

本文引用格式: 施凯悦,任志刚. LogoCLIP: 基于零样本学习的商标外观异常检测模型[J]. 自动化与信息工程, 2026, 47(3): 32-41.
SHI Kaiyue, REN Zhigang. LogoCLIP: A zero-shot method for trademark appearance anomaly detection[J]. Automation & Information Engineering, 2026, 47(3): 32-41.

LogoCLIP: 基于零样本学习的商标外观异常检测模型*

施凯悦¹ 任志刚^{1,2}

(1. 广东工业大学自动化学院, 广东 广州 510006

2. 广东省智能系统与优化集成重点实验室, 广东 广州 510006)

摘要: 商标作为产品的重要标识, 其外观质量对购物体验 and 品牌形象等具有深远影响。针对现有工业异常检测方法需使用大量数据训练模型, 难以应用于数据量少、种类繁多、缺陷模式不固定的商标外观异常检测的问题, 提出一种基于零样本学习的商标外观异常检测模型 LogoCLIP。该模型在视觉-语言预训练模型 CLIP 的基础上, 引入可学习的对象无关文本提示, 引导模型关注异常区域而非对象语义, 以实现对不同对象通用的外观异常检测。提出一种基于注意力机制的跨模态特征融合模块, 实现图像-文本特征对齐; 通过设计门控融合适配器解决预训练数据与目标数据的分布偏移; 借助辅助数据集, 模型可在无需商标数据训练样本的情况下实现商标外观异常检测。在自建商标外观缺陷数据集上的实验结果表明, LogoCLIP 的图像级 AUROC 达到 92.9%, 像素级 AUROC 达到 99.3%; 在 VisA、SDD、BTAD 3 个公开工业数据集上, 模型同样取得了具有竞争力的检测结果, 验证了其较强的泛化能力。

关键词: 商标外观; 异常检测; 零样本; 跨模态; 特征融合

中图分类号: TP271

文献标志码: A

文章编号: 1674-2605(2026)03-0005-10

DOI: 10.12475/aie.20260305

开放获取

LogoCLIP: A Zero-Shot Method for Trademark Appearance Anomaly Detection

SHI Kaiyue¹ REN Zhigang^{1,2}

(1. School of Automation Guangdong University of Technology, Guangzhou 510006, China

2. Guangdong Provincial Key Laboratory of Intelligent Systems and Optimization Integration Guangdong University of Technology, Guangzhou 510006, China)

Abstract: As an important identifier of products, the visual quality of trademarks has a profound impact on consumers' shopping experiences as well as brand reputation. However, existing appearance defect detection methods in the industry primarily rely on training models with large amounts of data, making it difficult to apply them to the field of trademark appearance anomaly detection, which involves a wide variety of products, limited data, and non-fixed defect patterns. To address this issue, this paper proposes LogoCLIP: A zero-shot learning-based model for trademark appearance anomaly detection. Based on the vision-language pre-training model CLIP, LogoCLIP introduces learnable object-agnostic text prompt to guide the model to focus on anomalous regions rather than object semantics, enabling generalized anomaly detection across different types of objects. Then, a cross-modal feature fusion method based on attention mechanisms is proposed to achieve image-text feature alignment, and a gated fusion adapter is employed to address the distribution shift between pre-trained data and target data. By utilizing auxiliary datasets, the model achieves anomaly detection for trademark appearances without requiring training samples from trademark data. Finally, Experimental results on trademark appearance defect dataset demonstrate that LogoCLIP achieves image-level AUROC and pixel-level AUROC scores of 92.9% and 99.3% respectively. The model also achieved competitive detection results on three public industrial datasets—VisA, SDD and BTAD, which verified its strong generalization ability.

0 引言

商标作为产品的核心视觉标识,其外观质量不仅影响消费者的购买决策与使用体验,更关乎企业的品牌形象与市场竞争力。在服饰鞋包、日化用品、电子产品等领域,商标外观缺陷(如模糊、颜色缺失、破损、污渍等)会降低产品质感,甚至引发品牌信任危机。因此,稳定、准确、低成本的商标外观异常检测方法具有重要的应用价值。

传统的商标外观异常检测方法大多依赖人工质检,效率低、成本高且主观性强,难以满足现代制造业对检测实时性和一致性的需求。随着智能制造与无人质检技术的发展,商标外观异常检测逐渐成为工业视觉领域的关键任务之一。但现有的工业异常检测方法通常建立在“对象中心”的隐含假设上,即每类检测对象(如瓶子、电缆、金属零部件等)具有稳定的语义类别、结构形态和外观分布。而商标主要由字体、图形、符号、颜色等抽象视觉元素组合而成,不存在明确、统一的对象语义表示,且其缺陷具有明显的跨对象共性特征。这种“弱对象语义、强局部异常”的特性,使商标外观异常检测难以直接套用现有的工业异常检测方法。同时,大多数工业异常检测方法依赖于大量标注数据或特定类别的数据进行训练,在面对数据量少、风格多变、缺陷模式不固定的商标时,检测性能有限^[1-3]。此外,小样本、多品类的商标生产需求,导致其缺陷样本数量稀少、类型多样且分布不均,致使传统监督学习模型难以获得稳定、可推广的检测性能。近年来,虽然已有研究探索基于重建^[4-6]或特征学习^[7-9]的异常检测方法,但这些方法普遍需要大量同类产品的数据进行训练,应用于商标外观异常检测时,检测性能显著下降。

然而,视觉-语言预训练模型,如对比语言-图像预训练(contrastive language-image pre-training, CLIP)模型^[10]在跨模态语义理解和零样本识别任务中展现出较强的泛化能力,为解决工业领域缺陷数据稀缺与

分布多变等问题提供了新方向^[10-12]。与传统依赖大规模专用数据集训练的模型不同,CLIP模型利用图文语料进行预训练,具备较强的语义泛化能力,能够直接推广至新类别。基于CLIP模型的零样本异常检测(zero-shot anomaly detection, ZSAD)方法,如WinCLIP^[13]、APRIL-GAN^[14]、CLIP-AD^[15]等,已在多个工业数据集上取得较好的检测性能。但CLIP模型应用于商标外观异常检测时,仍面临如下挑战:1) 现有基于CLIP模型的ZSAD方法普遍依赖对象语义作为跨模态对齐的核心锚点,通过构造与具体对象类别相关的文本提示来建立图像-文本映射关系,而不同商标间因缺乏明确、统一的对象语义表示,CLIP模型在预训练过程中学习的“物体-文本”对齐先验知识可能干扰异常区域识别;2) 商标缺陷通常表现为局部纹理破坏或细粒度结构偏移,而CLIP模型更倾向于全局语义建模,在缺乏显式对象语义约束的情况下,难以形成稳定的异常语义对齐。因此,如何在弱对象语义甚至无对象语义情况下,引导CLIP模型关注缺陷本身而非对象语义,已成为商标外观ZSAD亟待解决的关键问题。

为此,本文提出一种基于零样本学习的商标外观异常检测模型LogoCLIP。该模型通过引入可学习的对象无关文本提示,削弱CLIP模型对对象语义的依赖;设计门控融合适配器(gated fusion adapter, GFA),缓解视觉-语言预训练特征与工业图像特征之间的分布偏移;构建基于注意力机制的跨模态特征融合(cross-modal feature fusion, CMFF)模块,实现异常语义在图像与文本模态间的细粒度对齐。可在不使用商标训练样本的情况下,LogoCLIP仅借助辅助数据即可学习通用的异常表示,实现对多类型商标外观异常的零样本检测与分割。

1 相关工作

1.1 CLIP模型

CLIP模型的核心架构主要包括文本编码器和图

像编码器。其中，文本编码器通常基于 Transformer 结构，用于处理自然语言描述；图像编码器采用视觉 Transformer（vision Transformer, ViT）或 ResNet 等骨干网络，用于提取图像的视觉特征。CLIP 模型通过双编码器设计，可将图像、文本通过相应的编码器映射到统一的嵌入空间，实现跨模态的相似度计算。具体而言，对于给定的输入图像 x_i ，在类别集合 C 下，通过文本编码器可得到每个类别 c 对应的文本嵌入向量 f_c ；通过图像编码器得到其视觉嵌入向量 $f_i \in \mathbb{R}^D$ ，则图像 x_i 属于类别 c 的概率为

$$P(y=c|x_i) = \frac{\exp(\langle f_c, f_i \rangle / \tau)}{\sum_{c \in C} \exp(\langle f_c, f_i \rangle / \tau)} \quad (1)$$

式中： $\langle \cdot, \cdot \rangle$ 为余弦相似度， τ 为温度参数， y 为预测类别。

1.2 ZSAD

随着大型视觉-语言预训练模型的出现，CLIP 模型为 ZSAD 提供了全新的跨模态思路，并广泛应用于工业^[16]、医学影像^[17]、遥感^[18]等领域的异常检测任务中。文献[13]提出的 WinCLIP 框架首次利用 CLIP 模型进行零样本异常定位；文献[14]提出的 APRIL-GAN 模型通过跨模态映射与先验记忆，有效提升了 CLIP 模型在工业场景的适用性；文献[15]提出的分阶段双路径模型（staged dual-path model, SDP）解决了 CLIP 模型图像-文本特征不对齐的问题，改善了异常分割性能。

此外，还有研究探索利用提示工程和提示学习的方法来提升 CLIP 模型的分类与分割性能^[19-20]。在工业异常检测领域，文献[21]提出的 segment any anomaly + (SAA+) 方法在无需额外训练的条件下，通过引入领域专家知识与混合提示正则化策略，引导 segment anything model (SAM)^[22]等分割模型实现零样本异常分割。文献[16]提出的 AnomalyCLIP 模型，则通过引入可学习的通用文本提示，增强了 CLIP 模型的 ZSAD 性能。

ZSAD 通过引入视觉-语言预训练模型的跨模态先验知识，在无需目标域标注数据的情况下即可完成异常检测任务，在工业场景中具有较强的应用潜力。

2 LogoCLIP 模型

ZSAD 任务是在缺乏商标训练样本的情况下，实现对商标图像的异常分类与像素级分割。具体而言，LogoCLIP 模型通过双分支结构的图像编码器分别提取全局视觉嵌入向量与局部图像块嵌入向量，并结合文本编码器生成的文本嵌入向量，通过跨模态相似度计算分别得到图像级与像素级异常得分，从而有效区分正常区域与异常区域。

本文提出的 LogoCLIP 模型架构如图 1 所示，主要包括对象无关文本提示、图像编码器、文本编码器、GFA 和基于注意力机制的 CMFF 模块。对于给定的一幅测试图像，LogoCLIP 通过图像编码器的 ViT 分支获取其全局视觉嵌入向量，使用文本编码器获取对象无关文本提示的文本嵌入向量。随后采用 GFA 对文本嵌入向量和全局视觉嵌入向量进行轻量化调整，以实现 CLIP 模型先验知识向工业异常检测领域的迁移，并通过基于注意力机制的 CMFF 模块获取跨模态融合特征，进一步增强全局视觉嵌入。另外，LogoCLIP 模型遵循 AnomalyCLIP^[16]所采用的 ZSAD 设置，在图像编码器中使用 V-V 注意力分支获取局部图像块嵌入向量，在文本编码器中引入可学习词嵌入向量以细化文本空间。最后，LogoCLIP 模型通过计算文本嵌入向量、全局视觉嵌入向量及局部图像块嵌入向量的余弦相似度，获得图像级与像素级异常得分。

2.1 对象无关文本提示

针对传统 CLIP 模型更关注商标内容而非商标外观缺陷特征，且商标缺陷类型难以预先穷举等问题，LogoCLIP 模型引入了对象无关文本提示，并借助辅助数据集学习一种跨对象通用的正常、异常文本提示，以降低商标类别语义对检测性能的影响。

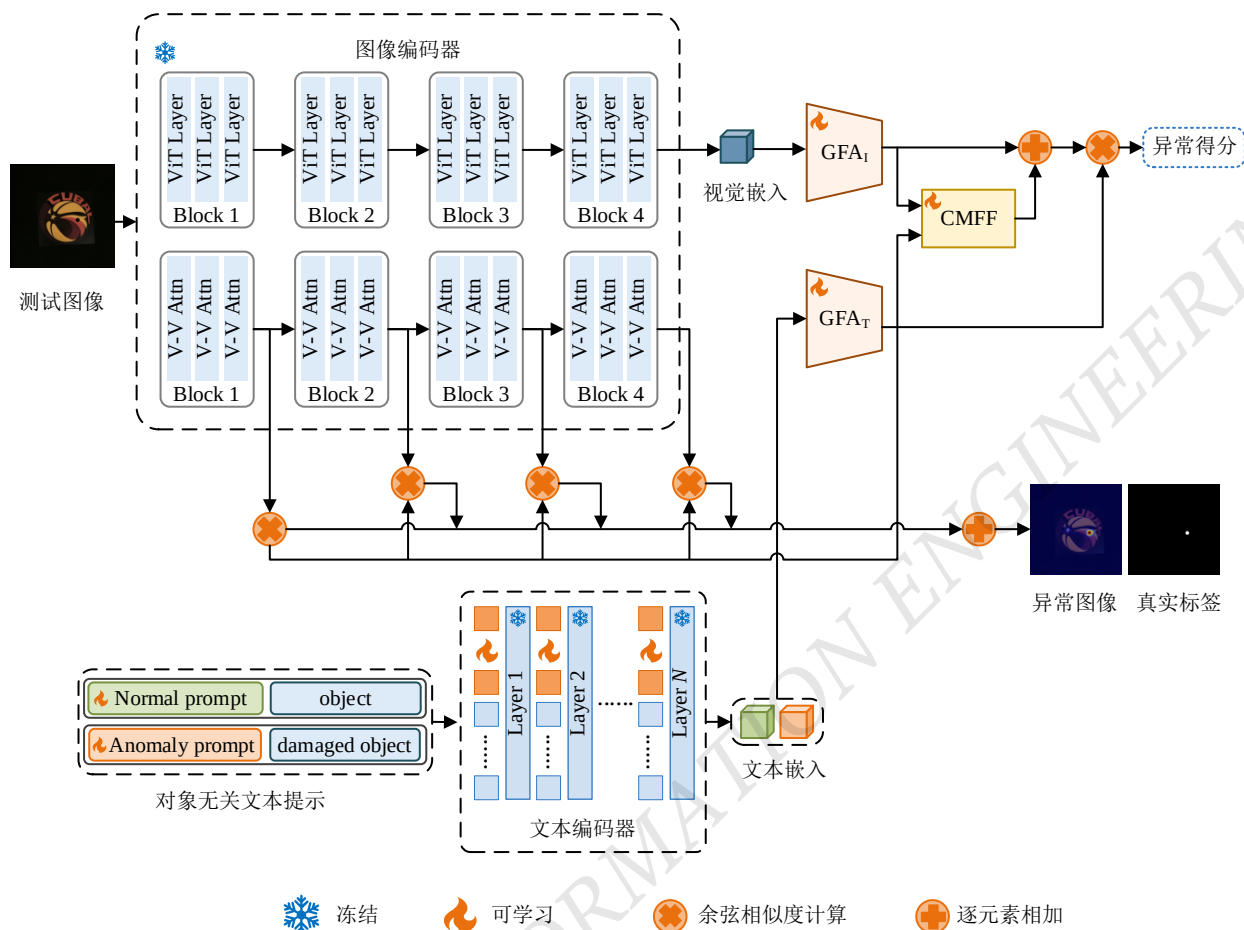


图1 LogoCLIP 模型架构

为适应零样本设定，LogoCLIP 模型构建了两类文本提示模板：

$$G_n = [N_1, N_2, \dots, N_E][\text{object}]$$

$$G_a = [A_1, A_2, \dots, A_E][\text{damaged object}]$$

式中： G_n 为正常文本提示； G_a 为异常文本提示； N_i 、 A_i 分别为正常、异常文本提示中第*i*个可学习词向量，在训练过程中用于表示“完好”与“损坏”等通用语义特征，从而缓解 CLIP 模型在预训练中偏向对象语义而忽视异常语义的问题。

2.2 GFA

在异常检测任务中，CLIP 模型的视觉嵌入与文本嵌入跨模态对齐受限（视觉嵌入偏向对象语义表示，文本嵌入依赖自然语言语义表达），两者之间缺乏针对异常检测的对齐机制。尤其在商标外观异常检测场

景中，图像结构抽象且缺乏自然语言中的实体概念，图像编码器生成的视觉嵌入容易混杂字体纹理、颜色变化和局部噪声等信息，导致其与文本提示难以形成稳定的语义对应关系。

为缓解这一问题，LogoCLIP 模型设计了 GFA。其核心思想为：通过轻量化适配模块，在不破坏 CLIP 模型预训练知识的前提下，对视觉嵌入与文本嵌入进行自适应调整，使其更适合商标外观异常检测。考虑到零样本设定下目标数据无法用于监督训练，该适配模块需在表达能力与过拟合风险之间取得平衡。因此，GFA 采用两层线性映射与门控机制的轻量化设计，以最小参数量实现对预训练特征的适度调整。GFA 主要由原嵌入分支、适配分支和 Sigmoid 函数构成的门控融合机制组成，其结构如图 2 所示。

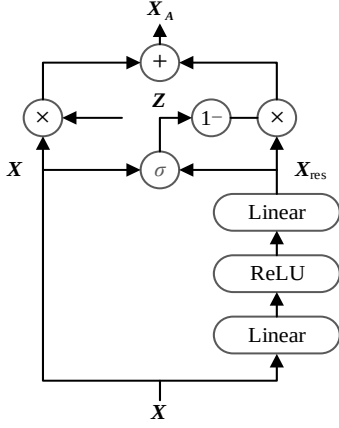


图2 GFA 结构图

对于给定的原始输入特征 $X \in \mathbb{R}^D$ (视觉嵌入向量或文本嵌入向量)，首先，通过两层线性变换与激活函数构成适配分支，用于提取其补充特征信息：

$$X_{\text{res}} = W_2(\phi(W_1(X))) \quad (2)$$

式中： X_{res} 为适配分支输出的补充特征， W_1 、 W_2 为线性变换函数， ϕ 为 ReLU 激活函数。

该适配分支旨在捕获异常检测相关的特征变化，如异常区域的重要局部模式或文本提示中代表“损坏”语义的方向信息。

随后，通过 Sigmoid 函数计算门控权重，并得到最终输出：

$$Z = \sigma(X + X_{\text{res}}) \quad (3)$$

$$X_A = Z \odot X + (1 - Z) \odot X_{\text{res}} \quad (4)$$

式中： σ 为 Sigmoid 函数， $\odot$ 为逐元素相乘， Z 为生成的门控权重， X_A 为融合后的输出特征。

门控融合机制接收原嵌入分支与适配分支的输入，并通过 Sigmoid 函数生成门控权重 Z ，用于动态调节原始特征与补充特征之间的融合比例，以避免过拟合或对预训练特征造成过度扰动。当某维度的异常特征模式明显时， Z 值较小，使 LogoCLIP 模型更多地采用补充特征 X_{res} 的信息；当特征维度具有足够的语义稳定性时，LogoCLIP 模型则保持对原始输入特征 X 的依赖。

GFA 有助于缓解 CLIP 模型视觉-语言预训练特

征与工业图像特征之间的分布偏移：从工业异常检测任务的角度对编码器输出进行轻量且高效的偏移修正，而非取代 CLIP 模型的编码能力；通过在图像编码器和文本编码器后分别引入 GFA，LogoCLIP 模型能够在保留预训练先验知识的同时适配工业异常检测任务。

2.3 基于注意力机制的 CMFF 模块

虽然图像编码器能够生成包含潜在异常模式的视觉表征，文本编码器也能提供正常、异常语义提示，但二者仍然分别驻留在各自的模态空间中，语义对齐程度有限，导致异常视觉模式与异常文本提示的映射关系不足。尤其在零样本设定下，由于缺乏商标训练样本，CLIP 模型的弱耦合图文空间结构更加难以有效建模异常类别。为增强视觉表示，挖掘其与正常、异常文本提示最相关的视觉信息，本文设计基于注意力机制的 CMFF 模块，用于建立视觉嵌入向量与文本嵌入向量之间更强的关联，从而在最终异常得分计算中引导模型更加关注异常视觉模式。基于注意力机制的 CMFF 模块结构如图3所示，

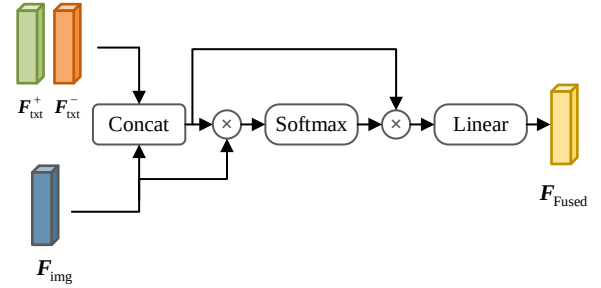


图3 基于注意力机制的 CMFF 模块结构图

具体而言，设通过 GFA 后的视觉嵌入向量为 $F_{\text{img}} \in \mathbb{R}^D$ ，正常文本提示向量为 $F_{\text{txt}}^+ \in \mathbb{R}^D$ 和异常文本提示向量为 $F_{\text{txt}}^- \in \mathbb{R}^D$ 。为构建跨模态关联，首先将视觉嵌入向量与文本嵌入向量在特征维度上进行拼接，得到拼接特征 F_{cat} ：

$$F_{\text{cat}} = [F_{\text{img}}, F_{\text{txt}}^+, F_{\text{txt}}^-] \quad (5)$$

通过计算图像特征 F_{img} 和拼接特征 F_{cat} 的交叉注

意力，得到融合特征为

$$\mathbf{F}_{\text{Fused}} = \text{Linear}(\text{Softmax}(\frac{\mathbf{F}_{\text{img}} \cdot \mathbf{F}_{\text{cat}}^T}{\sqrt{C}}) \mathbf{F}_{\text{cat}}) \quad (6)$$

将融合特征 $\mathbf{F}_{\text{Fused}}$ 与原始图像特征进行残差融合，从而在图像特征中引入跨模态交互信息：

$$\mathbf{F}_{\text{img}}^{\text{final}} = \mathbf{F}_{\text{img}} + \mathbf{F}_{\text{Fused}} \quad (7)$$

基于注意力机制的 CMFF 模块通过图像特征 $\mathbf{F}_{\text{img}}$ 和拼接特征 $\mathbf{F}_{\text{cat}}$ 的交叉注意力计算，构建了视觉嵌入向量与文本嵌入向量之间的细粒度交互，使 LogoCLIP 模型在没有商标训练样本的情况下，仍能充分利用辅助数据集及对象无关文本提示实现高精度的异常检测，为商标外观的 ZSAD 提供了关键支持。

2.4 损失函数

LogoCLIP 模型需要借助辅助数据集对对象无关文本提示、GFA 及基于注意力机制的 CMFF 模块进行训练优化。优化目标由分类损失 L_{cls} 和分割损失 L_{seg} 两部分组成：

$$L_{\text{total}} = L_{\text{cls}} + L_{\text{seg}} \quad (8)$$

分类损失 L_{cls} 是训练图像异常得分 A 与其对应标签 $Y \in \{0,1\}$ 之间的交叉熵损失（ L_{BCE} ）：

$$L_{\text{cls}} = L_{\text{BCE}}(A, Y) \quad (9)$$

由于图像中的异常区域通常远小于正常区域，分割损失 L_{seg} 利用 Focal loss^[23]和 Dice loss^[24]来解决类别不平衡问题，以提高模型的像素级分割性能。设输入图像的像素级分割结果为 S_a ，真实分割掩码为 S_{GT} ，则有

$$L_{\text{seg}} = \sum_{i=1}^N L_{\text{focal}}^i(S_a, S_{\text{GT}}) + L_{\text{dice}}^i(S_a, S_{\text{GT}}) \quad (10)$$

式中： N 为图像编码器中输出局部图像块嵌入的中间层层数。

通过引入图像级分类损失和像素级分割损失，LogoCLIP 模型既能从全局角度学习区分正常、异常语义的特征表示，又能关注到像素级的细粒度局部异

常区域，从而有效优化了 ZSAD 性能。

3 实验

本文通过主要实验、对比实验和消融实验对 LogoCLIP 模型的 ZSAD 性能进行评估，并验证该模型在商标外观异常检测、工业异常检测领域的有效性。

3.1 实验设置

3.1.1 数据集

本文在自建的商标外观缺陷数据集（Logo）上进行实验。该数据集包含 6 类商标对象（Hh、Cubal、Fila、Kappa、Kun、S），共计 600 幅样本图像（正常、异常样本图像各 300 幅）；涵盖丝织、注塑、镭射等不同工艺的商标；包含墨点、划痕、脏污、异物等外观缺陷。为衡量 LogoCLIP 模型的零样本推理能力，该数据集不参与模型训练，仅用于模型性能的测试。

此外，本文在 VisA 数据集^[25]、BeanTech 异常检测数据集（BTAD）、Kolektor 表面缺陷数据集（SDD）3 个公开工业数据集（共包含 16 种不同的工业产品）上进行 ZSAD 实验，以评估 LogoCLIP 模型在工业场景下的泛化能力。上述数据集同样仅用于测试阶段的性能验证。LogoCLIP 模型预训练采用的辅助数据集为 MVTec AD 数据集^[26]，包含 5 000 余幅正常、异常工业产品图像，涵盖 15 个不同对象和纹理类别。

3.1.2 评估指标

为全面评估 LogoCLIP 模型在商标外观 ZSAD 任务中的性能，本文从图像级异常识别能力、像素级异常定位与分割能力两个层面进行定量分析。具体而言，选取工业异常检测领域常用的受试者工作特征曲线下面积（area under the receiver operating characteristic curve, AUROC）、平均精度（average precision, AP）、区域重叠率（per region overlap, PRO）等评估指标，对 LogoCLIP 模型进行综合评估。

AUROC 指标通过真阳性率（true positive rate, TPR）与假阳性率 FPR（false positive rate, FPR）之间的关系，计算 ROC 曲线下的面积，其值越接近 1，表示模型对正常、异常样本的区分能力越强，可用于同

时评估模型的图像级、像素级异常检测性能。其计算公式为

$$S_A = \int y d(x) \quad (11)$$

$$y = \frac{TP}{TP + FN} \quad (12)$$

$$x = \frac{FP}{FP + TN} \quad (13)$$

式中： x 和 y 分别表示 ROC 曲线的横坐标和纵坐标，即 FPR 和 TPR； TP 、 FP 、 TN 、 FN 分别为真阳性、假阳性、真阴性、假阴性的样本数量。

AP 指标通过不同阈值下精确率-召回率（precision-recall, PR）曲线下面积，反映模型在召回率与精确率之间的平衡能力，用于评估模型的图像级异常识别能力。其计算公式为

$$S_P = \sum_{i=1}^n (r_i - r_{i-1}) p_i \quad (14)$$

式中： r_i 为召回率， p_i 为精确率。

PRO 指标通过预测异常区域与真实异常区域之间的重叠程度，反映模型对局部异常区域的分割质量，用于评估模型的像素级异常定位与分割能力。首先，将预测异常区域与真实异常区域划分为 N 个连通区域；然后，计算每个真实异常区域 G_n 与预测异常区域 P 的交集占 G_n 的比例；最后，对所有真实区域加权平均。其计算公式为

$$S_R = \frac{1}{N} \sum_{n=1}^N \frac{P \cap G_n}{G_n} \quad (15)$$

3.1.3 参数设置

本文遵循 ZSAD 的常见设置，即在训练过程中不使用任何目标数据。CLIP 模型采用 ViT-L/14@336px 作为骨干网络，并冻结其全部参数，使用 MVTec AD 数据集进行训练。另外，本文保留了文本编码器中的可学习词嵌入机制细化文本空间，相关参数设置与文献[16]保持一致。模型训练采用 Adam 优化器，初始学习率为 0.001，批量大小为 16。所有实验均基于 PyTorch-2.0.0 框架，在单个 NVIDIA RTX 3090Ti 24

GB GPU 上进行。

3.2 主要实验

在商标外观缺陷数据集上进行 ZSAD 实验，LogoCLIP 模型在 6 类商标上的可视化检测结果如图 4 所示，商标外观 ZSAD 性能如表 1 所示。



(a) 白斑 (b) 墨点 (c) 划痕 (d) 材料不良 (e) 脏污 (f) 异物

图 4 商标外观 ZSAD 可视化检测结果

由图 4 可知，LogoCLIP 模型能够准确分割出 6 个类型商标的异常区域，如白斑、墨点、划痕、脏污等区域。

表 1 商标外观 ZSAD 性能 %

商标类别	像素级指标		图像级指标	
	AUROC	PRO	AUROC	AP
Hh	99.2	96.5	93.1	94.2
Cubal	99.5	97.5	98.7	99.0
Fila	98.9	94.5	97.6	98.6
Kappa	99.6	93.7	76.2	81.0
Kun	99.1	93.7	95.8	97.2
S	99.4	97.3	95.9	96.9
平均值	99.3	95.5	92.9	94.5

由表 1 可知：在像素级指标方面，LogoCLIP 模型的平均 AUROC、PRO 分别为 99.3%、95.5%，表明其具备较好的异常区域定位与分割能力；在图像级指标方面，平均 AUROC、AP 分别为 92.9%、94.5%，表明其异常识别能力较好。LogoCLIP 模型在不同商标类别上的表现存在一定差异，如 Kappa 的图像级 AUROC 相对较低（76.2%），这可能与该商标的异常特征模式较为复杂或辅助数据与商标数据之间的分

布差异较大有关上述实验结果表明，LogoCLIP 模型在不依赖商标训练样本的前提下，能够通过 CMFF 与预训练特征适配，实现较好的图像级异常识别能力与像素级异常定位与分割能力。

3.3 对比实验

为全面评估 LogoCLIP 模型的 ZSAD 性能和泛化能力，将其与主流 ZSAD 方法（包括原生 CLIP、基

于手工文本提示的 VAND、采用对象无关文本提示学习的 AnomalyCLIP 及使用混合提示并结合语义融合的 AdaCLIP^[27]）进行对比实验。对比实验分别在自建的商标外观缺陷数据集（Logo）和 3 个公开工业数据集（VisA、SDD、BTAD）上进行，实验结果如表 2 所示。

表 2 5 种方法 ZSAD 性能对比

指标	数据集	类别数 C	CLIP	VAND	AnomalyCLIP	AdaCLIP	本文模型
图像级指标 (AUROC, AP)	Logo	6	(82.0, 86.2)	(93.0, 95.0)	(88.4, 92.6)	(97.1, 97.8)	(92.9, 94.5)
	VisA	12	(52.1, 60.9)	(78.0, 81.4)	(82.1, 85.4)	(81.4, 84.2)	(83.8, 86.1)
	SDD	1	(65.7, 45.2)	(79.8, 71.4)	(84.7, 80.0)	(82.1, 74.2)	(86.9, 85.1)
	BTAD	3	(32.7, 50.5)	(73.6, 68.6)	(88.3, 87.3)	(78.0, 88.5)	(90.1, 87.2)
像素级指标 (AUROC, PRO)	Logo	6	(33.9, 8.1)	(99.1, 58.3)	(99.0, 80.9)	(99.6, 77.9)	(99.3, 95.5)
	VisA	12	(34.5, 9.2)	(94.2, 86.8)	(95.5, 87.0)	(95.9, 83.7)	(95.5, 86.8)
	SDD	1	(39.0, 8.9)	(79.8, 65.1)	(90.6, 67.8)	(73.6, 26.9)	(89.9, 68.3)
	BTAD	3	(42.8, 9.0)	(60.8, 25.0)	(94.2, 74.8)	(88.7, 29.9)	(93.9, 75.7)

由表 2 可知：在图像级异常检测任务中，各类基于 CLIP 的方法性能均优于原生 CLIP，验证了引入包含异常语义的文本提示能有效提升 ZSAD 性能。LogoCLIP 模型在 VisA、SDD 和 BTAD 数据集上均取得了最优的图像级检测性能，表现出较强的跨数据集泛化能力；在 Logo 数据集上，AdaCLIP 在图像级 AUROC、AP 指标上取得了最高值，表明正常、异常文本提示在提升图像级异常识别方面具有优势。

在像素级异常定位与分割任务中，5 种方法之间的差异更加明显：CLIP 因缺乏对异常语义的建模，其像素级指标显著偏低；VAND 和 AnomalyCLIP 通过引入异常相关的语义信息，提升了其像素级指标；其中 AnomalyCLIP 借助对象无关文本提示，在多数数据集上表现稳定；AdaCLIP 虽在像素级 AUROC 上保持较高值，但在 PRO 等强调区域连贯性的指标上，尤其是在 SDD 和 BTAD 数据集上表现相对受限；LogoCLIP 模型在 Logo 数据集上，其像素级 AUROC

略低于 AdaCLIP，但在 PRO 指标上取得优势，表明该模型在视觉嵌入、文本嵌入上的适配设计与跨模态交互机制能够产生更稳定的异常响应，从而在区域重叠类指标上取得了更优结果；在 VisA、SDD 和 BTAD 数据集上，LogoCLIP 模型在 PRO 指标上取得了有竞争力的结果，表明其在适配商标外观缺陷的前提下，在其他工业产品异常检测任务上也展现出良好的泛化能力。

上述实验结果表明，LogoCLIP 模型在图像级异常识别与像素级定位与分割任务中均表现出较为均衡的性能，验证了其在弱对象语义、跨数据集场景下的有效性。

3.4 消融实验

为验证 LogoCLIP 模型各模块对 ZSAD 性能的贡献，本文在 4 个数据集上对基于注意力机制的 CMFF 模块（T1）和 GFA（T2）进行了消融实验，实验结果如表 3 所示。

表 3 消融实验结果				%
指标	数据集	Base	+T1	+T1+T2
图像级指标 (AUROC, AP)	Logo	(88.4, <u>92.6</u>)	(<u>89.4</u> , 92.0)	(92.9 , 94.5)
	VisA	(81.9, 85.3)	(<u>83.4</u> , <u>85.9</u>)	(83.8 , 86.1)
	SDD	(86.3, 82.8)	(<u>86.7</u> , <u>84.6</u>)	(86.9 , 85.1)
	BTAD	(87.4, 88.5)	(<u>89.8</u> , 86.2)	(90.1 , <u>87.2</u>)
像素级指标 (AUROC, PRO)	Logo	(99.0, 80.9)	(<u>99.2</u> , <u>93.6</u>)	(99.3 , 95.5)
	VisA	(95.3, 85.1)	(95.6 , <u>86.0</u>)	(<u>95.5</u> , 86.8)
	SDD	(<u>90.5</u> , 68.5)	(91.0 , 66.8)	(89.9, <u>68.3</u>)
	BTAD	(93.5, 72.1)	(94.2 , <u>75.2</u>)	(<u>93.9</u> , 75.7)

由表 3 可知: 引入基于注意力机制的 CMFF 模块 (T1) 后, LogoCLIP 模型在所有数据集上的大部分指标均有提升, 表明其能够有效增强视觉嵌入与文本嵌入之间的跨模态特征交互, 从而提升模型的图像级异常识别能力与像素级异常定位与分割能力; 当进一步加入 GFA (T2) 后, LogoCLIP 模型在大多数指标上均达到最佳性能, 表明引入 GFA 有助于在保持预训练特征泛化能力的同时, 更好地适配商标外观异常检测任务。

4 结论

本文针对商标外观异常检测中存在的缺陷样本稀缺、产品种类繁多及缺陷模式多变等问题, 提出一种基于零样本学习的商标外观异常检测模型 LogoCLIP。该模型以视觉-语言预训练模型 CLIP 为基础, 引入对象无关文本提示, 通过可学习的正常、异常文本提示, 引导模型聚焦于跨对象类别的通用异常表示, 而非具体商标的语义内容。同时, 为增强模型对局部异常区域的感知能力, 并缓解预训练数据与目标数据之间的分布偏移, 设计基于注意力机制的 CMFF 模块与 GFA, 实现了视觉-语言特征的细粒度对齐与分布自适应。在自建的商标外观缺陷数据集及 3 个公开工业数据集上的实验结果表明, LogoCLIP 模型在零样本设定下表现出均衡的图像级异常识别能力、像素级异常定位与分割能力, 验证了其较强的泛化能力。尽管不同商标产品的缺陷类型难以枚举, 但商标的正常样本往往较易获取。未来的研究可尝试探

索如何利用少量正常商标图像作为参照, 进一步增强模型对细微、高隐蔽性外观缺陷的检测精度。

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

参考文献

- [1] LIU Y, ZHANG C, DONG X. A survey of real-time surface defect inspection methods based on deep learning[J]. Artificial Intelligence Review, 2023,56(10):12131-12170.
- [2] CHEN Y, DING Y, ZHAO F, et al. Surface defect detection methods for industrial products: A review[J]. Applied Sciences, 2021,11(16):7657.
- [3] SABERIRONAGHI A, REN J, EL-GINDY M. Defect detection methods for industrial products using deep learning techniques: A review[J]. Algorithms, 2023,16(2):95.
- [4] ZHOU Q, MEI J, ZHANG Q, et al. Semi-supervised fabric defect detection based on image reconstruction and density estimation[J]. Textile Research Journal, 2021,91(9-10):962-972.
- [5] LIU J, WANG C, SU H, et al. Multistage GAN for fabric defect detection[J]. IEEE Transactions on Image Processing, 2019,29: 3388-3400.
- [6] SCHLEGL T, SEEBÖCK P, WALDSTEIN S M, et al. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks[J]. Medical Image Analysis, 2019,54: 30-44.
- [7] ROTH K, PEMULA L, ZEPEDA J, et al. Towards total recall in industrial anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 14318-14328.
- [8] WANG G, HAN S, DING E, et al. Student-teacher feature

- pyramid matching for anomaly detection[J]. arXiv preprint arXiv:2103.04257, 2021.
- [9] DEFARD T, SETKOV A, LOESCH A, et al. Padim: A patch distribution modeling framework for anomaly detection and localization[C]//International Conference on Pattern Recognition. Cham: Springer International Publishing, 2021:475-489.
- [10] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]//International Conference on Machine Learning. PMLR, 2021:8748-8763.
- [11] LI J, LI D, XIONG C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International Conference on Machine Learning. PMLR, 2022:12888-12900.
- [12] BAO H, WANG W, DONG L, et al. VLMO: Unified vision-language pre-training with mixture-of-modality-experts[J]. Advances in Neural Information Processing Systems, 2022,35: 32897-32912.
- [13] JEONG J, ZOU Y, KIM T, et al. WinCLIP: Zero-/few-shot anomaly classification and segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 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[J]. arXiv preprint arXiv:2305.17382, 2023.
- [15] CHEN X, ZHANG J, TIAN G, et al. CLIP-AD: A language-guided staged dual-path model for zero-shot anomaly detection [C]//International Joint Conference on Artificial Intelligence. Singapore: Springer Nature Singapore, 2024:17-33.
- [16] ZHOU Q, PANG G, TIAN Y, et al. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection[C]//International Conference on Learning Representations, 2024, 2024:49705-49737.
- [17] ZHANG X, XU M, QIU D, et al. MediCLIP: Adapting CLIP for few-shot medical image anomaly detection[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024:458-468.
- [18] XIAO A, XUAN W, WANG J, et al. Foundation models for remote sensing and earth observation: A survey[J]. IEEE Geoscience and Remote Sensing Magazine, 2025.
- [19] ZHOU K, YANG J, LOY C C, et al. Conditional prompt learning for vision-language models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022:16816-16825.
- [20] JIA M, TANG L, CHEN B C, et al. Visual prompt tuning[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022:709-727.
- [21] CAO Y, XU X, SUN C, et al. Segment any anomaly without training via hybrid prompt regularization[J]. arXiv preprint arXiv:2305.10724, 2023.
- [22] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023:4015-4026.
- [23] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017:2980-2988.
- [24] MILLETARI F, NAVAB N, AHMADI S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]//2016 fourth International Conference on 3D Vision (3DV). IEEE, 2016:565-571.
- [25] ZOU Y, JEONG J, PEMULA L, et al. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022:392-408.
- [26] BERGMANN P, FAUSER M, SATTLEGER D, et al. MVTec AD-A comprehensive real-world dataset for unsupervised anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019:9592-9600.
- [27] CAO Y, ZHANG J, FRITTOLI L, et al. AdaCLIP: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024:55-72.

作者简介:

施凯悦, 男, 2000 年生, 硕士研究生, 主要研究方向: 零/小样本异常检测。E-mail: shi_kaiyue@163.com

任志刚 (通信作者), 男, 1987 年生, 博士研究生, 副教授, 博士生导师, 主要研究方向: 工业智能、具身智能机器人、工业视觉检测等。E-mail: renzhigang@gdut.edu.cn