# HarmoniAD：协调局部结构与全局语义的异常检测

Naiqi Zhang[1§] Chuancheng Shi[2§] Jingtong Dou[2] Wenhua Wu[2] Fei Shen[3] Jianhua Cao[1*]1天津科技大学 2悉尼大学 3新加坡国立大学§ 共同贡献

* 通讯作者：caojh@tust.edu.cn

arXiv:2601.00327v1 [cs.CV] 1 Jan 2026

**摘要—异常检测在工业产品质量检测中至关重要。未能检测出极小缺陷往往会造成严重后果。现有方法面临结构与语义权衡：结构导向模型（如基于频率的滤波器）对噪声敏感，而语义导向模型（如基于 CLIP 的编码器）常常会遗漏精细细节。为此，我们提出 HarmoniAD，一种频率引导的双分支框架。首先由 CLIP 图像编码器提取特征，然后将其转换到频域，最后通过自适应截止阈值将其分解为高频和低频路径，以互补地建模结构与语义。高频分支配备了细粒度结构注意力模块（FSAM），以增强纹理和边缘，从而检测小尺寸异常；低频分支则采用全局结构上下文模块（GSCM），以捕获长程依赖并保持语义一致性。这两个分支共同兼顾精细细节与全局语义。HarmoniAD 还采用多类别联合训练策略，在 MVTec-AD、VisA 和 BTAD 上开展的实验表明，该方法在灵敏度和鲁棒性方面均达到最先进水平。**

索引词—异常检测、频率引导学习、结构注意力、语义一致性

## 一、引言

异常检测 [1]–[3] 是计算机视觉中的一项基础任务，在工业检测、医学影像及各种安全关键系统中有着广泛应用。然而，现有方法存在结构与语义失衡的问题：结构导向模型往往对噪声过于敏感 [4], ，而语义导向模型通常无法检测到细微缺陷 [5], [6]。如图 1 所示，对热力图的系统分析揭示了两个反复出现的问题：虚假激活经常出现在正常背景区域，导致误报；相邻的微小缺陷则常被错误地合并到模糊区域中，使响应中心偏离实际异常。这些问题反映出对结构边界建模的根本局限，限制了复杂场景中对异常与背景以及不同异常之间的有效区分。因此，协调结构敏感性与语义一致性是提升异常检测能力的关键。

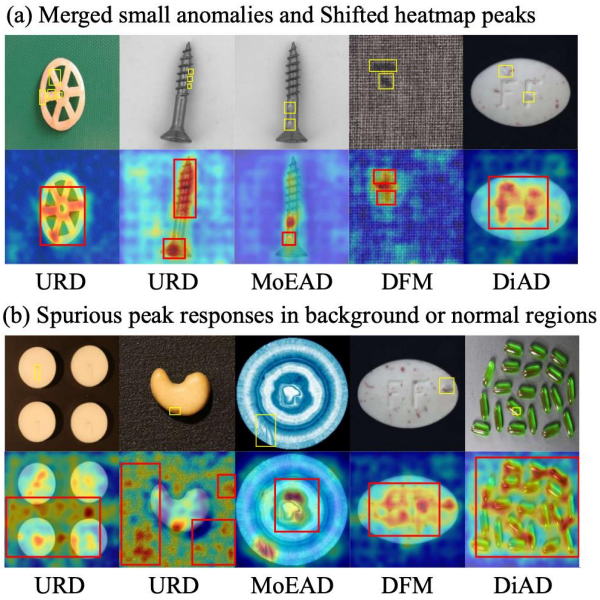



图 1：现有异常检测方法中的失败案例示例。黄色框表示真实异常，红色框表示假阳性。

现有的异常检测方法大致可分为两类：一类强调对局部结构的敏感性，另一类依赖高层语义表征。第一类侧重于捕捉细粒度线索，例如纹理、边缘和图块级的不规则性。典型方法包括对比式图块表征学习、连续记忆建模以及改进的变分自编码器变体。尽管这些技术能够有效突出局部异常，但它们往往会过拟合噪声，且缺乏全局上下文，从而在复杂场景中导致误报。

第二类方法建立在高层语义表征之上，通常借助 CLIP [7] 等大规模视觉语言模型。这些方法通过提示工程或语义对齐引入异常感知。例如，AA-CLIP [5] 通过构建异常感知文本锚点来优化跨模态对齐，从而提升零样本异常检测能力。类似地，KanoCLIP [6]融合知识驱动的提示学习与增强的跨模态融合，以更好地区分异常。尽管这些方法展现出出色的泛化能力和语义推理能力，却往往忽视细微的结构不规则性，例如小划痕或细粒度纹理偏差。

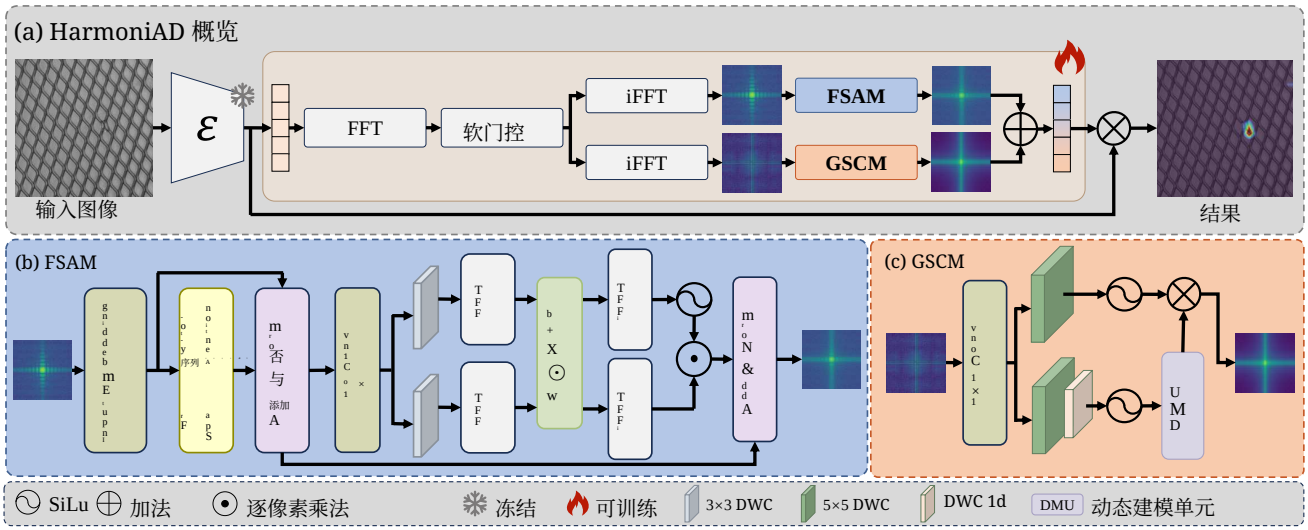



图 2：HarmoniAD 的整体框架。首先将 CLIP 图像嵌入转换到频域，并通过软门控将其分为高频流和低频流。随后，分别由细粒度结构注意力模块（FSAM）和全局结构上下文模块（GSCM）重建这些流，并将其作为最终表征进行感知。将感知表征与原始嵌入进行对比，以得出补丁级异常分数。

为解决这些局限性，我们提出 HarmoniAD，这是一种频率引导的双分支框架，可融合局部结构细节与全局语义上下文。频域划分将高频和低频成分分离，使结构纹理与语义依赖关系能够并行建模。两个分支协同工作，兼顾细粒度定位与语义连贯性，生成具有区分性且稳健的异常表征。这种结构与语义的协同作用生成了既精确又一致的热力图，并在多个异常检测基准数据集上取得了最先进的性能。

- 我们提出 HarmoniAD，这是一种频率引导的双分支框架，通过互补的高频和低频路径，将细粒度敏感性与全局语义一致性融为一体。
- 我们设计了细粒度结构注意力模块（FSAM），利用频率-空间交互增强对纹理和边缘的敏感性，以检测细微异常。
- 我们开发了全局结构上下文模块（GSCM），用于动态建模语义依赖关系，以确保全局一致性。

## II. 相关工作

### A. 传统异常检测

图像异常检测（IAD）旨在无监督或弱监督的设定下，识别偏离正常数据分布的样本或区域。以往工作主要遵循两种范式：一是表征或分布建模方法，如 PaDiM [8]，、SPADE [9], 和 PatchCore [10]，，通过分布估计或记忆库，在预训练视觉编码器的特征空间中建模正常性，并利用分布差异或最近邻距离定位异常；二是生成式或重建方法，包括自编码器和变分自编码器的各种变体，以及基于生成对抗网络的方法，它们将重建残差视为异常线索。其他互补方向包括结合自监督学习的合成异常增强，例如 CutPaste [11], ，以及融入重建与定位流程的基于扩散的去噪先验，例如 DiAD [12]。尽管这些方法在许多工业场景中表现出色，但它们往往依赖局部外观偏差，对于需要全局语义理解和上下文连贯性的异常可能不足，因此促使研究者采用 CLIP [7] 等高层语义先验。

### B. 基于*CLIP*的异常检测

CLIP 是一种大规模视觉语言模型，通过图文对齐进行预训练，可为异常检测提供具有迁移能力的语义表征。WinCLIP [13]和 PromptAD [14]等代表性方法通过引入类别特定文本提示，并利用相似度评分将其与图像特征进行匹配，从而实现无标签异常检测与定位；然而，它们的性能对提示的设计和措辞十分敏感。更重要的是，尽管基于 CLIP 的方法受益于语义对齐，从而提升了泛化能力，但它们通常不会显式地解耦或联合建模视觉编码器多个层中蕴含的局部结构线索与全局语义依赖。这往往导致一种明显的权衡：既要检测细微、细粒度的结构异常，又要保持语义一致性。

## III. 方法

### A. 概述

我们提出了 HarmoniAD，这是一种频率引导的双分支框架，可联合建模细粒度结构与

全局语义（图 2a）。冻结的 CLIP编码器提取高层特征，并将其投影到频域；可微分的软门控根据频率半径自适应地分离高频和低频分量，从而实现端到端的结构感知路由。与多尺度或多层级解耦方法不同，后者主要改变空间分辨率或感受野，同时让结构与语义纠缠在同一表示中；我们的频域划分则在单一嵌入中明确分离结构细节和全局语义成分，从而实现更可控的专门化。针对局部异常，细粒度结构注意力模块（FSAM）（图 2b）通过频率-空间注意力增强纹理和边缘敏感性；而配备 DMU 的全局结构上下文模块（GSCM）（图 2c）则捕获语义依赖，以维持全局一致性。

### *B.* 通过软门控自适应分离高频与低频

为避免硬阈值选择的不可微性，我们将截止点视为离散候选集 $\{r_m\}_{m=1}^{M}$ 上的潜在尺度变量，并定义吉布斯分布 $p_m = \mathrm{Softmax}_m\big(\kappa J(r_m)\big)$。随后，我们取其期望作为输入自适应边界，

$$c = \sum_{m=1}^{M} r_m\, p_m = \sum_{m=1}^{M} r_m\, \mathrm{Softmax}_m\big(\kappa J(r_m)\big). \tag{1}$$

此处 $\kappa > 0$ 充当逆温度，在近似点选择与分布平均之间进行权衡；因此，尺度划分成为数据条件推断，而非固定的设计选择。

### *C.* 细粒度结构注意力模块

细粒度结构注意力模块（FSAM）利用频域先验进行结构感知特征细化，从而增强局部异常检测能力。不同于仅在单一空间表征内重新加权特征的通用注意力，FSAM 通过频率到空间注意力（F2S注意力）和幅度调制注入由频率导出的结构先验，明确增强边界与纹理线索，以实现像素级定位。

**频率到空间注意力（F2S注意力）**我们通过四维偏移描述符$\Delta \mathbf{p}_{ij} = [x_i - x_j,\ y_i - y_j,\ |x_i - x_j|,\ |y_i - y_j|]$ 与$\mathbf{p}_i = (x_i, y_i)$，向注意力对数几率中添加一个标量相对偏置。设$\mathbf{e}_\theta \in \mathbb{R}^4$，$\beta_\theta(\Delta \mathbf{p}^{ij}) = \mathbf{e}_\theta^\top \Delta \mathbf{p}^{ij}$，以及$\mathbf{B} = [\beta_\theta(\Delta \mathbf{p}^{ij})]_{i,j} \in \mathbb{R}^{N_f \times N_s}$:

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{B}\Big)\mathbf{V}. \tag{2}$$

这里$\mathbf{Q} \in \mathbb{R}^{N_f \times d}$和$\mathbf{K},\ \mathbf{V} \in \mathbb{R}^{N_s \times d}$；$[\cdot]_{i,\,j}$对所有（$i,\ j$）配对进行索引。

**频域通道调制**。设 $X$ 为傅里叶变换的结果，其幅度为 $A = |X|$，相位为 $\Phi = \arg(X)$。非负掩码 $m \in \mathbb{R}_+^{C \times H \times W}$对幅度进行调制:

$$\widehat{A} = m \odot \sigma(A), \quad \widehat{X} = \widehat{A} \odot \tfrac{X}{A+\varepsilon}, \tag{3}$$

其中，$\sigma(\cdot)$为非线性函数，$\odot$ 表示阿达马积，$\varepsilon > 0$ 用于确保稳定性。通过对 $X$进行逆$\widehat{\ }$傅里叶变换得到输出，在保留相位的同时，通过幅度调制自适应地增强结构细节。

### *D.* 全局结构上下文模块

我们设计了一个轻量级模块，通过动态调制和坐标编码捕获长程依赖。与采用静态聚合的标准非局部模块或 Transformer 上下文块不同，GSCM 执行逐标记动态调制，并使用 DMU 将低频语义与上层流相结合，从而抑制伪激活，同时确保全局语义一致性。给定 $X = [x_1, \ldots, x_T] \in \mathbb{R}^{T \times C}$,$S \in \mathbb{R}^{T \times T}$和 $B^{\mathrm{rel}} \in \mathbb{R}^{T \times T}$，该模块首先计算动态亲和度:

$$S = \mathrm{Softmax}\Big(\frac{XW_q(XW_k)^\top}{\sqrt{r}} + B^{\mathrm{rel}}\Big), \tag{4}$$

其中 $W_q,\ W_k \in \mathbb{R}^{C \times r}$ 是可学习的，$r < C$，且 $B^{\mathrm{rel}} = \{b_{ij}^{\mathrm{rel}}\}$编码相对位置偏置。每个标记 $x_t$ 都会生成动态权重 $\Pi_t = \mathrm{diag}(\xi(W_\pi x_t))$ 和偏置 $\Gamma_t = W_\gamma x_t$，其中 $W_\pi,\ W_\gamma \in \mathbb{R}^{C \times C}$ 以及非负激活函数 $\xi(\cdot)$。随后得到聚合表示

$$Z = \mathcal{A}\big(S\,[(\Pi_1 x_1 + \Gamma_1), \ldots, (\Pi_T x_T + \Gamma_T)]\big), \tag{5}$$

其中 $\mathcal{A}(\cdot)$ 是非线性函数，且 $Z \in \mathbb{R}^{T \times C}$。我们引入动态建模单元（DMU），使下分支具备跨标记动态，并与上分支进行乘性耦合:

$$x_t = W_o\big(u_t \odot \mathrm{SiLU}(v_t + m_t)\big). \tag{6}$$

调制项通过门控进行控制:

$$m_t = g_t \odot s_{t-1} + (1 - g_t) \odot \phi\big(W_d[\,r_t \odot v_t\,]\big). \tag{7}$$

这里，$u_t$（上分支卷积特征）、$v_t$（下分支特征）和 $s_{t-1}$均属于 $\mathbb{R}^C$；$g_t = \sigma(W_g^{\mathrm{d}} v_t)$、$r_t = \sigma(W_r v_t)$；$W_{\cdot}$可学习，$\sigma$，$\phi$逐点，$\odot$ 为阿达马。最后，门控 $\mathrm{Gate} = \sigma(XW_{og}) \in \mathbb{R}^{T \times C}$和坐标编码 $\mathrm{Coord} = [W_c\rho(p_1)$。。。。，$W_c\rho(p_T)]$与 $Z$融合，以生成

$$Y = \mathrm{Gate} \odot Z + \mathrm{Coord}. \tag{8}$$

### *E.* 重建与损失监督

**重建**。令 $\hat{}\, F_{\mathrm{high}}$，$\hat{}F_{\mathrm{low}} \in \mathbb{R}^{C \times H \times W}$ 表示两个分支输出（第 III-B 节）。我们使用 $\mathcal{P}_h$ 和 $\mathcal{P}_l$ 对其进行融合，得到$[0, 1]$:

$$X_{\mathrm{recon}} = \mathcal{P}_h(\hat{F}_{\mathrm{high}}) + \mathcal{P}_l(\hat{F}_{\mathrm{low}}). \tag{9}$$

当 $\mathcal{P}_h = \mathcal{P}_l = \mathcal{I}$（恒等映射）时，融合退化为加权求和。

**损失监督**。总体目标由六个互补的损失项组成：针对正常图像和区域的余弦重建损失（$\mathcal{L}_n^{cos}$、$\mathcal{L}_{an}^{cos}$）、针对异常区域的余弦重建损失和推远损失（$\mathcal{L}_a^{cos}$、$\mathcal{L}_{far}$），以及空间感知对比损失（$\mathcal{L}_{con}$）和三元组损失（$\mathcal{L}_{tri}$）。其中，$\theta$ 表示模型参数，$\lambda$表示各损失项的加权系数。这些

组件共同增强了正常特征与异常特征的分离、聚类和定位。形式化地，

$$\mathcal{L}_{\text{total}} = \lambda_n \mathcal{L}_n^{\cos} + \lambda_a \mathcal{L}_a^{\cos} + \lambda_{con} \mathcal{L}_{con} + \lambda_{\text{an}} \mathcal{L}_{\text{an}}^{\cos} + \lambda_{\text{far}} \mathcal{L}_{\text{far}} + \lambda_{\text{tri}} \mathcal{L}_{\text{tri}} + \lambda_{\text{reg}} \|\theta\|_2^2. \tag{10}$$

## IV. 实验与分析

### A. 实现细节

**数据集**。 遵循 DiAD [12], ，我们采用相同的训练流程和数据划分策略进行多类别联合训练，以确保与先前方法的比较公平且一致。我们在三个广泛使用的工业异常检测基准数据集上评估 HarmoniAD：MVTec-AD [15]、VisA [16], 和 BTAD [17]。

**评估指标**。 我们使用图像级和像素级的 ROC 与精确率-召回率指标，将 HarmoniAD 与最先进方法进行比较。具体而言，图像级 ROC 和图像级 PR 用于评估图像级异常检测准确率，而像素级 ROC 和像素级 PR 用于评估像素级异常定位性能。

**超参数**。 我们使用 ViT-B/16 作为冻结的 CLIP骨干网络，输入为 $224 \times 224$，并在单张 NVIDIA A100- 40GB GPU 上使用 Adam 进行训练（批量大小为 36，学习率为 $1 \times 10^{-2}$）。

### B. 与最先进方法比较

我们在多类别联合训练设置下，将所提出的 HarmoniAD 与多种具有代表性的最先进方法进行定量和定性比较，包括 MoEAD [18],URD [19], DFM [20], DiAD [12], WinCLIP [13], Promp- tAD [14],KanoCLIP [6] , 而我们的方法在所有基准数据集上始终取得最佳结果。

**定量结果**。 为评估统一异常检测，我们在 MVTec-AD、VisA 和 BTAD 上进行了定量实验。如表 I 所示，我们的方法在所有数据集上始终表现最佳。具体而言，在 MVTec-AD 上，我们的方法取得了 58.3 的像素级 PR 分数，比最强竞争方法高出 9.3 分，展现出其在抑制错误激活和精确定位缺陷方面的卓越能力。在 VisA 上，我们的方法将图像级 ROC 从 93.1 提升至 94.3，表明其图像级异常检测更加可靠。在 BTAD 上也观察到类似的性能提升，我们的方法在全部四项指标上均取得最高分。这些提升表明，HarmoniAD 更好地平衡了结构敏感性与语义一致性。我们将性能提升归因于频率引导的双流专门化设计：该设计显式分离高频结构证据，以实现精确定位；同时分离低频语义连贯性，以减轻背景诱发的错误激活。多层级特征解耦或标准注意力/上下文模块通常不会强制实现这种分离。

**定性结果**。 如图 3 所示，我们将 HarmoniAD 与 MVTec-AD 上的多种最先进方法进行了比较。MoEAD 往往过度强调高频边缘和纹理，导致产生空心或环状响应；而 DFM 会产生集中的激活，却经常出现错位，并漏检细微缺陷。DiAD 提高了对小尺寸异常的敏感度，但存在大量假阳性；URD 虽然能够抑制背景噪声，却容易将相邻缺陷合并，并对复杂纹理产生误报。相比之下，HarmoniAD 的定位更加准确、紧凑，边界更清晰，对小缺陷的分离效果也更好。此外，其由粗到细的预测在图块级和像素级输出上均保持了良好的一致性。这些改进得益于自适应频率分离和双分支重建，能够更充分地利用 CLIP 语义特征。

### C. 消融研究

**软门控消融**。 为验证自适应高低频分离模块（SoftGate）的有效性，我们在 BTAD 数据集上开展消融研究，将其与采用固定阈值 $t \in \{0.3, 0.5, 0.7\}$ 的硬划分进行比较。如表 II 所示，固定阈值的性能对 $t$ 表现出明显的敏感性，而软门控在全部四项指标上均取得最佳结果，且始终优于表现最强的固定设置，尤其在精确率-召回率和 AUROC 上提升更为显著。这些结果表明，自适应软划分能够更稳健地分配高频和低频贡献，从而提升异常检测性能。

**FSAM 和 GSCM 消融**。 为验证细粒度结构注意力模块（FSAM）和全局结构上下文模块（GSCM）的有效性及互补性，我们在 BTAD 数据集上开展模块级消融研究（表 III）。结果突显了局部结构建模与全局语义约束的互补作用。在启用软门控的情况下，移除 FSAM 并保留 GSCM，可维持图像级判别能力（图像级 ROC = 93.4），但会显著降低像素级定位性能（像素级精确率-召回率 = 53.2），表明模型对细粒度异常的敏感性下降。相反，保留 FSAM 并移除 GSCM，可提升局部检测能力（像素级 ROC = 97.8），但会牺牲全局一致性（图像级 ROC = 92.6），说明背景干扰有所增加。两个模块同时启用时，模型取得最佳整体性能，像素级 ROC、图像级 ROC、像素级精确率-召回率和图像级 PR 分别达到 98.9、94.4、60.9 和 98.8。这些结果表明，FSAM 能增强细粒度结构敏感性；与此同时，GSCM 能抑制虚假激活并确保全局连贯性，二者结合构成了均衡且稳健的异常检测框架。

**F2S注意力的消融**。 我们在启用 FSAM 和 GSCM 的条件下，进一步通过 BTAD数据集上的消融研究评估频率到空间注意力（F2S注意力）的作用。如表IV和图4所示，停用 F2S注意力会导致所有指标的性能持续下降（像素级 ROC 98.0、图像级 ROC 93.5、像素级 PR 56.6、图像级 PR 95.7），并产生空间上弥散、噪声较多且边界模糊的异常响应。相比之下，启用 F2S注意力后性能得到明显提升，尤其是在像素级指标上，像素级 PR 和图像级 PR 分别提高了 4.3 和 3.1 个百分点，同时生成了更加紧凑、集中的激活图，与真实缺陷高度吻合。这些

表 I：**三种数据集上七种统一异常检测方法的比较**。每个数据集均通过四项指标进行评估：像素级 ROC、图像级 ROC、像素级 PR 和图像级 PR。所有指标均为越高越好。

| 方法 | MVTec-AD [15] | | | | VisA [16] | | | | BTAD [17] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 像素级 ROC | 图像级 ROC | P-PR | I-PR | 像素级 ROC | 图像级 ROC | P-PR | I-PR | 像素级 ROC | 图像级 ROC | P-PR | I-PR |
| MoEAD [18] | 97.0 | 97.7 | 43.8 | 97.9 | 98.7 | 93.1 | 34.2 | 93.7 | 97.1 | 92.3 | 51.3 | 98.2 |
| URD [19] | 95.8 | 90.8 | 47.4 | 96.7 | 97.0 | 91.5 | 33.9 | 93.7 | 98.5 | 92.4 | 59.3 | 98.1 |
| DFM [20] | 96.5 | 69.7 | 42.4 | 89.8 | 96.5 | 51.6 | 25.2 | 77.8 | 96.3 | 68.8 | 48.0 | 82.8 |
| DiAD [12] | 96.8 | 97.2 | 49.0 | 96.9 | 96.0 | 86.8 | 24.3 | 90.2 | 96.9 | 92.0 | 47.9 | 94.4 |
| WinCLIP [13] | 81.4 | 71.6 | 17.8 | 84.5 | 73.8 | 66.1 | 5.34 | 71.1 | 66.7 | 55.2 | 7.3 | 62.7 |
| PromptAD [14] | 95.4 | 91.4 | 49.3 | 95.8 | 96.7 | 85.5 | 27.8 | 87.5 | 96.5 | 90.0 | 55.5 | 94.2 |
| KanoCLIP [6] | 93.1 | 94.3 | – | – | 83.8 | 97.7 | – | – | 90.6 | 96.5 | – | – |
| **HarmoniAD（我们的方法）** | **98.0** | **98.0** | **58.3** | **99.5** | **98.9** | **94.3** | **44.2** | **95.6** | **98.9** | **94.4** | **60.9** | **98.8** |

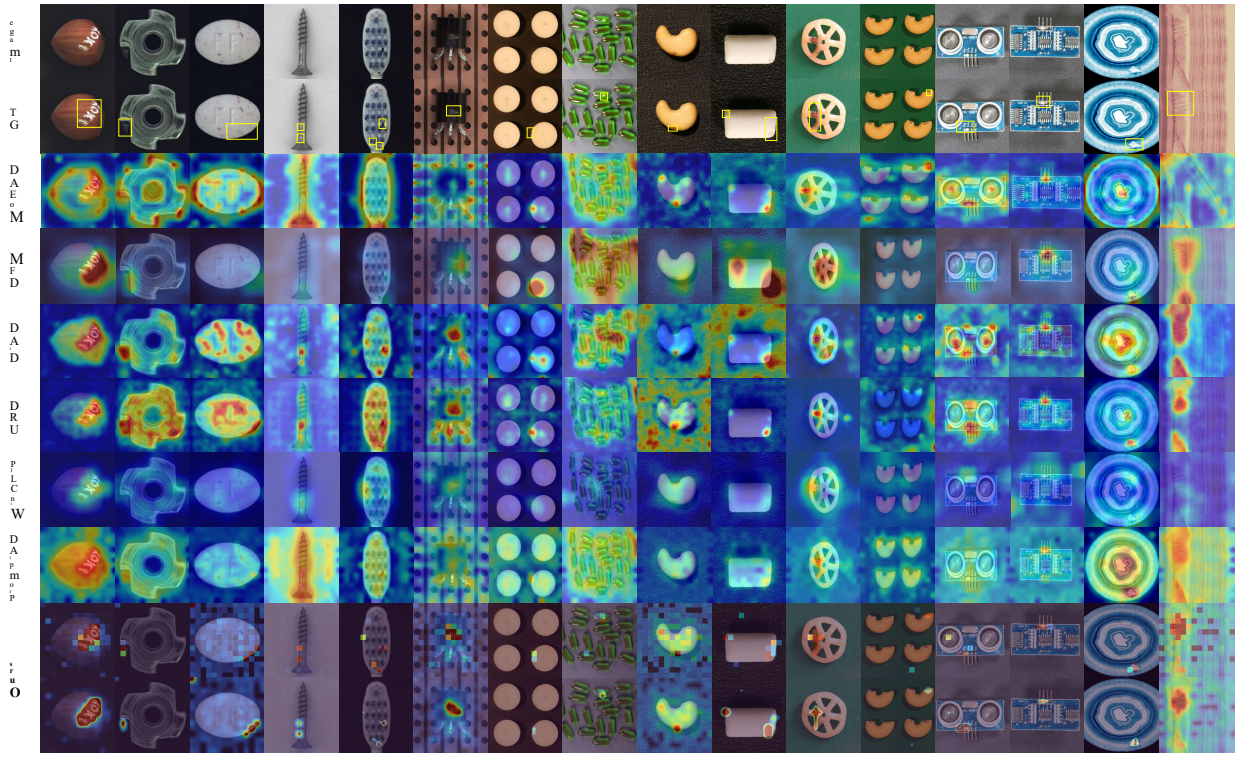


图 3：在多类别联合训练下，展示了从 MVTec-AD、VisA 和 BTAD 数据集中选取的类别的异常定位结果。每列代表一个测试样本。第一行显示真实缺陷（黄色框）。第 3–8 行为各对比方法生成的热力图。第 9–10 行展示我们方法生成的图块级和像素级热力图。我们的方法在异常区域产生较高响应，在其他区域产生较低响应，并且在具有挑战性的案例中优于对比方法。

表II：BTAD 上的软门控消融研究

| 阈值 | 像素级 ROC | 图像级 ROC | P-PR | I-PR |
|---|---|---|---|---|
| 0.3 | 96.8 | 90.1 | 52.5 | 91.8 |
| 0.5 | 98.2 | 91.7 | 54.4 | 94.0 |
| 0.7 | 97.7 | 91.3 | 52.9 | 92.5 |
| **软门控** | **98.9** | **94.4** | **60.9** | **98.8** |

表 III：**BTAD 上的消融**。FSAM 和 GSCM 的消融（FSAM 默认使用 F2SAttn）。

| FSAM | GSCM | 像素级 ROC | 图像级 ROC | P-PR | I-PR |
|---|---|---|---|---|---|
| ✗ | ✓ | 97.4 | 93.4 | 53.2 | 90.5 |
| ✓ | ✗ | 97.8 | 92.6 | 52.1 | 92.6 |
| ✓ | ✓ | **98.9** | **94.4** | **60.9** | **98.8** |

定量和定性结果表明，F2S注意力能够有效利用频域线索来引导空间注意力，在保持全局一致性的同时增强细粒度异常定位。

表IV：**BTAD 上的消融研究**。在保持 FSAM 和 GSCM 启用的同时启用 F2S 注意力的效果。

| F2S 注意力。 | 像素级 ROC | 图像级 ROC | P-PR | I-PR |
|---|---|---|---|---|
| ✗ | 98.0 | 93.5 | 56.6 | 95.7 |
| ✓ | **98.9** | **94.4** | **60.9** | **98.8** |

*D.* 参数敏感性分析

为评估我们的方法对不同损失项权重的鲁棒性，我们在 BTAD数据集上进行超参数敏感性分析：逐一改变各损失权重，同时保持其他权重不变。图5展示了图像级和像素级 AUROC 的性能变化趋势。结果表明，在较宽的损失权重范围内，性能均保持稳定，这说明所提出的框架无需仔细进行超参数调优。对于每个损失项，性能均在所选默认值附近呈现明显但

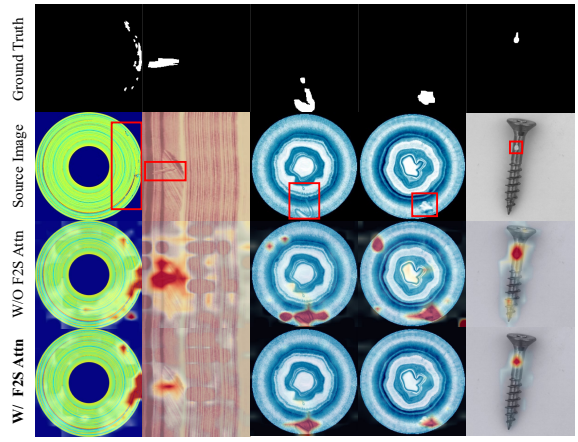



图 4：F2S注意力的消融研究。红框表示真实异常。

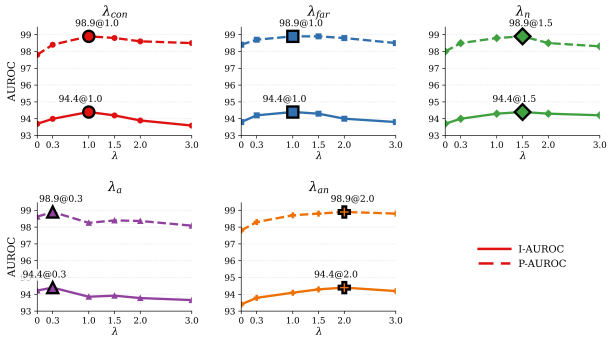



图 5：BTAD 上核心损失权重的敏感性分析。

平滑的峰值，而偏离这一最优值只会导致性能略有下降。尤其是，图像级和像素级 AUROC 曲线均呈现一致的趋势，这表明各损失组件相互补充，且权重配置均衡。

## V. 结论

总之，我们提出了 HarmoniAD，这是一种用于异常检测的频率引导双分支框架。通过自适应频率解耦，该框架能够联合捕获局部细节与全局语义，从而实现准确且高效的异常定位。大量基准数据集上的实验验证了其最先进的性能，也证实了频域结构建模的有效性。这些结果还表明，频域结构线索为异常表征提供了一种有理论依据且可解释的信号。未来，我们将为 HarmoniAD 加入时序建模，以用于视频异常检测。

## 参考文献

[1] Xiaolu Chen、Haote Xu、Chenghao Deng、Xiaotong Tu、Xinghao Ding 和 Yue Huang，“用于异常检测与定位的隐式前景引导网络”，载于 *ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2024，第 2970–2974 页。

[2] YeongHyeon Park、Sungho Kang、Myung Jin Kim、Hyeonho Jeong、Hyunkyu Park、Hyeong Seok Kim 和 Juneho Yi，《提升异常检测性能的神经网络训练策略：重构损失放大的视角》，载于 *ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2024年，第 5165–5169 页。[3] Debarpan Bhattacharya、Sumanta Mukherjee、Chandramouli Kamanchi、Vijay Ekambaram、Arindam Jati 和 Pankaj Dayama，《迈向对时间序列异常检测器的无偏评估》，载于*ICASSP 2025- 2025 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2025年，第 1–5 页。[4] Mingqing Wang、Jiawei Li、Zhenyang Li、Chengxiao Luo、Bin Chen、Shu-Tao Xia 和 Zhi Wang，《利用局部敏感 VQVAE 和全局敏感 Transformer 进行无监督异常检测》，载于 *2023*年 *IEEE* 图像处理国际会议 *(ICIP)*，2023年，第 1080–1084 页。[5] Wenxin Ma、Xu Zhang、Qingsong Yao、Fenghe Tang、Chenxu Wu、Ying- tai Li、Rui Yan、Zihang Jiang 和 S Kevin Zhou，《Aa-clip：通过异常感知 CLIP 增强零样本异常检测》，载于 计算机视觉与模式识别会议论文集，2025年，第 4744–4754 页。[6] Chengyuan Li、Suyang Zhou、Jieping Kong、Lei Qi 和 Hui Xue，《Kan- oclip：通过知识驱动的提示学习和增强的跨模态融合实现零样本异常检测》，载于 *ICASSP 2025- 2025 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2025年，第 1–5 页。[7] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等，《通过自然语言监督学习可迁移视觉模型》，载于 国际机器学习会议。PMLR，2021年，第 8748–8763 页。[8] 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。Springer，2021年，第 475–489 页。[9] Jinsung Yoon、Kihyuk Sohn、Chun-Liang Li、Sercan O Arik 和 Tomas Pfister，《SPADE：分布失配下的半监督异常检测》，*arXiv* 预印本 *arXiv:2212.00173*，2022年。[10] Karsten Roth、Latha Pemula、Joaquin Zepeda、Bernhard Sch¨olkopf、Thomas Brox 和 Peter Gehler，《迈向工业异常检测中的全面召回》，载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第 14318–14328 页。[11] Chun-Liang Li、Kihyuk Sohn、Jinsung Yoon 和 Tomas Pfister，《Cut-paste：用于异常检测与定位的自监督学习》，载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第 9664–9674 页。[12] Haoyang He、Jiangning Zhang、Hongxu Chen、Xuhai Chen、Zhishan Li、Xu Chen、Yabiao Wang、Chengjie Wang 和 Lei Xie，《一种基于扩散的多类别异常检测框架》，载于 *AAAI*人工智能会议论文集，2024年，第 38 卷，第 8472–8480 页。[13] Jongheon Jeong、Yang Zou、Taewan Kim、Dongqing Zhang、Avinash Ravichandran 和 Onkar Dabeer，《WinCLIP：零样本/少样本异常分类与分割》，载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，2023年，第 19606–19616 页。[14] Yiting Li、Adam Goodge、Fayao Liu 和 Chuan-Sheng Foo，《PromptAD：使用文本提示进行零样本异常检测》，载于 *IEEE/CVF*冬季计算机视觉应用会议论文集，2024年，第 1093–1102 页。[15] Paul Bergmann、Michael Fauser、David Sattlegger 和 Carsten Ste- ger，《MVTec AD——用于无监督异常检测的综合真实世界数据集》，载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，2019年，第 9592–9600 页。[16] Yang Zou、Jongheon Jeong、Latha Pemula、Dongqing Zhang 和 Onkar Dabeer，《用于异常检测与分割的找不同自监督预训练》，*arXiv* 预印本 *arXiv:2207.14315*，2022年。[17] Subarna Tripathi、Sandesh Shetty、Xiaoli Z Fern 和 Raviv Raich，《BTAD：用于工业检测的 BTech异常检测数据集》，*arXiv* 预印本 *arXiv:2012.10408*，2020年。[18] Shiyuan Meng、Wenchao Meng、Qihang Zhou、Shizhong Li、Weiye Hou 和 Shibo He，《MoEAD：一种用于多类别异常检测的参数高效模型》，载于欧洲计算机视觉会议。Springer，2024年，第 345–361 页。

[19] Xinyue Liu、Jianyuan Wang、Biao Leng 和 Shuo Zhang，《释放反向蒸馏在异常检测中的潜力》，载于 *AAAI*人工智能会议论文集，2025年，第 39 卷，第 5640–5648 页。[20] Ibrahima J. Ndiour、Nilesh A. Ahuja、Utku Genc 和 Omesh Tickoo，《Fre：一种快速异常检测与分割方法》，载于 英国机器视觉会议 *(BMVC)*，2023年。



这些附录提供了支持和拓展正文的补充细节。附录A展示了更多实验结果和消融研究。附录C讨论了常见问题。附录C介绍了我们工作的局限性。

### *A.* 不同异常区域尺度下的定性结果

本附录展示了不同异常区域尺度（极小 / 小 / 中 / 大）下的补充定性可视化结果。对于每个尺度组，我们提供输入图像（图像）、真实掩码（真实值）以及两种输出粒度（图块和像素）下预测的热力图。为确保公平比较，所有可视化结果均遵循与正文相同的流程（包括归一化、上采样和一致的颜色映射）。

### *B.* 跨域评估

我们进一步评估跨域泛化能力：使用在 BTAD、MVTec-AD 和 VisA 上联合训练的一组权重进行零样本推理。在不针对任何数据集进行适配的情况下，我们将训练好的模型应用于域外基准数据集（MPDD 和 WFDD），以及从真实工业生产中收集的异常样本，并额外纳入一个来自农业的新语义类别（蓝莓缺陷）。图7展示了具有代表性的跨域热力图：第一行展示了 MPDD 和 WFDD 上的结果，最右侧示例对应蓝莓缺陷；第二行则完全由我们收集的真实世界工业异常设备/组件组成。尽管存在明显的分布偏移、训练中未见过的类别，以及训练期间未观察到的异常模式，模型通常仍能生成合理的异常响应，其中较高的激活集中在疑似缺陷区域附近，并提供较为有效的定位线索。值得注意的是，在蓝莓样本中，模型在一定程度上仍能突出显示异常表面区域，这表明其具备超越训练期间所见工业领域的非平凡迁移能力。总体而言，这些定性结果表明，模型在纯零样本设置下具备一定程度的跨域稳健性，而更全面的定量分析将留待未来工作。

### *C.* 推理延迟

图8报告了延迟方面的推理吞吐量。我们的方法取得了与轻量高效基线方法相当的吞吐量，同时显著快于此次比较中的其他基于CLIP的方法，这表明我们的方法在保持高效率的同时，并未引入明显的速度开销。

▷***Q1.*** 为什么我们报告精确率-召回率和 ***AUROC***，而不是 ***AUPRO***？

AUROC 和精确率–召回率（PR）是不依赖阈值的基于排序的指标，可在统一协议下同时对图像级和像素级结果进行计算。AUROC 衡量正常分数分布与异常分数分布之间的整体可分性，而精确率–召回率（PR）

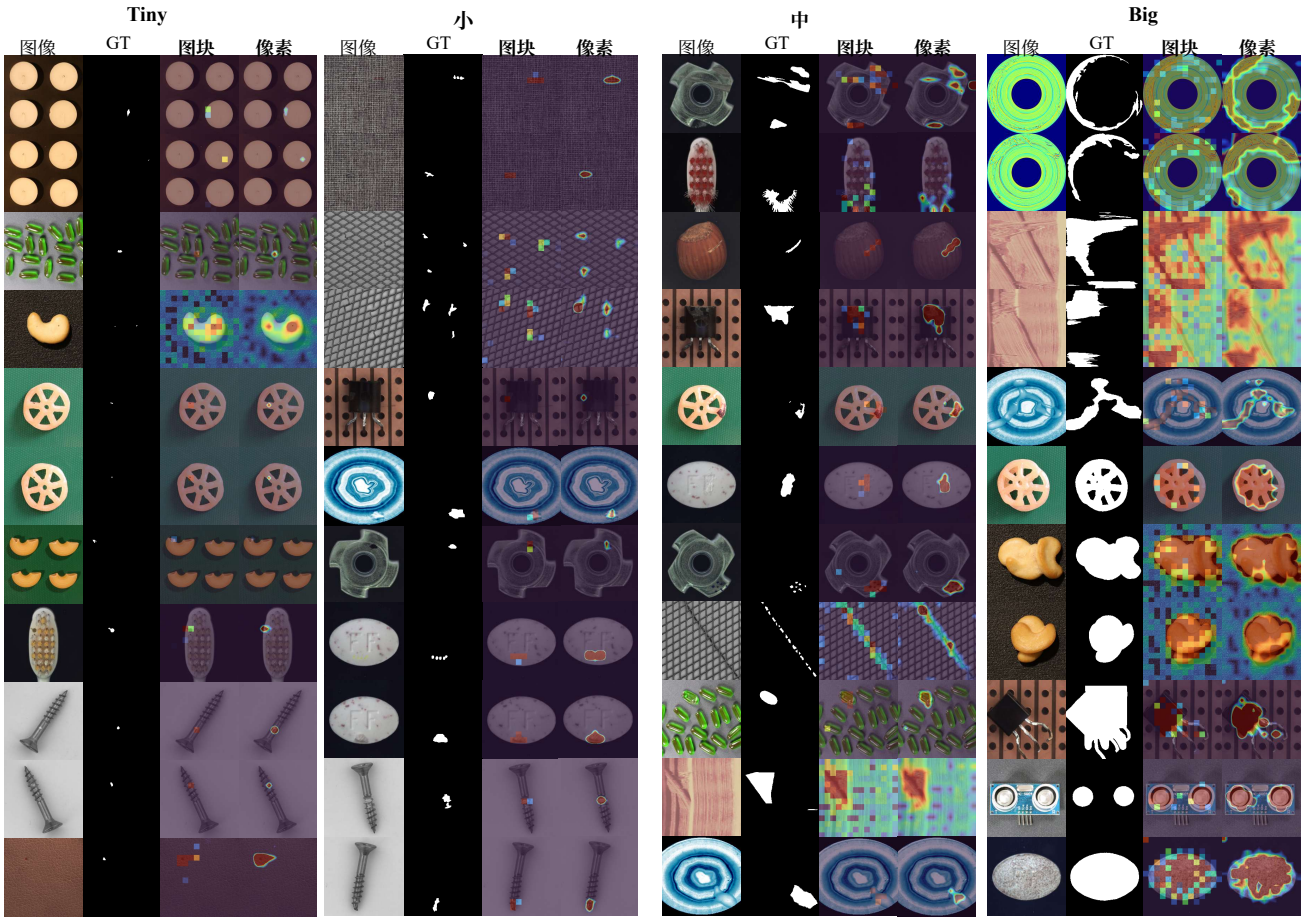



图6：不同异常区域尺度下的热力图。各列按异常区域尺度（极小、小、中、大）分组。每组内的列依次为图像 / 真实值 / 图块 / 像素。图块表示图块级输出热力图（上采样至图像空间），像素表示像素级输出热力图。热力图颜色使用与正文相同的颜色映射来表示异常分数。

在异常稀疏性严重时对假阳性更加敏感，能够更好地体现精确率–召回率权衡。相比之下，AUPRO 需要将异常图通过阈值化转换为连通区域，并在指定的假阳性率范围内对性能进行积分。它对阈值、连通区域的定义以及平滑等后处理操作高度敏感，同时也受区域形态影响；这些因素可能使评估设计选择与方法贡献相互纠缠，并低估方法在小尺度异常和边界异常上的改进。

***Q2.*** 为什么消融研究只在 ***BTAD*** 上进行，而不在其他数据集上进行？

我们在全部三个数据集上对 HarmoniAD 进行基准测试，但仅在 BTAD 上开展系统性消融，因为它更有助于识别各项设计选择的独立影响。MVTec AD 和 VisA 是规模更大、类别更多样的基准数据集，且都包含图像级和像素级标注。然而，近期在这些数据集上取得的工业异常检测结果往往已接近上限，这会压缩切换组件所造成的可观测差异，使归因变得不那么可靠。在我们的设置中，BTAD 提供了更大的提升空间，并且对架构变更表现出更清晰的敏感性，从而带来更高的信噪比和更可靠的组件级验证。

***Q3.*** 为什么我们冻结 ***CLIP*** 骨干网络而不对其进行微调？***HarmoniAD*** 对骨干网络选择的敏感程度如何？

我们将骨干网络视为稳定的通用特征提取器，并将主要性能提升归因于我们的频率分解和所提出的结构语义协调模块。冻结 CLIP ViT 有两个目的。这显著降低了训练成本，并减轻了在某些类别的有限数据上进行微调时经常出现的过拟合风险。此外，我们还测试了其他基于 ViT 的预训练骨干网络，例如 DINOv2，发现指标差异很小，这表明 HarmoniAD 对特定骨干网络并不十分敏感。综合考虑准确率、推理速度和内存占用，冻结的 CLIP 提供了最佳的整体成本效益，因此被用作默认配置。

所提出的方法针对单帧图像表示而设计，并未引入显式的时间建模。因此，我们没有开展视频异常检测实验，因为处理演变异常、正常运动和动态背景时，时间依赖关系和一致性

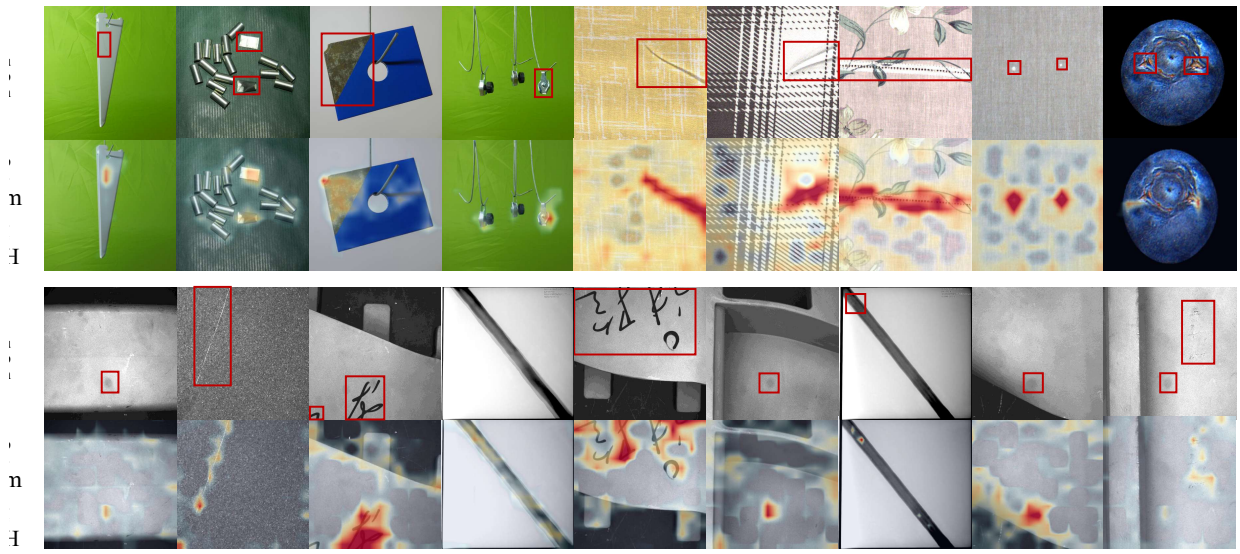


图 7：跨域零样本推理与定位。异常区域用红色边界框标出。对于每个样本，上方图像显示输入图像，下方图像显示零样本推理热力图。

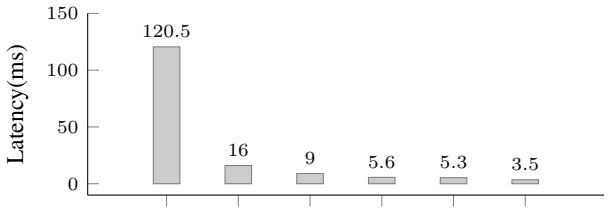



图 8：延迟比较。

往往至关重要。未来，我们将通过引入时间关系建模和时间一致性正则化，将基于频率的结构语义划分扩展至时空场景，旨在构建专用的视频异常检测框架并开展评估。