arXiv:2303.00601v2 [cs.CV] 7 Sep 2023

通过混合融合进行多模态工业异常检测

Yue Wang[1*], Jinlong Peng[2*], JiangningZhang[2],RanYi[1†],YabiaoWang[2],Chengjie Wang[2,1]上海交通大学，中国上海； [2]优图实验室，腾讯[1]{imwangyue,ranyi}@sjtu.edu.cn [2]{jeromepeng,vtzhang,caseywang,jasoncjwang}@tencent.com

摘要

基于二维的工业异常检测已得到广泛讨论，然而，基于三维点云和*RGB*图像的多模态工业异常检测仍有许多未被探索的领域。现有的多模态工业异常检测方法直接拼接多模态特征，这会导致特征之间产生强烈干扰，并损害检测性能。本文提出***Multi-3D-Memory (M3DM)***，一种采用混合融合方案的新型多模态异常检测方法：首先，我们设计了一种结合逐块对比学习的无监督特征融合方法，以促进不同模态特征之间的交互；其次，我们采用结合多个记忆库的决策层融合，以避免信息损失，并使用额外的新颖性分类器做出最终决策。我们进一步提出点特征对齐操作，以更好地对齐点云特征和*RGB*特征。大量实验表明，我们的多模态工业异常检测模型在*MVTec-3D AD*数据集上的检测和分割精度均优于最先进水平（*SOTA*）方法。代码见*github.com/nomewang/M3DM*。

1. 引言

工业异常检测旨在找出产品的异常区域，在工业质量检测中发挥着重要作用。在工业场景中，获取大量正常样本很容易，但缺陷样本十分稀少。目前的工业异常检测方法大多是无监督方法，即仅使用正常样本进行训练，并且只在推理阶段使用检测样本进行测试。此外，现有的大多数工业异常检测方法 [2, 10, 26, 35] 都基于二维图像。然而，在工业产品的质量检测中，人工检验员会利用三维形状和颜色特征来判断产品是否存在缺陷，

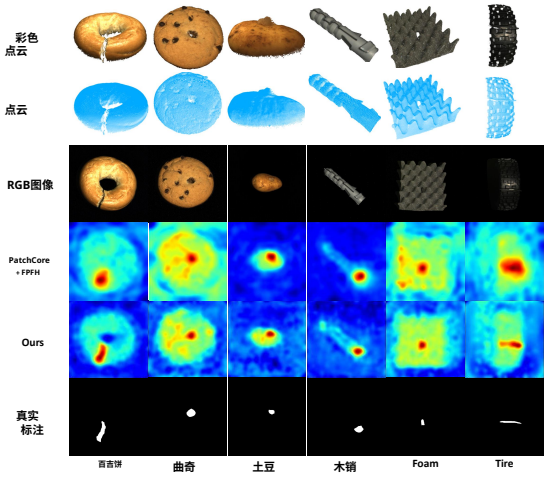


图 1. MVTec-3D AD数据集 [3]示意图。第二行和第三行是输入点云数据和 RGB 数据。第四行和第五行是预测结果。根据真实标注，与此前的方法相比，我们的预测结果更加准确。

其中，三维形状信息对于正确检测至关重要。如图1所示，对于曲奇和土豆，仅凭 RGB 图像很难识别缺陷。随着三维传感器的发展，近期发布了 MVTec-3D AD 数据集 [3] （图1），该数据集包含二维图像和三维点云数据，推动了多模态工业异常检测研究的发展。

无监督异常检测的核心思想是找出正常表征与异常之间的差异。目前，二维工业异常检测方法可分为两类：（1）基于重建的方法。图像重建任务广泛用于异常检测方法中 [2, 10, 15,23, 35, 36] ，以学习正常表征。对于单模态输入（二维图像或三维点云），基于重建的方法易于实现。但对于多模态输入，很难找到重建目标。（2）基于预训练特征提取器的方法。利用特征提取器的一种直观方式是将提取的特征映射到正态分布，并将分布外的样本判定为异常。基于归一化流的方法 [16, 28, 33]

*贡献均等。本工作完成于 Yue Wang 在腾讯优图实验室实习期间。

†通讯作者。

使用可逆变换直接构建正态分布，而基于记忆库的方法 [9, 26]则存储一些代表性特征，以隐式地构建特征分布。与基于重建的方法相比，直接使用预训练特征提取器无需设计多模态重建目标，因此是多模态任务的更好选择。此外，目前的多模态工业异常检测方法 [17, 28] 会将两种模态的特征直接拼接在一起。然而，当特征维度较高时，多模态特征之间的干扰会非常剧烈，导致性能下降。

为解决上述问题，我们提出了一种基于RGB图像和三维点云的新型多模态异常检测方案，称为 *Multi-3D-Memory*（*M3DM*）。不同于现有方法直接拼接两种模态的特征，我们提出了一种混合融合方案，以减少多模态特征之间的干扰，并促进特征交互。我们提出无监督特征融合（*UFF*）来融合多模态特征，并通过图像块级对比损失进行训练，以学习相同位置处多模态特征图像块之间的内在关系。为使异常检测模型保留单域推理能力，我们分别为RGB、三维和融合特征构建了三个记忆库。对于最终决策，我们构建了决策层融合（*DLF*），综合利用所有记忆库进行异常检测和分割。

异常检测需要同时包含全局信息和局部信息的特征，其中局部信息有助于检测细小缺陷，而全局信息则关注各个部分之间的关系。基于这一观察，我们使用Point Transformer [21, 37] 处理3D特征，并使用Vision Transformer [6, 12] 处理RGB 特征。我们进一步提出一种点特征对齐（*PFA*）操作，以更好地对齐三维特征和二维特征。

我们的贡献总结如下：

- 我们提出了 M3DM，这是一种采用混合特征融合的新型多模态工业异常检测方法，在 MVTec-3DAD 上的检测和分割精度均优于当前最先进的方法。
- 我们提出了采用逐块对比损失的无监督特征融合（UFF），以促进多模态特征之间的交互。
- 我们设计了决策层融合（DLF），利用多个记忆库实现稳健的决策。
- 我们探索了 Point Transformer 在多模态异常检测中的可行性，并提出点特征对齐（PFA）操作，将 Point Transformer 特征对齐到二维平面，以实现高性能三维异常检测。

2. 相关工作

对于大多数异常检测方法而言，核心思路是寻找正常数据的良好表征。传统异常检测已发展出多种不同途径。基于概率的方法利用正常样本的经验累积分布函数 [8, 18] 进行决策。表示空间邻居的位置也可用于检测，还可以采用多种聚类方法，例如 k近邻 [1, 25], 、相关积分 [22] 和直方图 [14]。离群点集成使用一系列决策模型来检测异常数据，其中最著名的方法是孤立森林 [19]。线性模型也可用于异常检测，例如简单地利用主成分分析 [30] 或单类支持向量机（OCSVM） [29]的特性。与深度学习相比，传统机器学习方法对训练数据的需求较少，因此我们利用这一优势，设计了一个基于 OCSVM 和随机梯度下降的决策层融合模块。

二维工业异常检测 工业异常检测通常是在无监督设置下进行的。MVTec AD 数据集广泛用于工业异常检测 [2]研究，其训练数据集仅包含正常样本，而测试数据集则同时包含正常和异常样本。工业异常检测需要提取图像特征以进行判断，这些特征可以隐式或显式地使用。隐式特征方法使用某种图像重建模型，例如自动编码器 [2, 15, 36] 和生成对抗网络 [23]；重建方法无法还原异常区域，通过比较生成图像和原始图像，可以定位异常并作出判断。一些数据增强方法 [35]被提出用于提升异常检测性能，研究人员会手动在正常样本中添加伪异常，训练目标则是定位伪异常。显式特征方法依赖预训练特征提取器，并由额外的检测模块利用已学习的特征或表征来定位异常区域。知识蒸馏方法 [10] 旨在训练学生网络重建图像或提取特征，教师网络与学生网络之间的差异可以表征异常。归一化流 [16, 33] 利用可逆变换将图像特征转换为正态分布，而异常特征会落在分布边缘。事实上，上述所有方法都试图将特征信息存储在深度网络的参数中；近期研究表明，仅使用记忆库 [26] 就能在异常检测中实现完全召回。二维与三维异常检测有许多相似之处，因此我们将记忆库方法扩展到三维和多模态设置，并取得了令人瞩目的结果。

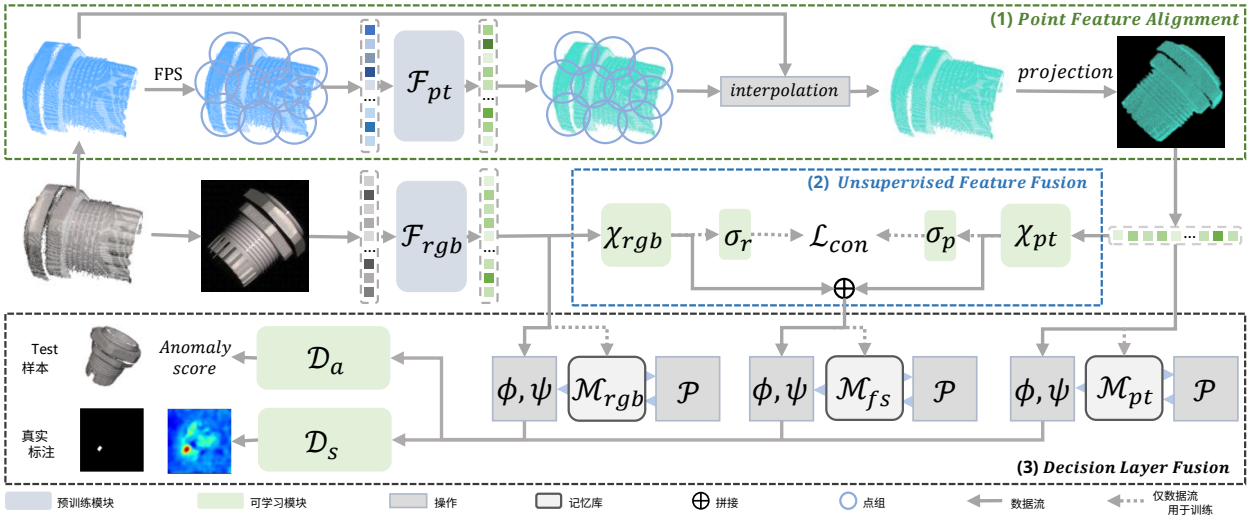


图 2. **Multi-3D-Memory（M3DM）的流程**。我们的 M3DM 包含三个重要部分：(1) 点特征对齐（PFA）通过插值和投影操作将点组特征转换为平面特征，FPS 表示最远点采样，$\mathcal{F}_{pt}$ 是一个预训练的 Point Transformer；(2) 无监督特征融合（UFF）通过图像块级对比损失将点特征和图像特征融合在一起 $\mathcal{L}_{con}$，其中 $\mathcal{F}_{rgb}$ 是一个 Vision Transformer，χ_{rgb}，χ_{pt} 是多层感知机层，σ_r，σ_p 是单个全连接层；(3) 决策层融合（DLF）利用多个记忆库融合多模态信息，并通过 2 个可学习模块作出最终决策 $\mathcal{D}_a$，$\mathcal{D}_s$ 用于异常检测和分割，其中 $\mathcal{M}_{rgb}$，$\mathcal{M}_{fs}$，$\mathcal{M}_{pt}$ 是记忆库，ϕ，ψ 是用于单个记忆库检测和分割的评分函数，而 $\mathcal{P}$ 是记忆库构建算法。

三维工业异常检测 首个公开的3D工业异常检测数据集是 MVTec-3D [3]AD 数据集，其中包含同一实例的 RGB信息和点位置信息。受医学异常检测中体素自编码器和生成对抗网络的启发 [3]，这些方法最初被用于三维工业异常检测，但它们丢失了大量空间结构信息，效果不佳。随后，有人提出了3D学生-教师网络 [4]，通过额外的预训练数据，专注于局部点云几何描述符。记忆库方法 [17] 也被用于三维异常检测，结合了几何点特征和简单的特征拼接。知识蒸馏方法 [28] 进一步利用深度信息，提升了纯 RGB 和多模态异常检测的结果。我们的方法基于记忆库；与以往方法不同，我们提出了一种新颖的流程，利用预训练的 Point Transformer 和混合特征融合方案，以实现更精准的检测。

3. 方法

3.1. 概述

我们的 Multi-3D-Memory（M3DM）方法以三维点云和 RGB图像作为输入，执行三维异常检测与分割。我们提出了一种混合融合方案，以促进跨域信息交互，同时保留每个单独域的原始信息。我们分别使用两个预训练特征提取器：用于 RGB 的 DINO [6] 和用于点云的 PointMAE [21]，以提取颜色和三维表征。

如图2所示，M3DM由三个重要部分组成：（1）点特征对齐（PFA，见第3.2节）：为解决颜色特征与三维特征之间的位置信息不匹配问题，我们提出点特征对齐，将三维特征对齐到2D空间，从而简化多模态交互并提升检测性能。（2）无监督特征融合（UFF，见第3.3节）：由于多模态特征之间的交互能够生成有助于异常检测的新表征，[17, 28], 我们提出无监督特征融合模块，以统一多模态特征的分布并学习它们之间的内在联系。（3）决策层融合（DLF，见第3.4节）：尽管UFF有助于提升检测性能，但我们发现信息损失不可避免，因此提出决策层融合，利用多个记忆库作出最终决策。

3.2. 点特征对齐

点特征提取。我们使用 Point Trans- former（$\mathcal{F}_{pt}$）[37]来提取点云特征。输入点云 p 是一个由 N个点组成的位置序列。经过最远点采样（FPS）[24], 后，点云被划分为 M 个组，每组包含 S 个点。接着，将每组中的点编码为一个特征向量，并将这些 M 向量输入 Point Trans-

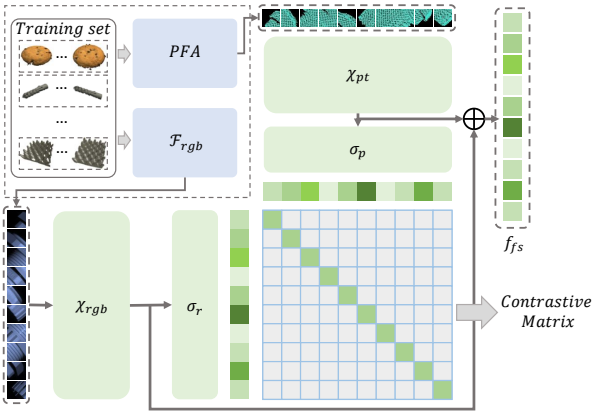


图3. UFF架构。UFF是一个统一模块，使用MVTec-3D AD的全部训练数据进行训练。逐块对比损失 $\mathcal{L}_{con}$ 促使位于相同位置的多模态补丁特征具有最大的互信息，即对比矩阵的对角元素具有最大值。

former。Point Transformer 的输出 g 是 M个点特征，随后将其组织为点特征组：每组包含一个点特征，可视为中心点的特征。

点特征插值。由于经过最远点采样（FPS）后，点中心在空间中的分布并不均匀，导致点特征的密度不平衡。我们提出将特征插值回原始点云。给定与 M 组中心点 c_i相关联的 M 点特征 g_i，我们使用反距离权重将特征插值到输入点云中的每个点 p_j（$j \in \{1,2,...,N\}$）。该过程可描述为：

$$p'_j = \sum_{i=1}^{M} \alpha_i g_i, \quad \alpha_i = \frac{\frac{1}{\|c_i - p_j\|_2 + \epsilon}}{\sum_{k=1}^{M}\sum_{t=1}^{N} \frac{1}{\|c_k - p_t\|_2 + \epsilon}}, \tag{1}$$

其中，ϵ 是一个相当小的常数，用于避免分母为 0。

点特征投影。插值后，我们根据点坐标和相机参数将 p'_j 投影到二维平面上，并将投影后的点表示为$\hat{p}$。我们注意到点云可能较为稀疏；如果二维平面上的某个位置没有对应的点，就将该位置设为 0。我们将投影特征图表示为$\{\hat{p}_{x,\ y}|(x,\ y) \in \mathbb{D}\}$（$\mathbb{D}$ 是RGB图像的二维平面区域），其尺寸与输入RGB图像相同。最后，我们通过平均池化操作，在二维平面特征图上获取图块特征。

3.3. 无监督特征融合

多模态特征之间的交互能够产生有助于检测工业异常的新信息。例如，在图 1 中，我们需要结合黑色这一颜色特征和形状上的凹陷来检测曲奇上的孔洞。为了学习训练数据中两种模态之间固有的关系，我们设计了无监督特征融合（UFF）模块。我们提出了一种逐块对比损失来训练特征融合模块：给定 RGB 特征 f_{rgb} 和点云特征 f_{pt}，我们的目标是促使不同模态在相同位置的特征包含更多对应信息，同时使不同位置的特征包含更少对应信息。

我们将图像块的特征表示为 $\{f_{(i,\ j)rgb},\ f_{pt}^{(i,j)}\}$，其中 i 表示训练样本的索引，j 表示图像块的索引。我们使用多层感知机（MLP）层 $\{\chi_{rgb}, \chi_{pt}\}$ 提取两种模态之间的交互信息，并使用全连接层 $\{\sigma_{rgb}, \sigma_{pt}\}$ 将处理后的特征映射为查询向量或键向量。我们将映射后的特征表示为 $\{h_{rgb}^{(i,j)}, h_{pt}^{(i,j)}\}$。随后，我们采用 InfoNCE [20] 损失进行对比学习：

$$\mathcal{L}_{con} = \frac{h_{rgb}^{(i,j)} \cdot h_{pt}^{(i,j)T}}{\sum_{t=1}^{N_b}\sum_{k=1}^{N_p} h_{rgb}^{(t,k)} \cdot h_{pt}^{(t,k)T}}, \tag{2}$$

其中，N_b 表示批大小，N_p 表示非零图像块数量。UFF 是一个统一模块，使用 MVTec-3D AD 所有类别的训练数据进行训练，其架构如图3所示。

在推理阶段，我们将多层感知机各层的输出拼接起来，得到融合图像块特征，记为 $f_{fs}^{(i,j)}$：

$$f_{fs}^{(i,j)} = \chi_{rgb}(f_{rgb}^{(i,j)}) \oplus \chi_{pt}(f_{pt}^{(i,j)}). \tag{3}$$

3.4. 决策层融合

如图 1所示，部分工业异常仅出现在单一域中（例如土豆的凸出部分），多模态特征之间的对应关系可能并不十分明显。此外，尽管特征融合促进了多模态特征之间的交互，但我们仍然发现，融合过程中有部分信息丢失。

为了解决上述问题，我们提出使用多个记忆库来存储原始颜色特征、位置特征和融合特征。我们分别将这三种记忆库记为 $\mathcal{M}_{rgb}$，$\mathcal{M}_{pt}$，$\mathcal{M}_{fs}$。我们参照 PatchCore [26] 构建这三个记忆库；在推理过程中，每个记忆库都会用于预测异常分数和分割图。随后，我们使用两个可学习的单类支持向量机（OCSVM）[29] $\mathcal{D}_a$ 和 $\mathcal{D}_s$，分别针对异常分数 a 和分割图S做出最终决策。我们将上述过程称为决策层融合（DLF），可描述如下：

$$a = \mathcal{D}_a(\phi(\mathcal{M}_{rgb}, f_{rgb}), \phi(\mathcal{M}_{pt}, f_{pt}), \phi(\mathcal{M}_{fs}, f_{fs})), \tag{4}$$

$$S = \mathcal{D}_s(\psi(\mathcal{M}_{rgb}, f_{rgb}), \psi(\mathcal{M}_{pt}, f_{pt}), \psi(\mathcal{M}_{fs}, f_{fs})), \tag{5}$$

其中， ϕ， ψ 是 [26],引入的评分函数，其形式如下：

$$\phi(\mathcal{M}, f) = \eta \| f^{(i,j),*} - m^* \|_2, \tag{6}$$

$$\psi(\mathcal{M}, f) = \{ \min_{m \in \mathcal{M}} \| f^{(i,j)} - m \|_2 \Big| f^{(i,j)} \in f \}, \tag{7}$$

$$f^{(i,j),*}, m^* = \arg \max_{f^{(i,j)} \in f} \arg \min_{m \in \mathcal{M}} \| f^{(i,j)} - m \|_2, \tag{8}$$

其中 $\mathcal{M} \in \{\mathcal{M}_{rgb}, \mathcal{M}_{pt}, \mathcal{M}_{fs}\}$、 $f \in \{f_{rgb}, f_{pt}, f_{fs}\}$ 和η 是重加权参数。

我们提出了一个两阶段训练流程：第一阶段构建记忆库，第二阶段训练决策层。DLF的伪代码如算法1所示。

算法1： 决策层融合训练

输入： 记忆库构建算法 $\mathcal{P}$ [26]、决策层 $\{\mathcal{D}_a, \mathcal{D}_s\}$，OCSVM损失函数 $\mathcal{L}_{oc}$ [29]
数据： 训练集特征 $\{\mathbb{F}_{rgb}, \mathbb{F}_{pt}, \mathbb{F}_{fs}\}$。
输出： 多模态记忆库 $\{\mathcal{M}_{rgb}, \mathcal{M}_{pt}, \mathcal{M}_{fs}\}$，决策层参数 $\{\Theta_{\mathcal{D}_a}, \Theta_{\mathcal{D}_s}\}$。
对于 $modal \in \{rgb, pt, fs\}$ **执行**
  对于 $f_{modal} \in \mathbb{F}_{modal}$ **执行**
    $\mathcal{M}_{modal} \leftarrow f_{modal}$
  end
  $\mathcal{M}_{modal} \leftarrow \mathcal{P}(\mathcal{M}_{modal})$
end
对于 $f_{rgb} \in \mathbb{F}_{rgb}$， $f_{pt} \in \mathbb{F}_{pt}$， $f_{fs} \in \mathbb{F}_{fs}$ **执行**
  $\Theta_{\mathcal{D}_a} \overset{optim}{\longleftarrow} \mathcal{L}_{oc}(\mathcal{D}_a; \Theta_{\mathcal{D}_a})$
  $\Theta_{\mathcal{D}_s} \overset{optim}{\longleftarrow} \mathcal{L}_{oc}(\mathcal{D}_s; \Theta_{\mathcal{D}_s})$
end

4. 实验

4.1. 实验细节

数据集。 三维工业异常检测尚处于起步阶段。MVTec-3D AD数据集是首个三维工业异常检测数据集。我们的实验在MVTec-3D数据集上进行。

MVTec-3D AD [3]数据集包含10个类别，共有2656个训练样本和1137个测试样本。三维扫描由工业传感器使用结构光采集，位置信息存储在3个通道张量中，分别表示 x、 y 和 z 坐标。这些3个通道张量可以一一映射到相应的点云。此外，还记录了每个点的RGB信息。由于数据集中的所有样本都从相同角度观察，因此每个样本的RGB信息可以存储在单张图像中。总之，MVTec-3D AD数据集中的每个样本都包含一个带颜色的点云。

数据预处理。 与二维数据不同，三维数据更容易去除背景信息。接下来[17], 我们使用 RANSAC 估计背景平面， [13]并移除距离该平面小于 0.005 的所有点。同时，将 RGB 图像中被移除点对应的像素设为 0。这一操作不仅加快了训练和推理过程中的三维特征处理，还减少了异常检测中的背景干扰。最后，我们将位置张量和 RGB 图像都调整为 224×224 大小，使其与特征提取器的输入尺寸相匹配。

特征提取器。 我们使用两个基于 Transformer 的特征提取器，分别提取 RGB 特征和点云特征：1）对于 RGB 特征，我们使用视觉 Transformer（ViT） [12] 直接提取每个图像块的特征。为了适应异常检测任务，我们采用 ViT-B/8架构，以兼顾效率和检测粒度；为获得更高性能，我们使用在 ImageNet 上预训练的 ViT-B/8 [11]，并结合 DINO [6], ，该预训练模型接收一张 224×224 图像，并为每张图像总共输出784 个图像块特征；由于先前研究表明，ViT 在每一层都能关注全局和局部信息，因此我们使用最后一层的输出，其具有768 维特征，用于异常检测。2）对于点云特征，我们使用一个 Point Transformer [21,37]，该模型在 ShapeNet [7]数据集上进行了预训练，作为我们的三维特征提取器，并使用其 $\{3, 7, 11\}$ 层的输出作为三维特征；Point Transformer 首先将点云编码为点组，这些点组与 ViT 的图像块类似，每个点组都包含一个用于表示位置的中心点，以及用于确定组大小的邻居数量。如第3.2节所述，我们分别测试设置 $M = 784$、 $S = 64$ 和 $M = 1024$， $S = 128$ 以开展我们的实验。在 PFA操作中，我们分别将点特征池化至 28×28 和 56×56 进行测试。

可学习模块详情。 M3DM 有 2 个可学习模块：无监督特征融合模块和决策层融合模块。1）对于 UFF， χ_{rgb}, χ_{pc}是两个双层多层感知机，以 $4\times$ 隐藏维度作为输入特征；我们使用 AdamW 优化器，将学习率设为 0.003，采用 250 步余弦预热，批大小设为 16，并报告 UFF 训练 750 步时取得的最佳异常检测结果。2）对于 DLF，我们使用两个线性 OCSVM，并采用随机梯度下降优化器；学习率设为 1×10^{-4}，每个类别训练 1000 步。

评估指标。 所有评估指标与 [3] 中的完全相同。我们使用接收者操作特征曲线下面积（I-AUROC）评估图像级异常检测性能，I-AUROC 越高表示图像级异常检测性能越好。对于分割评估，我们采用区域重叠率（AUPRO）指标，该指标定义为二值预测与真实标注中每个连通分量之间相对重叠率的平均值。与 I-AUROC 类似，可以利用像素级预测的接收者操作特征曲线计算 P-AUROC，以评估分割性能。

	方法	百吉饼	电缆腺体	胡萝卜	曲奇	木销	Foam	桃子	土豆	Rope	Tire	Mean
3D	深度生成对抗网络 [3]	0.530	0.376	0.607	0.603	0.497	0.484	0.595	0.489	0.536	0.521	0.523
	深度自编码器 [3]	0.468	0.731	0.497	0.673	0.534	0.417	0.485	0.549	0.564	0.546	0.546
	深度体积模型 [3]	0.510	0.542	0.469	0.576	0.609	0.699	0.450	0.419	0.668	0.520	0.546
	体素生成对抗网络 [3]	0.383	0.623	0.474	0.639	0.564	0.409	0.617	0.427	0.663	0.577	0.537
	体素自编码器 [3]	0.693	0.425	0.515	0.790	0.494	0.558	0.537	0.484	0.639	0.583	0.571
	体素体积模型 [3]	0.750	**0.747**	0.613	0.738	0.823	0.693	0.679	0.652	0.609	0.690	0.699
	3D-ST [4]	0.862	0.484	0.832	0.894	0.848	0.663	0.763	0.687	**0.958**	0.486	0.748
	FPFH [17]	0.825	0.551	0.952	0.797	0.883	0.582	0.758	0.889	0.929	0.653	0.782
	AST [28]	0.881	0.576	**0.965**	0.957	0.679	**0.797**	**0.990**	0.915	0.956	0.611	0.833
	Ours	**0.941**	0.651	**0.965**	**0.969**	**0.905**	0.760	0.880	**0.974**	0.926	**0.765**	**0.874**
RGB	DifferNet [27]	0.859	0.703	0.643	0.435	0.797	0.790	0.787	0.643	0.715	0.590	0.696
	PADiM [9]	**0.975**	0.775	0.698	0.582	0.959	0.663	0.858	0.535	0.832	0.760	0.764
	PatchCore [26]	0.876	0.880	0.791	0.682	0.912	0.701	0.695	0.618	0.841	0.702	0.770
	STFPM [32]	0.930	0.847	0.890	0.575	0.947	0.766	0.710	0.598	0.965	0.701	0.793
	CS-Flow [16]	0.941	**0.930**	0.827	0.795	**0.990**	0.886	0.731	0.471	0.986	0.745	0.830
	AST [28]	0.947	0.928	0.851	**0.825**	0.981	**0.951**	0.895	0.613	**0.992**	**0.821**	**0.880**
	Ours	0.944	0.918	**0.896**	0.749	0.959	0.767	**0.919**	**0.648**	0.938	0.767	0.850
RGB + 3D	深度生成对抗网络 [3]	0.538	0.372	0.580	0.603	0.430	0.534	0.642	0.601	0.443	0.577	0.532
	深度自编码器 [3]	0.648	0.502	0.650	0.488	0.805	0.522	0.712	0.529	0.540	0.552	0.595
	深度体积模型 [3]	0.513	0.551	0.477	0.581	0.617	0.716	0.450	0.421	0.598	0.623	0.555
	体素生成对抗网络 [3]	0.680	0.324	0.565	0.399	0.497	0.482	0.566	0.579	0.601	0.482	0.517
	体素自编码器 [3]	0.510	0.540	0.384	0.693	0.446	0.632	0.550	0.494	0.721	0.413	0.538
	体素体积模型 [3]	0.553	0.772	0.484	0.701	0.751	0.578	0.480	0.466	0.689	0.611	0.609
	PatchCore + FPFH [17]	0.918	0.748	0.967	0.883	0.932	0.582	0.896	0.912	0.921	**0.886**	0.865
	AST [28]	0.983	0.873	**0.976**	0.971	0.932	0.885	**0.974**	**0.981**	**1.000**	0.797	0.937
	Ours	**0.994**	**0.909**	0.972	**0.976**	**0.960**	**0.942**	0.973	0.899	0.972	0.850	**0.945**

表 1. MVTec-3D AD 所有类别的异常检测 I-AUROC 分数。我们的方法在纯 3D 设置和 3D +RGB 设置下均明显优于其他方法；在纯 3D 设置下，我们的方法达到 0.874 的平均 I-AUROC 分数；在 3D +RGB 设置下，平均 I-AUROC 分数达到 0.945。基线方法的结果来自 [3, 17, 28, 38]。

4.2. MVTec-3D AD 上的异常检测

我们在 MVTec-3D 上将我们的方法与若干三维、RGB 以及 RGB+ 3D 多模态方法进行比较。表 1 展示了采用 I-AUROC 记录的异常检测结果，表 2展示了采用 AUPRO 报告的分割结果，而 P-AUROC 则在补充材料中报告。1）在纯三维异常检测方面，我们取得了最高的 I-AUROC，并超过 AST [28] 4.1%，这表明我们的方法相比以往方法具有更出色的检测性能；借助 PFA，Point Transformer 成为了该任务中更好的三维特征提取器。在分割方面，我们以 0.906 的 AUPRO 取得了第二好的结果。由于我们的三维域分割基于点云，我们发现点云与真实标签之间存在偏差，并在第 4.7 节讨论了这一问题。2）在 RGB异常检测方面，我们的方法与 Patchcore [26] 的区别在于，我们使用基于 Transformer 的特征提取器，而不是 Wide-ResNet，并且在构建记忆库之前移除了池化操作；我们在 RGB 域中的 I-AUROC 比原始 PatchCore 结果高 8.0%，并在分割方面取得了最高的 AUPRO 分数，比第二名高 7.6%。3）在 RGB + 3D 多模态异常检测方面，我们的方法在 I-AUROC 和 AUPRO 分数上均取得了最佳结果；I-AUROC 比 AST 高 0.8%，AUPRO 比 PatchCore + FPFH [17] 高 0.5%。这些结果得益于我们的融合策略以及高性能的三维异常检测结果。此前如表3所示，以往的方法无法同时取得出色的检测和分割性能。由于 AST [28] 未报告 AUPRO 结果，因此我们在此使用 P-AUROC 比较分割性能。尽管 PatchCore + FPFH 方法取得了较高的 P-AUROC 分数，但其 I-AUROC 远低于另外两种方法。此外，AST 的 P-AUROC 分数低于另外两种方法，这意味着 AST 在定位异常方面较弱。

4.3. 消融研究

我们在多模态设置下进行了消融研究。为展示我们对多模态融合的贡献，我们在表4 中对方法进行了如下分析：1）仅使用点云（$\mathcal{M}_{pt}$）信息；2）仅使用 RGB（$\mathcal{M}_{rgb}$）信息；3）使用单个记忆库（$\mathcal{M}_{fs}$），直接将 Point Transformer 特征与 RGB 特征拼接；4）使用单个记忆库（$\mathcal{M}_{fs}$），利用 UFF 融合多模态特征；5）分别构建两个记忆库（$\mathcal{M}_{rgb}$，$\mathcal{M}_{pt}$），然后直接将分数相加；6）分别构建两个记忆库（$\mathcal{M}_{rgb}$，$\mathcal{M}_{pt}$），并使用 DLF 得到最终结果；7）构建三个记忆库（$\mathcal{M}_{rgb}$，$\mathcal{M}_{pt}$，$\mathcal{M}_{fs}$）（我们的方法）。比较表4中的第 3 行和第 4 行可以发现，加入 UFF 后，三个指标的结果均得到显著提升（图像级 AUROC 提升 4.1% ↑，AURPO 提升 1.2% ↑，P-AUROC 提升 0.3% ↑），这表明 UFF 在多模态交互中发挥着重要作用，并有助于统一特征分布；比较第 5 行和第 6 行可以证明，DLF 模型有助于提升异常检测和分割性能（图像级

	方法	百吉饼	电缆腺体	胡萝卜	曲奇	木销	Foam	桃子	土豆	Rope	Tire	Mean
3D	深度生成对抗网络 [3]	0.111	0.072	0.212	0.174	0.160	0.128	0.003	0.042	0.446	0.075	0.143
	深度自编码器 [3]	0.147	0.069	0.293	0.217	0.207	0.181	0.164	0.066	0.545	0.142	0.203
	深度体积模型 [3]	0.280	0.374	0.243	0.526	0.485	0.314	0.199	0.388	0.543	0.385	0.374
	体素生成对抗网络 [3]	0.440	0.453	0.875	0.755	0.782	0.378	0.392	0.639	0.775	0.389	0.583
	体素自编码器 [3]	0.260	0.341	0.581	0.351	0.502	0.234	0.351	0.658	0.015	0.185	0.348
	体素体积模型 [3]	0.453	0.343	0.521	0.697	0.680	0.284	0.349	0.634	0.616	0.346	0.492
	3D-ST [4]	0.950	0.483	**0.986**	**0.921**	**0.905**	0.632	0.945	**0.988**	**0.976**	0.542	0.833
	FPFH [17]	**0.973**	**0.879**	0.982	0.906	0.892	0.735	**0.977**	0.982	0.956	**0.961**	**0.924**
	Ours	0.943	0.818	0.977	0.882	0.881	**0.743**	0.958	0.974	0.95	0.929	0.906
RGB	CFlow [16]	0.855	0.919	0.958	0.867	**0.969**	0.500	0.889	0.935	0.904	0.919	0.871
	PatchCore [26]	0.901	0.949	0.928	0.877	0.892	0.563	0.904	0.932	0.908	0.906	0.876
	PADiM [9]	**0.980**	0.944	0.945	0.925	0.961	0.792	0.966	0.940	0.937	0.912	0.930
	Ours	0.952	**0.972**	**0.973**	**0.891**	0.932	**0.843**	**0.97**	**0.956**	**0.968**	**0.966**	**0.942**
RGB + 3D	深度生成对抗网络 [3]	0.421	0.422	0.778	0.696	0.494	0.252	0.285	0.362	0.402	0.631	0.474
	深度自编码器 [3]	0.432	0.158	0.808	0.491	0.841	0.406	0.262	0.216	0.716	0.478	0.481
	深度体积模型 [3]	0.388	0.321	0.194	0.570	0.408	0.282	0.244	0.349	0.268	0.331	0.335
	体素生成对抗网络 [3]	0.664	0.620	0.766	0.740	0.783	0.332	0.582	0.790	0.633	0.483	0.639
	体素自编码器 [3]	0.467	0.750	0.808	0.550	0.765	0.473	0.721	0.918	0.019	0.170	0.564
	体素体积模型 [3]	0.510	0.331	0.413	0.715	0.680	0.279	0.300	0.507	0.611	0.366	0.471
	图像块 + FPFH [17]	**0.976**	0.969	**0.979**	**0.973**	0.933	0.888	0.975	**0.981**	0.950	0.971	0.959
	Ours	0.970	**0.971**	**0.979**	0.950	**0.941**	**0.932**	**0.977**	0.971	**0.971**	**0.975**	**0.964**

表 2. MVTec-3D 所有类别异常分割的 AUPRO 分数。我们的方法在 RGB 和 RGB + 3D 设置下均优于其他方法。在 RGB 设置下，我们的方法达到 0.942 的平均 AUPRO 分数；在 RGB + 3D 设置下，我们的方法达到 0.964 的平均 AUPRO 分数。基线方法的结果来自 [3, 17, 38]。

方法	图像级 AUROC	P-AUROC
图像块 + FPFH [17]	0.865	0.992
AST [28]	0.937	0.976
Ours	**0.945**	**0.992**

表3. MVTec-3D 所有类别异常检测的平均值 I-AUROC 和 P-AUROC 得分。我们的方法在异常检测和分割方面均表现出色。

方法	记忆库	图像级 AUROC	AUPRO	P-AUROC
仅点云	$\mathcal{M}_{pt}$	0.874	0.906	0.970
仅 RGB	$\mathcal{M}_{rgb}$	0.850	0.942	0.987
不使用UFF	$\mathcal{M}_{fs}$	0.857	0.944	0.987
使用UFF	$\mathcal{M}_{fs}$	0.898	0.956	0.990
不使用DLF	$\mathcal{M}_{rgb}, \mathcal{M}_{pt}$	0.929	0.953	0.987
使用DLF	$\mathcal{M}_{rgb}, \mathcal{M}_{pt}$	0.932	0.959	0.990
Ours	$\mathcal{M}_{rgb}, \mathcal{M}_{pt}, \mathcal{M}_{fs}$	**0.945**	**0.964**	**0.992**

表4。融合模块的消融研究。M 表示所使用的记忆库数量。与直接拼接特征相比，采用 UFF 的单记忆库方法性能更好。采用 DLF 后，异常检测和分割性能均得到显著提升。

AUROC 提升 0.3% ↑，AURPO 提升 0.6% ↑，P-AUROC 提升 0.3% ↑）。我们的完整设置如表4中的第 7 行所示。与第 6 行相比，我们的图像级 AUROC 提升了 1.3%，AUPRO 提升了 0.5%，这进一步证明 UFF 激活了两种模态之间的交互，并为异常检测创建了新的特征。

4.4. PFA 超参数分析

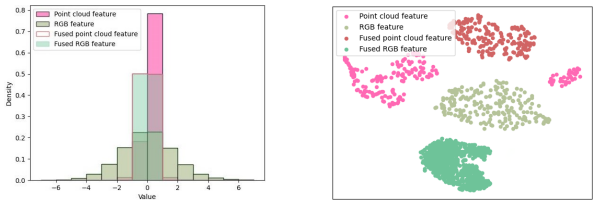
由于我们是首个将 Point Transformer 用于三维异常检测的团队，因此围绕 Point Transformer 的设置开展了一系列探索性实验。1）我们首先探究 Point Transformer 的两个重要超参数：最远点采样时的点组数量和每组的点数。点组数量决定了 Point Transformer 将提取多少个特征，而每组的点数则相当于感受野的概念。如表5所示，采用 1,024 个点组、每组 128 个点的模型在此任务上表现更好。我们认为，这是因为更多的特征向量有助于模型找到更精细的表征，而合适的邻居数量则能提供更多局部位置信息。2）为验证 PFA 操作，我们使用原始点组特征开展了另一项三维异常检测实验：一个点组可视为一个图像块，此处记忆库存储点组特征；检测方法与基于图像块的方法相同。为获得分割预测，我们先将点组特征投影到二维平面，再使用反距离插值计算每个像素的值。如表5 第 2 行所示，基于点组的方法优于第 4 行中的对应方法；然而，当图像块尺寸变小时，基于 PFA 的方法在三项指标上均取得最佳结果。

4.5. 多模态特征分布分析

我们使用直方图和 t-SNE 可视化特征分布 [31]。原始点云特征在 t-SNE图中存在两个不相连的区域（图4b），这是由非点云区域与点云区域边界处的池化操作造成的。两个融合特征具有相似的分布（见图4a），这些特性使拼接后的特征更适合用于构建记忆库和进行特征距离计算。原始特征具有更复杂的分布，这

S.G	N.G	采样	图像级 AUROC	AUPRO	P-AUROC
64	784	点组	0.793	0.813	0.922
128	1024	点组	0.841	0.896	0.960
64	784	28×28图像块	0.805	0.879	0.963
128	1024	28×28图像块	0.819	0.896	0.967
128	1024	56×56图像块	**0.874**	**0.906**	**0.970**

表5. 在纯三维数据集上探索 Point Transformer 的设置。S.G 表示每个点组中的点数，N.G 表示点组总数。每个样本使用 1024 个点组、每个点组包含 128 个点时，性能最佳；与直接计算点组上的异常分数和分割分数相比，基于二维平面图像块的方法需要较小的图像块大小才能获得较高性能。



(a) 统计分布。 (b) T-SNE 分布。

图4. 百吉饼多模态特征的分布。融合后的点特征分布方差较小，且与融合后的 RGB 特征分布更接近。

方法	图像级 AUROC	AUPRO	P-AUROC
5样本	0.796	0.927	0.981
10样本	0.821	0.939	0.985
50样本	0.903	0.953	0.988
完整数据集	0.945	0.964	0.992

表 6. 少样本设置下的结果。在 10样本或 5样本设置下，我们的方法仍具有良好的分割性能，并优于大多数非少样本方法。

有助于进行单域异常检测。我们的混合融合方案结合了原始特征和融合特征的优势，因此性能优于单记忆库方法。

4.6. 少样本异常检测

我们在少样本设置下评估了我们的方法，结果如表 6 所示。我们从每个类别中随机选取 10 张和 5 张图像作为训练数据，并在完整的测试数据集上测试少样本模型。我们发现，在 10样本或 5样本设置下，我们的方法仍比一些非少样本方法具有更好的分割性能。

4.7. 关于 MVTec-3D AD 的讨论

本节讨论 MVTec-3D AD 数据集的一些特性。1）3D信息有助于检测更多种类的异常。在图 5的第一行中，我们可以看到，模型无法利用 RGB信息检测出异常，但使用点云后，便能准确预测该异常。这表明 3D信息确实在该数据集中发挥着重要作用。2）标签偏差会导致分割不准确。如图 5所示，

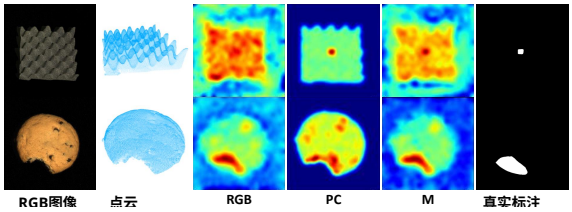


图 5. MVTec-3D AD 的特性。PC 是点云预测的简称，M 是多模态预测的简称。借助点云中的三维信息，第 1 行中的异常得到了准确定位。标签偏差会导致分割不准确，因为模型难以聚焦于缺失区域。

第 2 行的曲奇有一些切口，异常标签标注在缺失区域上。然而，对于纯点云方法，不会报告非点云区域，而是将切割边缘报告为异常区域。在图 5中曲奇的 PC 预测结果可以看到这一现象。由于三维点云与二维真实标注之间存在这种偏差，表 2中的三维版本 AUPRO 分数低于 RGB 版本。使用 RGB信息时，缺失区域会被更准确地报告为异常。尽管我们利用多模态数据成功预测出更多异常区域，但预测图与真实标注之间仍存在差距。未来的研究中，我们将致力于解决这一问题。

5. 结论

本文提出了一种结合点云和 RGB 图像的多模态工业异常检测方法。该方法基于多个记忆库，并提出了一种用于处理多模态数据的混合特征融合方案。具体而言，我们提出了基于图像块级对比损失的无监督特征融合，以促进多模态交互并统一分布；随后提出决策层融合，以融合多个记忆库的输出。此外，我们采用预训练的 Point Transformer 和 Vision Transformer 作为特征提取器。为将上述两个特征提取器对齐到相同的空间位置，我们提出了点特征对齐，将 3D特征转换到二维平面。我们的方法在 MVTec-3D AD 数据集上取得了优于最先进水平的结果，并希望我们的工作能为进一步研究提供帮助。

致谢。 本研究得到国家自然科学基金（72192821、61972157、62272447）、上海市科技重大项目（2021SHZDZX0102）、上海市科学技术委员会（21511101200）、上海市扬帆计划（22YF1420300、23YF1410500）、CCF-腾讯开源研究基金（RAGR20220121）以及中国科学技术协会青年科技人才托举工程（2022QNRC001）的资助。

参考文献

[1] Fabrizio Angiulli 和 Clara Pizzuti。高维空间中的快速离群点检测。载于《欧洲数据挖掘与知识发现原理会议论文集》，第15–27页。施普林格，2002年。2[2] Paul Bergmann、Michael Fauser、David Sattlegger 和 Carsten Steger。MVTec AD——用于无监督异常检测的综合性真实世界数据集。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第9592–9600页，2019年。1, 2[3] Paul Bergmann、Xin Jin、David Sattlegger 和 Carsten Steger。用于无监督三维异常检测与定位的 MVTec 3D-AD数据集。载于 Giovanni Maria Farinella、Petia Radeva 和 Kadi Bouatouch 编，《第17届国际计算机视觉、图像处理与计算机图形理论联合会议论文集》，VISIGRAPP 2022，第5卷：VISAPP，在线直播，2022年2月6日至8日，第202–213页。SCITEPRESS，2022年。1, 3, 5, 6, 7, 12, 14[4] Paul Bergmann 和 David Sattlegger。使用深度几何描述符进行三维点云异常检测。arXiv 预印本 arXiv:2202.11660，2022年。3, 6, 7[5] Luca Bonfiglioli、Marco Toschi、Davide Silvestri、Nicola Fioraio 和 Daniele De Gregorio。用于无监督多模态异常检测与定位的 Eyecandies 数据集。载于《亚洲计算机视觉会议论文集》，第3586–3602页，2022年。12, 15[6] Mathilde Caron、Hugo Touvron、Ishan Misra、Herv´e J´egou、Julien Mairal、Piotr Bojanowski 和 Armand Joulin。自监督视觉变换器中涌现的特性。载于《IEEE/CVF国际计算机视觉会议论文集》，第9650–9660页，2021年。2, 3, 5, 12[7] Angel X Chang、Thomas Funkhouser、Leonidas Guibas、Pat Hanrahan、Qixing Huang、Zimo Li、Silvio Savarese、Manolis Savva、Shuran Song、Hao Su 等。ShapeNet：信息丰富的三维模型库。arXiv 预印本 arXiv:1512.03012，2015年。5[8] Teri Crosby。如何检测和处理离群点，1994年。2[9] Thomas Defard、Aleksandr Setkov、Angelique Loesch 和 Romaric Audigier。PADiM：用于异常检测与定位的图像块分布建模框架。载于《国际模式识别会议论文集》，第475–489页。施普林格，2021年。2, 6, 7, 15[10] Hanqiu Deng 和 Xingyu Li。通过单类嵌入反向蒸馏进行异常检测。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第9737–9746页，2022年。1, 2[11] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大规模分层图像数据库。载于《2009年IEEE计算机视觉与模式识别会议论文集》，第248–255页。IEEE，2009年。5, 12[12] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、J akob Uszkoreit 和 Neil Houlsby。一张图像胜过 16×16 个词：用于图像识别的变换器达到规模。在第九届国际学习表征会议（ICLR 2021）论文集，第7卷，2021年，奥地利线上活动，5月。 2, 5 OpenReview.net。[13]

Martin A Fischler 和 Robert C Bolles。随机采样一致性：一种用于模型拟合的范式及其在图像分析和自动制图中的应用。《ACM通讯》，24(6):381–395，1981年。5[14] Markus Goldstein 和 Andreas Dengel。基于直方图的离群值评分（hbos）：一种快速的无监督异常检测算法。KI-2012：海报与演示环节，第9卷，2012年。2[15] Dong Gong、Lingqiao Liu、Vuong Le、Budhaditya Saha、Moussa Reda Mansour、Svetha Venkatesh 和 Anton van den Hengel。记忆正常模式以检测异常：用于无监督异常检测的记忆增强型深度自编码器。载于 IEEE/CVF国际计算机视觉会议论文集，第1705–1714页，2019年。1，2。[16] Denis Gudovskiy、Shun Ishizaka 和 Kazuki Kozuka。Cflow-ad：通过条件归一化流实现实时无监督异常检测与定位。载于 IEEE/CVF冬季计算机视觉应用会议论文集，第98–107页，2022年。1，2，6，7[17] Eliahu Horwitz 和 Yedid Hoshen。三维异常检测与分割的实证研究。arXiv预印本 arXiv:2203.05550，2022年。2，3，5，6，7，11，12[18] Zheng Li、Yue Zhao、Xiyang Hu、Nicola Botta、Cezar Ionescu 和 George Chen。ECOD：使用经验累积分布函数进行无监督离群点检测。《IEEE知识与数据工程汇刊》，2022年。2[19] Fei Tony Liu、Kai Ming Ting 和 Zhi-Hua Zhou。孤立森林。载于2008年第八届 IEEE国际数据挖掘会议论文集，第413–422页。IEEE，2008年。2[20] Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。基于对比预测编码的表征学习，2018年。4[21] Yatian Pang、Wenxiao Wang、Francis E. H. Tay、Wei Liu、Yonghong Tian 和 Li Yuan。用于点云自监督学习的掩码自编码器，2022年。2，3，5，12[22] Spiros Papadimitriou、Hiroyuki Kitagawa、Phillip B Gibbons 和 Christos Faloutsos。LOCI：利用局部相关积分进行快速离群点检测。载于第19届国际数据工程会议论文集（目录编号 *03 37405* 315 326 2003 2 CH），第–页。[23] IEEE，。Pramuditha Perera、Ramesh Nallapati 和 Bing Xiang。OC-GAN：使用具有受限潜在表征的生成对抗网络进行单类新颖性检测。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第2898–2906页，2019年。1，2[24] Charles Ruizhongtai Qi、Li Yi、Hao Su 和 Leonidas J Guibas。PointNet++：度量空间中点集的深度层次特征学习。《神经信息处理系统进展》，第30卷，2017年。3[25] Sridhar Ramaswamy、Rajeev Rastogi 和 Kyuseok S him。大型数据集离群点挖掘的高效算法。载于2000年 ACM SIGMOD数据管理国际会议论文集，第427–438页，2000年。2

[26] Karsten Roth、Latha Pemula、Joaquin Zepeda、Bernhard Sch¨olkopf、Thomas Brox 和 Peter Gehler。迈向工业异常检测中的完全召回。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第14318–14328页，2022年。1，2，4，5，6，7，11，12[27] Marco Rudolph、Bastian Wandt 和 Bodo Rosenhahn。看似相同，实则不同：利用归一化流进行半监督缺陷检测。载于 IEEE冬季计算机视觉应用会议（WACV 2021），怀科洛亚，夏威夷州，美国，2021年1月3日至8日，第1906–1915页。IEEE，2021年。6[28] Marco Rudolph、Tom Wehrbein、Bodo Rosenhahn 和 Bastian Wandt。用于工业异常检测的非对称学生-教师网络。载于 IEEE/CVF冬季计算机视觉应用会议论文集，第2592–2602页，2023年。1，2，3，6，7，11，12[29] Bernhard Sch¨olkopf、John C Platt、John Shawe-Taylor、Alex J Smola 和 Robert C Williamson。高维分布支撑集的估计。《神经计算》，13(7):1443–1471，2001年。2，4，5，11[30] Mei-Ling Shyu、Shu-Ching Chen、Kanoksri Sarinnapakorn 和 LiWu Chang。一种基于主成分分类器的新型异常检测方案。技术报告，迈阿密大学科勒尔盖布尔斯分校电气与计算机工程系，2003年。2[31] Laurens Van der Maaten 和 Geoffrey Hinton。使用 t-SNE 可视化数据。《机器学习研究期刊》，9(11)，2008年。7[32] Guodong Wang、Shumin Han、Errui Ding 和 Di Huang。用于异常检测的学生-教师特征金字塔匹配。载于第32届英国机器视觉会议（BMVC 2021），线上，2021年11月22日至25日，第306页。BMVA出版社，2021年。6，15[33] Jiawei Yu、Ye Zheng、Xiang Wang、Wei Li、Yushuang Wu、Rui Zhao 和 Liwei Wu。FastFlow：通过二维归一化流实现无监督异常检测与定位。arXiv预印本 arXiv:2111.07677，2021年。1，2[34] Xumin Yu、Lulu Tang、Yongming Rao、Tiejun Huang、Jie Zhou 和 Jiwen Lu。Point-Bert：通过掩码点建模对三维点云变换器进行预训练。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第19313–19322页，2022年。12[35] Vitjan Zavrtanik、Matej Kristan 和 Danijel Skoˇcaj。DRAEM——一种通过判别式训练获得的表面异常检测重建嵌入。载于 IEEE/CVF国际计算机视觉会议论文集，第8330–8339页，2021年。1，2[36] Vitjan Zavrtanik、Matej Kristan 和 Danijel Skoˇcaj。通过图像修复进行重建以实现视觉异常检测。《模式识别》，112:107706，2021年。1，2[37] Hengshuang Zhao、Li Jiang、Jiaya Jia、Philip HS Torr 和 Vladlen Koltun。Point Transformer。载于 IEEE/CVF国际计算机视觉会议论文集，第16259–16268页，2021年。2，3，5[38] Ye Zheng、Xiang Wang、Yu Qi、Wei Li 和 Liwei Wu。无监督异常检测与定位基准测试，2022年。6，7

概述

本补充材料包括：

- 有关硬件和软件包的实现细节（附录 A）；
- 异常分割的 P-AUROC 分数（附录 B）；
- MVTec-3D AD 所有类别的消融研究详细分数（附录 C）；
- MVTec-3D AD 所有类别的 Point Transformer 设置详细分数（附录D）；
- MVTec-3D AD 所有类别少样本设置的详细分数（附录 E）；
- 关于骨干网络选择的讨论（附录 F）；
- Eyecandies 所有类别的结果（附录 G）；
- MVTec-3D AD 所有类别的可视化结果（附录H）。

A. 实现细节

我们使用 Pytorch[1] 和 Scikit-Learn 软件包实现 M3DM[2]。特征提取器和记忆库算法基于 Pytorch 实现，我们使用 Scikit-Learn 软件包进行 OCSVM [29]。AUROC 的计算也依赖于 Scikit-Learn 软件包。所有实验均在单台 Nvidia Tesla V100 上运行，在完整设置下最多占用 50 GB 内存。

B. 用于分割的 P-AUROC

在主论文中，我们报告了异常分割的 AUPRO 分数。本节中，我们报告 P-AUROC 分数，以进一步验证我们方法的分割性能，如表I所示。我们主要将结果与 FPFH [17], PatchCore [26]和 AST[28][3]进行比较。对于多模态输入，我们取得了与 PatchCore + FPFH 方法相同的分数，并比 AST 高 1.6%。对于单一 RGB 输入，我们仍比 PatchCore 提高了 2%。对于三维分割，与主论文报告的 AUPRO 结果类似，我们的三维分割结果略低于基于 FPFH 的方法。我们认为，这同样是由主论文第 4.7 节讨论的标签与点云之间的偏差造成的。P-AUROC 是异常分割中的一项饱和指标，各方法之间的差异小于 AUPRO 上的差异。

[1]https://pytorch.org/ [2]https://scikit-learn.org/ [3]由于 AST [28] 在论文中只提供了平均分数，因此我们仅展示 AST 的平均分数。

C. 消融研究的详细结果

在主论文第 4.3 节中，我们针对 UFF、DLF 和多个记忆库开展了消融研究。本节报告 MVTec-3D AD 所有类别的详细消融研究结果。表II 和表III 分别展示了以下设置下的 I-AUPRO 和 AUPRO 分数：1）仅使用点云（$\mathcal{M}_{pt}$）信息；2）仅使用 RGB（$\mathcal{M}_{rgb}$）信息；3）使用单个记忆库（$\mathcal{M}_{fs}$），直接拼接 Point Transformer 特征和 RGB 特征；4）使用单个记忆库（$\mathcal{M}_{fs}$），通过 UFF 融合多模态特征；5）分别构建两个记忆库（$\mathcal{M}_{rgb}$、$\mathcal{M}_{pt}$），并直接将分数相加；6）分别构建两个记忆库（$\mathcal{M}_{rgb}$、$\mathcal{M}_{pt}$），并使用 DLF 得出最终结果；7）构建三个记忆库（$\mathcal{M}_{rgb}$、$\mathcal{M}_{pt}$、$\mathcal{M}_{fs}$）（我们的方法）。使用 UFF 后，泡沫、曲奇和桃子的表现相较于单一领域输入和不使用UFF的版本均有显著提升，这说明 UFF 促进了多模态特征之间的交互，并为异常检测和分割生成了有用信息。使用双记忆库后，胡萝卜、曲奇和土豆的分数有所提升，而 DLF 有助于提升电缆接头和轮胎等较难类别的表现。使用三个记忆库后，DLF 和 UFF 的大部分优势得以保留；我们的完整设置取得了最佳结果，这表明 DLF 和 UFF 相辅相成，共同实现了最佳性能。

D. PFA 分析的详细结果

在主论文第 4.4 节中，我们对 PFA 设置进行了实验。此处报告主论文表5的详细结果，其中包含各类别的分数，见表IV和表 V。PFA 设置如下：

- Point Transformer 的两个重要超参数：组数以及最远点采样期间每组的大小；组数决定了 Point Transformer 将提取多少个特征，而组大小等同于感受野的概念；
- 3D 异常检测实验采用原始点组特征：一个点组可以视为一个图像块，记忆库存储点组特征；检测方法与基于图像块的方法相同。为了获得分割预测，我们首先将点组特征投影到二维平面，再使用反距离插值计算每个像素的值。

我们发现，在某些类别（例如百吉饼、电缆接头、泡沫和绳子）中，直接在点组上计算异常具有一定优势，原因是

	方法	百吉饼	电缆腺体	胡萝卜	曲奇	木销	Foam	桃子	土豆	Rope	Tire	Mean
3D	FPFH [17]	0.994	0.966	0.999	0.946	0.966	0.927	0.996	0.999	0.996	0.990	**0.978**
	Ours	0.981	0.949	0.997	0.932	0.959	0.925	0.989	0.995	0.994	0.981	0.970
RGB	PatchCore [26]	0.983	0.984	0.980	0.974	0.972	0.849	0.976	0.983	0.987	0.977	0.967
	Ours	0.992	0.990	0.994	0.977	0.983	0.955	0.994	0.990	0.995	0.994	**0.987**
RGB+3D	AST [28]	-	-	-	-	-	-	-	-	-	-	0.976
	图像块 + FPFH [17]	0.996	0.992	0.997	0.994	0.981	0.974	0.996	0.998	0.994	0.995	**0.992**
	Ours	0.995	0.993	0.997	0.985	0.985	0.984	0.996	0.994	0.997	0.996	**0.992**

表 I. MVTec-3D AD [3] 数据集所有类别的异常分割 P-AUROC 得分。P-AUROC 是异常分割的一项饱和指标，不同方法之间的差异小于 AUPRO。

方法	记忆库	百吉饼	电缆腺体	胡萝卜	曲奇	木销	Foam	桃子	土豆	Rope	Tire	Mean
仅点云	$\mathcal{M}_{pt}$	0.941	0.651	0.965	0.969	0.905	0.760	0.880	0.974	0.926	0.765	0.874
仅 RGB	$\mathcal{M}_{rgb}$	0.944	0.918	0.896	0.749	0.959	0.767	0.919	0.648	0.938	0.767	0.850
不使用UFF	$\mathcal{M}_{fs}$	0.920	0.900	0.914	0.727	0.963	0.795	0.946	0.656	0.954	0.792	0.857
使用UFF	$\mathcal{M}_{fs}$	0.976	0.895	0.922	0.912	0.949	0.868	0.978	0.723	0.960	0.798	**0.898**
不使用DLF	$\mathcal{M}_{pt}, \mathcal{M}_{rgb}$	0.981	0.831	0.980	0.985	0.960	0.905	0.936	0.964	0.967	0.780	0.929
使用DLF	$\mathcal{M}_{pt}, \mathcal{M}_{rgb}$	0.980	0.880	0.975	0.965	0.947	0.910	0.943	0.927	0.958	0.840	**0.932**
Ours	$\mathcal{M}_{pt}, \mathcal{M}_{rgb}, \mathcal{M}_{fs}$	0.994	0.909	0.972	0.976	0.960	0.942	0.973	0.899	0.972	0.850	**0.945**

表 II. MVTec-3D AD 所有类别异常检测消融实验的详细 I-AUROC 分数。

经过最远点采样（FPS）后，原始点组特征包含了更多细微缺陷的信息。随着图像块变小，我们的二维平面点特征在检测细微缺陷方面表现更好。

E. 少样本设置的详细结果

在论文正文第 4.6 节中，我们评估了方法在少样本设置下的表现，所有类别的详细结果如表 VII 和表 VI所示。我们从每个类别中随机选取 10 张和 5 张图像作为训练数据，并在完整测试数据集上测试少样本模型。我们发现，即使在 10样本或 5样本设置下，我们的方法仍比一些非少样本方法具有更好的分割性能。在 50 样本设置下，我们发现某些类别的表现优于完整数据集版本（例如百吉饼和土豆），这表明记忆库构建算法仍有改进空间，我们将在未来的研究中讨论这一问题。

F. 骨干网络选择

特征提取器在异常检测中发挥着重要作用。本节探讨点云和 RGB图像的不同骨干网络设置。对于 RGB，我们比较了四种特征提取器设置：1）使用 ImageNet [11] 1K 预训练的 ViT-B/8 有监督骨干网络；2）使用 ImageNet 21K 预训练的 ViT-B/8 有监督骨干网络；3）通过 DINO[6]预训练的 ViT-S/8 自监督骨干网络；4）通过 DINO 预训练的 ViT-B/8 自监督骨干网络。对于点云 Transformer，我们比较了两种自监督预训练骨干网络：1）Point-Bert [34]；2）Point-MAE [21]。检测和分割结果分别展示在表 VIII 和表 IX中。结果表明，自监督预训练方法的表现优于监督方法；使用自监督方法预训练的小型骨干网络，表现也优于使用监督方法预训练的大型骨干网络。Point-MAE 的表现优于 Point-Bert。我们认为，原因在于 Point-MAE 需要重建比 Point-Bert 更多的点云细节，因此能够在异常检测中捕捉到微小缺陷。

G. Eyecandies 结果

我们注意到，最近出现了一个名为 Eyecan- dies 的新数据集 [5]，它提供了 10 类糖果的多模态信息，每类包含 1000 张训练图像和 50 张公开测试图像。源数据集为每个样本提供 6 张不同光照条件下的 RGB图像、一张深度图和一张法线图。本节中，我们将 Eyecandies 数据集转换为 M3DM 支持的格式。具体而言，我们使用环境光图像作为输入 RGB 数据；对于三维数据，我们首先利用内部参数将深度图像转换为点云，然后根据点坐标移除背景点。为提高计算效率，我们每类仅使用不到 400 个样本进行训练。我们尝试构建不同大小的记忆库（范围为 10 至 400 个样本），以找出在该数据集上效果最佳的记忆库。如表 X 和表 XI所示，我们报告了最佳的图像级 AUROC 和像素级 AUCROC 分数。与基线方法相比，我们在这两项指标上均取得了显著提升

方法	记忆库	百吉饼	电缆腺体	胡萝卜	曲奇	木销	Foam	桃子	土豆	Rope	Tire	Mean
仅点云	$\mathcal{M}_{pt}$	0.943	0.818	0.977	0.882	0.881	0.743	0.958	0.974	0.950	0.929	0.906
仅 RGB	$\mathcal{M}_{rgb}$	0.952	0.972	0.973	0.891	0.932	0.843	0.970	0.956	0.968	0.966	0.942
不使用UFF	$\mathcal{M}_{fs}$	0.951	0.971	0.974	0.893	0.935	0.855	0.972	0.958	0.969	0.967	0.944
使用UFF	$\mathcal{M}_{fs}$	0.963	0.964	0.978	0.930	0.946	0.896	0.974	0.966	0.972	0.972	**0.956**
不使用DLF	$\mathcal{M}_{pt}, \mathcal{M}_{rgb}$	0.968	0.925	0.979	0.914	0.909	0.948	0.975	0.976	0.967	0.965	0.953
使用DLF	$\mathcal{M}_{pt}, \mathcal{M}_{rgb}$	0.965	0.968	0.978	0.933	0.933	0.927	0.976	0.967	0.971	0.973	**0.959**
Ours	$\mathcal{M}_{pt}, \mathcal{M}_{rgb}, \mathcal{M}_{fs}$	0.970	0.971	0.979	0.950	0.941	0.932	0.977	0.971	0.971	0.975	**0.964**

表 III. MVTec-3D AD 所有类别的消融异常分割 AUPRO 分数详情。

S.G	N.G	采样	百吉饼	电缆腺体	胡萝卜	曲奇	木销	Foam	桃子	土豆	Rope	Tire	Mean
64	784	点组	0.933	0.579	0.854	0.843	0.874	0.748	0.761	0.863	0.989	0.483	0.793
128	1024	点组	0.945	0.690	0.905	0.925	0.897	0.809	0.854	0.888	0.991	0.509	0.841
64	784	8×8图像块	0.905	0.508	0.939	0.923	0.817	0.725	0.857	0.916	0.897	0.561	0.805
128	1024	8×8图像块	0.886	0.560	0.925	0.971	0.832	0.711	0.873	0.909	0.897	0.624	0.819
128	1024	4×4图像块	0.941	0.651	0.965	0.969	0.905	0.760	0.880	0.974	0.926	0.765	**0.874**

表 IV. 在纯 3D 设置下探索 Point Transformer 设置的详细图像级 AUROC 结果。S.G 表示每个点组的点数，N.G 表示点组总数。每个样本使用 1,024 个点组且每个点组包含 128 个点时，我们取得了最佳性能；与直接计算点组上的异常分数相比，基于二维平面图像块的方法需要较小的图像块大小才能获得较高性能。

RGB 设置和 RGB+3D 设置。以往的工作未报告 Eyecandies 数据集上的 AUPRO 分数。为供进一步研究参考，我们在表中提供了 M3DM 的这一分割性能指标分数。XII。

H. 可视化结果

本节展示了 MVTec-3D AD 数据集所有类别的更多异常分割结果。如图 I所示，我们展示了采用多模态输入时，本文方法和 PatchCore + FPFH 的热力图结果。与 PatchCore + FPFH 的结果相比，本文方法获得了更好的分割图。

S.G	N.G	采样	百吉饼	电缆腺体	胡萝卜	曲奇	木销	Foam	桃子	土豆	Rope	Tire	Mean
64	784	点组	0.906	0.709	0.942	0.854	0.869	0.681	0.871	0.906	0.943	0.447	0.813
128	1024	点组	0.956	0.812	0.964	0.895	0.892	0.644	0.965	0.974	0.966	0.891	0.896
64	784	8×8图像块	0.899	0.789	0.970	0.848	0.871	0.718	0.931	0.951	0.939	0.874	0.879
128	1024	8×8图像块	0.934	0.808	0.977	0.856	0.877	0.745	0.949	0.970	0.948	0.894	0.896
128	1024	4×4图像块	0.943	0.818	0.977	0.882	0.881	0.743	0.958	0.974	0.950	0.929	**0.906**

表 V. 在纯 3D 设置下探索 Point Transformer 设置的详细 AUPRO 结果。S.G 表示每个点组中的点数，N.G 表示点组总数。当每个样本包含 1024 个点组且每个点组包含 128 个点时，我们取得了最佳性能；与直接在点组上计算分割分数相比，基于二维平面图像块的方法需要采用较小的图像块大小才能获得较高性能。

方法	百吉饼	电缆腺体	胡萝卜	曲奇	木销	Foam	桃子	土豆	Rope	Tire	Mean
5样本	0.974	0.645	0.833	0.942	0.636	0.798	0.820	0.781	0.914	0.615	0.796
10样本	0.987	0.662	0.854	0.969	0.643	0.799	0.908	0.771	0.931	0.682	0.821
50样本	0.997	0.745	0.957	0.966	0.910	0.915	0.937	0.910	0.946	0.744	0.903
完整数据集	0.994	0.909	0.972	0.976	0.960	0.942	0.973	0.899	0.972	0.850	**0.945**

表VI. MVTec-3D AD所有类别的少样本I-AUROC。我们的方法在少样本设置下仍具有良好的异常检测性能。

方法	百吉饼	电缆腺体	胡萝卜	曲奇	木销	Foam	桃子	土豆	Rope	Tire	Mean
5样本	0.959	0.879	0.974	0.906	0.879	0.848	0.968	0.957	0.963	0.935	0.927
10样本	0.972	0.910	0.976	0.923	0.905	0.870	0.972	0.956	0.967	0.939	0.939
50样本	0.969	0.955	0.977	0.940	0.906	0.912	0.971	0.965	0.968	0.959	0.952
完整数据集	0.970	0.971	0.979	0.950	0.941	0.932	0.977	0.971	0.971	0.975	**0.964**

表VII。少样本AUPRO o 在 MVTec-3D AD 的所有类别中。我们的方法在少样本设置下仍具有更好的异常分割性能 or-，优于大多数非少样本 方法。

	方法	百吉饼	电缆腺体	胡萝卜	曲奇	木销	Foam	桃子	土豆	Rope	Tire	Mean
RGB	监督式 ImageNet 1K	0.793	0.729	0.774	0.709	0.723	0.601	0.607	0.606	0.605	0.556	0.670
RGB	监督式 ImageNet 21K	0.814	0.658	0.788	0.630	0.784	0.582	0.615	0.459	0.674	0.621	0.662
RGB	DINO ViT-S/8	0.933	0.865	0.898	0.786	0.878	0.759	0.902	0.520	0.898	0.748	0.819
RGB	DINO ViT-B/8	0.944	0.918	0.896	0.749	0.959	0.767	0.919	0.648	0.938	0.767	**0.850**
3D	Point-Bert	0.900	0.632	0.932	0.915	0.851	0.659	0.826	0.899	0.894	0.530	0.803
3D	Point-MAE	0.941	0.651	0.965	0.969	0.905	0.760	0.880	0.974	0.926	0.765	**0.874**

表VIII. MVTec-3D AD 异常检测的 I-AUROC分数 [3] 数据集采用不同骨干网络时的结果。对于 RGB 特征提取器，自监督骨干网络优于有监督骨干网络。

	方法	百吉饼	电缆腺体	胡萝卜	曲奇	木销	Foam	桃子	土豆	Rope	Tire	Mean
RGB	监督式 ImageNet 1K	0.844	0.842	0.892	0.681	0.842	0.568	0.765	0.865	0.915	0.871	0.808
RGB	监督式 ImageNet 21K	0.805	0.878	0.927	0.712	0.888	0.62	0.785	0.909	0.919	0.930	0.837
RGB	DINO ViT-S/8	0.948	0.973	0.971	0.906	0.947	0.788	0.972	0.954	0.964	0.949	0.937
RGB	DINO ViT-B/8	0.952	0.972	0.973	0.891	0.932	0.843	0.970	0.956	0.968	0.966	**0.942**
3D	Point-Bert	0.895	0.775	0.972	0.841	0.871	0.680	0.918	0.964	0.938	0.877	0.873
3D	Point-MAE	0.943	0.818	0.977	0.882	0.881	0.743	0.958	0.974	0.950	0.929	**0.906**

表 IX. 异常分割的 AUPRO 分数 不同骨干网络下 MVTec-3D AD [3]数据集的异常分割。对于 RGB 特征提取器，t he 自监督骨干网络优于 有监督骨干网络。

	方法	糖果 Cane	巧克力曲奇	巧克力果仁糖	糖衣杏仁	软糖 Bear	榛子松露	甘草糖沙子 wich	棒棒糖	棉花糖 棉花糖	胡椒-mint 糖果	Mean
3D	Ours	0.482	0.589	0.805	0.845	0.780	0.538	0.766	0.827	0.800	0.822	0.725
RGB	RGB [5]	0.527	0.848	0.772	0.734	0.590	0.508	0.693	0.760	0.851	0.730	0.701
	STFPM [32]	0.551	0.654	0.576	0.784	0.737	**0.790**	0.778	0.620	0.840	0.749	0.708
	PaDiM [9]	0.531	0.816	0.821	0.856	0.826	0.727	0.784	0.665	0.987	0.924	0.794
	Ours	**0.648**	**0.949**	**0.941**	**1.000**	**0.878**	0.632	**0.933**	**0.811**	**0.998**	**1.000**	**0.879**
RGB+3D	RGB-D [5]	0.529	0.861	0.739	0.752	0.594	0.498	0.679	0.651	0.838	0.75	0.689
	RGB-cD-N [5]	0.596	0.843	0.819	0.846	0.833	0.550	0.750	**0.846**	0.940	0.848	0.787
	Ours	**0.624**	**0.958**	**0.958**	**1.000**	**0.886**	**0.758**	**0.949**	0.836	**1.000**	**1.000**	**0.897**

表 X. Eyecandies [5] 数据集中所有类别的异常检测 I-AUROC分数。基线方法的结果来自 [5]。

	方法	糖果 Cane	巧克力曲奇	巧克力果仁糖	糖衣杏仁	软糖 Bear	榛子松露	甘草糖沙- wich	棒棒糖	棉花糖 -允许	胡椒-mint 糖果	Mean
3D	Ours	0.977	0.903	0.902	0.93	0.875	0.832	0.909	0.968	0.868	0.918	0.908
RGB	RGB [5]	**0.972**	0.933	**0.960**	0.945	0.929	0.815	0.855	0.977	0.931	0.928	0.925
	Ours	**0.956**	**0.979**	0.958	**0.998**	**0.976**	**0.941**	**0.977**	**0.986**	**0.997**	**0.988**	**0.976**
RGB+3D	RGB-D [5]	0.973	0.927	0.958	0.945	0.929	0.806	0.827	0.977	0.931	0.928	0.920
	RGB-cD-N [5]	**0.980**	0.979	**0.982**	0.978	0.951	0.853	0.971	0.978	0.985	0.967	0.962
	Ours	0.974	**0.987**	0.962	**0.998**	**0.966**	**0.941**	**0.973**	**0.984**	**0.996**	**0.985**	**0.977**

表 XI. Eyecandies 数据集所有类别的异常检测 P-AUROC 得分 [5] 。基线方法的结果来自 [5]。

方法	糖果 Cane	巧克力曲奇	巧克力果仁糖	糖衣杏仁	软糖 Bear	榛子松露	甘草糖沙子- wich	棒棒糖	棉花糖 -allow	胡椒-mint 糖果	Mean
点云	0.911	0.645	0.581	0.748	0.748	0.484	0.608	0.904	0.646	0.750	0.702
RGB	0.867	0.904	0.805	0.982	0.871	0.662	0.882	0.895	0.970	0.962	0.880
点云 + RGB	0.906	0.923	0.803	0.983	0.855	0.688	0.880	0.906	0.966	0.955	0.882

表 XII. Eyecandies [5] 数据集所有类别的异常检测 AUPRO 分数。

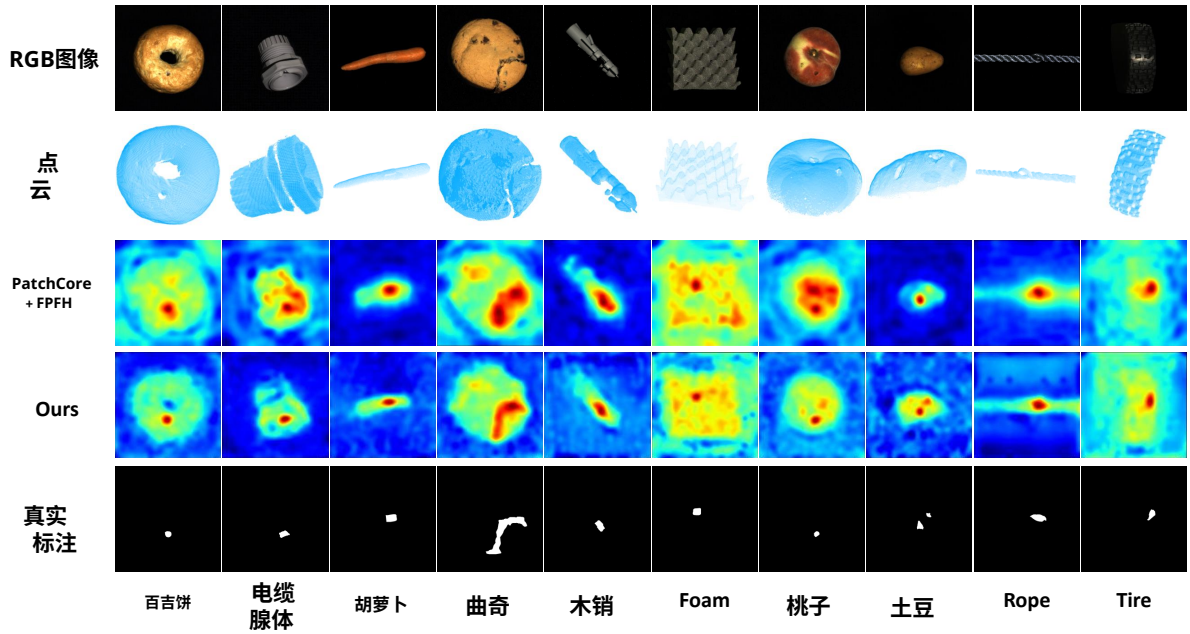


图 I. 我们的异常分割结果热力图（多模态输入）。与 PatchCore + FPFH 相比，我们的方法输出的分割区域更加准确。