南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

AdaptCLIP：适用于通用视觉异常检测的CLIP适配方法

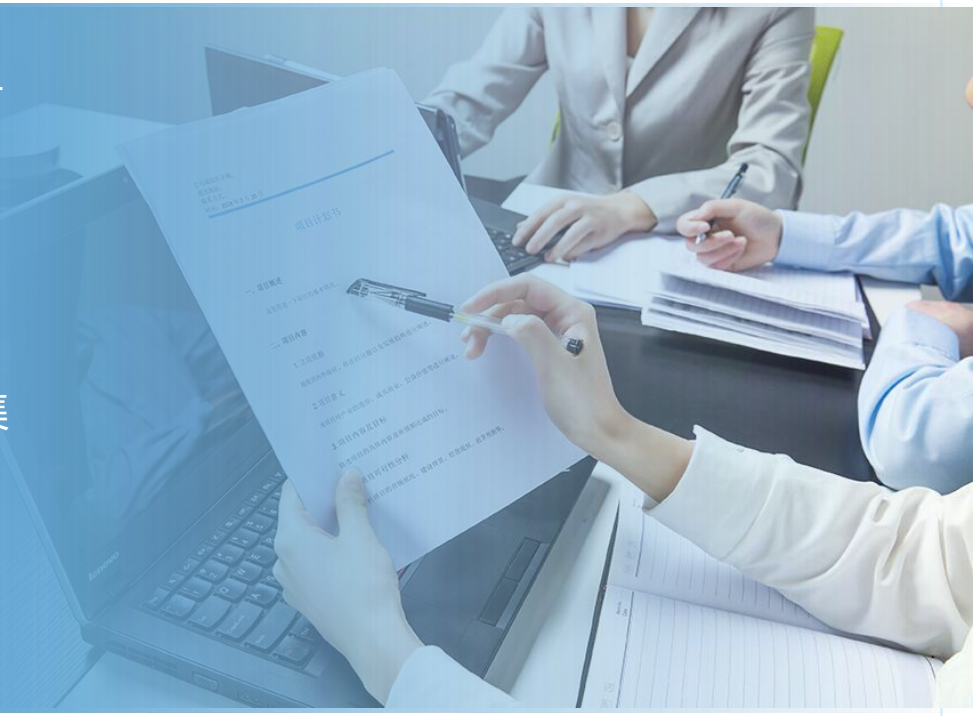
高彬彬[1]周悦2,3闫江涛1蔡月智2
张伟喜2孟王军2刘军1刘勇1王磊2王成杰[1,4]1腾讯优图实验室[2]西门子股份公司[3]慕尼黑工业大学[4]上海交通大学

汇报人：蒋明忠

时间：2025.06

Background

通用视觉异常检测 (AD) 旨在在一个基础数据集或已知数据集上学习单个模型后，识别异常图像并分割来自新颖或未见视觉对象的异常像素。 这是一个更具挑战性的任务，因为它在面对跨域数据集时需要强大的泛化能力。 同时，它也是一个更实际的话题，因为人们更关注现实场景中的快速适应性，尤其是在低数据状态下（即少样本甚至零样本）。 例如，在医学图像诊断和工业视觉质量检测中，由于固有的稀缺性和隐私保护，难以收集大规模数据集。 最近，开发通用的视觉异常检测 (AD) 越来越受到关注，因为现有的无监督 AD，无论是分离的模型还是统一模型，尽管在已知对象上表现良好，但在未知对象上的性能都很差。



Background

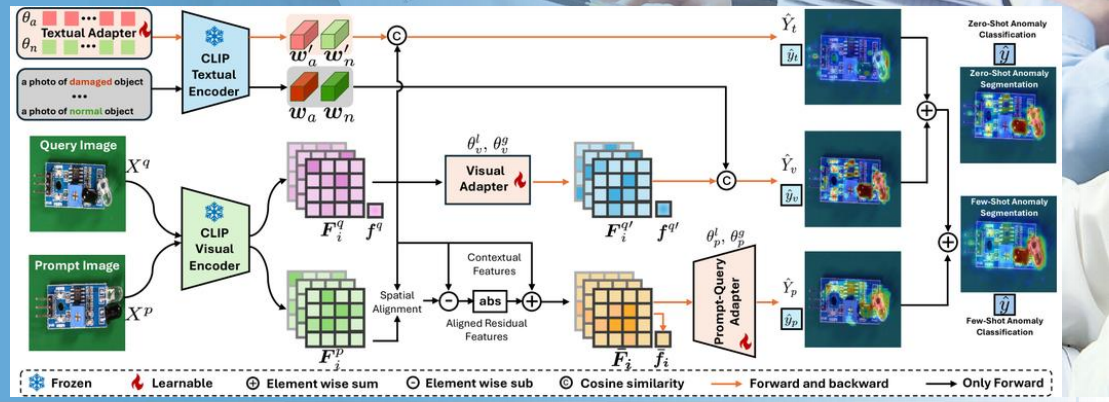
(1) 无监督异常检测的目标是在给定足够的正常训练图像的情况下识别异常。 大多数无监督AD方法大致可以分为三类：基于嵌入、基于判别和基于重建的方法。 所有这些方法都局限于识别已知类别的异常，但在未知类别上的性能往往较差。 对于新的场景，人们必须首先收集足够的正常图像，然后重新训练模型。 这是低效的，并且缺乏实际应用所需的快速适应性。

(2) 零样本异常检测 (Zero-Shot ADs) 通过利用大型视觉语言模型（例如，CLIP ）取得了令人印象深刻的性能。 AnomalyCLIP 学习与类别无关的提示嵌入，以对齐逐块符元，从而避免密集窗口操作。 此外，AnomalyCLIP 通过在 CLIP 的中间层附加一些可学习的符元来改进原始 CLIP 表示。我们认为，这些额外的操作使模型更加复杂，并可能损害 CLIP 的原始能力

(3) 少样本异常检测 (Few-Shot AD) 主要关注学习或仅使用数量有限的正常图像。 WinCLIP是首个将 CLIP 模型应用于少样本异常检测的工作，它将正常符元存储到内存库中，然后使用余弦相似度检索每个查询符元最近的符元，最后使用最近距离计算异常图。但是，当应用于目标数据集时，它需要重新训练模型。

Background

通用视觉异常检测旨在无需额外微调即可识别来自新颖或未见视觉领域中的异常，这在开放场景中至关重要。 最近的研究表明，像CLIP这样的预训练视觉语言模型仅使用零样本或少量正常图像就能展现出强大的泛化能力。 然而，现有方法在设计提示模板、复杂的符元交互或需要额外微调方面存在困难，导致灵活性有限。 在这项工作中，我们提出了一种简单而有效的方法，称为AdaptCLIP，它基于两个关键见解。 首先，应该交替学习自适应视觉和文本表示，而不是联合学习。 其次，查询和正常图像提示之间的对比学习应该结合上下文和对齐的残差特征，而不仅仅依赖于残差特征。 AdaptCLIP将CLIP模型视为基础服务，在其输入或输出端仅添加三个简单的适配器：视觉适配器、文本适配器和提示-查询适配器。 AdaptCLIP支持跨领域的零样本/少样本泛化，并在基于数据集训练后，能够在目标域上实现免训练的方式。 AdaptCLIP在来自工业和医学领域的12个异常检测基准测试中取得了最先进的性能，显著优于现有的竞争方法。



Method

$$\hat{Y} = [\frac{\exp(\langle \vec{w}_a, \vec{F}_i^q \rangle)}{\exp(\langle \vec{w}_a, \vec{F}_i^q \rangle) + \exp(\langle \vec{w}_n, \vec{F}_i^q \rangle)}],$$

所有基于图像块的预测分数根据其空间位置重新排列并插值到原始输入分辨率。

$$\hat{y} = \frac{\exp(\langle \vec{w}_a, \vec{f}^q \rangle)}{\exp(\langle \vec{w}_a, \vec{f}^q \rangle) + \exp(\langle \vec{w}_n, \vec{f}^q \rangle)}.$$

视觉适配器：使用固定的文本嵌入(Wa和Wn)来适应视觉标记(Fqi和fq))。它由global和local两个分支组成，分别变换全局图像token和局部token。架构上，全局和局部分支使用简单的残差多层感知(MLP)实现。

$$\vec{F}_i^{q'} = \vec{F}_i^q + \mathrm{MLP}(\vec{F}_i^q; \theta_v^l); \vec{f}^{q'} = \vec{f}^q + \mathrm{MLP}(\vec{f}^q; \theta_v^g),$$

文本适配器：直接学习两类提示 θ→a,θ→n∈Rr×d，无需提示模板，其中 r>0 是提示的长度。 我们将其输入到 CLIP 的冻结文本编码器中，并获得相应的embeddings

$$\vec{w}_a' = \mathcal{T}(\vec{\theta}_a), \vec{w}_n' = \mathcal{T}(\vec{\theta}_n).$$

Method

基于对比学习的AdaptCLIP：与静态或可学习的文本提示相比，使用普通图像作为视觉提示更直观。 因此，我们期望学习查询图像 Xq 及其对应的普通提示 Xp 之间的比较能力，该能力能够很好地泛化到未见过的物体。 我们发现，应用多层特征可以产生更好的结果。

1、空间对齐

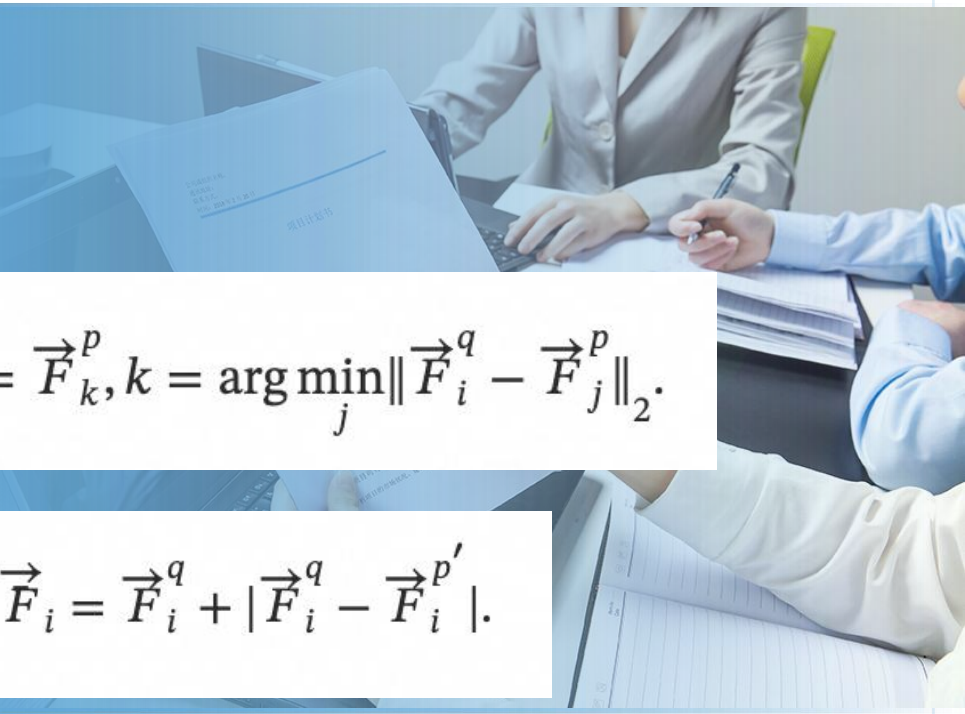
2、联合上下文和对齐残差特征

3、提示-查询适配器

最终目标是实现像素级异常分割和图像级异常分类。

因此，我们提出一个轻量级的分割头 $\mathcal{G}$ [×](· ;θpl)

基于联合特征学习异常分割 F¯



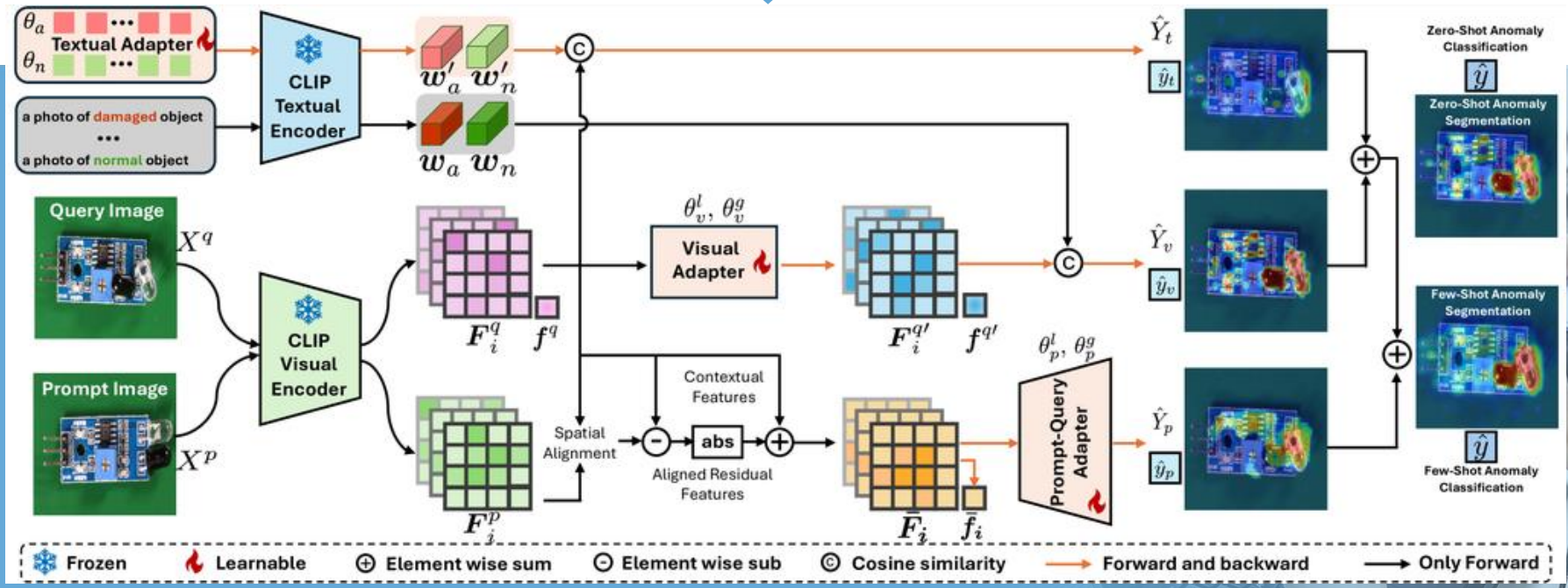
$$\vec{F}_i^{p'} = \vec{F}_k^p, k = \arg\min_j \|\vec{F}_i^q - \vec{F}_j^p\|_2.$$

$$\vec{F}_i = \vec{F}_i^q + |\vec{F}_i^q - \vec{F}_i^{p'}|.$$

$$\hat{Y}_p = \mathcal{G}(\bar{F}; \theta_p^l),$$

$$\hat{y}_p = \mathrm{MLP}\left((\mathrm{AvgPool}(\bar{F}) + \mathrm{MaxPool}(\bar{F}))/2; \theta_p^g\right),$$

Method

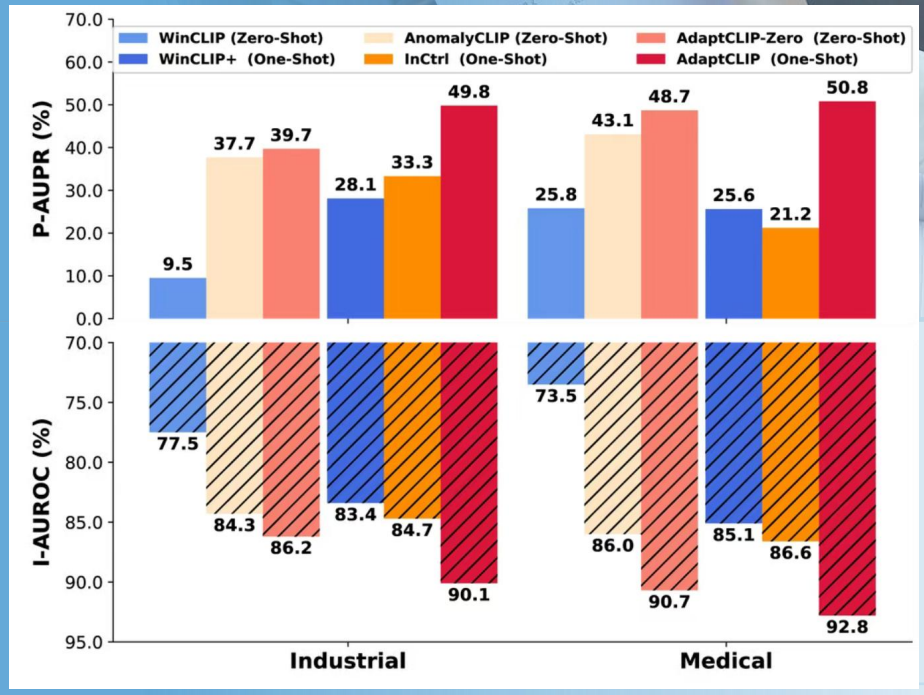


Adaptclip：由三个可插的适配器组成，即视觉适配器，文本适配器和及时引用适配器。 首先，前两个适配器交替学习用于零样本异常检测的视觉和文本表示。 提示查询适配器进一步学习查询图像与其对应的正常提示之间的比较能力，用于少样本异常检测。 一旦训练完成，它只需少量样本甚至零样本的正常图像提示即可分割任何异常。

Experiments

最先进方法和我们的AdaptCLIP的比较。 ✓ 表示满足， ✗ 表示不满足。 我们的方法支持在不同领域进行零样本/少样本 (ZS 和 FS) 视觉异常检测，无需在目标数据集上进行微调 (FT)。 它只在 CLIP 的输入或输出端添加简单的适配器，无需复杂的符元交互，从而保留 CLIP 的原始能力 (OA)。 使用单个普通图像提示的 AdaptCLIP 在来自工业和医学领域的 12 个 AD 基准测试中，在图像级异常分类 (I-AUROC) 和像素级异常分割 (P-AUPR) 上取得了最佳性能。 此外，零样本 AdaptCLIP 也明显优于现有的零样本方法，甚至一些单样本方法。

Methods	ZS	FS	OA	w/o FT
WinCLIP [16]	✓	✓	✓	✓
AdaCLIP [6]	✓	✗	✗	✓
InCtrl [54]	✗	✓	✓	✓
AnomalyCLIP [53]	✓	✗	✗	✓
PromptAD [23]	✗	✓	✓	✗
AdaptCLIP	✓	✓	✓	✓



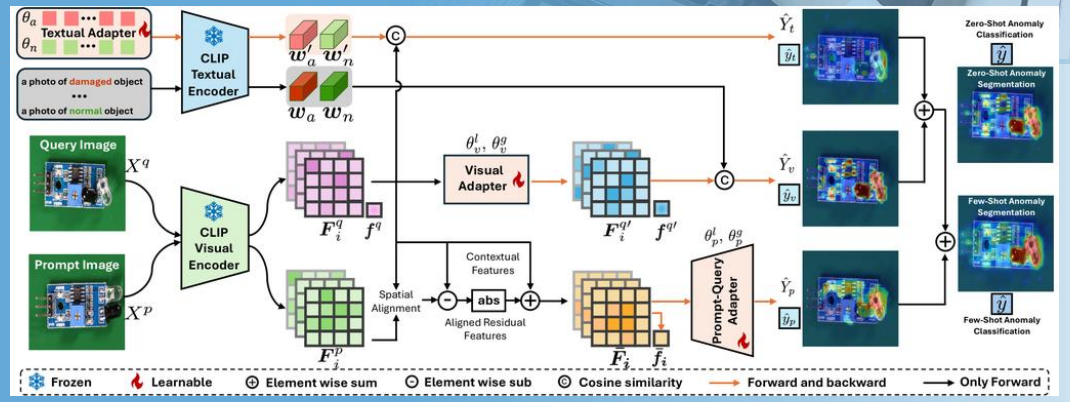
Experiments

为了证明AdaptCLIP中提出的三个适配器（TA：文本适配器，VA：视觉适配器和 PQA：提示-查询适配器）以及两个主要见解（交替学习和基于联合上下文和对齐残差特征的比较学习）的有效性，我们在MVTec和VisA上进行了实验

No.	Methods	Shots	TA	VA	PQA	MVTec	VisA
0	baselines	0	✗	✗	✗	91.1 / 33.0	82.1 / 18.0
1		0	✓	✗	✗	92.2 / 31.4	82.9 / 19.7
2		0	✗	✓	✗	90.5 / 39.4	81.0 / 22.1
3	joint	0	✓	✓	✗	89.3 / 36.2	81.6 / 21.5
4	alternating	0	✓	✓	✗	93.5 / 38.3	84.8 / 26.1
5	w/o context	1	✗	✗	✓	62.6 / 7.0	85.3 / 28.7
6	w context	1	✗	✗	✓	88.1 / 50.2	88.9 / 38.1
7	AdaptCLIP	1	✓	✓	✓	94.2 / 52.5	92.0 / 38.8

Experiments

在本文中，我们引入了一个通用的异常检测任务，该任务侧重于跨领域（例如工业和医疗领域）以及在开放场景（例如零样本或少样本设置）中泛化异常检测模型。 一旦训练了通用的异常检测模型，它就不需要对目标数据集进行任何微调。 与单个零样本或少样本AD模型相比，通用异常检测模型更灵活，支持通过固定或可学习的文本提示和一些正常图像提示进行零样本/少样本推理，同时提供图像级和像素级的异常预测。 我们提出了一种通用的异常检测框架AdaptCLIP，它交替学习自适应视觉表示和文本提示嵌入，并根据查询图像的上下文信息以及查询和提示之间的对齐残差特征共同学习比较。 在8个标准工业数据集和4个医疗数据集上进行的大量实验表明，AdaptCLIP在多种设置下显著优于当前的竞争模型。



Thanks