虽然更贴合数据集, 但可能导致过拟合. 并且在零样本设定下, 测试集与训练集领域可能存在差异, 特定描述反而限制了模型的跨域泛化能力. 使用通用描述可以取得最佳性能, 原因在于通用描述不局限于具体的物体, 而是聚焦于正常和异常这一核心语义. 在工业数据集上预训练后, 迁移到医学数据集时仍能保持良好性能语义稳定性.

4.4.3 异常特征增强策略的有效性

根据表 9 的消融实验结果分析, 异常特征增强策略改善了模型的检测性能. 在引入该策略后, 模型在工业与医疗领域的评估指标均有明显提升, 尤其在医学影像数据上效果突出. Head_CT 和 BrainMRI 数据集的 AP 指标分别提升了 2.8% 和

表 7 不同原型数量的实验结果
Table 7 Experimental results for different numbers of prototypes

	$C = 16$	$C = 12$	$C = 8$	$C = 4$
CVC-ClinicDB	(84.9, 69.4)	(84.9, 69.4)	(85.0, 69.2)	(84.9, 69.4)
ISIC	(90.5, 82.2)	(90.6, 82.0)	(90.5, 82.2)	(90.5, 82.2)
VisA	(95.3, 90.9)	(95.4, 90.9)	(95.3, 90.9)	(95.3, 90.9)
BTAD	(94.3, 79.8)	(94.3, 80.0)	(94.2, 80.1)	(94.5, 80.0)

表 8 不同构造策略下的实验结果
Table 8 Experimental results under different construction strategies

	随机初始化	特定描述	通用描述
CVC-ClinicDB	(84.1, 66.7)	(85.0, 69.3)	(84.9, 69.4)
ISIC	(80.9, 80.8)	(90.5, 81.9)	(90.5, 82.2)
VisA	(94.3, 90.2)	(95.3, 90.7)	(95.3, 90.9)
BTAD	(93.4, 79.6)	(94.0, 79.5)	(94.5, 80.0)

表 9　异常特征增强策略消融实验结果
Table 9　Ablation study results of anomaly feature enhancement strategy

特征增强	医学领域数据集		工业领域数据集	
	Head_CT	BrainMRI	VisA	BTAD
×	(92.3, 92.1)	(92.7, 93.5)	(82.5, 84.5)	(91.6, 94.1)
✓	(**94.1**, **94.7**)	(**95.0**, **95.9**)	(**83.7**, **85.8**)	(**91.8**, **94.8**)

注: 单元格中的数据是图像级指标 (AUROC, AP).

2.6%, 显示出该策略能够有效增强模型对医学图像中细微病变特征的感知能力. 总体而言, 通过强化模型对异常区域的敏感度, 该策略提升了异常检测的精度.

表 10 展示了在不同融合权重下的实验结果. 由结果可知, 当融合权重为 0.1 时, 在工业数据集和医学数据集上的平均性能最佳. 当融合权重为 0.3 和 0.5 时, 在工业数据集上性能保持较好, 但在医学数据集上出现明显下降. 其原因是本文使用工业数据集进行预训练, 模型在工业数据上的注意力分布更适配工业图像的异常模式. 当融合权重较高时, 在两类数据集上性能均显著下降. 其原因是过高的权重导致过度依赖异常区域特征, 可能引入噪声或丢失全局上下文信息.

4.5　参数量与推理速度对比

表 11 展示了不同方法的参数量与推理速度对比. 从中可以看出, 本文方法的可训练参数量约为 62.9 万个, 仅为 AnomalyCLIP 的 11.4%, 显著降低了训练开销. 推理速度为 268 ms/幅, 较 AnomalyCLIP 提升 21%. 综合来看, 本文方法在取得最优检测与定位性能的同时, 与 AnomalyCLIP 相比可训练参数量更少、推理速度更快, 实现了性能与效率的良好平衡.

表 10　不同融合权重下的实验结果
Table 10　Experimental results under different fusion weights

	融合权重				
	0.1	0.3	0.5	0.7	0.9
Head_CT	(94.1, 94.7)	(93.3, 93.7)	(92.1, 93.4)	(90.4, 92.3)	(83.8, 87.8)
BrainMRI	(95.0, 95.9)	(94.9, 95.5)	(94.9, 95.4)	(94.3, 95.0)	(92.9, 94.3)
VisA	(83.7, 85.8)	(83.8, 85.9)	(83.9, 85.8)	(79.7, 82.8)	(76.4, 80.3)
BTAD	(91.8, 94.8)	(92.4, 95.1)	(92.7, 95.4)	(92.4, 94.7)	(93.7, 93.8)
平均值	(91.1, 92.8)	(91.1, 92.6)	(90.9, 92.5)	(89.2, 91.2)	(86.7, 89.1)

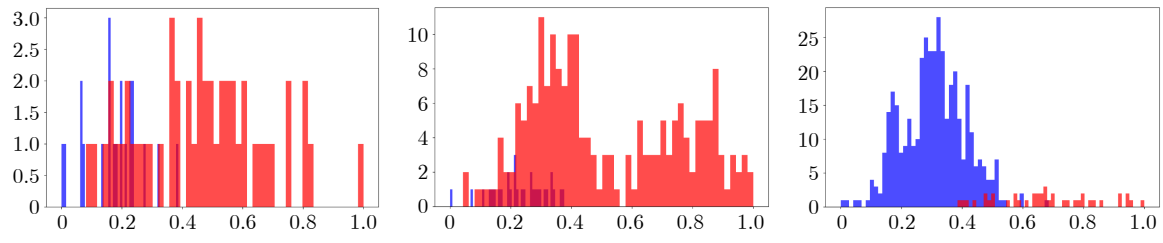
4.6　可视分析

针对异常检测任务, 图 3 和图 4 分别展示了来自 BTAD 数据集和 Head_CT、BrainMRI、Br35H 医学数据集的异常得分分布直方图. 其中, BTAD 数据集进一步细分为 01、02、03 三个子类别. 图中横轴为模型输出的异常得分, 纵轴为对应得分区间的图像数量, 红色和蓝色分别代表异常图像和正常图像的预测得分. 从分布中可清晰观察到, 异常图像的得分集中分布于高位区间, 而正常图像的得分普遍处于低位区间. 这一结果表明, 本文模型能够有效区分正常与异常样本, 展现出优异的异常检测能力.

图 5 ~ 8 展示了不同方法在多个数据集上的异常定位效果对比. 对于每一幅图像, 从左到右依次是输入图像、真值、本文方法、AnomalyCLIP 和 WinCLIP. 可视结果显示, 由于未使用辅助训练集进行参数优化, WinCLIP 的定位结果存在明显的区域不完整问题. 相比之下, 本文提出的方法能够对不同尺寸, 形态的异常区域实现更完整且准确的定位, 体现了模型优异的泛化能力. 图 5 第四行展

表 11　不同方法的参数量与推理速度对比
Table 11　Comparison of parameters and inference speed of different methods

方法	可训练的参数量 (个)	推理速度 (ms/幅)
WinClip	0	253
AnomalyCLIP	5 525 500	339
本文方法	629 376	268



(a) 01　(b) 02　(c) 03

图 3　BTAD 数据集异常得分直方图
Fig. 3　Histogram of anomaly scores on BTAD

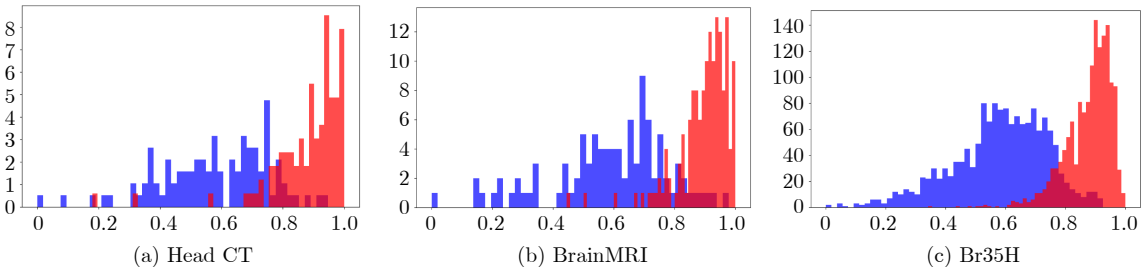


图 4 医学数据集异常得分直方图

Fig. 4 Histogram of anomaly scores on medical datasets

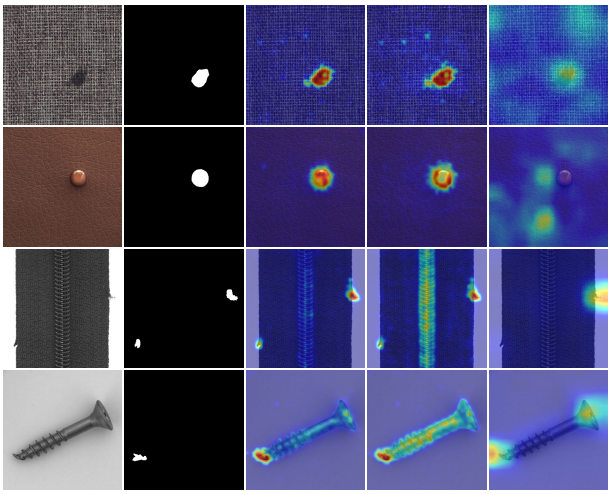


图 5 不同方法在 MVtec AD 数据集上的可视分析

Fig. 5 The visual analysis of different methods on the MVtec AD dataset

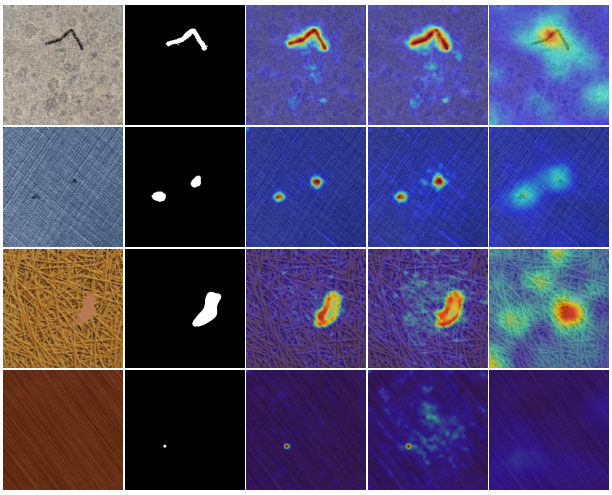


图 7 不同方法在 DTD-Synthetic 数据集上的可视分析

Fig. 7 The visual analysis of different methods on the DTD-Synthetic dataset

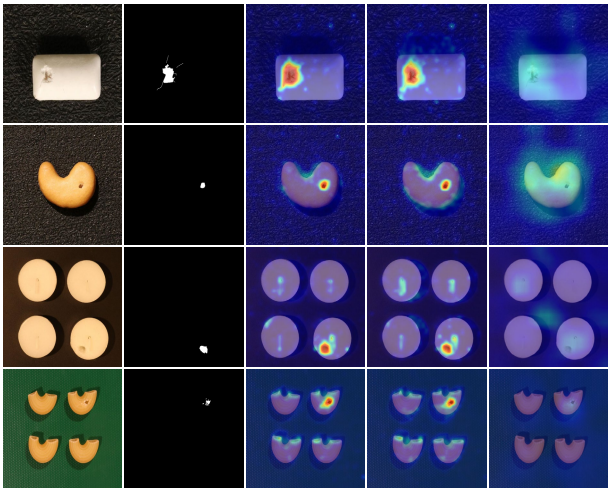


图 6 不同方法在 VisA 数据集上的可视分析

Fig. 6 The visual analysis of different methods on the VisA dataset

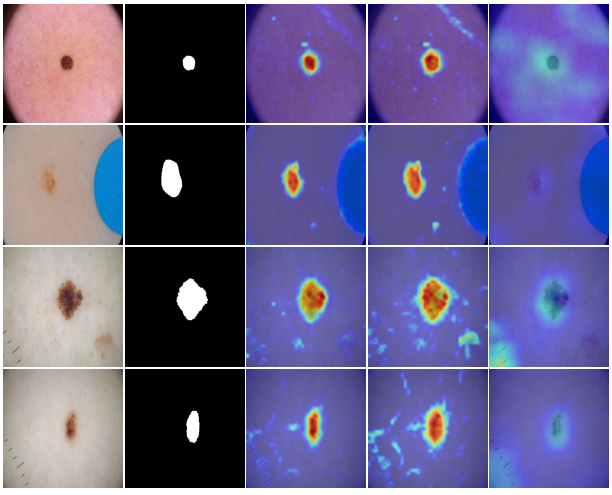


图 8 不同方法在 ISIC 数据集上的可视分析

Fig. 8 The visual analysis of different methods on the ISIC dataset

示了螺丝类别的异常定位结果. 可视对比表明, 本文方法对螺丝尾端异常的定位更加精确. 相比之下, AnomalyCLIP 将整个的螺丝钉区域都误判为异常, 而 WinCLIP 则将螺丝的顶部和尾部都识别为异常. 图 8 展示了在 ICIS 数据集上的皮肤异常定位结果, 本文方法准确定位了病变区域, 而 AnomalyCLIP

将部分正常皮肤识别为异常, 反映出其在部分场景下存在显著的过检问题.

5 结束语

本文提出一种基于文本解耦的零样本异常检测方法. 该方法核心在于分别为异常检测与异常定位任务设计独立的可学习文本提示, 以优化任务间的目标冲突. 同时, 设计原型对齐模块, 通过自监督方式使图像块特征和原型语义进行对齐, 从而增强模型在零样本场景下的跨域泛化能力. 此外, 引入的异常特征增强策略能够有效聚焦潜在异常区域. 强化模型对细微异常特征的感知与判别能力. 在多个数据集上的实验结果验证了所提方法的有效性, 其异常检测与定位性能均优于现有基准方法.

参考文献

1 Mao Guo-Jun, Wu Xing-Zhen, Xing Shu-Li. Research on visual anomaly detection based on multi-scale normalizing flow. *Acta Automatica Sinica*, 2024, **50**(3): 640−648
(毛国君, 吴星臻, 邢树礼. 基于多尺度流模型的视觉异常检测研究. 自动化学报, 2024, **50**(3): 640−648)

2 Lv Cheng-Kan, Shen Fei, Zhang Zheng-Tao, Zhang Feng. Review of image anomaly detection. *Acta Automatica Sinica*, 2022, **48**(6): 1402−1428
(吕承侃, 沈飞, 张正涛, 张峰. 图像异常检测研究现状综述. 自动化学报, 2022, **48**(6): 1402−1428)

3 Wu Lin, Hao Hong-Yu, Song You. A review of metal surface defect detection based on computer vision. *Acta Automatica Sinica*, 2024, **50**(7): 1261−1283
(伍麟, 郝鸿宇, 宋友. 基于计算机视觉的工业金属表面缺陷检测综述. 自动化学报, 2024, **50**(7): 1261−1283)

4 Luo Dong-Liang, Cai Yu-Xuan, Yang Zi-Hao, Zhang Zhe-Yan, Zhou Yu, Bai Xiang. Survey on industrial defect detection with deep learning. *Scientia Sinica Informations*, 2022, **52**(6): 1002−1039
(罗东亮, 蔡雨萱, 杨子豪, 章哲彦，周瑜，白翔. 工业缺陷检测深度学习方法综述. 中国科学: 信息科学, 2022, **52**(6): 1002−1039)

5 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. Long Beach, USA: IEEE, 2019. 9592–9600

6 Zhao M Y, Guo Q. Reconstruction-based distillation for anomaly detection. *Computers and Graphics*, 2025, **116**: Article No. 104328

7 Xiang T, Zhang Y, Lu Y, Yuille A L, Zhang C, Cai W, et al. Squid: Deep feature in-painting for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023. 23890–23901

8 Hansen S, Gautam S, Jenssen R, Kampffmeyer M. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. *Medical Image Analysis*, 2022, **78**: Article No. 102385

9 Qiao Q, Qu M, Wang W, Jiang B, Guo Q. Effective global context integration for lightweight 3D medical image segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025, **35**: 4661−4674

10 Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 6479–6488

11 Prieto-Martín M, Comino-Trinidad M, Casas D. Detecting anomalies in dense 3D crowds. *Computers and Graphics*, 2025, **130**: Art-icle No. 104267

12 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 International Conference on Machine Learning. Virtual Event: PMLR, 2021. 8748–8763

13 Jeong J, Zou Y, Kim T, Zhang D, 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. Vancouver, Canada: IEEE, 2023. 19606–19616

14 Chen X, Han Y, Zhang J. 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.

15 Huang C, Jiang A, Feng J, Zhang Y, Wang X, Wang Y. Adapting visual-language models for generalizable anomaly detection in medical images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2024. 11375–11385

16 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 International Conference on Learning Representations. Vienna, Austria: ICLR, 2024. 49705–49737

17 Guo J, Lu S, Zhang W H, Chen F, Li H Q, Liao H E. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2025. 20405–20415

18 Beizaee F, Lodygensky G A, Desrosiers C, Dolz J. Correcting deviations from normality: A reformulated diffusion model for multi-class unsupervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2025. 19088–19097

19 Yang H, Chen Y, Song K, Yin Z. Multiscale feature-clustering-based fully convolutional autoencoder for fast accurate visual inspection of texture surface defects. *IEEE Transactions on Automation Science and Engineering*, 2019, **16**(3): 1450−1467

20 Avola D, Cascio M, Cinque L, Fagioli A, Foresti G L, Marini M R, et al. Real-time deep learning method for automated detection and localization of structural defects in manufactured products. *Computers & Industrial Engineering*, 2022, **172**: Article No. 108512

21 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. New Orleans, USA: IEEE, 2022. 14318–14328

22 Liu Z, Zhou Y, Xu Y, Wang Z. Simplenet: A simple network for image anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023. 20402–20411

23 Bergmann P, Fauser M, Sattlegger D, Steger C. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Event: IEEE, 2020. 4183–4192

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

25 Deng H, Li X. Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 9737–9746

26 Wang X L, Wang X Y, Bai H H, Lim E G, Xiao J M. CNC: Cross-modal normality constraint for unsupervised multi-class anomaly detection. In: Proceedings of the AAAI Conference on

Artificial Intelligence. Philadelphia, USA: AAAI, 2025. 7943–7951

27 Li B, Weinberger K Q, Belongie S, Koltun V, Ranftl R. Language-driven semantic segmentation. In: Proceedings of the International Conference on Learning Representations. Virtual Event: ICLR, 2022.

28 Gu X, Lin T Y, Kuo W, Cui Y. Open-vocabulary object detection via vision and language knowledge distillation. In: Proceedings of the International Conference on Learning Representations. Virtual Event: ICLR, 2021.

29 Vinker Y, Pajouheshgar E, Bo J Y, Bachmann R C, Bermano A H, Cohen-Or D, et al. CLIPasso: Semantically-aware object sketching. *ACM Transactions on Graphics*, 2022, **41**(4): 1–11

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

31 Rao Y, Zhao W, Chen G, Tang Y, Zhu Z, Huang G, et al. DenseCLIP: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 18082–18091

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

33 Cao Y, Zhang J, Frittoli L, Cheng Y, Shen W, Boracchi G. AdaCLIP: Adapting CLIP with hybrid learnable prompts for zero-shot anomaly detection. In: Proceedings of the European Conference on Computer Vision. Milan, Italy: Springer, 2024. 55–72

34 Gu Z, Zhu B, Zhu G, Chen 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

35 Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations. Virtual Event: ICLR, 2021.

36 Lin T Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2980–2988

37 Milletari F, Navab N, Ahmadi S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the 2016 Fourth International Conference on 3D Vision. Stanford, USA: IEEE, 2016. 565–571

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

39 Jezek S, Jonak M, Burget R, Dvorak P, Skotak M. Deep learning-based defect detection of metal parts: Evaluating current methods in complex conditions. In: Proceedings of the 2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops. Brno, Czech Republic: IEEE, 2021. 66–71

40 Mishra P, Verk R, Fornasier D, Piciarelli C, Foresti G L. VT-ADL: A vision transformer network for image anomaly detection and localization. In: Proceedings of the 30th IEEE International Symposium on Industrial Electronics. Kyoto, Japan: IEEE, 2021. 1–6

41 Wieler M, Hahn T. Weakly supervised learning for industrial optical inspection. In: DAGM Symposium. Germany: Springer, 2007.

42 Aota T, Tong L T T, Okatani T. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE, 2023. 5564–5572

43 Gutman D, Codella N C F, Celebi E, Helba B, Marchetti M, Mishra N, et al. Skin lesion analysis toward melanoma detection: A challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). arXiv preprint arXiv: 1605.01397, 2016.

44 Bernal J, Sánchez F J, Fernández-Esparrach G, Gil D, Rodríguez C, Vilariño F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 2015, **43**: 99–111

45 Tajbakhsh N, Gurudu S R, Liang J M. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging*, 2015, **35**(2): 630–644

46 Kitamura F C. Head CT——Hemorrhage [Online], available: https://www.kaggle.com/dsv/152137, September 6, 2025.

47 Kanade P B, Gumaste P P. Brain tumor detection using MRI images. *Brain*, 2015, **3**(2): 146–150

48 Hamada A. Br35H: : Brain Tumor Detection 2020. *IEEE Dataport*, March 10, 2025, DOI: 10.21227/tbkk-q937



赵梦阳 山东财经大学计算机与人工智能学院硕士研究生. 主要研究方向为异常检测与零样本学习.
E-mail: zhaomengyang@mail.sdufe.edu.cn
(**ZHAO Meng-Yang** Master student at the School of Computing and Artificial Intelligence, Shandong University of Finance and Economics. His research interests include anomaly detection and zero-shot learning.)



郭 强 山东财经大学计算机与人工智能学院教授. 2010 年获得上海大学博士学位. 主要研究方向为计算机视觉与机器学习. 本文通信作者.
E-mail: guoqiang@sdufe.edu.cn
(**GUO Qiang** Professor at the School of Computing and Artificial Intelligence, Shandong University of Finance and Economics. He received his Ph.D. degree from Shanghai University in 2010. His research interests include computer vision and machine learning. Corresponding author of this paper.)