一种基于滑动窗口的零/少量样本异常检测方法复现

蔡政革 2410815001

2024 年 12 月

摘要

本文主要采用最新版本的 Pytorch 复现了一种基于滑动窗口的零/少量样本异常检测方法。该方法利用预训练的视觉语言模型 CLIP，通过窗口化处理图像，提取局部特征，并将其与文本描述相关联，以识别图像中的异常。WinCLIP+ 作为 WinCLIP 的扩展，进一步结合了少量正常样本的视觉信息，以增强异常识别能力。本工作在 MVTec-AD、VisA 和 MPDD 三个数据集上进行了广泛的实验验证，结果表明，最终的复现效果基本达到了原论文的水平。同时部分类别的评价指标要优于原论文的结果指标。

关键词：异常检测；零样本学习；少样本学习；视觉语言模型

1 引言

在现代工业生产中，自动化质量检测已成为保证产品质量和提高生产效率的关键环节。然而，传统的质量检测方法往往需要大量的标注数据和针对特定任务的模型训练，这在实际应用中面临着高成本和低灵活性的挑战。

视觉异常分类（AC）和分割（AS）分别对工业制造中的缺陷进行分类和定位，预测某图像或某像素是正常还是异常。视觉检测是一个长尾问题，因为物体及其缺陷在颜色、质地和尺寸上差异很大，且涉及广泛的工业领域，包括电子、航空航天、汽车、制药等。在这些领域存在两个主要挑战：首先，缺陷是少见的，且变化范围大，这导致训练数据中缺乏有代表性的异常样本。其次，先前工作集中在为每项视觉检测任务训练一个定制的模型，这在长尾任务中是不可扩展的。因此传统的需要大量的标注数据和针对特定任务的模型训练质量检测方法已经难以满足工业自动化对提高生产效率和产品质量的要求，开发一种通用且有效的异常检测方法对于提高生产效率和产品质量至关重要。

我要复现的这篇论文 [11] 提出了基于窗口的 CLIP(WinCLIP)。WinCLIP 采用了窗口化的方式对图像进行处理，提取出局部特征并将其与文本描述相关联；WinCLIP 还引入了一个基于状态词和提示模板的复合集成方法，以及一种有效的特征提取和聚合方式，以更好地捕捉图像中的异常情况。此外本文还提出了一个针对少量正常样本的扩展版本 WinCLIP+，它能够利用少量的参考图像的信息来进行训练来进一步提高在异常分类和异常分割任务上的表现，WinCLIP 和 WinCLIP+ 的整体工作流程如图 1所示. 实验结果表明，在 MVTec-AD [2] 和 VisA [29] 数据集上，WinCLIP 在零/少量样本下取得了显著的性能提升，超过了现有最好的方法。

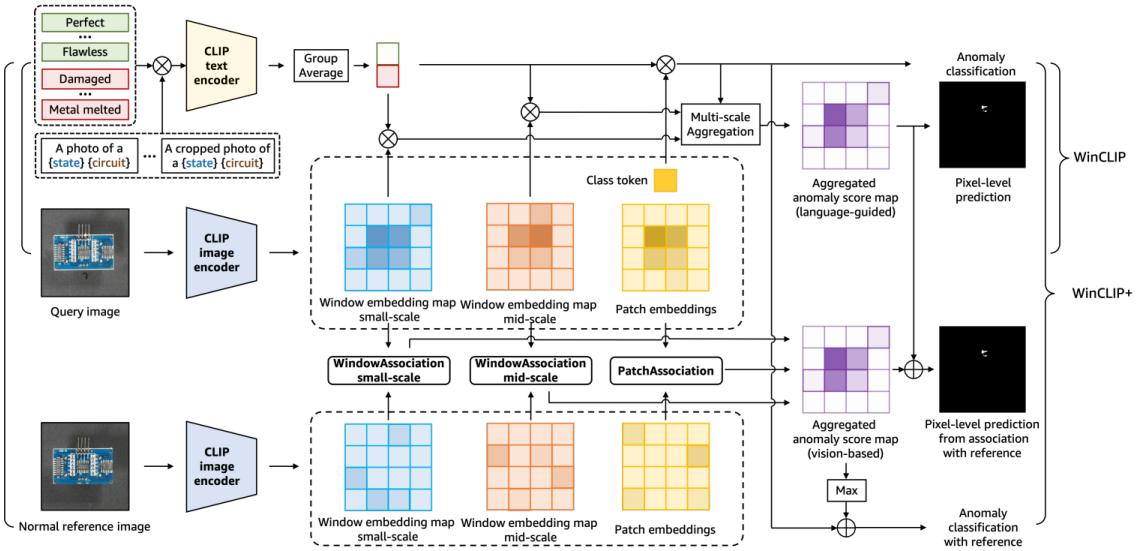

图 1. WinCLIP/WinCLIP+ 的方法示意图

2 相关工作

2.1 对于工业图像无监督异常检测算法

随着工业 4.0 的发展，表面缺陷检测/异常检测成为工业领域的一个热门课题。近年来，基于深度学习的算法在提高效率的同时节省人力成本，已逐渐成为一个备受关注的问题，其性能优于传统的视觉检测方法。而现有的基于深度学习的算法偏向于有监督学习，这不仅需要大量的标记数据和人工劳动，而且会带来效率低下和局限性。相比之下，最近的研究表明，无监督学习在解决视觉工业异常检测的上述缺点方面具有很大的潜力。

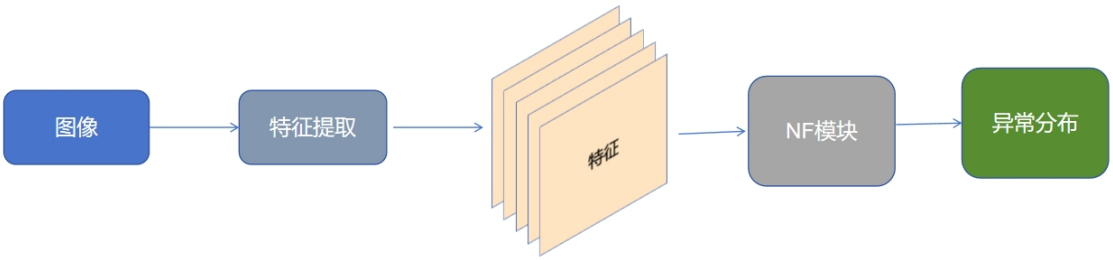

图 2. 基于归一化流（NF）的方法流程

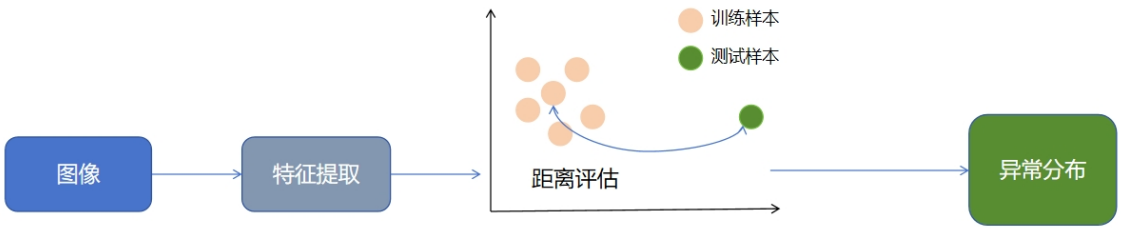

图 3. 基于表示的方法流程

现有的研究可以大概分为五种类型：基于重构的方法，规范化流的方法，基于表示的方法，基于数据增强的方法，算法增强。基于重构的方法主要有:CAVGA [22] 结合 VAE、GAN 等手段，首次将注意力机制引入异常检测领域、UTAD [15] 提出了一种结合互信息、GAN 和自动编码器的自然图像无监督视觉异常检测方法、MOCCA [16] 利用不同深度的深度模型的输出来检测一类设置中的异常输入等；规范化流的方法有:Differnet [20] 和 CFlow [9] 采用归一化流来通过最大化正常图像特征的对数似然的可训练过程来估计分布。将正态图像特征嵌入到标准正态分布中，利用概率对异常进行识别和定位。基于归一化流（NF）的方法的基本流程如 2所示；基于表示的方法主要有 SPADE [6]、PaDIM [7]、SVDD [23] 等，其基本流程如图 3所示；基于数据增强的方法主要有 DRAEM [24]、NSA [21] 等，其流程图如图 4所示；算法增强主要是改进的损失函数或可解释性，比如 IGD [5] 等。

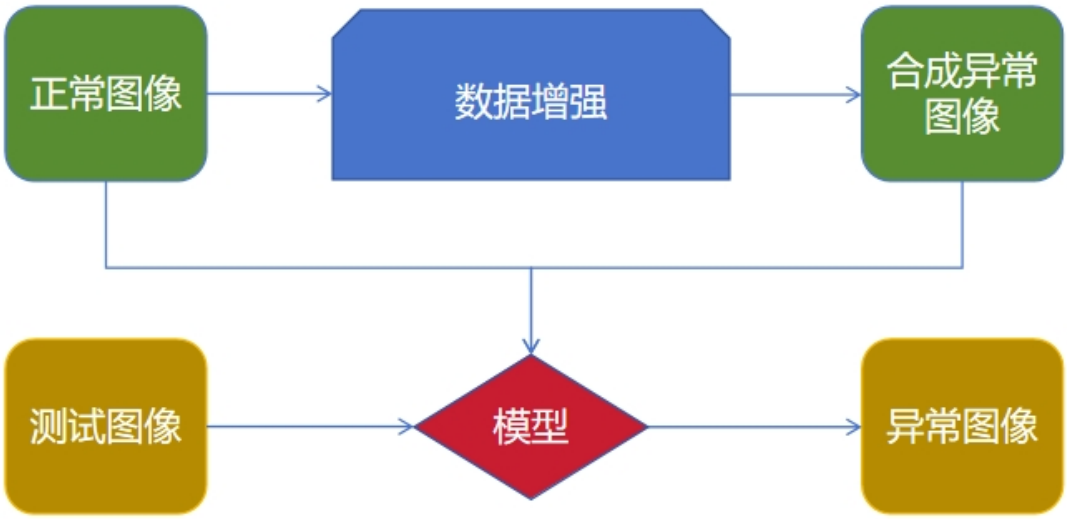

图 4. 基于数据增强的方法流程

2.2 视觉语言模型

视觉语言模型（Vision-Language Models, VLMs）旨在实现图像与文本之间的跨模态理解与生成。近年来，随着深度学习的发展，VLMs 取得了显著进展。早期工作如 Flickr30K [17] 和 COCO 数据集推动了图像描述生成任务的发展，随后基于注意力机制的模型（如 Show, Attend and Tell）进一步提升了性能。Transformer 架构的引入（如 ViLT [12]、CLIP [18]）使得 VLMs 能够更好地捕捉跨模态关联，CLIP 通过对比学习实现了图像与文本的联合表示，成为里程碑式工作。多模态预训练模型（如 UNITER [4]、Oscar [14]、VinVL [25]）通过大规模数据预训练，显著提升了下游任务性能。近期，基于生成式预训练的方法（如 BLIP [13]、Flamingo [1]）在图像描述、视觉问答等任务中表现出色。此外，VLMs 在医疗、自动驾驶等领域的应用也逐步扩展。未来，VLMs 的研究方向可能包括更高效的跨模态对齐、少样本学习以及多模态生成能力的进一步提升。利用预训练的视觉语言模型例如 CLIP 等来进行语言引导检测和分割 [19, 26]，也取得了一些不错的成果。

3 本文方法

3.1 本文方法概述

该论文 [11] 提出了一种新颖的框架，用于通过细粒度的文本定义和正常参考图像来全面进行异常分类和分割。研究中首先展示了在大规模网络数据上预训练的 CLIP 模型，该模型在文本和图像之间提供了强大的表示能力，并具有良好的对齐性，适用于异常识别任务。通过组合式提示集成（Compositional Prompt Ensemble），定义了文本中的正常和异常状态，并帮助从预训练的 CLIP 中提取知识，以实现更好的零样本异常识别。WinCLIP 通过从窗口和图像级别高效聚合与图像-文本对齐的多尺度特征，执行零样本分割。此外，给定少量正常样本，基于视觉的参考关联提供了关于两种状态的补充信息，以补充语言定义，从而形成了少样本 WinCLIP+。在最近的基准测试中，WinCLIP 和 WinCLIP+ 在零样本/少样本设置中均超越了最先进的方法，并取得了显著的优势。

3.2 组合式提示集成（CPE）

这里的组合式提示集成其实就是一组特别设计的图像描述模板，其主要目的是为了更好地定义物体的两种抽象状态。CPE 包括 [a. 每个标签的状态词] 和 [b. 文本模板的预定义列表] 的所有组合。状态词包括大多数对象共有的常见状态，例如，“flawless” 表示正常/ “damaged” 表示异常。此外，我们还可以根据缺陷的先验知识选择性地添加特定于任务的状态词，例如，木头上的 “anomalous”、PCB 上的 “bad soldering” 等等。此外，我们专门为异常任务策划了一个模板列表，例如，“a photo of a [c] for visual inspection”。具体组成如图 5所示。

(a) *State*-level (normal)

- c := "[o]"
- c := "flawless [o]"
- c := "perfect [o]"
- c := "unblemished [o]"
- c := "[o] without flaw"
- c := "[o] without defect"
- c := "[o] without damage"

(b) *State*-level (anomaly)

- c := "damaged [o]"
- c := "[o] with flaw"
- c := "[o] with defect"
- c := "[o] with damage"

(c) *Template*-level

- "a cropped photo of the [c]."
- "a cropped photo of a [c]."
- "a close-up photo of a [c]."
- "a close-up photo of the [c]."
- "a bright photo of a [c]."
- "a bright photo of the [c]."
- "a dark photo of the [c]."
- "a dark photo of a [c]."
- "a jpeg corrupted photo of a [c]."
- "a jpeg corrupted photo of the [c]."
- (cont'd) "a blurry photo of the [c]."
- "a blurry photo of a [c]."
- "a photo of a [c]."
- "a photo of the [c]."
- "a photo of a small [c]."
- "a photo of the small [c]."
- "a photo of a large [c]."
- "a photo of the large [c]."
- "a photo of the [c] for visual inspection."
- "a photo of a [c] for visual inspection."
- "a photo of the [c] for anomaly detection."
- "a photo of a [c] for anomaly detection."

图 5. 多层次提示列表，以构建 CPE

3.3 WinCLIP

WinCLIP 是一种语言引导的结合了滑动窗口和 CLIP 的用于零样本异常检测的方法，它通过从图像中提取与语言对齐的多尺度特征，并高效聚合这些特征以进行像素级别的异常预测。WinCLIP 的核心在于其能够从图像的不同窗口中提取密集的视觉特征，这些特征与通

过组合式提示集成（Compositional Prompt Ensemble）生成的文本嵌入保持对齐。具体来说，WinCLIP 首先在输入图像上滑动窗口，生成一系列局部窗口特征，然后利用预训练的 CLIP 图像编码器提取这些窗口的特征嵌入。这些特征嵌入随后与通过 CLIP 文本编码器得到的表示正常和异常状态的文本嵌入进行相关性比较，生成局部异常分数，其流程如图 6所示。WinCLIP 通过在所有重叠窗口上应用谐波平均1（harmonic averaging）来聚合这些分数，从而得到最终的异常分割图。

$$\overline{M}^{u}_{0,ij} := \left(\frac{1}{\sum_{u,v} \langle \mathbf{w}_{uv} \rangle_{ij}} \sum_{u,v} \frac{\langle \mathbf{w}_{uv} \rangle_{ij}}{M^{u}_{0,uv}} \right)^{-1}, \tag{1}$$

这种方法不仅能够捕捉局部细节，还能够通过多尺度特征聚合来识别不同尺寸的缺陷，从而实现精确的异常分割。WinCLIP 的优势在于其无需针对特定任务进行调整，即可广泛应用于各种视觉检查任务，且在零样本学习设置下无需任何分割标注。

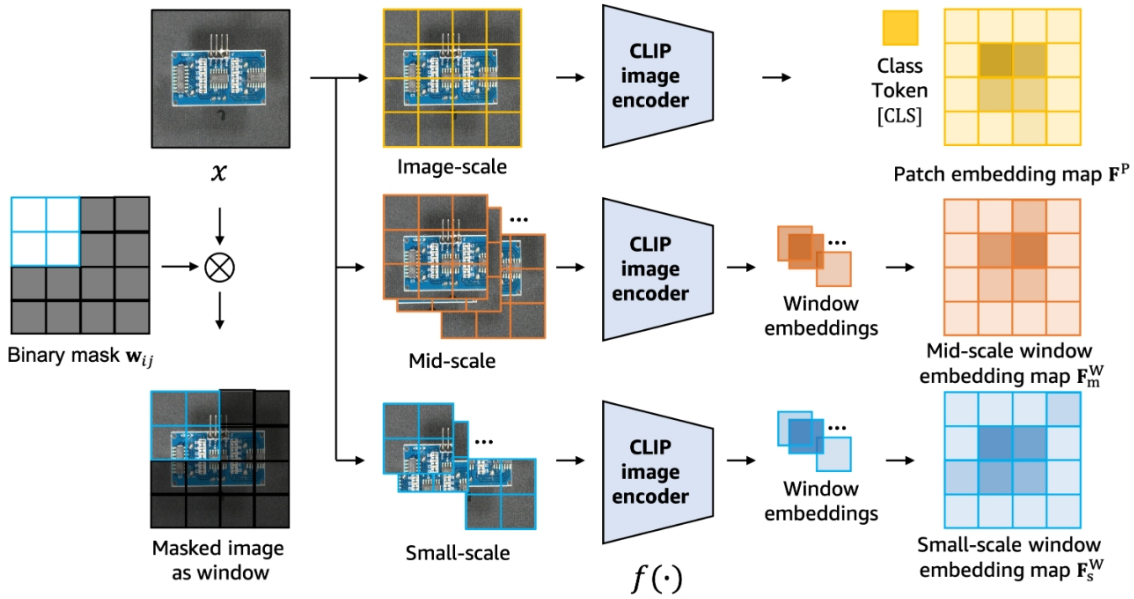

图 6. 通过 CLIP 图像编码器在多尺度窗口中提取 WinCLIP 特征

3.4 WinCLIP+

为了进行全面的异常分类和分割，仅靠语言引导的零样本方法是不够的，因为某些缺陷只能通过视觉参考而不是仅通过文本来定义。例如，MVTecAD [2] 中的 “Metal-nut” 有一种异常类型标记为 “flipped upside-down”，这只能通过正常图像相对识别。为了更精确地定义和识别异常，原文 [11] 提出了 WinCLIP+。WinCLIP+ 是 WinCLIP 的扩展，它结合了语言引导和基于视觉的互补预测，以实现更好的异常分类和分割。WinCLIP 是面向零样本的异常检测，而 WinCLIP+ 则是在 WinCLIP+ 的基础上，给模型提供 1 张或几张正常图像作为参考，WinCLIP+ 利用从少量正常参考图像中学习到的特征来增强异常识别的能力。这些参考图像提供了视觉上下文，有助于定义和识别那些仅通过文本难以描述的异常类型。首先提出了一个参考关联模块作为引入给定参考图像的关键模块，该模块可以简单地基于余弦相似度存储和检索 $\mathcal{D}$ 的记忆特征 $\mathbf{R}$。给定这样的模块和从查询图像中提取的相应（例如，patch-level[2]）

特征 $\mathbf{F} \in \mathbb{R}^{h \times w \times d}$，可以通过以下方式生成异常分割的预测 $\mathbf{M} \in [0,1]^{h \times w}$：

$$\mathbf{M}_{ij} := \min_{r \in \mathbf{R}} \frac{1}{2}(1 - \langle \mathbf{F}_{ij}, r \rangle), \tag{2}$$

然后，在从 WinCLIP 获得的多尺度特征图上应用此关联模块（如图 1所示）。具体来说，给定少量样本，然后从三种不同的特征构建独立的参考记忆：1 · 小尺度的 WinCLIP 特征 $\mathbf{F}_{\mathbf{s}}$；2. 中尺度的 WinCLIP 特征 $\mathbf{F}_{\mathbf{m}}^{\mathsf{W}}$；3. 具有全局上下文的特征 $\mathbf{F}^{\mathsf{P}}$（例如，ViT [8] 中的 patch tokens 由于自注意力机制捕获图像上下文）。尽管 $\mathbf{F}^{\mathsf{P}}$ 不与语言对齐，但它仍然有助于定义正常和异常。

接着，WinCLIP+ 获得了三个参考记忆：$\mathbf{R}^{\mathsf{P}}$、$\mathbf{R}_{\mathbf{s}}^{\mathsf{W}}$ 和 $\mathbf{R}_{\mathbf{m}}^{\mathsf{W}}$ 。然后，对它们的多尺度预测进行平均，以对给定查询进行异常分割，

$$\mathbf{M}^{\mathsf{W}} := \tfrac{1}{3}(\mathbf{M}^{\mathsf{P}} + \mathbf{M}_{\mathbf{s}}^{\mathsf{W}} + \mathbf{M}_{\mathbf{m}}^{\mathsf{W}}), \tag{3}$$

然后与之前获得的语言引导预测 $\widetilde{\mathbf{M}}_0^{\mathsf{W}}$ 进行融合。

为了执行异常分类，将 $\mathbf{M}^{\mathsf{W}}$ 的最大值（来自少量参考样本的空间特征）与 WinCLIP 零样本分类分数（通过语言检索的 CLIP 知识）结合起来获得 $\text{ascore}_{\mathsf{W}}(\mathbf{x})$：

$$\text{ascore}_{\mathsf{W}}(\mathbf{x}) := \tfrac{1}{2}\left(\text{ascore}_0(f(\mathbf{x})) + \max_{ij} \mathbf{M}_{ij}^{\mathsf{W}}\right), \tag{4}$$

这种结合语言引导和视觉基础方法的策略，使得 WinCLIP+ 能够在少样本情况下提供更全面的异常识别能力，显著提高了异常检测的性能。WinCLIP+ 的设计强调了在有限的标注数据下，如何有效地利用正常图像作为参考，以及如何通过语言提示来增强模型的泛化能力。WinCLIP+ 和 WinCLIP 的整体流程图如图 1所示。

4 复现细节

4.1 与已有开源代码对比

截至作者复现时，该论文 [11] 的研究团队暂未开源其实验代码。因此作者结合论文中的描述以及其他研究者复现的工作 [3, 27, 28]，来完成本次的复现内容。并且相比于原论文，在更广阔的数据集上验证了方法的可行性，在一些类别上的实验结果要优于原论文。

4.2 实验环境搭建

本代码采用最新版的 Pytorch 进行实现，预训练模型选用的是 OpenCLIP [10] 的 ViT-B/16+，实验设备为单张 NVIDIA GeForce RTX 4090，评价指标如表 1所示。最后在 MVTec、VisA、MPDD 三个数据集上验证了复现效果，其中前两个数据集为原论文的实验数据集，第三个数据集为作者扩展验证的数据集。

评价指标	作用
AUROC	图像级别，评估方法异常分类性能
F1	图像级别，评估方法异常分类性能
AP	图像级别，评估方法异常分类性能
AUROC-px	像素级别，评估方法异常分割性能
F1-px	像素级别，评估方法异常分割性能
AP-px	像素级别，评估方法异常分割性能

表 1. 实验结果评价指标

4.3 复现内容和创新点

- 对 MVTec-AD 和 VisA 基准数据集进行数据处理和分析，以便进行后续实验。
- 研究论文中提到的 CLIP 模型和其他相关模型的结构，构建了相同或类似的模型来进行复现。
- 根据论文中提到的实验设计，设计了相似的实验来验证模型是否具有相似的效果。
- 原论文只在 MVTec-AD 和 VisA 数据集上进行了实验验证，本文将其拓展到 MPDD 数据集上进行测试。
- 编写了一个组合提示集合类，结合数据集中的类名和缺陷类型用来对所有测试数据批量生成文本描述。
- 采用最新版本的 Pytorch 对该工作进行代码复现，还原设计其多级特征提取的过程。
- 采用热力图的形式对实验结果进行可视化。

5 实验结果分析

本文在三个数据集上分别进行了 zero-shot、1-shot、2-shots、4-shots 的实验，并保存其定性和定量的实验结果，其中 zero-shot 采用的是 WinCLIP 方法，其他提供了正常样本作为参考的如 1-shot、2-shots、4-shots 则采用的是 WinCLIP+ 方法。

如表 2所示, 这张表展示的是在 MVTec 数据集上 AUROC 指标复现结果和原论文结果的比较，可以看到从整体上来看，复现的效果是比较理想的。对于零样本的情况，结果略逊于原论文，但是在 1-shot 和 4-shots 的情况下，结果均要好于原论文。从个人理解来看，导致这种情况的根本原因在于，预训练的 CLIP 虽然在多级特征提取和 CPE 的帮助下，可以有效改善其零样本检测的能力，但是针对某些特定的异常类别，仅靠语言的描述是无法准确定义异常的，比如 PCB 板上零件焊接的正反，这种异常往往需要和正常的情况进行类比来进行异常的定义，因此模型对于这种异常的零样本异常检测能力就较差。此时如果我们再给模型提供少量的正常图片作为参考，可以看到在 few-shots 的情况下实验结果就有了明显的提升。由于篇幅的限制，在此仅展示这一个指标的定量结果表格，更多的定量实验结果可以在提交的辅助材料压缩包中查看。

	k=0		k=1		k=4	
AUROC	Paper	Reprod	Paper	Reprod	Paper	Reprod
Bottle	99.2	98.6	98.2±0.9	99.4↑	99.3±0.4	99.4
Cable	86.5	85	88.9±1.9	89.3↑	90.9±0.9	89.6
Capsule	72.9	68.6	72.3±6.8	83.5↑	82.3±8.9	86.6↑
Carpet	100	99.3	99.8±0.3	100↑	100.0±0.0	99.9
Grid	98.8	99.2↑	99.5±0.3	99.6↑	99.6±0.1	99.7↑
Hazelnut	93.9	92.3	97.5±1.4	98↑	98.4±0.4	97.6
Leather	100	100↑	99.9±0.0	100↑	100.0±0.0	100↑
Metal nut	97.1	96.2	98.7±0.8	98.2	99.5±0.2	99.3
Pill	79.1	81.5↑	91.2±2.1	89.6	92.8±1.0	92.5
Screw	83.3	71.6	86.4±0.9	81.5	87.9±1.2	81.3
Tile	100	99.9	99.9±0.0	100↑	87.9±1.2	100↑
Toothbrush	87.5	85.3	92.2±4.9	91.4	96.7±2.6	98.1↑
Transistor	88	89.1↑	83.4±3.8	89.7↑	85.7±2.5	90.3↑
Wood	99.4	97.6	99.9±0.1	99	99.8±0.3	99.3
Zipper	91.5	91.2	88.8±5.9	86.3	94.5±0.5	95.5↑
Mean	91.8	90.4	93.1	**93.7↑**	94.4	**95.3↑**

表 2. MVTec 数据集上 AUROC 指标复现结果和原论文结果的比较

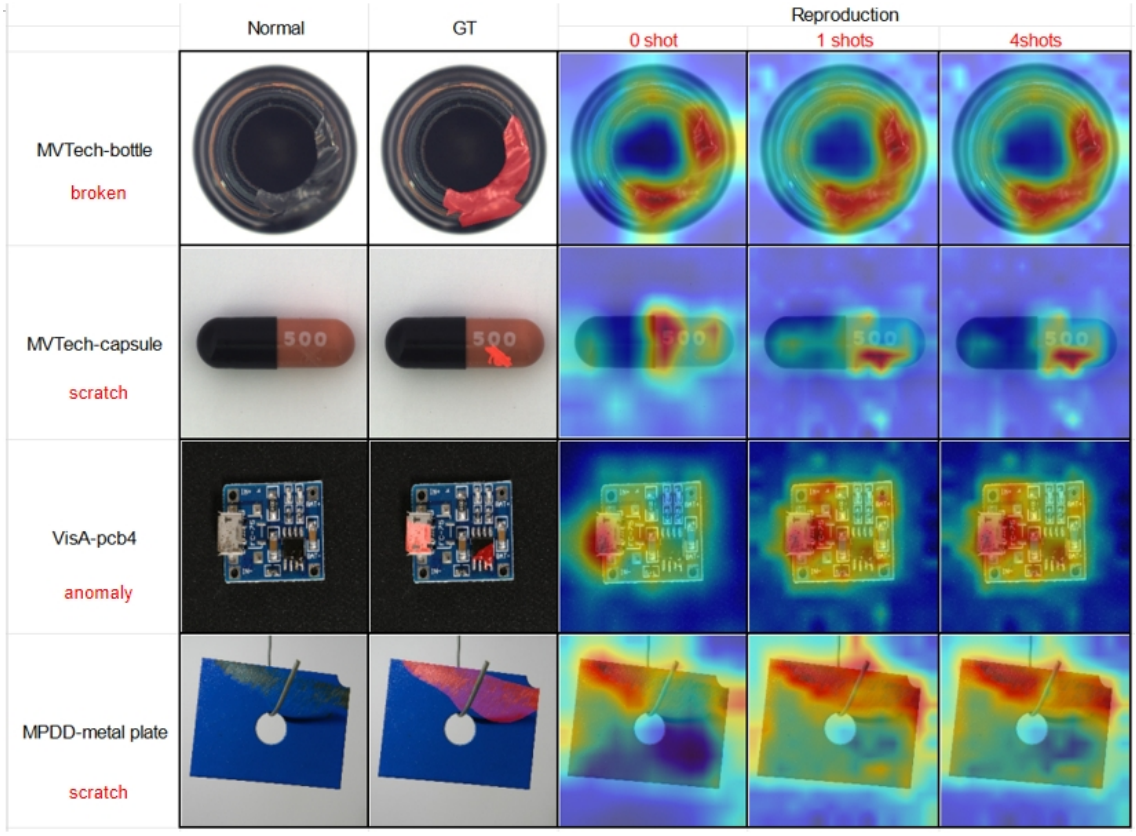

图 7. 不同数据集上部分类别的可视化结果比较

图 7则是不同数据集上部分类别的可视化比较，我们从不同数据集中选取了部分类别，将其可视化结果与 GT 进行对比。可以看到，通过提供少量的正常图像给模型作为参照，其对于异常的检测正确率会有很大的改善。在扩展的 MPDD 数据集上，该方法仍然表现出了较好的鲁棒性。

6 总结与展望

本文通过对 WinCLIP 这篇工作 [11] 进行复现，成功验证了基于滑动窗口的零/少量样本异常检测方法 WinCLIP 和 WinCLIP+ 的有效性。研究复现了其提出的 WinCLIP 框架，并在 MVTec-AD、VisA 和 MPDD 三个数据集上进行了广泛的实验验证。实验结果表明，最终的复现效果基本达到了原论文的水平。尽管复现结果在多个数据集上取得了理想的性能，但还是有指标略低于原论文，这表明模型在完全无监督的情况下可能存在一定的局限性。未来研究可进一步探索更高效的跨模态对齐、少样本学习和多模态生成能力的提升，以及如何更有效地利用正常图像作为参考，增强模型的泛化能力。此外，研究还可扩展到更广泛的工业领域，以验证模型的普适性和鲁棒性。

参考文献

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Proc. Conf. on Neural Information Processing Systems*, 35:23716–23736, 2022.

[2] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *Int. J. Computer Vision*, 129(4):1038–1059, April 2021.

[3] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization, 2023.

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Proc. Euro. Conf. on Computer Vision*, pages 104–120. Springer, 2020.

[5] Yuanhong Chen, Yu Tian, Guansong Pang, and Gustavo Carneiro. Deep one-class classification via interpolated gaussian descriptor. In *Proc. AAAI Conf. on Artificial Intelligence*, volume 36, pages 383–392, 2022.

[6] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020.

[7] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *Proc. IEEE Int. Conf. on Pattern Recognition*, pages 475–489. Springer, 2021.

[8] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF winter Conf. on applications of computer vision*, pages 98–107, 2022.

[10] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021.

[11] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 19606–19616, 2023.

[12] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conf. on machine learning*, pages 5583–5594. PMLR, 2021.

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conf. on machine learning*, pages 12888–12900. PMLR, 2022.

[14] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proc. Euro. Conf. on Computer Vision*, pages 121–137. Springer, 2020.

[15] Yunfei Liu, Chaoqun Zhuang, and Feng Lu. Unsupervised two-stage anomaly detection. *arXiv preprint arXiv:2103.11671*, 2021.

[16] Fabio Valerio Massoli, Fabrizio Falchi, Alperen Kantarci, Şeymanur Akti, Hazim Kemal Ekenel, and Giuseppe Amato. Mocca: Multilayer one-class classification for anomaly detection. *IEEE Trans. on neural networks and learning systems*, 33(6):2313–2323, 2021.

[17] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proc. Int. Conf. on Computer Vision*, pages 2641–2649, 2015.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conf. on machine learning*, pages 8748–8763. PMLR, 2021.

[19] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 18082–18091, 2022.

[20] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF winter Conf. on applications of computer vision*, pages 1907–1916, 2021.

[21] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *Proc. Euro. Conf. on Computer Vision*, pages 474–489. Springer, 2022.

[22] Shashanka Venkataramanan, Kuan-Chuan Peng, Rajat Vikram Singh, and Abhijit Mahalanobis. Attention guided anomaly localization in images. In *Proc. Euro. Conf. on Computer Vision*, pages 485–503. Springer, 2020.

[23] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proc. of the Asian conference on computer vision*, 2020.

[24] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proc. Int. Conf. on Computer Vision*, pages 8330–8339, 2021.

[25] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 5579–5588, 2021.

[26] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 16793–16803, 2022.

[27] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *Proc. Int. Conf. on Learning Representations*, 2024.

[28] Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, 2024.

[29] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *Proc. Euro. Conf. on Computer Vision*, 2022.