MGCLIP：Multi-grained Prompt Learning for Industrial Zero-shot Anomaly Detection

# Abstract

工业制造场景中，在无可用训练样本的情况下实现对目标图片的零样本异常检测，对于保证零件的质量至关重要。由于数据隐私以及一些定制化新工业产品生产过程中，难以获取目标域图片数据时，ZSAD是一项重要的任务。此外由于工业零件复杂多样，前景物体，背景特征以及异常区域的类型与尺度等都具有很大差异，因此建立能够推广到不同应用领域的模型至关重要。最近，大型预训练的视觉语言模型（VLM），例如CLIP，在异常检测等视觉任务中表现出了强大的零样本识别能力。但仍存在较多局限，原始CLIP更注重于对图像中前景物体的类别语义进行建模，而不是图片中的正常或者异常特征。当前许多研究通过手工设计或采用可学习的与图像中对象无关的表示正常or异常的文本提示来驱动CLIP模型捕获图像中的正常/异常视觉特征，而不是表示对象类别的语义特征。然而，这些方法多集中于设计针对图像级别的文本提示，并没有充分驱动VLM模型更集中地针对图片中每个patch的局部区域内对视觉编码器中提取到的异常特征进行抓取。另外，在原始CLIP中，图像和文本嵌入是被分别映射到一个联合空间内进行语义对齐，文本和视觉模态之间的信息并没有进行交互，未能使文本嵌入能够适应更细粒度以及不同尺度的异常特征。为此，我们提出了MGCLIP，通过设计多粒度的可学习文本提示框架与多粒度的跨模态特征融合，来实现更准确的，泛化性更强的ZSAD框架。首先，引入了了多粒度可学习文本提示框架MGPL：通过Integrating Spatial Information-Embedded Patch-Scale and Global-Scale Multi-grained Prompt Learning for zero-shot anomaly detection. （整合带有空间信息的补丁级与全局级多粒度提示学习）驱动模型更好地从视觉编码器中获取局部异常的视觉上下文特征，使得模型既能够整体地检测图像中是否有异常，又能在局部的Patch级别的区域内检测一些不易被发现的复杂，细小的异常。另外一方面，考虑将视觉编码器中不同尺度大小的异常特征与相对应的Patch-scale级别的文本特征进行交互。设计了多粒度的语言-视觉嵌入交互模块（MGLV，Multi-grained language-visual embedding interaction module），在视觉编码器中间层的视觉特征上应用不同尺度大小的卷积核来提取全局和局部异常特征，进一步使得文本嵌入能够驱动捕获多粒度的图像特征。此外，我们对视觉嵌入和文本嵌入应用多头的交叉注意力机制，获得更丰富的视觉特征用于微调文本提示，使模型自适应地关注图像中的重要区域和语言提示中的关键信息。最后，为了避免全局和局部patch级别的文本提示出现不一致的情况，以及避免两种提示独立学习时提示不足的问题。针对patch-scale的文本提示嵌入特征，计算其与全局文本嵌入特征之间的损失，有效地将全局提示和 patch-scale 提示进行对齐。在常用的6个工业缺陷检测数据集MVtec AD, VisA等进行了实验，与SOTA的模型，包含了基于手工设计和可学习的文本提示的方法进行对比，均取得了优秀的ZSAD performance.

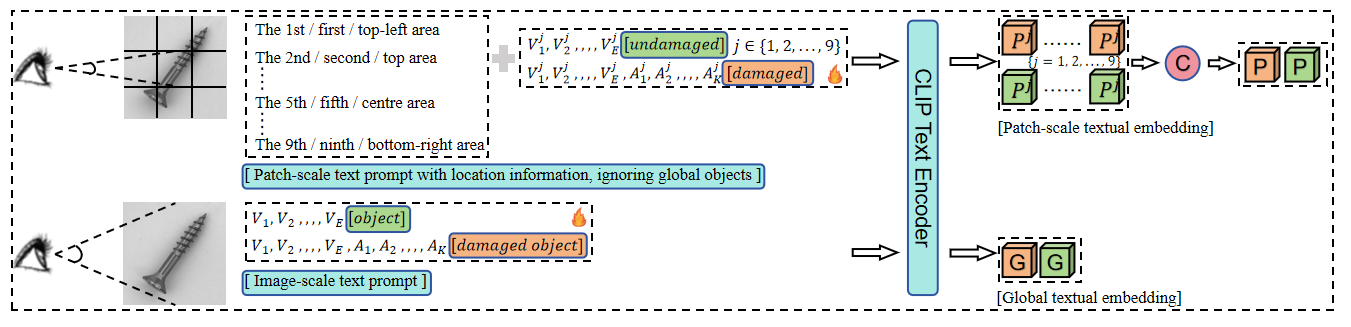
# Introduction

产品的异常检测，特别是在工业制造领域，异常检测对于保证零件的质量至关重要。针对产品图像的视觉异常检测技术获得了广泛的关注[1]。在工业生产过程中，异常是小概率事件，缺陷样本图片少，且缺陷类型具有不可控性，复杂多样。先前的研究基于大量的正常图片进行无监督的训练[2][3][4]，当待测零件出现不同于正常图片中的特征，即判定异常。然而，传统的异常检测方式存在一些不可忽视的弊端：First，传统解决方法是为特定场景或者某一产品训练定制模型，模型可移植性差，资源消耗高[6]。Second，在一些保密场景，由于数据隐私等问题，难以获取目标域图片数据[7]。在一些多品种，小批量的定制化工业产品生产过程中，难以获取到特定产品的训练样本[8]。因此在目标数据集中没有任何训练样本的情况下实现零样本异常检测[9]成为解决上述问题的办法。

最近，大型预训练视觉语言模型（VLMs）[10][11][12]在各种视觉任务中，包括图像目标分类，目标分割，展现出了强大的零样本识别能力，在不同的工业产线应用场景中，针对不同种类的产品，复杂多样的缺陷类型，有一定的零样本异常检测能力。特别是基于数十亿图像-文本对进行预训练的CLIP[13]，已经学习到了图像中各种通用的特征表达信息，能够在无需微调或者少量微调的情况下，仅依靠文本提示即可泛化到各种下游任务中。但CLIP最初是训练用于与图像中前景物体的类别语义对齐[14]，而不是检测图像中的正常或异常特征，因此CLIP在检测视觉正常或异常状态的泛化能力较弱。WinCLIP[15]首先提出语言描述引导模型识别图像中的“正常”和“异常”状态，第一个提出使用“正常”和“异常”的文本提示作为类别来替代图像中物体的类别语义。并设计了状态级的单词组合，“Perfect, Flawless, Damaged, Metal melted”等来更好地表示物体的正常和异常状态，通过CLIP，提取正常或异常的文本语义，与视觉异常特征进行对齐。但该文本提示方法通常会导致文本提示嵌入选择的是与图像级别的全局特征中的对象语义进行对齐，忽视了CLIP视觉编码器中所提取到的细粒度的局部特征[Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15254–15264, 2023.];[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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16793–16803, 2022.]。AnomalyCLIP[16]设计了一组简单但统一且有效的与对象无关的可学习文本提示模板，分别用于提示正常和异常两类，捕捉图像中的通用正常和异常状态。避免了WinCLIP中需要严重依赖于数百个手动定义提示的缺点。另外，AnomalyCLIP中将文本提示嵌入与视觉编码器中的全局嵌入特征和局部嵌入特征同时计算损失函数，更好地学习到了图像中异常的细粒度特征。FAPrompt[17]考虑学习细粒度的异常提示，认为每一个异常图案都是由一些缺陷图案叠加在正常图案之上。其提出了可学习的复合异常提示模块，由通用的可学习的正常提示标记和一组互补且相互分离的可学习异常标记复合而成。通过该模块学习可以建模各种细粒度异常语义的提示，实现了对同一产品中不同类型缺陷的检测。但是无论是WinCLIP中手动定义的文本提示，AnomalyCLIP中可学习的正常/异常提示，亦或者是FAPrompt中细粒度的可学习异常提示，上述方法都集中在针对整张图像进行图像级别的文本提示。在工业产品制造过程中由于生产环境以及零件的复杂及多样性，缺陷形貌复杂多样，尺度大小不一，不乏尺寸较小且易出现在零件中不易被观察到的边缘区域的缺陷，上述文本提示方法并没有聚焦到图像的局部区域。可能会出现一些问题：从图像的全局来看，物体并没有缺陷，但如果把视野集中到图像的某一个Patch内，物体是存在缺陷的。即上述文本提示方法并没有充分驱动VLM模型更集中地针对图片中的每个patch的局部区域内对视觉编码器中提取到的异常特征进行抓取。在本文中，我们提出了一种新的提示方法，即MGCLIP（Multi-grained Prompt Learning for Industrial Zero-shot Anomaly Detection），考虑将图像划分为多个Patch，在包含了方位信息以及位置编码的Patch层级内对异常检测进行文本提示，提示CLIP模型从视觉编码器中更精准地获取图像中每个patch内的异常特征。首先设计了多粒度提示模块MGPL(Multi-grained Prompt Learning)，如图1所示：不关注物体的包含局部空间信息的patch级别的局部文本提示，以及关注物体但与对象类别无关的图像级别的全局文本提示。实现将文本提示聚焦到图像中的局部Patch区域内，同时再结合图像级别的全局文本提示，多粒度的文本提示驱动模型更好地从视觉编码器中获取局部异常的视觉上下文特征，使得模型既能够整体地检测图像中是否有异常，又能在局部的Patch级别的区域内检测一些不易被发现的复杂，细小的异常。

另外一方面，在原始CLIP中，图像和文本嵌入是被分别映射到一个联合空间内进行语义对齐，在此之前，图像和文本嵌入并没有进行交互。上述方法[15, 16, 17]在原始CLIP的基础上，对视觉嵌入和文本嵌入进行了简单的交互，增强了不同模态之间的相互理解。但上述方法对于不同模态之间的交互仍然不够充分，而且容易导致图像对于某些文本提示的过拟合，未能使文本嵌入能够适应更细粒度以及不同尺度的异常特征。最近，视觉-语言特征跨模态的交互与融合对于缺陷区域的特征提取非常重要，已被证明有利于帮助异常区域的定位[29Filo,30,31]。跨模态特征在特征层维度拼接之后应用卷积来实现多模态的信息融合是一种广泛而有效的方式[26,Hu]。Chen等[27]设计了视觉文本联合嵌入的兼容性度量组件See-through-Text Embedding Pixelwise来实现视觉特征逐像素地与文本特征进行对齐。此外，注意力机制同样也是跨模态交互中对齐语言描述与视觉编码器所提取到的视觉特征的有效方法。Xu[27]等提出了一种新的视觉语言交互模块bridger对视觉编码器提取到的中间层特征与语言特征应用注意力机制进行交互，实现跨模态的信息交换。然而，上述所有方法并没有考虑将视觉编码器中不同尺度大小的异常特征与相对应的Patch-scale级别的文本特征进行交互。我们将针对图像中的每个Patch区域内的文本提示嵌入进行融合，然后为了实现将视觉编码器中不同尺度的视觉特征与文本特征能够进行良好的交互与融合，以帮助CLIP对局部区域异常语义的理解。设计了多粒度的语言-视觉嵌入交互模块（MGLV，Multi-grained language-visual embedding interaction module），在中间层的视觉特征上应用不同尺度大小的卷积核来提取全局和局部异常特征，进一步使得文本嵌入能够驱动捕获多粒度的图像特征。此外，我们对视觉嵌入和文本嵌入应用多头的交叉注意力机制，获得更丰富的视觉特征用于微调文本提示，使模型自适应地关注图像中的重要区域和语言提示中的关键信息。

最后，针对Patch-scale的文本嵌入，利用像素级别的损失函数，在每一个Patch区域内计算损失。实现结合异常的细粒度特征，学习针对每一个Patch区域内的忽略物体的通用正常/异常提示。针对Image-scale的文本嵌入，利用全局级别的损失函数，使不关注物体类别的文本嵌入与图像的全局视觉嵌入进行对齐，学习针对图像全局的通用正常/异常提示。最后，引入Image-scale与Patch-scale的文本提示之间的对比损失，以解决可学习的图像级别的文本提示与Patch-scale局部区域内的文本提示之间出现不一致或者独立提示不足的问题。



我们做的主要贡献：

1. 设计了多粒度可学习文本提示框架MGPL：将图像划分为多个patch，设计不关注物体的包含局部空间信息的patch-scale的可学习文本提示，以及关注物体但与对象类别无关的图像级别的全局提示。实现将文本提示聚焦到图像中的局部Patch区域内，通过在Patch层级增加文本提示，同时再结合图像级别的全局文本提示，多粒度的文本提示驱动模型更好地从视觉编码器中获取局部异常的视觉上下文特征。使得模型既能够整体地检测图像中是否有异常，又能在局部的Patch级别的区域内检测一些不易被发现的复杂，细小的异常。

Integrating Spatial Information-Embedded Patch-Scale and Global-Scale Multi-grained Prompt Learning for zero-shot anomaly detection. （整合带有空间信息的补丁级与全局级多粒度提示学习）

1. 设计多粒度的视觉语言特征跨模态交互模块MGLV。具体为首先将带有位置编码的分别指代图像中不同局部区域内的patch-scale的文本嵌入特征进行融合。然后通过在视觉编码器中间层的视觉特征上应用不同尺度大小的卷积核来提取不同尺度大小的异常特征，进一步使得文本嵌入能够驱动捕获多粒度的图像特征。同时对视觉嵌入和文本嵌入应用多头的交叉注意力机制，获得更丰富的视觉特征用于微调文本提示，使模型自适应地关注图像中的重要区域和语言提示中的关键信息。
2. 设计多粒度损失函数框架：MGLF：Multi-Grained Loss Function Framework。由融合后的patch-scale的文本提示嵌入与视觉编码器中不同层得到的图像嵌入（embedding）进行计算得到的异常检测图分别与ground-truth在每一个Patch内计算损失，帮助针对图像中局部区域的patch-scale的文本提示学习建模更细粒度的异常提示。其次，将图像级别的全局提示与全局图像嵌入进行相似度得分计算，并与ground-truth计算损失。最后，将patch级别的损失与全局的损失结合到一起，实现同时结合图像级别的全局和局部patch级别的异常提示进行异常检测。为了避免全局和局部patch级别的文本提示出现不一致的情况，以及避免两种提示独立学习时提示不足的问题。针对融合多个patch级别的文本提示嵌入的特征，计算其与全局文本嵌入特征之间的损失，有效地将全局提示和patch-scale提示进行对齐。

章节安排：

# Related Work

## Industrial Zero-shot Anomaly Detection With CLIP

依托于预训练VLM视觉语言大模型的发展，针对新产品目标数据集，无需训练新模型即可实现零样本异常检测任务。特别是基于大规模图像文本对进行预训练的CLIP[21]模型，它的由语言提示驱动的零样本推理以及鲁棒性都有明显提高[22](its language-driven zero-shot inference, improved effective robustness, as well as showing a better perceptual alignment.(《Measuring robustness to natural distribution shifts in image classification》))。CLIP-AD[30]和ZOC[31]是最早使用CLIP进行零样本异常检测研究的，但是主要集中在异常分类任务上。另外，针对不同的目标数据集，ACR[31]需要在与目标域相关的数据集上进行调优，然后才能用于异常检测。WinCLIP[15]是第一个利用CLIP来实现语言引导的零样本异常检测，通过设计表示物体正常或异常状态的提示词与提示模板集合来引导CLIP模型对物体的正常或异常状态进行分类，另外将输入图像裁剪为不同尺度的patch，利用CLIP的视觉编码器，提取多尺度的图像特征与文本特征进行对齐来定位异常区域。AprilGAN[23]利用可学习的线性投影层将CLIP视觉编码器中间层提取到的patch-level的图像特征与文本特征进行对齐来实现异常的定位，解决了WinCLIP中由于在图像输入阶段输入大量裁剪后的图像窗口所导致的inefficiency的问题。AnomalyCLIP[16]中通过将手工设计的文本提示模板改为由可学习的文本向量[24]组成的提示模板，以增强文本提示面对不同测试数据时的表达能力，更好地泛化到不同的场景中。AnoVL[25]在视觉编码器中提取values，提出了无需训练的v-v注意力机制，来将中间层的patch-level的视觉特征与文本特征进行对齐。FAPrompt[17]通过设计一种由通用的可学习的正常提示与一组complementary, decomposed的异常提示标记复合而成的细粒度的提示模块，来学习自适应的细粒度的异常语义，以实现对不同异常类型的提示。然而，上述所有方法中的文本提示均是针对图像级别的全局描述，并没有针对图像中的局部位置区域进行文本提示。与现有方法相比，我们提出了一种多粒度提示框架，融合了图像级别的全局正常/异常提示与Patch-scale的局部正常/异常提示。将图像划分为多个patch，设计不关注物体的包含局部空间信息的patch-scale的可学习文本提示，以及关注物体但与对象类别无关的图像级别的全局提示。通过多粒度的文本提示，驱动模型更好地从视觉编码器中获取局部异常的视觉上下文特征。在局部的Patch-scale的区域内检测一些不易被发现的复杂，细小的异常。

## Language-Visual Modal Interaction

在视觉-语言多模态模型学习领域，跨模态交互对于信息融合，特征提取至关重要。早期研究中，Hu等[26]对拼接之后的语言与视觉模态特征应用卷积来实现多模态的信息融合。Chen等[27]设计了视觉文本联合嵌入的兼容性度量组件See-through-Text Embedding Pixelwise来实现视觉特征逐像素地与文本特征进行对齐。AprilGAN[23]利用可学习的线性投影层将CLIP视觉编码器中间层提取到的patch-level的图像特征与文本特征进行对齐来实现异常的定位。近年来，注意力机制已经成为跨模态交互中对齐语言描述与视觉编码器所提取到的视觉特征的有效方法。Xu[27]等提出了一种新的视觉语言交互模块bridger对视觉编码器提取到的中间层特征与语言特征应用注意力机制进行交互，实现跨模态的信息交换。然而，上述所有方法并没有考虑将视觉编码器中不同尺度大小的异常特征与相对应的Patch-scale级别的文本特征进行交互。多尺度卷积是检测图像中不同大小目标的常用的有效方法，MixConv[28]对不同通道组应用不同大小的卷积核来捕获不同分辨率的视觉特征。我们通过在中间层的视觉特征上应用不同尺度大小的卷积核来提取全局和局部异常特征，进一步使得文本嵌入能够驱动捕获多粒度的图像特征。此外，我们对视觉嵌入和文本嵌入应用多头的交叉注意力机制，获得更丰富的视觉特征用于微调文本提示，使模型自适应地关注图像中的重要区域和语言提示中的关键信息。

# MGCLIP：

## Preliminaries

## Problem formulation

Firstly, we define the problem setting for ZSAD within this framework. 给定一个包含正常和异常图像以及其真实掩码的训练数据集，包含了正常图像和异常图像，以及每张图像对应的真实掩码，表示正常图像，表示异常图像。表示每张图像对应的像素级真实掩码，表示正常像素，表示异常像素，，分别表示图像的宽度和高度。目标测试集为。目的是训练一个与目标无关的模型框架，在目标数据集中没有任何训练样本的情况下实现零样本异常检测。具体来说，对于测试集中的图像，图像级异常分数表示对图像是否异常的全局评估，值越大，异常的概率越高，表示像素级异常分数。

CLIP for Zero-Shot Anomaly Detection

本文使用在网络级规模的图像文本对上进行预训练的视觉语言模型CLIP。原始CLIP由文本和视觉编码器组成，通过将文本嵌入与视觉嵌入进行对齐，用以实现对图像中目标类的识别。文本提示模版以及物体类别名被输入到文本编码器中，得到文本嵌入。原始CLIP中的文本提示示例如：A photo of a [c]，得到文本嵌入。待测图像被输入到视觉编码器中，得到全局视觉嵌入。CLIP通过计算文本嵌入和视觉嵌入的相似性来预测图像属于类别的概率：

|  |  |
| --- | --- |
|  | (1) |

为温度超参数，operator表示余弦相似度计算。与原始用于物体对象类别识别不同，AnomalyCLIP用于异常检测时是与对象无关的，即认为物体只区分正常或者异常，输入正常文本提示与异常文本提示到文本编码器中，得到文本嵌入和，同样用式(1)计算物体为正常或异常的概率和。

|  |  |
| --- | --- |
|  | (2) |

另外，针对视觉编码器中间层所提取到的Patch视觉嵌入表示不同的视觉编码器中间层。表示第层中的对应于原图中第个Patch区域的视觉嵌入。我们可以获得针对每个Patch的正常与异常度得分与：

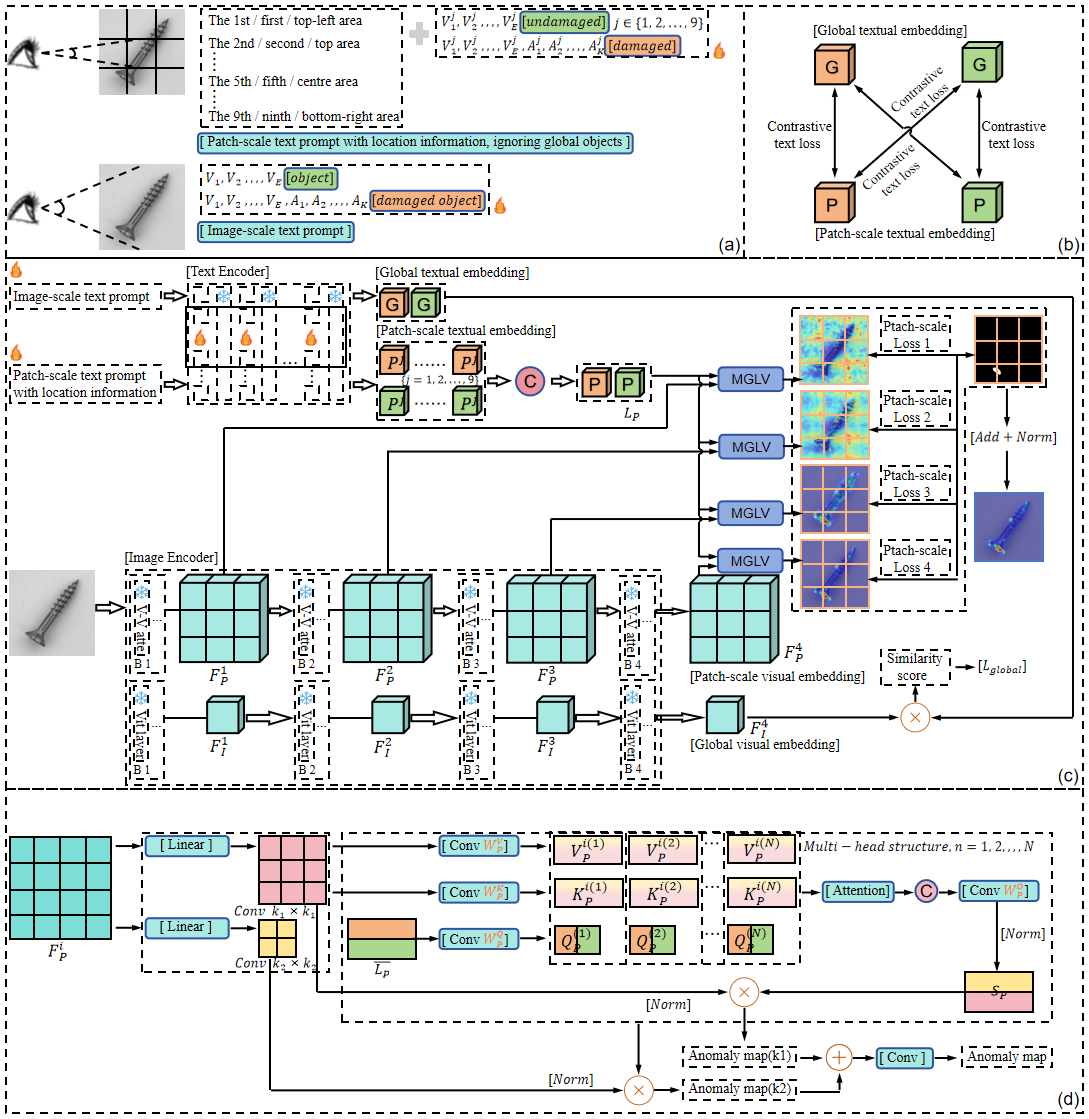
|  |  |
| --- | --- |
|  | (3) |

## Overview of Framework:

在本文中，我们提出了一个用于零样本异常检测的框架MGCLIP来自适应地学习针对图像级别的全局文本提示以及包含局部空间方位信息的Patch-scale的文本提示，建模多粒度的异常语义，通过多粒度的文本提示驱动模型更好地从视觉编码器中获取不同大小异常的上下文特征。其次，通过多粒度的视觉语言特征跨模态交互模块MGLV，实现将视觉编码器中间层获取到的不同尺度的视觉特征与文本特征能够进行良好的交互与融合，以帮助CLIP对不同尺度大小异常特征的理解，学习多粒度的异常语义。最后，引入Image-scale与Patch-scale的文本提示之间的对比损失，帮助不同粒度之间的文本提示实现互补与一致。

Integrating Spatial Information-Embedded Patch-Scale and Image-Scale Multi-grained Prompt Learning for zero-shot anomaly detection.

图2展示了MGCLIP的整体框架，它包含了两个创新性的模块，包括MGPL：结合Patch-scale和Image-scale的多粒度文本提示模块，以及MGLV：多粒度的语言-视觉特征跨模态交互模块。具体来说，MGPL模块，考虑首先将图像划分为多个patch，设计不关注物体的包含局部空间信息的patch-scale的可学习文本提示，其次设计关注物体但与对象类别无关的图像级别的全局提示。多粒度的文本提示驱动模型更好地从视觉编码器中获取局部异常的视觉上下文特征。MGLV模块考虑从多尺度方面更好地将文本嵌入和视觉嵌入进行更好的交互，首先通过在中间层的视觉特征上应用不同尺度大小的卷积核来提取不同尺度的异常特征，使得文本嵌入能够驱动捕获多粒度的图像特征。此外，对视觉嵌入和文本嵌入应用多头的交叉注意力机制，获得更丰富的视觉特征用于微调文本提示，使模型自适应地关注图像中的重要区域和语言提示中的关键信息。最后设计Image-scale与Patch-scale的文本提示之间的交叉对比损失，将两种不同粒度的提示进行对齐，来保证提示一致性，以及解决单一粒度文本提示不足的问题。



## Multi-Grained Prompt Learning

**Integrating Spatial Information-Embedded Patch-Scale and Image-Scale Multi-grained Prompt Learning for zero-shot anomaly detection.** Previous文本提示方法均是针对图像级别的全局描述，并没有针对图像中的局部位置区域进行文本提示，对一些尺寸较小且易出现在工业零件中不易被观察到的边缘区域缺陷的异常语义的捕获效果较差（这里可以做一个可视化的实验来说明）。与现有方法相比，我们提出了一种多粒度提示框架表示如图2a所示，融合了图像级别的全局正常/异常提示与Patch-scale的局部正常/异常提示。具体为：将图像划分为多个patch，设计不关注物体的包含局部空间信息的patch-scale的可学习文本提示，以及关注物体但与对象类别无关的图像级别的全局提示。通过多粒度的文本提示，驱动模型更好地从视觉编码器中获取局部异常的视觉上下文特征。在局部的Patch-scale的区域内检测一些不易被发现的复杂，细小的异常。

图像级别的全局正常提示表示为式(4)，与前述研究[16]类似，正常提示包含了一组可学习的文本提示标记。原始CLIP中的文本提示模板是为了侧重于与对象语义对齐，并不是物体中的正常与异常语义，通过引入可学习的文本提示模板，并在大量的工业异常图像数据集上进行训练，能够很好地学习到如何建模通用正常与异常语义的文本嵌入。其次，尽管不同图像中物体的类别是不同的，但异常信息的底层模式可能是相似的，比如轴承与金属螺母上的划痕，钉子与PCB板上的凹坑与裂缝。因此引入不关注物体类别的文本提示模板，帮助文本提示更准确地捕获到异常本身的特征，排除掉对象语义，以[object]替换原始CLIP中表示物体类别信息的[cls]。借助于先前的研究[17]，物体中的异常模式表现为在正常图片的基础上叠加异常特征，因此图像级别的全局异常提示如式(5)所示，同样表示可学习的正常提示标记。由于工业零件的缺陷类型复杂多样，如同一物体类别上的不同缺陷类型，如地毯上的色斑，切口，洞和裂缝。因此只以[damaged object]来表示物体的异常状态并不能够满足对于异常模式的细粒度识别。所以在正常提示标记的基础上添加可学习的异常提示标记来建模一种能够表示物体对象缺陷类型的文本嵌入，来学习更细粒度的异常提示。

|  |  |
| --- | --- |
|  | (4) |
|  | (5) |

包含局部空间信息的patch-scale的可学习文本提示表示如式(6)(7)，与图像级别的全局正常/异常提示不同，Patch-scale文本提示将图像划分为多个Patch，以划分为9个patch区域为例，针对图像中的局部区域进行正常/异常提示，更有效地驱动VLM模型针对图片中的每个Patch局部区域内对视觉编码器中提取到的局部异常特征进行捕获，来识别工业零件中出现的尺寸较小且易分布在零件边缘的缺陷。另外，原始CLIP是基于大量文本-图像对进行预训练的，文本数据集中包含了复杂多样的语言类型，因此这里分别采用了数字编码如1st/2nd，语言编码first/secod以及空间方位信息描述top-left/centre等几种不同的针对每个patch的表述方式，来验证CLIP对不同表述方式的敏感性。Patch-scale的正常提示如式(6)，表示针对每一个patch的可学习正常标记，，为超参数，表示将图像划分为不同个数的Patch，并设计与之匹配数量的Patch-scale级别的文本提示。其次，由于异常的视觉特征相对于图像中的前景物体的特征是有明显区别的，图像中缺陷区域的特征与周围像素区域的视觉特征有明显的割裂，因此即使物体类别不同或者甚至在不关注全局物体的情况下，也能判断出异常。所以在设计针对物体局部区域内的可学习文本提示时，不再设置[object]标记，只设置[undamaged]标记。参照前述的Image-scale的异常文本提示，Patch-scale的异常提示如式(7)所示，同样表示针对每个Patch的可学习正常提示标记，针对每一个Patch建模一种能够表示物体对象缺陷类型的文本嵌入，来学习更细粒度的异常提示。

|  |  |  |
| --- | --- | --- |
|  |  | (6) |
|  | (7) |

其次，参考文献[13]，原始CLIP的输入文本长度受到极大限制，并且训练数据集主要由简短文本组成，文本编码器采用了限制在77个tokens的绝对位置嵌入，对输入标记数进行了严格的限制。考虑到针对输入图像设计Patch-scale的文本提示，会超出原始CLIP文本编码器所能承载的可学习标记数量，在输入阶段，将针对每一个Patch设计的正常与异常文本提示依次输入到文本编码器中，然后再进行融合，得到式(8)与式(9)。

|  |  |
| --- | --- |
|  | (8) |
|  | (9) |

## Multi-grained Language-visual embedding interaction module

**Cross-modal and multi-grained feature interaction**如前所述，跨模态的特征交互与融合已被证明对于异常特征的捕获与理解有效。因此，为了更好地获取到不同尺度异常上下文特征，设计了多粒度的语言-视觉嵌入交互模块，针对视觉编码器中间层所提取的Patch视觉特征上应用不同大小的卷积核来提取全局和局部异常特征，进一步使得文本嵌入能够驱动获取多粒度的图像特征。在该路径中，首先将针对每一个Patch设计的正常与异常文本提示依次输入到文本编码器中，然后再进行融合，得到的文本特征向量表示正常和异常两类，图像输入到视觉编码器中得到的图像特征经过投影后得到，分别表示所对应的视觉编码器不同的中间层。然后我们使用两个核大小分别为和的卷积层来掌握不同尺度的视觉特征。

|  |  |
| --- | --- |
|  | (10) |
|  | (11) |

其中，，表示不同尺度的视觉特征。以为例。文本特征被投影到具有维度的联合空间中，。然后，我们对视觉嵌入和文本嵌入应用多头的交叉注意力机制，更新每个头内的文本嵌入。获得更丰富的视觉特征用于微调文本提示，使模型自适应地关注图像中的重要区域和语言提示中的关键信息。

|  |  |
| --- | --- |
|  | (12) |
|  | (13) |
|  | (14) |
|  | (15) |

，，分别为对应的投影矩阵。是头的个数，，，，分别表示沿嵌入维度进行操作后得到的每个头部内的特征。表示每个头部内通过视觉特征更新的注意力图，表示每个头部内更新后的文本特征。之后，将所有头内更新后的文本特征在嵌入维度上进行拼接，使用权重矩阵进行最后的映射操作，得到最终经过注意力机制更新后的文本嵌入。最后，针对卷积核尺度得到的正常得分图与异常得分图：

|  |  |
| --- | --- |
|  | (16) |

作为一个可学习的温度系数。同样地，针对卷积核尺度得到的正常与异常得分图：

|  |  |
| --- | --- |
|  | (17) |

最后，针对在不同尺度卷积核下得到的正常/异常得分图，分别表示不同尺度的异常特征。我们利用卷积和双线性插值层融合他们的结果，并还原到图像的原始尺度。

|  |  |
| --- | --- |
|  | (18) |

## Training and Inference

Multi-Grained Loss Function Framework

设计多粒度损失函数框架MGLF，训练损失分为三个部分：(1)图像级别的全局提示与全局图像特征进行相似度得分计算，并与ground-truth计算损失，定义为；(2)由融合后的patch-scale的文本提示嵌入与视觉编码器中不同层得到的图像特征（embedding）进行计算得到的异常检测图分别与ground-truth在每一个Patch内计算损失，定义为；(3)为了避免全局和局部patch级别的文本提示出现不一致的情况，以及避免两种提示独立学习时提示不足的问题。针对融合多个patch级别的文本提示嵌入的特征，计算其与全局文本嵌入特征之间的损失，有效地将全局提示和patch-scale提示进行对齐，定义为。总的训练损失计算为：

|  |  |
| --- | --- |
|  | (19) |

和表示为控制不同损失部分的权重系数。图像级别的全局损失借助二元交叉熵计算为：

|  |  |
| --- | --- |
|  | (20) |

|  |  |
| --- | --- |
|  | (21) |

与表示Image-level的全局正常与异常文本嵌入。表示Image-level的全局视觉嵌入。表示每张图像对应的真实掩码，表示正常图像，表示异常图像。表示由式2计算得到的对图像全局的异常度得分，值越大，图像被预测为异常的概率越高。

其次，针对patch-scale的文本提示嵌入与视觉编码器中不同层得到的图像特征（embedding）进行计算得到的异常检测图分别与ground-truth在每一个Patch内计算损失：

|  |  |
| --- | --- |
|  | (22) |

因为异常区域通常只占整个目标的小部分区域，因此我们采用Focal loss和来解决异常检测中正负样本数量极不平衡的问题，驱动模型更关注对目标区域的挖掘，同时为一些重要的样本分配更高的权重。另外结合Dice loss来减小预测的和实际的异常区域之间的差异。

损失计算如下：

|  |  |
| --- | --- |
|  | (23) |
|  | (24) |
|  | (25) |

是由式18得到的异常度得分，即表示由视觉编码器中第层的视觉嵌入与Patch-scale文本嵌入对齐所得到的相对应于图像中第个Patch区域内的第个像素点的异常度得分。表示每张图像对应的像素级真实掩码，表示正常像素，表示异常像素，，分别表示图像的宽度和高度。表示调制因子，用来聚焦难检测的样本。

最终，针对融合多个patch级别的文本提示嵌入的特征，计算其与Image-scale全局文本嵌入特征之间的损失，用以将全局提示和patch-scale提示进行对齐。在自然语言处理过程中，三元组损失Triplet Loss[29]应用于文本匹配领域时，可训练模型使得两种不同粒度的文本提示在做出针对目标图片的图像级别与Patch-scale的正常/异常提示保持一致性。通过建立三个关键文本提示样本：锚点(Anchor)、正样本(Positive)、负样本(Negative)。针对同一目标图像，通过学习相似度度量，在文本嵌入特征空间中将同时表示正常/异常的文本嵌入到彼此接近的位置，而表示不一致的文本嵌入到相距较远的位置，从而提高文本提示的准确性和鲁棒性，使得两种不同粒度的文本提示在互补的同时保持一致性。损失计算如下：

|  |  |
| --- | --- |
|  | (26) |

表示anchor，表示与anchor同一类别的文本提示，表示与anchor不同类别的文本提示。是一个大于0的常数，最终的优化目标是拉近与之间的距离，拉远与之间的距离。

Inference

在推理阶段，针对测试图像，Image-scale的异常分数由式2与式20得到。对于异常图，由视觉编码器中第层的视觉嵌入与Patch-scale文本嵌入对齐所得到的异常分数组合得到，然后进行插值和平滑操作，：

|  |  |
| --- | --- |
|  | (26) |

表示高斯滤波器，为平滑系数。

# Experiments

## Experiment Setup

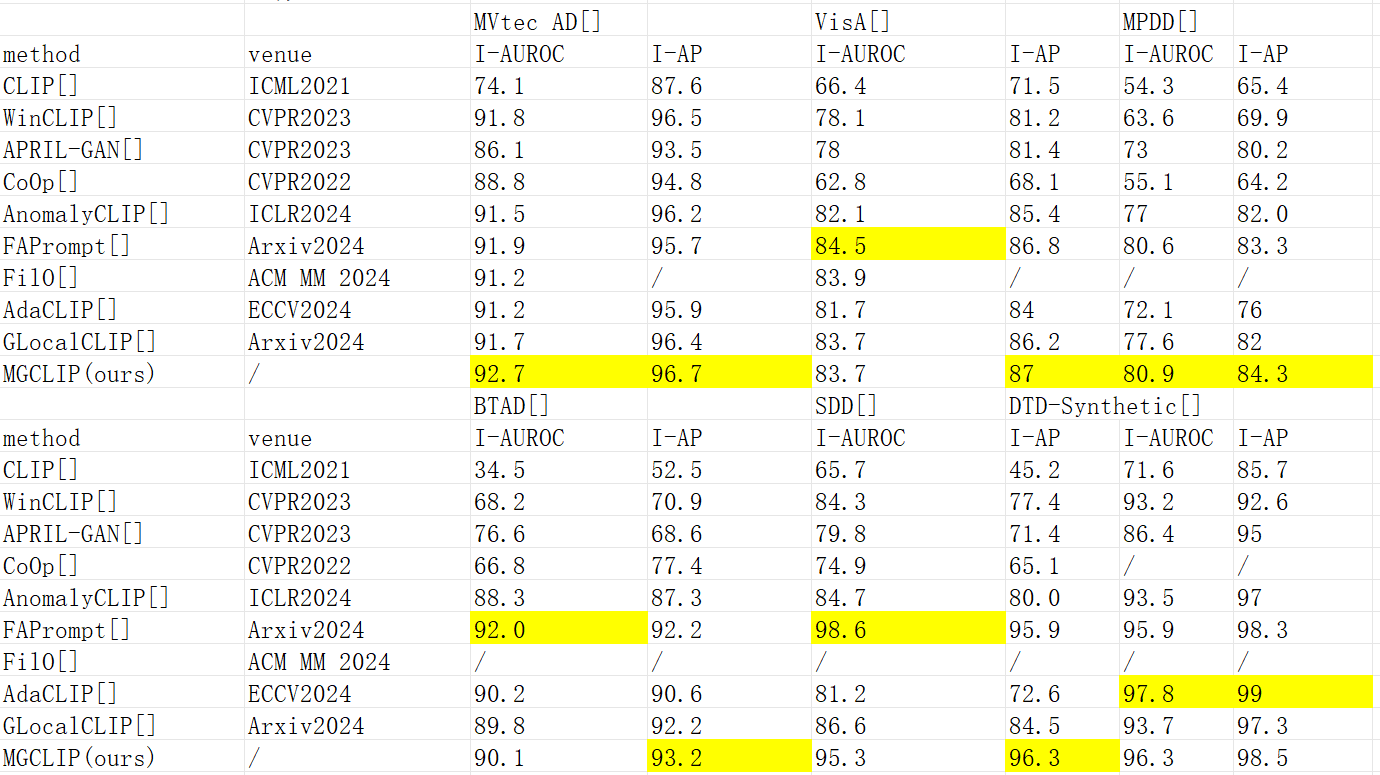
Datasets 为了评估MGCLIP模型在工业异常检测领域的有效性，我们在常用的6个工业缺陷检测数据集进行了实验，具体包括MVTec AD，VisA，MPDD，BTAD，SDD，DTD-Synthetic.以最广泛使用的工业异常检测数据集MVTec AD和VisA为例，MVTec包含如Cable，Transistor和Metal\_nut共15个不同物体类别的5354张正常和异常样本图像。VisA是一个新兴的工业异常检测数据集，包含如Candle，Capsules和Pcb共12个图像类别。为了评估模型的零样本遗产检测性能，与大多数研究[ ]一样，模型默认在MVTecAD数据集上进行训练，然后在不经任何进一步的训练或者微调下，对其他数据集上的测试集进行异常检测。至于MVTecAD数据集，我们通过在VisA数据集上进行训练，然后对MVTecAD数据集中的测试集进行测试。

Comparison Methods and Evaluation Metrics 为了评估MGCLIP模型在工业异常检测领域的先进性，我们与SOTA的模型，包含了三种基于手工设计的文本提示的方法：CLIP，WinCLIP和APRIL-GAN，六种可学习的文本提示的方法：CoOp，AnomalyCLIP，FAPrompt，FilO，GlocalCLIP和AdaCLIP。我们使用area under the receiver operating characteristic curve (AUROC)和average precision (AP)来评估Image-level的异常检测性能。对于Pixel-level的异常检测性能评估，we employ AUROC and Area under per region overlap (AUPRO)作为评价指标。

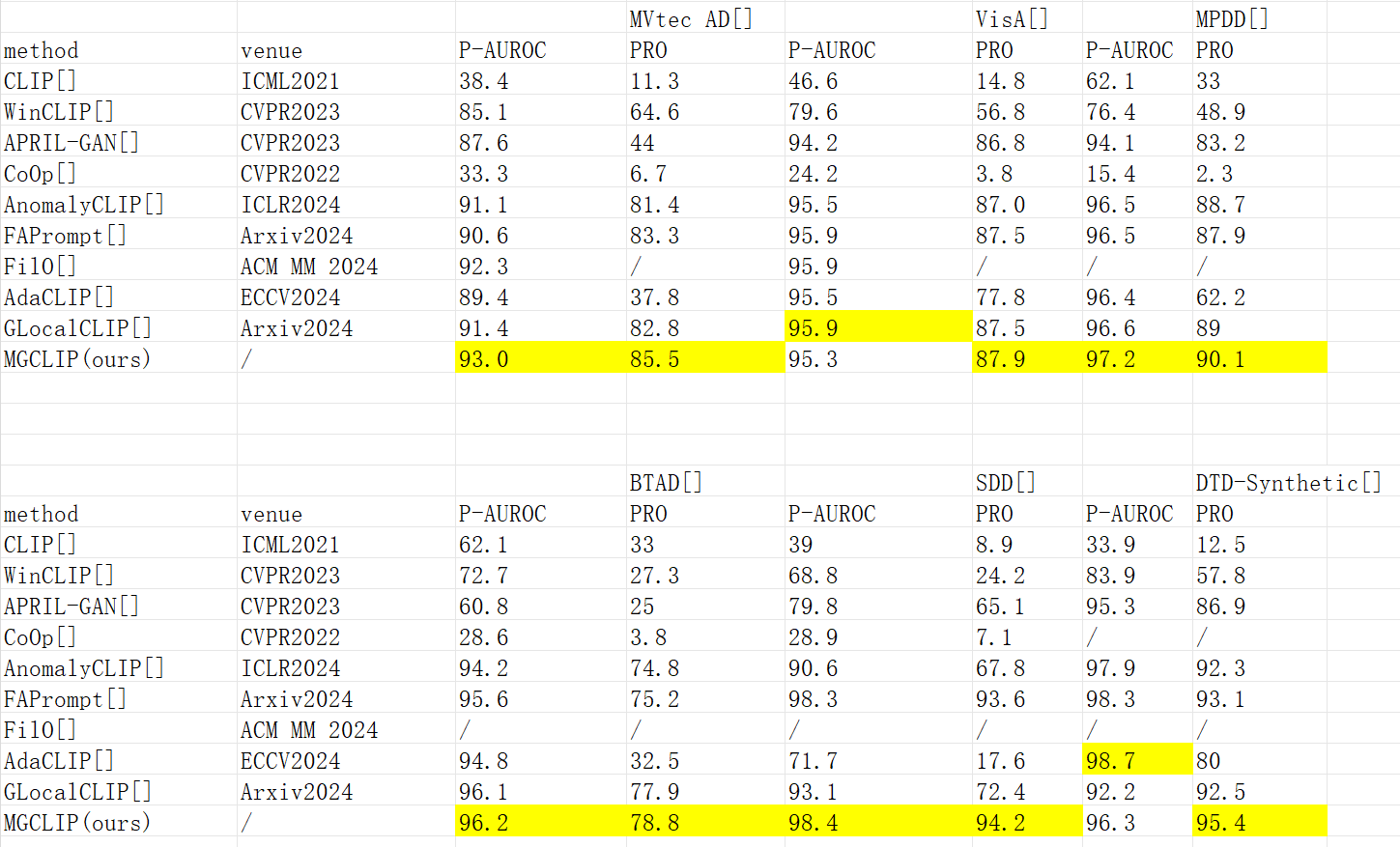
Implenmentation details遵循AnomalyClip中的设置，我们采用公开的CLIP模型VIT-L/@336px作为我们的backbone。CLIP的模型参数全部冻结。可学习的Image-level的正常文本提示的长度设置为7，异常文本提示由长度为7的代表正常的token与额外的4个代表异常的文本提示组合而成。可学习的Patch-level的正常文本提示的长度设置为10，异常文本提示的长度设置为11，共九组。对于文本编码器，可学习的token embeddings被附加在前9层，每层的可学习token的数量设置为4个，用于微调文本空间。对于视觉编码器，我们首先提取第2，6，12，20层输出的Patch tokens，然后应用V-V attention，最后将其与Patch-level的文本嵌入进行跨模态的特征交互与融合。所有实验均在Pytorch-2.0.0的框架下进行，硬件平台使用单个3090 24GB GPU。

## Main results

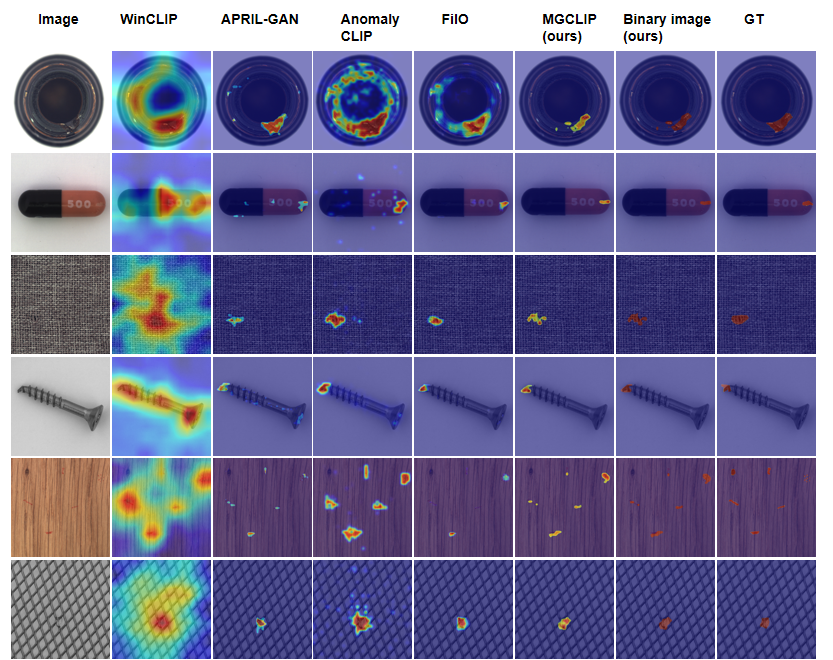
Image-level ZSAD Performance 表一展示了在六个工业缺陷检测数据集下，MGCLIP与其他9种SOTA方法的图像级的零样本异常检测的结果。结果表明，针对AUROC评价指标，MGCLIP在MVTecAD和MPDD数据集上的表现优于现有的SOTA方法，针对AP，其在MVTecAD，VisA，MPDD，BTAD和SDD数据集上取得了最高的分数。具体地，CLIP的检测效果较差归因于其最初是训练用于检测图像中前景物体的类别，而不是物体中复杂多样的缺陷类型，以及其过于简单的文本提示设计。WinCLIP通过设计数百个更加细致的状态级的表示异常/正常的文本提示，另外，通过将输入图像裁剪为不同大小的Patch，提取多尺度的视觉特征与文本特征进行对齐来定位异常区域。由此，WinCLIP获得了比原始CLIP较好的结果。AprilGAN利用可学习的线性投影层将 CLIP 视觉编码器中间层提取到的 patch-level 的图像特征与文本特征进行对齐来实现异常的定位，并且在文本提示中使用额外的可学习层来改进WinCLIP，也取得了不错的检测结果。然而，上述方法均依赖于手工定义的文本提示，在面对与预定义文本描述不太符合的异常时检测性能欠佳。CoOp采用了可学习的文本提示方法，但其是为检测图像中的物体类别进行训练与设计的，在捕捉同一对象的表示正常与异常的上下文特征方面能力较弱。AnomalyCLIP通过设计一组简单但统一且有效的与对象无关的分别表示正常与异常的可学习文本提示模版，在多个数据集上显著提高了检测效果，但其受限于粗粒度的文本提示设计，在面对同一物体的不同缺陷类型时，检测效果较差，如针对同一地毯，其上的缺陷类型可能存在色斑，切口，洞和裂缝等多种缺陷类型。针对这一问题，FAPrompt通过在可学习的表示正常的文本提示设计的基础上叠加可学习的异常提示tokens来建模能够表示物体对象缺陷类型的文本提示设计，来学习更细粒度的异常提示。FiLo与AdaCLIP同样通过建模自适应学习的细粒度异常描述来提高复杂多样的异常检测的准确率。GlocalCLIP在进行文本提示设计时，通过在文本提示模板前增加Local/Global单词组合前缀，来表示对整体和局部区域的异常进行提示，检测效果更进一步增强。但是上述研究中的文本提示方法均集中在针对整张图像进行图像级别的文本提示，并没有聚焦到图像的局部区域。在本文中，我们设计了多粒度的文本提示框架MGPL，通过设计多组针对图像中的每一个Patch的正常/异常提示，区别于GlocalCLIP中在图像级别的文本提示模板前增加”Local”前缀的文本提示方法，MGPL模块通过将图像划分为多个Patch，设计针对每个Patch的正常/异常文本提示，将文本提示聚焦到图像中的局部patch区域内，结合图像级别的全局文本提示，驱动模型更好地从视觉编码器中获取局部异常的视觉上下文特征。



Pixel-level ZSAD Performance 我们同样将MGCLIP与其他9种SOTA方法在6个工业缺陷检测数据集上比较了像素级的零样本异常检测结果，如表二所示。针对AUROC评价指标，MGCLIP在MVTecAD，MPDD，BTAD和SDD上表现出了优于其它所有方法的效果。针对PRO，MGCLIP在所有数据集上均取得了最高的分数。具体地，CLIP，WinCLIP，APRIL-GAN由于手工制作的文本提示，以及这三种方法中对于跨模态特征的交互不充分，容易导致图像对于某些文本提示的过拟合。通过设计针对图像中表示物体正常与异常状态的可学习的文本提示，结合Patch级别的视觉嵌入与文本嵌入之间的融合，AnomalyCLIP取得了较好的像素级的异常检测效果。更进一步地，FAPrompt通过设计更细粒度的异常提示，在VisA，BTAD和SDD数据集上取得了最高或者第二高的分数。Filo中通过设计自适应学习的文本描述，通过获取测试图像的视觉特征来辅助进行文本提示，以及设计了多尺度与多形状的特征提取模块，来提取视觉嵌入特征与文本特征进行跨模态的交互，同样获得了较好的异常检测结果。AdaCLIP同样通过建模自适应学习的细粒度异常描述来提高复杂多样的异常检测的准确率。GlocalCLIP在进行文本提示设计时，通过在文本提示模板前增加Local/Global单词组合前缀，来表示对整体和局部区域的异常进行提示，在BTAD，VisA数据集上获得了更好的像素级异常检测效果。然而，上述所有方法并没有考虑将视觉编码器中不同尺度大小的异常特征与相对应的 Patch-scale 级别的文本特征进行交互，以帮助 CLIP 对局部区域异常语义的理解。因此，我们设计了多粒度的语言-视觉嵌入交互模块（MGLV，Multi-grained language-visual embedding interaction module），在中间层的视觉特征上应用不同尺度大小的卷积核来提取全局和局部异常特征，进一步使得文本嵌入能够驱动捕获多粒度的图像特征。此外，我们对视觉嵌入和文本嵌入应用多头的交叉注意力机制，获得更丰富的视觉特征用于微调文本提示，使模型自适应地关注图像中的重要区域和语言提示中的关键信息。在所有数据集上均获得了最高或者第二高的检测分数。



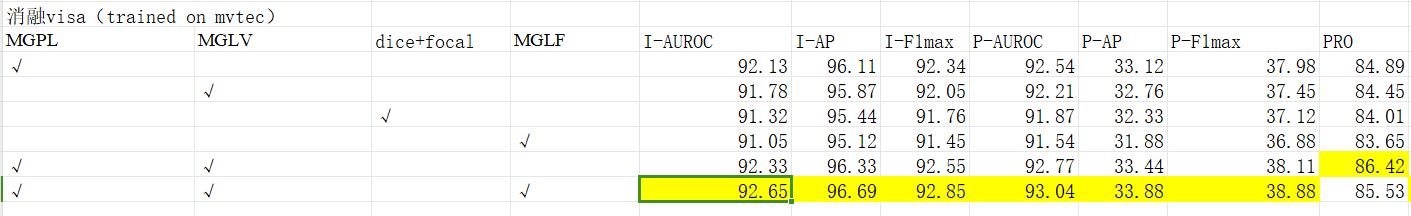
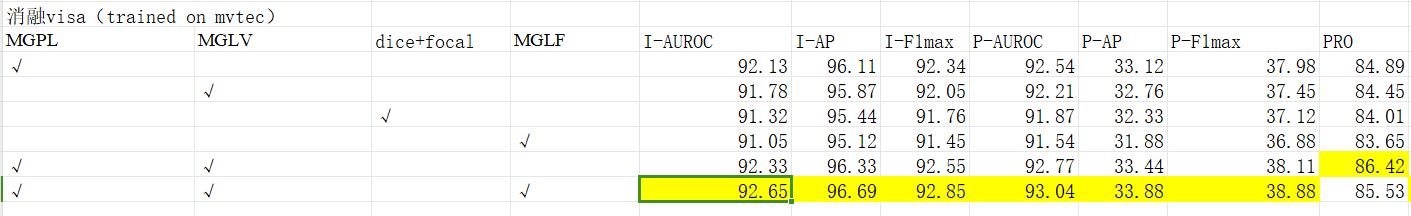
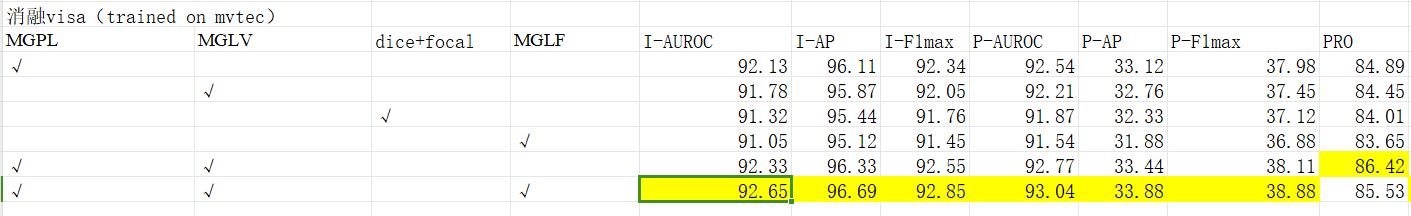
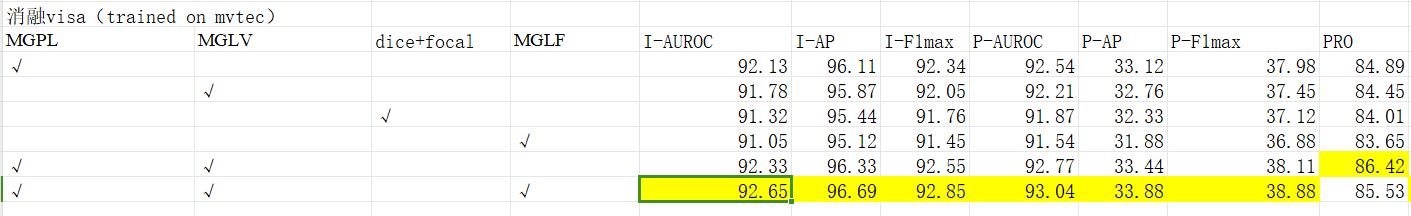
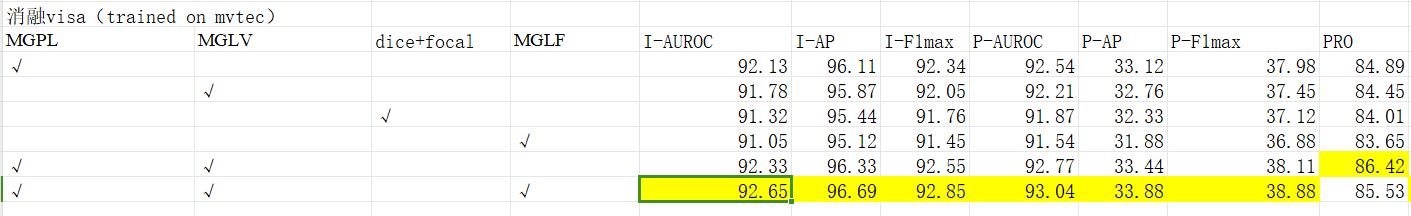
Comparison of anomaly localization maps 为了更直观地比较MGCLIP与其它SOTA方法在工业领域中的零样本异常检测效果，我们在MVTecAD数据集中的bottle，capsule，carpet，screw，wood，grid子集上可视化了不同方法的异常检测结果，如图3所示。工业产品类型复杂多样。缺陷分布在图像中不同区域内，尺度大小不一。可视化的异常区域定位结果表明多粒度的可学习文本提示框架MGPL与跨模态的视觉语言特征交互模块MGLV在检测全局异常与局部异常方面表现出了优异的性能。其它基于手工设计文本提示方法，基于图像级的可学习文本提示以及未充分进行跨模态的特征信息融合的异常定位方法难以定位到异常区域或者检测结果不完整。具体地，使用手工设计正常/异常状态词文本提示的方法WinCLIP在定位grid产品中的异常缺陷时能够粗略定位到异常区域，难以精确定位。同时其在多数产品中如bottle，capsule，carpet，screw，wood中无法区分正常与异常区域。APRIL-GAN在WinCLIP的基础上，通过添加额外的线性层将图像特征映射到联合嵌入空间与文本特征进行对齐来生成异常图，已经能够粗略地定位到不同类型产品中复杂的缺陷类型。但对于异常区域的定位精度及局部区域内细粒度的异常定位仍不够精确。Anomaly通过学习与对象无关的文本提示，专注于使用全局和局部的上下文优化来学习通用正常/异常语义提示。能够粗略定位异常所在，但Anomaly忽略了细粒度的异常建模，在区分不同的缺陷类型时表现欠佳。FiLO通过构建自适应学习的细粒度异常描述，以及位置增强的定位模块，针对目标图片中异常区域的定位与分割精度有较大的提高，但其仍局限于学习针对Image-level的异常提示，对于一些尺寸较小，分布区域杂乱的缺陷的检测效果较差。相较于上述方法，MGCLIP通过构建两个创新性的模块：MGPL：结合Patch-scale 和 Image-scale 的多粒度文本提示模块，以及 MGLV：多粒度的语言-视觉特征跨模态交互模块，在识别全局级与存在局部区域内的多尺度，不同类型的缺陷方面均表现出最优的结果。



## Ablation Study

Module Ablation. 为了评估不同模块对模型在异常检测方面的有效性，我们在MVtecAD 数据集上进行了实验。AnomalyCLIP用作我们的基础模型，表3中的结果显示了单独使用MGPL，MGLV，同时使用MGPL和MGLV，以及所有模块全部使用时模型的检测性能。具体地，通过设计多粒度提示框架MGPL，实现将文本提示聚焦到图像中的局部 Patch 区域内，通过在 Patch 层级增加文本提示，同时再结合图像级别的全局文本提示，多粒度的文本提示驱动模型更好地从视觉编码器中获取局部异常的视觉上下文特征。使得模型既能够整体地检测图像中是否有异常，又能在局部的 Patch 级别的区域内检测一些不易被发现的复杂，细小的异常。相比于其它两个模块，单独使用MGPL，模型在图像级和像素级异常检测性能提升都是最明显的。其次，通过设计多粒度的视觉语言特征跨模态交互模块MGLV，通过在视觉编码器中间层的视觉特征上应用不同尺度大小的卷积核来提取不同尺度大小的异常特征，进一步使得文本嵌入能够驱动捕获多粒度的图像特征。同时对视觉嵌入和文本嵌入应用多头的交叉注意力机制，获得更丰富的视觉特征用于微调文本提示，使模型自适应地关注图像中的重要区域和语言提示中的关键信息。单独使用MGLV时，在图像级的AP指标下，模型的准确率略有下降，考虑原因是由于MGLV单独使用时，缺乏Patch级别的文本提示嵌入。依托于充分的视觉语言特征的跨模态的交互与融合，在图像级AUROC，像素级的AUROC和PRO指标下模型的检测准确率有所提升。同时使用MGLV和MGPL模块时，模型在图像级和像素级异常检测的性能明显提升，归因于多粒度可学习文本提示框架与多粒度的视觉语言特征跨模态交互模块充分获取到了图片中的全局与局部异常的上下文特征。Furthermore, 通过设计多粒度的损失函数框架，针对融合多个Patch-scale的文本提示嵌入特征，计算其与全局文本嵌入特征之间的损失，有效地将Image-scale和Patch-scale的提示进行对齐，模型在图像级和像素级的异常检测效果达到最优。

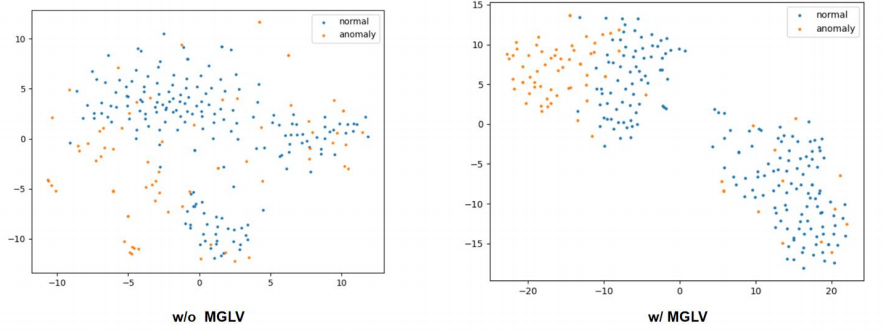
AnomalyCLIP: (I-AUROC, I-AP): (91.5, 96.2); (P-AUROC, P-PRO): (91.1, 81.4)



Hyperparameter Sentivity Analysis. 我们基于MVTecAD数据集分析了MGCLIP中的关键超参数对图像级和像素级的异常检测效果的影响，即多粒度可学习文本提示框架MGPL中的Patch-scale的文本提示的数量。具体地，J等于4,9,16分别表示将图像划分为对应数量的Patch，并设计对应于每个patch的文本提示。模型的异常检测准确率随着所划分的Patch的数量J的增加而上升，当J=9时，模型在图像级和像素级的异常检测效果达到最优，当J=16时，模型性能有所下降，考虑原因为划分过多的Patch时，局部提示过多时，引入不必要的噪音和冗余，导致模型的注意力过于分散，当集中在过小的局部Patch区域时，由于缺少全局特征，容易将物体的一些正常物理结构识别为异常，导致准确率下降。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | I-AUROC | I-AP | P-AUROC | PRO |
| 1（AnomalyCLIP） | 91.5 | 96.2 | 91.1 | 81.4 |
| 4 | 92.2 | 96.5 | 92.6 | 85.3 |
| 9 | 92.7 | 96.7 | 93 | 85.5 |
| 16 | 91.6 | 96 | 91.3 | 82 |

Performance of Multi-grained Language-visual Embedding Interaction Module 为了更好地评估通过多头交叉注意力机制实现跨模态的文本与视觉特征交互，进而通过视觉特征微调文本提示模块的效果。我们分别实现了使用与不使用MGLV模块进行特征交互，然后将交互后的特征与正常/异常提示进行对齐，最后将每个样本的正常与异常得分进行t-SNE可视。结果如图4所示，具体地，在未使用注意力机制进行文本与视觉特征交互时（w/o MGLV），正常与异常特征分类表现出明显的耦合，正常和异常样本之间的区别仍然较为模糊，这种情况下特征信息的融合使得异常的分类与定位变得复杂，限制了模型准确识别异常样本的能力。相反，由交叉注意力引导的文本与视觉特征的交互则表现出了明显的提升，分别表示正常与异常的融合后的特征表现出了明显的边界。这种结果表示注意力机制有效分离了正常与异常特征表示，从而创建了一个更加结构化且分离良好的特征空间。这种清晰化的正常/异常区分，不仅增强了异常检测，而且促进了训练过程中损失函数的收敛。结果证明了通过注意力机制引导的视觉文本跨模态特征交互与融合实现了正常/异常特征分离，有助于识别与定位不同尺度大小的复杂缺陷类型，最终实现零样本跨场景的更加准确的异常检测。



Why adopt the multi - grained prompt learning mechanism to enhance anomaly detection? 基于多粒度提示文本结构背后的核心思想在于针对图像中分布区域杂乱且尺度大小不一的异常，提供整合带有空间信息的补丁级与全局级的多粒度提示学习（Integrating Spatial Information-Embedded Patch-Scale and Global-Scale Multi-grained Prompt Learning for zero-shot anomaly detection.）。传统方法，例如 AnomalyCLIP，FAPrompt和 AdaCLIP，未能考虑局部的尺寸较小的异常对检测结果的影响。这些方法缺乏包含缺陷尺寸及分布区域信息的特征。同样，虽然 WinCLIP 强调了Multi-scale的patch，但其主要是指输入块大小的变化，而并没有针对每个patch提供带有空间信息的patch-level的可学习文本提示。

另一方面，基于多粒度的文本提示结构提供了一种简单而有效的方法来区分不同尺度的异常。其主要目标是探索分布在图形不同区域的异常以及异常的尺寸如何影响异常检测，这是以前的研究中很大程度上被忽视的一个方面。值得注意的是，许多研究报告称，细粒度的文本提示可以提高异常检测性能。然而，这些研究往往无法准确地指出细粒度细节的哪些方面起着至关重要的作用，因此很难通过实验定量评估它们的贡献。这种基于多粒度的方法从根本上为异常分布区域及异常尺寸提供了细粒度的指导。它引入了一个结合Patch-level与Image-level的由异常分布区域及异常尺度信息构建的新型异常检测框架。通过在图像的每一个patch区域内施加额外的关注，该结构填补了先前研究的一个关键空白，能够更准确地检测和定位分布在图中各种局部区域及不同尺寸的异常。

如图5所示，我们使用了MVTec AD数据集中针对Bottle和Carpet类别的训练数据。图5 a, c为未加入MGPL的训练过程（w/o MGPL），即训练时未引入Patch-scale的文本提示，图5 b, d为加入MGPL后的结果。最佳阈值由F1得分最大化的点决定，可以更好地衡量类别不平衡。加入MGPL后，最佳阈值对应的precision和recall值都有所提升，如图所示。说明在添加MGPL模块的策略下，模型能够更好地区分正常和异常样本。

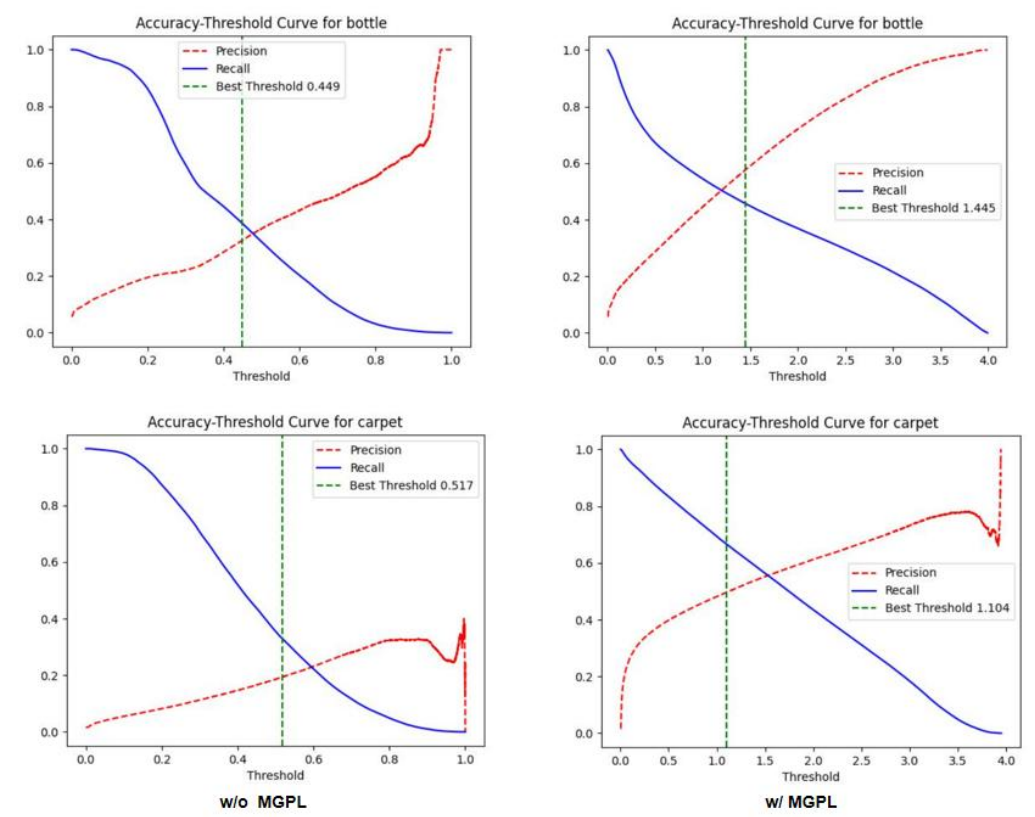
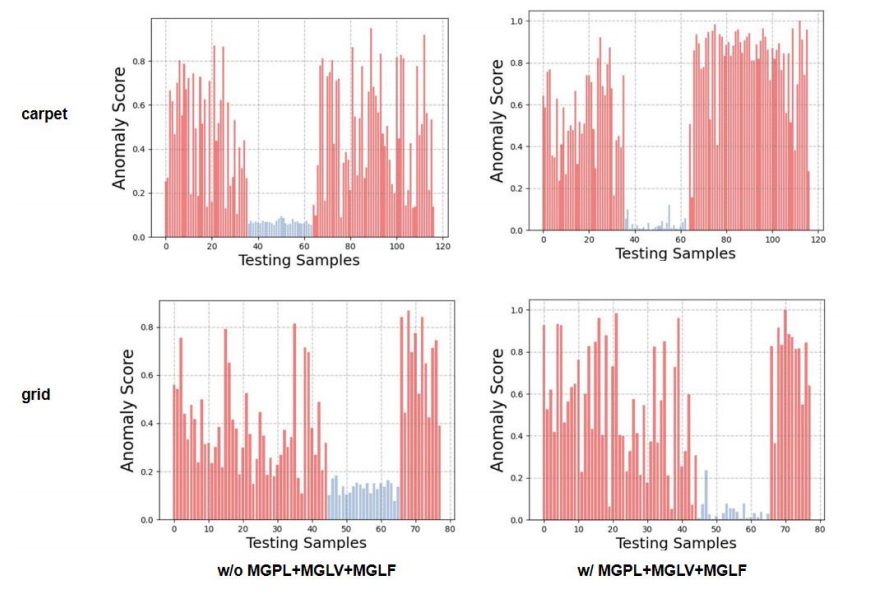


图5:

Validating the Effectiveness of the Architecture Through Anomaly Scores 为了更好地评估整体架构的有效性，强调所提出的多粒度的文本提示框架，视觉语言特征跨模态交互模块以及多粒度的损失函数框架协同作用下持续改善了异常分数的分布。尽管单个模块本身也有助于这种改进，重点在于证明所提出的架构作为一个整体有效地增强了异常分数分布中正常样本和异常样本之间的分离界限。在我们的设计中，MGPL 通过Integrating Spatial Information-

Embedded Patch-Scale and Global-Scale Multi-grained Prompt Learning for zero-shot anomaly detection. （整合带有空间信息的补丁级与全局级多粒度提示学习）从而实现了对分布在图像中不同区域内多尺度的异常的上下文视觉特征的增强。在此基础上，MGLV 进一步丰富了文本提示和图像特征的融合，以加强多模态表示。此外，MGLF 设计了一种针对 Patch-Scale and Global-Scale多粒度场景量身定制的新型损失机制，提供了更强大的指导。这种新机制超越了基线损失，专门解决了多粒度提示中由于引入Patch-scale的Multi-Patch的文本提示的挑战。上述这些模块共同增强了多粒度可学习文本提示框架的有效性。然而，在模块消融分析中，很明显没有一个模块可以完全实现引入了Patch-scale的多粒度提示的优势。如图6所示，我们比较了 MVTec AD 数据集中地毯和网格类别的异常分数分布。图6 a,c表示不添加任何模块下基线模型的异常分数，而b,d表示添加了所有模块下应用完整架构时的分数。MGCLIP 框架展示了其显著改善异常分数分布的能力，从而拉大了正常和异常样本之间的差距。这表明该模型整体上实现了优异的识别性能。此外，关于异常检测，由图中可知，在不添加任何模块的情况下，有缺陷图像的异常分数相对较低，导致正常和异常样本之间的混淆。当所有模块都集成在一起时，有缺陷图像的异常分数会显著增加，从而减少错误分类。虽然异常分数分布图不能直接反映像素级异常检测性能，但它们清楚地表明该模型可以在图像级别区分正常和异常样本。这凸显了所提出的架构在改进语义级异常检测方面的有效性。



# Conclusion

在本文中，我们提出了一种用于工业场景中的ZSAD的新型框架：MGCLIP，通过在包含了方位信息以及位置编码的Patch层级内对异常检测进行文本提示，提示CLIP模型从视觉编码器中更精准地获取图像中每个patch内的异常特征。具体地，多粒度提示模块MGPL，驱动模型更好地从视觉编码器中获取局部异常的视觉上下文特征，使得模型既能够整体地检测图像中是否有异常，又能在局部的Patch级别的区域内检测一些不易被发现的复杂，细小的异常。此外，多粒度的语言-视觉特征嵌入交互模块MGLV，通过在中间层的视觉特征上应用不同尺度大小的卷积核来提取全局和局部异常特征，进一步使得文本嵌入能够驱动捕获多粒度的图像特征。对视觉嵌入和文本嵌入应用多头的交叉注意力机制，获得更丰富的视觉特征用于微调文本提示，使模型自适应地关注图像中的重要区域和语言提示中的关键信息。最后，多粒度损失函数MGLF，通过将patch级别的损失与全局的损失结合到一起，实现同时结合图像级别的全局和局部 patch 级别的异常提示进行异常检测。为了避免全局和局部patch级别的文本提示出现不一致的情况，以及避免两种提示独立学习时提示不足的问题。针对融合多个 patch 级别的文本提示嵌入的特征，计算其与全局文本嵌入特征之间的损失，有效地将全局提示和 patch-scale 提示进行对齐。在常用的6个工业缺陷检测数据集进行了实验，与SOTA的模型，包含了基于手工设计和可学习的文本提示的方法进行对比，均取得了优秀的ZSAD performance.

逻辑异常检测  
GAOC模块：Graph Structure Guided Attention between Objects and Channels