DiAD: A Diffusion-based Framework for Multi-class Anomaly Detection

Haoyang He[1*] **Jiangning Zhang**[2*], **Hongxu Chen**[1], **Xuhai Chen**[1], **Zhishan Li**[1], **Xu Chen**[2], **Yabiao Wang**[2], **Chengjie Wang**[2], **Lei Xie**[1†]

[1]Zhejiang University [2]Youtu Lab, Tencent

Abstract

Reconstruction-based approaches have achieved remarkable outcomes in anomaly detection. The exceptional image reconstruction capabilities of recently popular diffusion models have sparked research efforts to utilize them for enhanced reconstruction of anomalous images. Nonetheless, these methods might face challenges related to the preservation of image categories and pixel-wise structural integrity in the more practical multi-class setting. To solve the above problems, we propose a ***Di***fusion-based ***A***nomaly ***D***etection (**DiAD**) framework for multi-class anomaly detection, which consists of a pixel-space autoencoder, a latent-space *Semantic-Guided* (SG) network with a connection to the stable diffusion's denoising network, and a feature-space pre-trained feature extractor. Firstly, The SG network is proposed for reconstructing anomalous regions while preserving the original image's semantic information. Secondly, we introduce *Spatial-aware Feature Fusion* (SFF) block to maximize reconstruction accuracy when dealing with extensively reconstructed areas. Thirdly, the input and reconstructed images are processed by a pre-trained feature extractor to generate anomaly maps based on features extracted at different scales. Experiments on MVTec-AD and VisA datasets demonstrate the effectiveness of our approach which surpasses the state-of-the-art methods, ***e.g.***, achieving 96.8/52.6 and 97.2/99.0 (AUROC/AP) for localization and detection respectively on multi-class MVTec-AD dataset. Code will be available at https://lewandofskee.github.io/projects/diad.

Introduction

Anomaly detection is a crucial task in computer vision and industrial applications (Tao et al. 2022; Salehi et al. 2022; Liu et al. 2023), which goal of visual anomaly detection is to determine anomalous images and locate the regions of anomaly accurately. Existing anomaly detection models (Liznerski et al. 2021; Yi and Yoon 2020; Yu et al. 2021) mostly correspond to one class, which requires a large amount of storage space and training time as the number of classes increases. Therefore, there is an urgent need for an unsupervised multi-class anomaly detection model that is robust and stable.

The current mainstream unsupervised anomaly detection methods can be divided into three categories: synthesizing-based (Zavrtanik, Kristan, and Skočaj 2021a; Li et al. 2021), embedding-based (Defard et al. 2021; Roth et al. 2022; Xie et al. 2023) and reconstruction-based (Liu et al. 2022; Liang et al. 2023) methods. The central concept of the reconstruction-based method is that during the training phase, the model only learns from normal images. During the testing phase, the model reconstructs abnormal images into normal ones using the trained model. Therefore, by comparing the reconstructed image with the input image, we can determine the location of anomalies. Traditional reconstruction-based methods, including AEs (Zavrtanik, Kristan, and Skočaj 2021b), VAEs (Kingma and Welling 2022), and GANs (Liang et al. 2023; Yan et al. 2021) can learn the distribution of normal samples and reconstruct abnormal regions during the testing phase. However, these models have limited reconstruction capabilities and cannot reconstruct complicated textures and objects well, especially large-scale defects or disappearances as shown in Figure 1. Hence, models with stronger reconstruction capability are required to effectively tackle multi-class anomaly detection.

Recently, the diffusion models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Zhang and Agrawala 2023) have demonstrated their powerful image-generation capability. However, directly using current mainstream diffusion

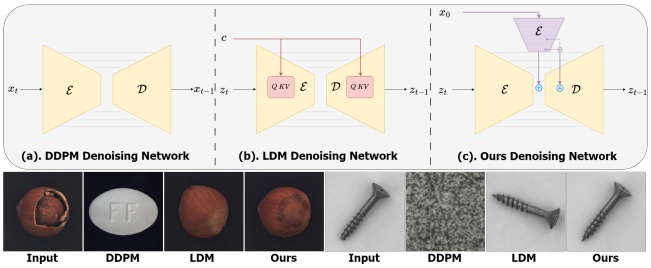

Figure 1: A analysis of different diffusion models for multi-class anomaly detection. The image above shows various denoising network architectures, while the images below demonstrate the results reconstructed by different methods for the same input image. ***a)*** DDPM suffers from categorical errors. ***b)*** LDM exhibits semantic errors. ***c)*** Our approach effectively reconstructs the anomalous regions while preserving the semantic information of the original image.

*Equal contribution.

†Corresponding author.

DiAD：一种基于扩散的多类别异常检测框架

何昊阳[1*] 张江宁[2*]、陈鸿旭[1]、陈旭海[1]、李志山[1]、陈旭[2]、王亚彪[2]、王成杰[2]、谢磊[1†] [1]浙江大学 [2]腾讯优图实验室

摘要

基于重建的方法在异常检测领域取得了显著成果。近期流行的扩散模型因其卓越的图像重建能力，引发了利用其增强异常图像重建的研究热潮。然而，这些方法在更实际的多类别场景中，可能面临图像类别保持和像素级结构完整性的挑战。为解决上述问题，我们提出了一种基于***Di***融合的***A***异常***D***检测（DiAD）框架用于多类别异常检测，该框架包含像素空间自编码器、与稳定扩散去噪网络相连的潜在空间*Semantic-Guided* (SG)网络，以及特征空间的预训练特征提取器。首先，SG网络旨在重建异常区域的同时保持原始图像的语义信息。其次，我们引入*Spatial-aware Feature Fusion* (SFF)模块以在处理大范围重建区域时最大化重建精度。第三，通过预训练特征提取器处理输入图像与重建图像，基于多尺度提取的特征生成异常图。在MVTec-AD和VisA数据集上的实验证明了我们方法的有效性，其在多类别MVTec-AD数据集上以96.8/52.6和97.2/99.0（AUROC/AP）的指标分别超越现有最优方法，实现了定位与检测性能的提升。代码将在https://lewandofskee.github.io/projects/diad发布。

引言

异常检测是计算机视觉和工业应用中的一项关键任务（Tao等人，2022；Salehi等人，2022；Liu等人，2023），其目标在于准确识别异常图像并定位异常区域。现有的异常检测模型（Liznerski等人，2021；Yi和Yoon，2020；Yu等人，2021）大多针对单一类别设计，随着类别数量的增加，会占用大量存储空间和训练时间。因此，迫切需要一种鲁棒且稳定的无监督多类别异常检测模型。

当前主流的无监督异常检测方法可分为三类：合成-

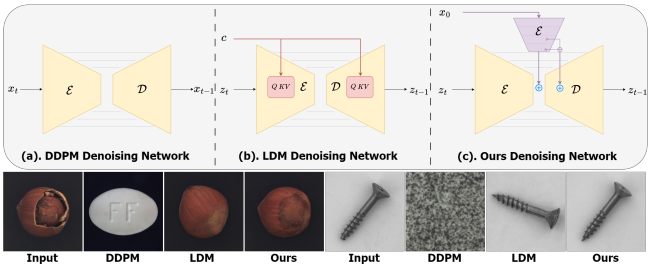

图1：针对多类异常检测的不同扩散模型分析。上图展示了各种去噪网络架构，而下图则展示了不同方法对同一输入图像的重建结果。***a)*** DDPM存在类别错误。***b)*** LDM表现出语义错误。***c)*** 我们的方法有效重建了异常区域，同时保留了原始图像的语义信息。

基于（Zavrtanik、Kristan和Skočaj 2021a；Li等人2021）、基于嵌入（Defard等人2021；Roth等人2022；Xie等人2023）以及基于重建（Liu等人2022；Liang等人2023）的方法。基于重建的方法的核心概念是：在训练阶段，模型仅从正常图像中学习；在测试阶段，模型利用训练好的模型将异常图像重建为正常图像。因此，通过比较重建图像与输入图像，我们可以确定异常的位置。传统的基于重建方法，包括自编码器（AEs）（Zavrtanik、Kristan和Skočaj 2021b）、变分自编码器（VAEs）（Kingma和Welling 2022）以及生成对抗网络（GANs）（Liang等人2023；Yan等人2021），能够学习正常样本的分布并在测试阶段重建异常区域。然而，这些模型的重建能力有限，无法很好地重建复杂的纹理和物体，特别是如图1所示的大规模缺陷或缺失情况。因此，需要具备更强重建能力的模型来有效应对多类别异常检测任务。

最近，扩散模型（Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Zhang and Agrawala 2023）已展现出其强大的图像生成能力。然而，直接使用当前主流的扩散

*Equal contribution.
†Corresponding author.

models cannot effectively address multi-class anomaly detection problems. 1) For the Denoising Diffusion Probabilistic Model (DDPM) (Ho, Jain, and Abbeel 2020) in Fig. 1-(a), when performing the multi-class setting, this method may encounter issues with misclassifying generated image categories. The reason is that after adding T timesteps noise to the input image, the image has lost its original class information. During inference, denoising is performed based on this Gaussian noise-like distribution, which may generate samples belonging to different categories. 2) Latent Diffusion Model (LDM) (Rombach et al. 2022) has an embedder as a class condition as shown in Fig. 1-(b), which does not exist the problem of misclassification found in DDPM. However, LDM still cannot address the issue of semantic loss in generated images. LDM is unable to simultaneously preserve the semantic information of the input image while reconstructing the anomalous regions. For example, they may fail to maintain direction consistency with the input image in terms of objects like screws and hazelnuts, as well as exhibit substantial differences from the original image in terms of texture class images.

To address the aforementioned problems, we propose a diffusion-based framework, DiAD, for multi-class anomaly detection and localization, illustrated in Fig. 2, which comprises three components: a pixel space autoencoder, a latent space denoising network and a feature space ImageNet pre-trained model. To effectively maintain consistent semantic information with the original image while reconstructing the location of anomalous regions, we propose the **S**emantic-**G**uided (SG) network with a connection to the Stable Diffusion (SD) denoising network in LDM. To further enhance the capability of preserving fine details in the original image and reconstructing large defects, we propose the **S**patial-aware **F**eature **F**usion (SFF) block to integrate features at different scales. Finally, the reconstructed and input images are passed through a pre-trained model to extract features at different scales and compute anomaly scores. We summarize our contributions as follows:

- We propose a novel diffusion-based framework DiAD for multi-class anomaly detection, which firstly tackles the problem of existing denoising networks of diffusion-based methods failing to correctly reconstruct anomalies.
- We construct an SG network connecting to the SD denoising network to maintain consistent semantic information and reconstruct the anomalies.
- We propose an SFF block to integrate features from different scales to further improve the anomaly reconstruction ability.
- Abundant experiments demonstrate the sufficient superiority of DiAD over SOTA methods, ***e.g.,*** we surpass the multi-class anomaly detection diffusion-based method by 20.6↑/ 11.7↑ in pixel/image AUROC and non-diffusion method by 9.2↑ in pixel-AP and 0.7↑ in image-AUROC on MVTec-AD dataset.

Related work

Diffusion model. The diffusion model has gained widespread attention and research interest since its remarkable reconstruction ability. It has demonstrated excellent performance in various applications such as image generation (Zhang and Agrawala 2023), video generation (Ho et al. 2022), object detection (Chen et al. 2022), image segmentation (Amit et al. 2022) and etc. LDM (Rombach et al. 2022) introduces conditions through cross-attention to control generation. However, it fails to accurately reconstruct images that contain the original semantic information.

Anomaly detection. AD contains a variety of different settings, *e.g.*, open-set (Ding, Pang, and Shen 2022), noisy learning (Tan et al. 2021; Yoon et al. 2022), zero-/few-shot (Huang et al. 2022; Jeong et al. 2023; Cao et al. 2023; Chen, Han, and Zhang 2023; Chen et al. 2023b; Zhang et al. 2023b), 3D AD (Wang et al. 2023; Chen et al. 2023a), *etc*. This paper studies general unsupervised anomaly detection, which can primarily be categorized into three major methodologies:

1) Synthesizing-based methods synthesize anomalies on normal image samples. During the training phase, both normal images and synthetically generated abnormal images are input into the network for training, which aids in anomaly detection and localization. DRAEM (Zavrtanik, Kristan, and Skočaj 2021a) consists of an end-to-end network composed of a reconstruction network and a discriminative sub-network, which synthesizes and generates just-out-distribution phenomena. However, due to the diversity and unpredictability of anomalies in real-world scenarios, it is impossible to synthesize all types of anomalies.

2) Embedding-based methods encode the original image's three-dimensional information into a multidimensional feature space (Roth et al. 2022; Cao et al. 2022; Gu et al. 2023). Most methods employ networks (He et al. 2016; Tan and Le 2019; Zhang et al. 2022, 2023c; Wu et al. 2023) pre-trained on ImageNet (Deng et al. 2009) for feature extraction. RD4AD (Deng and Li 2022) utilizes a WideResNet50 (Zagoruyko and Komodakis 2016) as the teacher model for feature extraction and employs a structurally identical network in reverse as the student model, computing the cosine similarity of corresponding features as anomaly scores. However, due to significant differences between industrial images and the data distribution in ImageNet, the extracted features might not be suitable for industrial anomaly detection purposes.

3) Reconstruction-based methods aim to train a model on a dataset without anomalies. The model learns to identify patterns and characteristics in the normal data. OCR-GAN (Liang et al. 2023) decouples images into different frequencies and uses GAN for reconstruction. EdgRec (Liu et al. 2022) achieves good reconstruction results by first synthesizing anomalies and then extracting grayscale edge information from images, which is ultimately input into a reconstruction network. However, there are certain limitations in the reconstruction of large-area anomalies. Moreover, the accuracy of anomaly localization is also not sufficient.

Recently, some studies have applied diffusion models to anomaly detection. AnoDDPM (Wyatt et al. 2022) is the first approach to employ a diffusion model for medical anomaly detection. DiffusionAD (Zhang et al. 2023a) utilizes an

模型无法有效解决多类别异常检测问题。1) 对于图1-(a)中的去噪扩散概率模型(DDPM)(Ho, Jain, and Abbeel 2020)，在多类别设置下，该方法可能遇到生成图像类别误判的问题。原因在于对输入图像添加T时间步的噪声后，图像已丢失原始类别信息。在推理过程中，基于这种类高斯噪声分布进行去噪时，可能生成属于不同类别的样本。2) 潜在扩散模型(LDM)(Rombach et al. 2022)如图1-(b)所示采用嵌入器作为类别条件，虽不存在DDPM中的误分类问题，但仍无法解决生成图像的语义丢失问题。LDM在重建异常区域时，无法同时保留输入图像的语义信息。例如在螺钉、榛子等物体方向上可能无法与输入图像保持一致性，同时在纹理类图像方面也可能与原始图像存在显著差异。

针对上述问题，我们提出了一种基于扩散的多类别异常检测与定位框架DiAD，其结构如图2所示，包含三个组成部分：像素空间自编码器、潜在空间去噪网络和特征空间ImageNet预训练模型。为在重建异常区域位置的同时有效保持与原始图像一致的语义信息，我们提出了语义引导网络，该网络与LDM中的稳定扩散去噪网络相连接。为进一步增强保留原始图像精细细节及重建大型缺陷的能力，我们提出了空间感知特征融合模块，以整合多尺度特征。最终，重建图像与输入图像经过预训练模型提取多尺度特征并计算异常分数。我们的贡献总结如下：

- 我们提出了一种新颖的基于扩散的多类别异常检测框架DiAD，该框架首次解决了现有基于扩散方法的去噪网络无法正确重建异常的问题。
- 我们构建了一个连接到SD去噪网络的SG网络，以保持一致的语义信息并重建异常。
- 我们提出了一个SFF模块，用于整合不同尺度的特征，以进一步提升异常重建能力。
- 大量实验证明，DiAD相较于当前最优方法具有显著优势，在MVTec-AD数据集上，我们的方法在像素/图像AUROC指标上超越基于扩散的多类异常检测方法20.6↑/11.7↑，在像素-AP指标上超越非扩散方法9.2↑，在图像-AUROC指标上超越0.7↑。

相关工作

扩散模型。扩散模型自其显著成就以来，已获得广泛关注和研究兴趣。出色的重建能力。它在图像生成（Zhang和Agrawala 2023）、视频生成（Ho等人2022）、目标检测（Chen等人2022）、图像分割（Amit等人2022）等多种应用中展现了卓越性能。LDM（Rombach等人2022）通过交叉注意力引入条件以控制生成过程，但该方法无法准确重建包含原始语义信息的图像。

异常检测。AD包含多种不同的设定，*e.g.*、开放集（Ding, Pang, and Shen 2022）、噪声学习（Tan et al. 2021; Yoon et al. 2022）、零样本/少样本（Huang et al. 2022; Jeong et al. 2023; Cao et al. 2023; Chen, Han, and Zhang 2023; Chen et al. 2023b; Zhang et al. 2023b）、3D异常检测（Wang et al. 2023; Chen et al. 2023a）、*etc*。本文研究通用的无监督异常检测，其主要可分为三大方法体系：

1) 基于合成的方法在正常图像样本上合成异常。在训练阶段，将正常图像和合成生成的异常图像输入网络进行训练，这有助于异常检测和定位。DRAEM（Zavrtanik, Kristan, and Skočaj 2021a）由一个重建网络和一个判别子网络组成的端到端网络构成，该网络合成并生成分布外现象。然而，由于现实场景中异常的多样性和不可预测性，不可能合成所有类型的异常。

2) 基于嵌入的方法将原始图像的三维信息编码到多维特征空间中（Roth等人，2022；Cao等人，2022；Gu等人，2023）。大多数方法采用在ImageNet（Deng等人，2009）上预训练的网络（He等人，2016；Tan和Le，2019；Zhang等人，2022，2023c；Wu等人，2023）进行特征提取。RD4AD（Deng和Li，2022）使用WideResNet50（Zagoruyko和Komodakis，2016）作为教师模型进行特征提取，并采用结构相同但反向的网络作为学生模型，通过计算对应特征的余弦相似度作为异常分数。然而，由于工业图像与ImageNet中的数据分布存在显著差异，提取的特征可能并不适用于工业异常检测任务。

3) 基于重构的方法旨在在无异常的数据集上训练模型。该模型学习识别正常数据中的模式和特征。OCR-GAN（Liang等人，2023年）将图像解耦为不同频率，并使用GAN进行重构。EdgRec（Liu等人，2022年）通过先合成异常，再从图像中提取灰度边缘信息，最终将其输入重构网络，实现了良好的重构效果。然而，在大面积异常的重构方面存在一定局限性。此外，异常定位的准确性也不够充分。

最近，一些研究将扩散模型应用于异常检测。AnoDDPM（Wyatt等人，2022年）是首个采用扩散模型进行医学异常检测的方法。DiffusionAD（Zhang等人，2023a）则利用

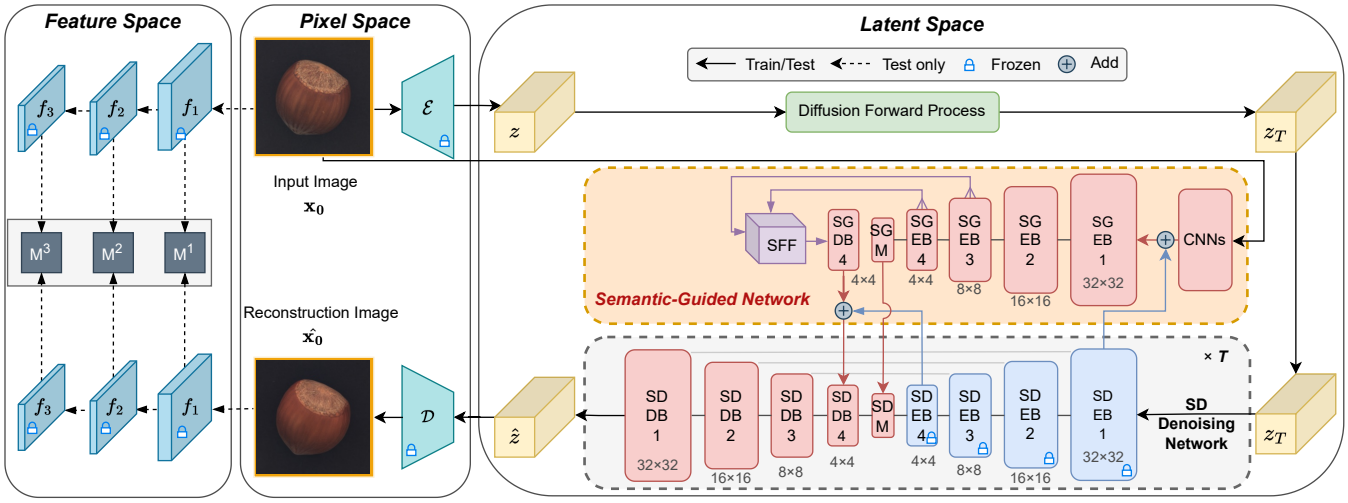

Figure 2: **Framework of the proposed DiAD that contains three parts**: ***1)*** a pixel-space autoencoder $\{\mathcal{E}, \mathcal{D}\}$; ***2)*** a latent-space Semantic-Guided (SG) network with a connection to Stable Diffusion (SD) denoising network; and ***3)*** a feature-space pre-trained feature extractor Ψ. During training, the input x_0 and the latent variable z_T are inputted into the SG network and the SD denoising network, respectively. The output noise and input noise are calculated for MSE loss and gradient optimization is computed. During testing, x_0 and the reconstructed image $\hat{x_0}$ are inputted into the same pre-trained feature extraction network to obtain feature maps $\{f_1, f_2, f_3\}$ of different scales, and their anomaly scores $\mathcal{S}$ are calculated.

anomaly synthetic strategy to generate anomalous samples and labels, along with two sub-networks dedicated to the tasks of denoising and segmentation. DDAD (Mousakhan, Brox, and Tayyub 2023) employs a score-based pre-trained diffusion model to generate normal samples while fine-tuning the pre-trained feature extractor to achieve domain transfer. However, these approaches only add limited steps of noise and perform few denoising steps, which makes them unable to reconstruct large-scale defects.

To overcome the aforementioned problems, We propose a diffusion-based framework DiAD for multi-class anomaly detection, which firstly tackles the problem of existing diffusion-based methods failing to correctly reconstruct anomalies.

Preliminaries

Denoising Diffusion Probabilistic Model. Denoising Diffusion Probabilistic Model (DDPM) consists of two processes: the forward diffusion process and the reverse denoising process. During the forward process, a noisy sample x_t is generated using a Markov chain that incrementally adds Gaussian-distributed noise to an initial data sample x_0. The forward diffusion process can be characterized as follows:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (1)$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i = \prod_{i=1}^{T}(1-\beta_i)$ and β_i represents the noise schedule used to regulate the quantity of noise added at each timestep.

In the reverse denoising process, x_T is first sampled from equation 1 and x_{t-1} is reconstructed from x_t and the model prediction $\epsilon_\theta(x_t, t)$ with the formulation:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z, \quad (2)$$

where $z \sim \mathcal{N}(\mathbf{0}, \mathrm{I})$, σ_t is a fixed constant related to the variance schedule, $\epsilon_\theta(x_t, t)$ is a U-Net (Ronneberger, Fischer, and Brox 2015) network to predict the distribution and θ is the learnable parameter which could be optimized as:

$$\min_\theta \mathbb{E}_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} \left\| \epsilon - \epsilon_\theta(\boldsymbol{x}_t, t) \right\|_2^2. \quad (3)$$

Latent Diffusion Model. Latent Diffusion Model (LDM) focuses on the low-dimensional latent space with conditioning mechanisms. LDM consists of a pre-trained autoencoder model and a denoising U-Net-like attention-based network. The network compresses images using an encoder, conducts diffusion and denoising operations in the latent representation space, and subsequently reconstructs the images back to the original pixel space using a decoder. The training optimization objective is:

$$\mathcal{L}_{LDM} = \mathbb{E}_{z_0, t, c, \epsilon \sim \mathcal{N}(0,1)}\left[\left\|\epsilon - \epsilon_\theta(z_t, t, c)\right\|_2^2\right], \quad (4)$$

where c represents the conditioning mechanisms which can consist of multimodal types such as text or image, connected to the model through a cross-attention mechanism. z_t represents the latent space variable,

Method

The proposed pipeline DiAD is shown in Fig. 2. First, the pre-trained encoder downsamples the input image into a latent-space representation. Then, noise is added to the latent representation, followed by the denoising process using an SD denoising network with a connection to the SG network. The denoising process is repeated for the same timesteps as the diffusion process. Finally, the reconstructed latent representation is restored to the original image level using the

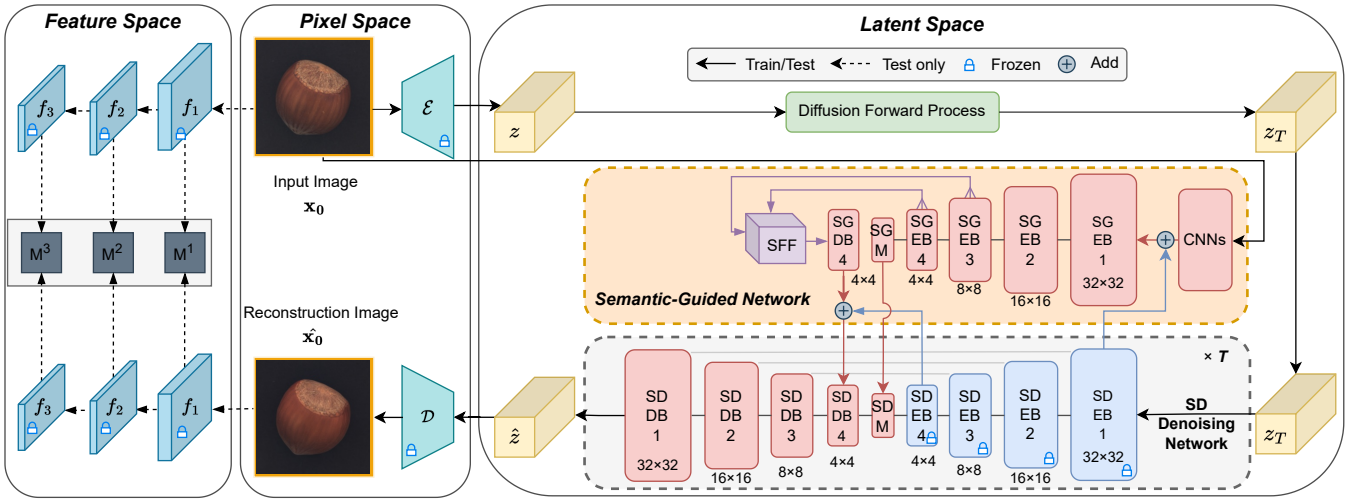

图2：所提出的DiAD框架包含三个部分：***1)***一个像素空间自编码器$\{\mathcal{E},\mathcal{D}\}$；***2)***一个与稳定扩散（SD）去噪网络相连的潜在空间语义引导（SG）网络；以及***3)***一个特征空间的预训练特征提取器Ψ。训练过程中，输入x_0和潜在变量z_T分别输入到SG网络和SD去噪网络中。计算输出噪声与输入噪声的均方误差损失，并进行梯度优化。测试过程中，x_0和重建图像$\hat{x}_0$输入到相同的预训练特征提取网络中，以获得不同尺度的特征图$\{f_1、f_2、f_3\}$，并计算其异常分数$\mathcal{S}$。

异常合成策略用于生成异常样本和标签，并配备两个专门负责去噪和分割任务的子网络。DDAD（Mousakhan, Brox, and Tayyub 2023）采用基于分数的预训练扩散模型生成正常样本，同时微调预训练的特征提取器以实现领域迁移。然而，这些方法仅添加了有限的噪声步骤且执行较少的去噪步骤，导致其无法重建大规模缺陷。

为了解决上述问题，我们提出了一种基于扩散的多类别异常检测框架DiAD，该框架首次解决了现有基于扩散的方法无法正确重构异常的问题。

预备知识

去噪扩散概率模型。去噪扩散概率模型（DDPM）包含两个过程：前向扩散过程与反向去噪过程。在前向过程中，通过马尔可夫链对初始数据样本x_0逐步添加高斯分布噪声，生成含噪样本x_t。前向扩散过程可表述如下：

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \epsilon_t \sim \mathcal{N}(\mathbf{0},\mathbf{I}), \quad (1)$$

其中 $\alpha_t = 1-\beta_t$、$\bar{\alpha}_t = \prod_{i=1}^{T}\alpha_i = \prod_{i=1}^{T}(1-\beta_i)$ 和 β_i 表示用于调控每个时间步所添加噪声量的噪声调度表。

在反向去噪过程中，首先从方程1中采样得到x_T，然后通过公式从x_t和模型预测$\epsilon_\theta(x_t,t)$重建出x_{t-1}：

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t,t)\right) + \sigma_t z, \quad (2)$$

其中 $z \sim \mathcal{N}(\mathbf{0},\mathrm{I})$，$\sigma_t$ 是与方差调度相关的固定常数，$\epsilon_\theta(x_t,t)$ 是用于预测分布的 U-Net（Ronneberger, Fischer, and Brox 2015）网络，θ 是可学习参数，可通过以下方式优化：

$$\min_\theta \mathbb{E}_{\boldsymbol{x}_0\sim q(\boldsymbol{x}_0),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}\left\|\epsilon-\epsilon_\theta(\boldsymbol{x}_t,t)\right\|_2^2. \quad (3)$$

潜在扩散模型。潜在扩散模型（LDM）专注于具有条件机制的低维潜在空间。LDM由一个预训练的自编码器模型和一个基于注意力机制的去噪U-Net类网络组成。该网络使用编码器压缩图像，在潜在表示空间中进行扩散和去噪操作，随后通过解码器将图像重建回原始像素空间。其训练优化目标为：

$$\mathcal{L}_{LDM} = \mathbb{E}_{z_0,t,c,\epsilon\sim\mathcal{N}(0,1)}\left[\left\|\epsilon-\epsilon_\theta(z_t,t,c)\right\|_2^2\right], \quad (4)$$

其中c代表条件机制，它可以包含文本或图像等多模态类型，通过交叉注意力机制与模型连接。z_t代表潜在空间变量，

方法

提出的DiAD流程如图2所示。首先，预训练的编码器将输入图像下采样为潜在空间表示。随后，向潜在表示添加噪声，并通过与SG网络连接的SD去噪网络进行去噪处理。去噪过程的迭代步数与扩散过程保持一致。最后，重构的潜在表示通过解码器恢复到原始图像层级。

pre-trained decoder. In terms of anomaly detection and localization, the input and reconstructed images are fed into the same pre-trained model to extract features at different scales and calculate the differences between these features.

Semantic-Guided Network

As discussed earlier, DDPM and LDM each have specific problems when addressing multi-class anomaly detection tasks. In response to these issues and the multi-class task itself, we propose an SG network to address the problem of LDM's inability to effectively reconstruct anomalies and preserve the semantic information of the input image.

Given an input image $x_0 \in \mathbb{R}^{3\times H\times W}$ in pixel space, the pre-trained encoder $\mathcal{E}$ encodes x_0 into a latent space representation $z = \mathcal{E}(x_0)$ where $z \in \mathbb{R}^{c\times h\times w}$. Similar to Eq. 1 where the original pixel space x is replaced by latent representation z, the forward diffusion process now can be characterized as follows:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{5}$$

The perturbed representation z_T and input x_0 are simultaneously fed into the SD denoising network and SG network, respectively. After T steps of the reverse denoising process, the final variable $\hat{z}$ is restored to the reconstructed image $\hat{x_0}$ from the pre-trained decoder $\mathcal{D}$ giving $\hat{x_0} = \mathcal{D}(\hat{z})$. The training objective of DiAD is:

$$\mathcal{L}_{DiAD} = \mathbb{E}_{z_0,t,c_i,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, c_i\right)\|_2^2\right]. \tag{6}$$

The denoising network consists of a pre-trained SD denoising network and an SG network that replicates the SD parameters for initiation as shown in Fig. 2. The pre-trained SD denoising network comprises four encoder blocks, one middle block and four decoder blocks. Here, 'block' means a frequently utilized unit in the construction of the neural network layer, ***e.g.***,, 'resnet' block, transformer block, multi-head cross attention block, *etc*.

The input image $x_0 \in \mathbb{R}^{3\times H\times W}$ is transformed into $\mathrm{x} \in \mathbb{R}^{d\times h\times w}$ by a set of 'conv-silu' layers $\mathcal{C}$ in SG network in order to keep the same dimension with the latent representations in SD Encoder Block 1 $\mathcal{E}_{SD1}$. Then, the result of the summation of x and z are input into the SG Encoder Blocks (SGEBs). After continuous downsampling by the encoder $\mathcal{E}_{SG}$, the results are finally added to the output of the SD middle block $\mathcal{M}_{SD}$ after its completion in the middle block $\mathcal{M}_{SG}$. Additionally, to address multi-class tasks of different scenarios and categories, the results of the SG Decoder Blocks (SGDBs) $\mathcal{D}_{SG}$ are also added to the results of the SD decoder $\mathcal{D}_{SD}$ with an SFF block combined which will be particularly explained in the next section. The output $\mathcal{G}$ of the denoising network is characterized as:

$$\begin{aligned}\mathcal{G} = \mathcal{D}_{SD}\left(\mathcal{M}_{SD}\left(\mathcal{E}_{SD}\left(z_t\right)\right) + \mathcal{M}_{SG}\left(\mathcal{E}_{SD}\left(z + \mathcal{C}\left(x_0\right)\right)\right)\right) \\ + \mathcal{D}_{SGj}(\mathcal{M}_{SG}\left(\mathcal{E}_{SD}\left(z + \mathcal{C}\left(x_0\right)\right)\right)),\end{aligned} \tag{7}$$

where z represents the latent representation with noise perturbed, x_0 represents the input image, $\mathcal{C}(\cdot)$ represents a set of 'conv-silu' layers in SG network, $\mathcal{E}_{SD}(\cdot)$ represents all the SD encoder blocks (SDEBs), $\mathcal{E}_{SG}(\cdot)$ represents all the SGEBs, $\mathcal{M}_{SG}(\cdot)$ and $\mathcal{M}_{SD}(\cdot)$ represent SG and SD middle blocks respectively, $\mathcal{D}_{SD}(\cdot)$ represent all the SDDBs and $\mathcal{D}_{SGj}(\cdot)$ represents SGDBs for j-th blocks.

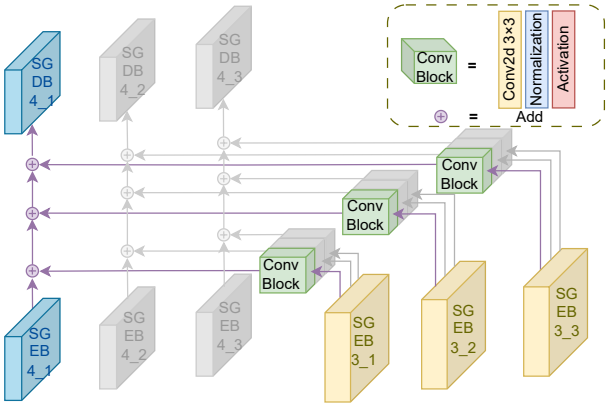

Figure 3: **Schematic diagram of SFF block**. Each layer in SGDB4 is obtained by adding the corresponding SGEB4 to every SGEB3 with Conv Block performed.

Spatial-aware Feature Fusion Block

When adding several layers of decoder blocks from SGEBs to SDDBs during the experiment as shown in Table 7, we found it to be challenging to solve the multi-class anomaly detection. This is because the dataset contains various types, such as objects and textures. For texture-related cases, the anomalies are generally smaller, so it is necessary to preserve their original textures. On the other hand, the defects often cover larger areas for object-related cases, requiring stronger reconstruction capabilities. Therefore, it is extremely challenging to simultaneously preserve the normal information of the original samples and reconstruct the abnormal locations in different scenarios.

Hence, we proposed a Spatial-aware Feature Fusion (SFF) block with the aim of integrating high-scale semantic information into the low-scale. This ultimately enables the model to both preserve the information of the original normal samples and reconstruct large-scale abnormal regions. The structure of the SFF block is shown in Fig. 3. Each SGEBs consists of three sub-layers. Therefore, the SFF block integrates the features of each layer in SGEB3 into each layer in SGEB4 and adds the fused features to the original features. The final output of each layer of the SGEB4 is:

$$\mathcal{Q}_i = \mathcal{P}_i + \sum_{j=1}^{J}\mathcal{F}(\mathcal{H}_j), \tag{8}$$

where $\mathcal{P}_i$ represents the low-scale output features of the i-th layer of SGEB4, $\mathcal{Q}_i$ represents the final low-scale output features of the i-th layer of SGDB4, $\mathcal{H}_j$ represents the high-scale output features of the j-th layer of SGEB3, $J = 3$ indicates three layers of SGEB3 used in the experiment and $\mathcal{F}(\cdot)$ represent a basic convolutional block which consists of a 3x3 convolution layer followed by a normalization layer and an activation layers.

预训练解码器。在异常检测和定位方面，输入图像与重建图像被输入同一预训练模型，以提取多尺度特征并计算这些特征间的差异。

语义引导网络

如前所述，DDPM和LDM在处理多类别异常检测任务时各自存在特定问题。针对这些问题及多类别任务本身，我们提出了一种SG网络，以解决LDM无法有效重建异常并保留输入图像语义信息的问题。

给定像素空间中的输入图像$x_0 \in \mathbb{R}^{3\times H\times W}$，预训练编码器$\mathcal{E}$将$x_0$编码为潜在空间表示$z = \mathcal{E}(x_0)$，其中$z \in \mathbb{R}^{c\times h\times w}$。类似于公式1中将原始像素空间$x$替换为潜在表示$z$的方式，前向扩散过程现可表征如下：

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{5}$$

扰动后的表示 z_T 和输入 x_0 分别同时输入到 SD 去噪网络和 SG 网络中。经过 T 步反向去噪过程后，最终变量 $\hat{z}$ 通过预训练解码器 $\mathcal{D}$ 恢复为重建图像$\hat{x_0}$，得到 $\hat{x_0} = \mathcal{D}(\hat{z})$。DiAD 的训练目标为：

$$\mathcal{L}_{DiAD} = \mathbb{E}_{z_0,t,c_i,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, c_i\right)\|_2^2\right]. \tag{6}$$

去噪网络由一个预训练的SD去噪网络和一个复制SD参数用于初始化的SG网络组成，如图2所示。预训练的SD去噪网络包含四个编码器块、一个中间块和四个解码器块。此处的“块”指神经网络层构建中常用的单元，***e.g.***,，如“残差网络”块、变换器块、多头交叉注意力块，*etc*。

输入图像$x_0 \in \mathbb{R}^{3\times H\times W}$通过SG网络中的一组'conv-silu'层$\mathcal{C}$转换为x$\in \mathbb{R}^{d\times h\times w}$，以保持与SD编码器块1$\mathcal{E}_{SD1}$中潜在表示相同的维度。随后，x与$z$的求和结果被输入到SG编码器块（SGEBs）中。经过编码器$\mathcal{E}_{SG}$的连续下采样后，最终结果被添加到SD中间块$\mathcal{M}_{SD}$在中间块$\mathcal{M}_{SG}$完成处理后的输出中。此外，为应对不同场景和类别的多类任务，SG解码器块（SGDBs）$\mathcal{D}_{SG}$的结果也通过结合SFF块（具体将在下一节详细说明）添加到SD解码器$\mathcal{D}_{SD}$的结果中。去噪网络的输出$\mathcal{G}$表征为：

$$\begin{aligned}\mathcal{G} = \mathcal{D}_{SD}\left(\mathcal{M}_{SD}\left(\mathcal{E}_{SD}\left(z_t\right)\right) + \mathcal{M}_{SG}\left(\mathcal{E}_{SD}\left(z + \mathcal{C}\left(x_0\right)\right)\right)\right) \\ + \mathcal{D}_{SGj}(\mathcal{M}_{SG}\left(\mathcal{E}_{SD}\left(z + \mathcal{C}\left(x_0\right)\right)\right)),\end{aligned} \tag{7}$$

其中z表示受噪声扰动的潜在表示，x_0表示输入图像，$\mathcal{C}(\cdot)$表示SG网络中的一组'conv-silu'层，$\mathcal{E}_{SD}(\cdot)$表示所有SD编码器块（SDEBs），$\mathcal{E}_{SG}(\cdot)$表示所有

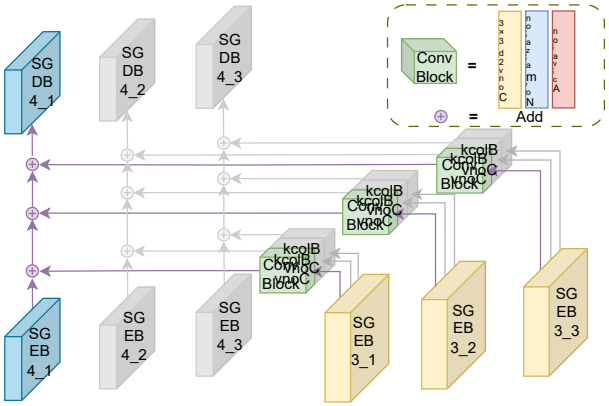

图3：SFF块示意图。SGDB4中的每一层都是通过将相应的SGEB4添加到每个已执行Conv Block的SGEB3上而得到的。

SGEBs、$\mathcal{M}_{SG}(\cdot)$和$\mathcal{M}_{SD}(\cdot)$分别代表SG和SD中间块，$\mathcal{D}_{SD}(\cdot)$代表所有SDDB，而$\mathcal{D}_{SGj}(\cdot)$代表第j个块的SGDBs。

空间感知特征融合块

在实验中，如表7所示，当从SGEBs向SDDBs添加多层解码器块时，我们发现解决多类别异常检测具有挑战性。这是因为数据集包含多种类型，例如物体和纹理。对于与纹理相关的情况，异常通常较小，因此需要保留其原始纹理。另一方面，对于与物体相关的情况，缺陷往往覆盖较大区域，需要更强的重建能力。因此，在不同场景中同时保留原始样本的正常信息并重建异常位置极具挑战性。

因此，我们提出了一个空间感知特征融合（SFF）模块，旨在将高尺度的语义信息整合到低尺度中。这最终使模型既能保留原始正常样本的信息，又能重建大尺度的异常区域。SFF模块的结构如图3所示。每个SGEB由三个子层组成。因此，SFF模块将SGEB3中每一层的特征整合到SGEB4的每一层中，并将融合后的特征添加到原始特征上。SGEB4每一层的最终输出为：

$$\mathcal{Q}_i = \mathcal{P}_i + \sum_{j=1}^{J}\mathcal{F}(\mathcal{H}_j), \tag{8}$$

其中$\mathcal{P}_i$代表SGEB4第i层的低尺度输出特征，$\mathcal{Q}_i$代表SGDB4第i层的最终低尺度输出特征，$\mathcal{H}_j$代表SGEB3第j层的高尺度输出特征，J =表示实验中使用的三层SGEB3，而$\mathcal{F}(\cdot)$代表由3x3卷积层、归一化层和激活层组成的基础卷积块。

Category		Non-Diffusion Method					Diffusion-based Method		
		PaDiM	MKD	DRAEM	RD4AD	UniAD	DDPM	LDM	Ours
Objects	Bottle	97.9/-	98.7/-	97.5/99.2/96.1	99.6/99.9/98.4	**99.7/100./100.**	63.6/71.8/86.3	93.8/98.7/93.7	99.7/96.5/91.8
	Cable	70.9/-	78.2/-	57.8/74.0/76.3	84.1/89.5/82.5	**95.2**/95.9/88.0	55.6/69.7/76.0	55.7/74.8/77.7	94.8/**98.8/95.2**
	Capsule	73.4/-	68.3/-	65.3/92.5/90.4	**94.1**/96.9/**96.9**	86.9/**97.8**/94.4	52.9/82.0/90.5	60.5/81.4/90.5	89.0/97.5/95.5
	Hazelnut	85.5/-	97.1/-	93.7/97.5/92.3	60.8/69.8/86.4	**99.8/100./99.3**	87.0/90.4/88.1	93.0/95.8/89.8	99.5/99.7/97.3
	Metal Nut	88.0/-	64.9/-	72.8/95.0/92.0	**100./100./99.5**	99.2/99.9/**99.5**	60.0/74.4/89.4	53.0/80.1/89.4	99.1/96.0/91.6
	Pill	68.8/-	79.7/-	82.2/94.9/92.4	**97.5/99.6/96.8**	93.7/98.7/95.7	55.8/84.0/91.6	62.1/93.1/91.6	95.7/98.5/94.5
	Screw	56.9/-	75.6/-	92.0/95.7/89.9	**97.7**/99.3/95.8	87.5/96.5/89.0	53.6/71.9/85.9	58.7/81.9/85.6	90.7/**99.7/97.9**
	Toothbrush	95.3/-	75.3/-	90.6/96.8/90.0	97.2/99.0/94.7	94.2/97.4/95.2	57.5/68.0/83.3	78.6/83.9/83.3	**99.7/99.9/99.2**
	Transistor	86.6/-	73.4/-	74.8/77.4/71.1	94.2/95.2/90.0	99.8/98.0/93.8	57.8/44.6/57.1	61.0/57.8/59.1	**99.8/99.6/97.4**
	Zipper	79.7/-	87.4/-	98.8/**99.9**/99.2	**99.5/99.9/99.2**	95.8/99.5/97.1	64.9/77.4/88.1	73.6/89.5/90.6	95.1/99.1/94.4
Textures	Carpet	93.8/-	69.8/-	98.0/99.1/96.7	98.5/99.6/97.2	**99.8/99.9/99.4**	95.5/98.7/91.0	99.4/99.8/**99.4**	99.4/**99.9**/98.3
	Grid	73.9/-	83.8/-	**99.3**/99.7/**98.2**	98.0/99.4/96.5	98.2/99.5/97.3	83.5/93.9/86.9	67.3/82.6/84.4	98.5/**99.8**/97.7
	Leather	99.9/-	93.6/-	98.7/99.3/95.0	**100./100./100.**	**100./100./100.**	98.4/99.5/96.3	97.4/99.0/96.3	99.8/99.7/97.6
	Tile	93.3/-	89.5/-	**99.8/100./100.**	98.3/99.3/96.4	99.3/99.8/98.2	93.697.5/92.0	97.1/98.7/94.1	96.8/99.9/98.4
	Wood	98.4/-	93.4/-	**99.8/100./100.**	99.2/99.8/98.3	98.6/99.6/96.6	98.6/99.6/97.5	97.8/99.4/95.9	99.7/**100./100.**
Mean		84.2/-	81.9/-	88.1/94.7/92.0	94.6/96.5/95.2	96.5/98.8/96.2	71.9/81.6/86.6	76.6/87.8/88.1	**97.2/99.0/96.5**

Table 1: Comparison with SOTA methods on MVTec-AD dataset for multi-class anomaly detection with $AUROC_{cls}/AP_{cls}/F1max_{cls}$ metrics.

As Batch Normalization (BN) (Ioffe and Szegedy 2015) considers the normalization statistics of all images within a batch, it leads to a loss of unique details in each sample. BN is suitable for a relatively large mini-batch scenario with similar data distributions. However, for multi-class anomaly detection where there are significant differences in data distributions among different categories, normalizing the entire batch is not suitable for tasks in the multi-class setting. Since the results generated by using SD mainly depend on the input image instance, using Instance Normalization (IN) (Ulyanov, Vedaldi, and Lempitsky 2017) can not only accelerate model convergence but also maintain the independence between each image instance. In addition, in terms of choosing the activation function, we use the SiLU (Elfwing, Uchibe, and Doya 2018) instead of the commonly used ReLU (Hahnloser et al. 2000), which can preserve more input information. Experimental results in Table 7 show that the performance is improved by using IN and SiLU simultaneously instead of the combination of BN and ReLU.

Anomaly localization and detection

During the inference stage, the reconstruction image is obtained through the diffusion and denoising process in the latent space. For anomaly localization and detection, We use the same ImageNet pre-trained feature extractor Ψ to extract features from both the input image x_0 and the reconstructed image $\hat{x_0}$ and calculate the anomaly map on different scale feature maps $\mathcal{M}^n$ using cosine similarity:

$$\mathcal{M}^n(x_0, \hat{x_0}) = 1 - \frac{(\Psi^n(x_0, \hat{x_0}))^T \cdot \Psi^n(x_0, \hat{x_0})}{\|\Psi^n(x_0, \hat{x_0})\| \, \|\Psi^n(x_0, \hat{x_0})\|}, \quad (9)$$

where n represents the n-th feature layer f_n and the anomaly score $\mathcal{S}$ for an input-pair of anomaly localization is:

$$\mathcal{S} = \sum_{n \in N} \sigma_n \mathcal{M}^n(x_0, \hat{x_0}), \quad (10)$$

Metrics	Non-Diffusion		Diffusion-based		
	DRAEM	UniAD	DDPM	LDM	Ours
$AUROC_{cls}$	79.1	85.5	54.5	56.7	**86.8**
AP_{cls}	81.9	85.5	57.9	61.4	**88.3**
$F1max_{cls}$	78.9	84.4	72.3	73.1	**85.1**
$AUROC_{seg}$	91.3	95.9	79.7	86.6	**96.0**
AP_{seg}	23.5	21.0	2.2	6.0	**26.1**
$F1max_{seg}$	29.5	27.0	4.5	9.9	**33.0**
PRO	58.8	**75.6**	46.8	55.0	75.2

Table 2: Quantitative comparisons on VisA dataset.

where σ_n indicates the upsampling factor in order to keep the same dimension of the pixel space image and N indicates the number of feature layers used during inference.

Experiment

Datasets and evaluation metrics

MVTec-AD dataset. MVTec-AD (Bergmann et al. 2019) dataset simulates real-world industrial production scenarios, filling the gap in unsupervised anomaly detection. It consists of 5 types of textures and 10 types of objects, in 5,354 high-resolution images from different domains. The training set contains 3,629 images with only anomaly-free samples. The test set consists of 1,725 images, including both normal and abnormal samples. Pixel-level annotations are provided for the anomaly localization evaluation.

VisA dataset. VisA (Zou et al. 2022) dataset consists of a total of 10,821 high-resolution images, including 9,621 normal images and 1,200 anomaly images with 78 types of anomalies. The VisA dataset comprises 12 subsets, each corresponding to a distinct object. 12 objects could be categorized into three different object types: Complex structure, Multiple instances, and Single instance.

MVTec-3D dataset. MVTec-3D (Bergmann et al. 2022) dataset comprises 4,147 scans obtained using a high-resolution industrial 3D sensor. It consists of 10 categories

Category		Non-Diffusion Method					Diffusion-based Method		
		PaDiM	MKD	DRAEM	RD4AD	UniAD	DDPM	LDM	Ours
Objects	Bottle	97.9/-	98.7/-	97.5/99.2/96.1	99.6/99.9/98.4	**99.7/100./100.**	63.6/71.8/86.3	93.8/98.7/93.7	99.7/96.5/91.8
	Cable	70.9/-	78.2/-	57.8/74.0/76.3	84.1/89.5/82.5	**95.2**/95.9/88.0	55.6/69.7/76.0	55.7/74.8/77.7	94.8/**98.8/95.2**
	Capsule	73.4/-	68.3/-	65.3/92.5/90.4	**94.1**/96.9/**96.9**	86.9/**97.8**/94.4	52.9/82.0/90.5	60.5/81.4/90.5	89.0/97.5/95.5
	Hazelnut	85.5/-	97.1/-	93.7/97.5/92.3	60.8/69.8/86.4	**99.8/100./99.3**	87.0/90.4/88.1	93.0/95.8/89.8	99.5/99.7/97.3
	Metal Nut	88.0/-	64.9/-	72.8/95.0/92.0	**100./100./99.5**	99.2/99.9/**99.5**	60.0/74.4/89.4	53.0/80.1/89.4	99.1/96.0/91.6
	Pill	68.8/-	79.7/-	82.2/94.9/92.4	**97.5/99.6/96.8**	93.7/98.7/95.7	55.8/84.0/91.6	62.1/93.1/91.6	95.7/98.5/94.5
	Screw	56.9/-	75.6/-	92.0/95.7/89.9	**97.7**/99.3/95.8	87.5/96.5/89.0	53.6/71.9/85.9	58.7/81.9/85.6	90.7/**99.7/97.9**
	Toothbrush	95.3/-	75.3/-	90.6/96.8/90.0	97.2/99.0/94.7	94.2/97.4/95.2	57.5/68.0/83.3	78.6/83.9/83.3	**99.7/99.9/99.2**
	Transistor	86.6/-	73.4/-	74.8/77.4/71.1	94.2/95.2/90.0	99.8/98.0/93.8	57.8/44.6/57.1	61.0/57.8/59.1	**99.8/99.6/97.4**
	Zipper	79.7/-	87.4/-	98.8/**99.9**/99.2	**99.5/99.9/99.2**	95.8/99.5/97.1	64.9/77.4/88.1	73.6/89.5/90.6	95.1/99.1/94.4
Textures	Carpet	93.8/-	69.8/-	98.0/99.1/96.7	98.5/99.6/97.2	**99.8/99.9/99.4**	95.5/98.7/91.0	99.4/99.8/**99.4**	99.4/**99.9**/98.3
	Grid	73.9/-	83.8/-	**99.3**/99.7/**98.2**	98.0/99.4/96.5	98.2/99.5/97.3	83.5/93.9/86.9	67.3/82.6/84.4	98.5/**99.8**/97.7
	Leather	99.9/-	93.6/-	98.7/99.3/95.0	**100./100./100.**	**100./100./100.**	98.4/99.5/96.3	97.4/99.0/96.3	99.8/99.7/97.6
	Tile	93.3/-	89.5/-	**99.8/100./100.**	98.3/99.3/96.4	99.3/99.8/98.2	93.697.5/92.0	97.1/98.7/94.1	96.8/99.9/98.4
	Wood	98.4/-	93.4/-	**99.8/100./100.**	99.2/99.8/98.3	98.6/99.6/96.6	98.6/99.6/97.5	97.8/99.4/95.9	99.7/**100./100.**
Mean		84.2/-	81.9/-	88.1/94.7/92.0	94.6/96.5/95.2	96.5/98.8/96.2	71.9/81.6/86.6	76.6/87.8/88.1	**97.2/99.0/96.5**

表1：在MVTec-AD数据集上采用$AUROC_{cls}/AP_{cls}/F1max_{cls}$指标进行多类别异常检测的SOTA方法对比。

由于批归一化（BN）（Ioffe和Szegedy，2015）考虑了一个批次内所有图像的归一化统计量，这导致每个样本独特细节的丢失。BN适用于数据分布相似、小批量相对较大的场景。然而，在多类别异常检测任务中，不同类别间的数据分布存在显著差异，对整个批次进行归一化并不适用于多类别设定下的任务。由于使用SD生成的结果主要依赖于输入图像实例，采用实例归一化（IN）（Ulyanov, Vedaldi和Lempitsky，2017）不仅能加速模型收敛，还能保持各图像实例间的独立性。此外，在激活函数的选择上，我们使用SiLU（Elfwing, Uchibe和Doya，2018）替代常用的ReLU（Hahnloser等人，2000），这能保留更多的输入信息。表7中的实验结果表明，同时使用IN和SiLU替代BN与ReLU的组合，能有效提升性能。

异常定位与检测

在推理阶段，重建图像是通过在潜在空间中的扩散和去噪过程获得的。对于异常定位和检测，我们使用相同的ImageNet预训练特征提取器Ψ，从输入图像x_0和重建图像$\hat{x_0}$中提取特征，并利用余弦相似度在不同尺度的特征图$\mathcal{M}^n$上计算异常图：

$$\mathcal{M}^n(x_0,\hat{x_0}) = 1 - \frac{(\Psi^n(x_0,\hat{x_0}))^T \cdot \Psi^n(x_0,\hat{x_0})}{\|\Psi^n(x_0,\hat{x_0})\| \, \|\Psi^n(x_0,\hat{x_0})\|}, \quad (9)$$

其中 n 表示第 n 个特征层 f_n，而异常定位输入对的异常分数 $\mathcal{S}$ 为：

$$\mathcal{S} = \sum_{n\in N} \sigma_n \mathcal{M}^n(x_0,\hat{x_0}), \quad (10)$$

Metrics	Non-Diffusion		Diffusion-based		
	DRAEM	UniAD	DDPM	LDM	Ours
$AUROC_{cls}$	79.1	85.5	54.5	56.7	**86.8**
AP_{cls}	81.9	85.5	57.9	61.4	**88.3**
$F1max_{cls}$	78.9	84.4	72.3	73.1	**85.1**
$AUROC_{seg}$	91.3	95.9	79.7	86.6	**96.0**
AP_{seg}	23.5	21.0	2.2	6.0	**26.1**
$F1max_{seg}$	29.5	27.0	4.5	9.9	**33.0**
PRO	58.8	**75.6**	46.8	55.0	75.2

表2：VisA数据集上的定量比较。

其中σ_n表示上采样因子，以保持像素空间图像的维度不变，而N表示推理过程中使用的特征层数量。

实验

数据集和评估指标

MVTec-AD数据集。MVTec-AD（Bergmann等人，2019）数据集模拟了真实世界的工业生产场景，填补了无监督异常检测领域的空白。该数据集包含5种纹理类型和10种物体类型，共计5,354张来自不同领域的高分辨率图像。训练集包含3,629张仅含正常样本的图像。测试集由1,725张图像组成，同时包含正常和异常样本。数据集提供了像素级标注，用于异常定位评估。

VisA数据集。VisA（Zou等人，2022）数据集共包含10,821张高分辨率图像，其中9,621张为正常图像，1,200张为包含78种异常类型的异常图像。该数据集由12个子集构成，每个子集对应一种独立物体。这12种物体可分为三类不同对象类型：复杂结构、多实例和单实例。

MVTec-3D数据集。MVTec-3D（Bergmann等人，2022年）数据集包含使用高分辨率工业3D传感器获取的4,147次扫描。它涵盖10个类别

Category		Non-Diffusion Method					Diffusion-based Method		
		PaDiM	MKD	DRAEM	RD4AD	UniAD	DDPM	LDM	Ours
Objects	Bottle	96.1/-	91.8/-	87.6/62.5/56.9	97.8/**68.2**/67.6	98.1/66.0/**69.2**	59.9/ 4.9/11.7	86.9/49.1/50.0	**98.4**/52.2/54.8
	Cable	81.0/-	89.3/-	71.3/14.7/17.8	85.1/26.3/33.6	97.3/39.9/45.2	66.5/ 6.7/10.6	89.3/18.5/26.2	**96.8/50.1/57.8**
	Capsule	96.9/-	88.3/-	50.5/ 6.0/10.0	**98.8/43.4/50.0**	98.5/42.7/46.5	63.1/ 6.2/ 9.7	90.0/ 7.9/27.3	97.1/42.0/45.3
	Hazelnut	96.3/-	91.2/-	96.9/70.0/60.5	97.9/36.2/51.6	98.1/55.2/56.8	91.2/24.1/28.3	95.1/51.2/53.5	**98.3/79.2/80.4**
	Metal Nut	84.8/-	64.2/-	62.2/31.1/21.0	93.8/**62.3**/65.4	94.8/55.5/**66.4**	62.7/14.6/29.2	70.5/19.3/30.7	**97.3**/30.0/38.3
	Pill	87.7/-	69.7/-	94.4/59.1/44.1	**97.5/63.4/65.2**	95.0/44.0/53.9	55.3/ 4.0/ 8.4	74.9/10.2/15.0	95.7/46.0/51.4
	Screw	94.1/-	92.1/-	95.5/33.8/40.6	**99.4**/40.2/44.6	98.3/28.7/37.6	91.1/ 1.8/ 3.8	91.7/ 2.2/ 4.6	97.9/**60.6/59.6**
	Toothbrush	95.6/-	88.9/-	97.7/55.2/55.8	**99.0**/53.6/58.8	98.4/34.9/45.7	76.9/ 4.0/ 7.7	93.7/20.4/ 9.8	**99.0/78.7/72.8**
	Transistor	92.3/-	71.7/-	64.5/23.6/15.1	85.9/42.3/45.2	**97.9/59.5/64.6**	53.2/ 5.8/11.4	85.5/25.0/30.7	95.1/15.6/31.7
	Zipper	94.8/-	86.1/-	98.3/**74.3/69.3**	**98.5**/53.9/60.3	96.8/40.1/49.9	67.4/ 3.5/ 7.6	66.9/ 5.3/ 7.4	96.2/60.7/60.0
Textures	Carpet	97.6/-	95.5/-	98.6/**78.7/73.1**	99.0/58.5/60.4	98.5/49.9/51.1	89.2/18.8/44.3	**99.1**/70.6/66.0	98.6/42.2/46.4
	Grid	71.0/-	82.3/-	98.7/44.5/46.2	**99.2**/46.0/47.4	96.5/23.0/28.4	63.1/ 0.7/ 1.9	52.4/ 1.1/ 1.9	96.6/**60.0/64.1**
	Leather	84.8/-	96.7/-	97.3/**60.3**/57.4	**99.3**/38.0/45.1	98.8/32.9/34.4	97.3/38.9/43.2	99.0/45.9/44.0	98.8/56.1/**62.3**
	Tile	80.5/-	85.3/-	**98.0/93.6/86.0**	95.3/48.5/60.5	91.8/42.1/50.6	87.0/35.2/36.6	90.1/43.9/51.6	92.4/65.7/64.1
	Wood	89.1/-	80.5/-	**96.0/81.4/74.6**	95.3/47.8/51.0	93.2/37.2/41.5	84.7/30.9/37.3	92.3/44.1/46.6	93.3/43.3/43.5
Mean		89.5/-	84.9/-	87.2/52.5/48.6	96.1/48.6/53.8	**96.8**/43.4/49.5	75.6/13.3/19.5	85.1/27.6/31.0	**96.8/52.6/55.5**

Table 3: Comparison with SOTA methods on MVTec-AD dataset for multi-class anomaly localization with $AUROC_{seg}$/AP_{seg}/$F1max_{seg}$ metrics.

Method	Non-Diffusion		Diffusion-based		
	DRAEM	UniAD	DDPM	LDM	Ours
PRO	71.1	90.4	49.0	66.3	**90.7**

Table 4: Multi-class anomaly localization results with PRO metric on MVTec-AD datasets.

with both RGB images and 3D point clouds respectively. The training set contains 2,656 images with only anomaly-free samples. The test set consists of 1,197 images, including both normal and abnormal samples. Only RGB images are used in this experiment.

Medical dataset. We also merge three types of medical datasets BraTS2021 (Baid et al. 2021), BTCV (Landman et al. 2015) and LiTs (Bilic et al. 2023) into one *Medical* dataset for multi-class anomaly detection. The training set contains 9,042 slices and the test set consists of 5,208 slices.

Evaluation Metrics. Following prior works, Area Under the Receiver Operating Characteristic Curve (AUROC), Average Precision (AP) and F1-score-max (F1max) are used in both anomaly detection and anomaly localization, where cls represents the image level anomaly detection and seg represents the pixel level anomaly localization. Also, Per-Region-Overlap (PRO) is used in anomaly localization. The DICE score is commonly used in the medical field.

Implementation Details

All images in MVTec-AD and VisA are resized to 256 $\times$ 256. For the denoising network, we adopt the 4-th block of SGDB for connection to SDDB. In this experiment, we adopt ResNet50 as the feature extraction network and choose $n \in \{2, 3, 4\}$ as the feature layers used in calculating the anomaly localization. We utilized the KL method as the Auto-encoder and fine-tune the model before training the denoising network. We train for 1000 epochs on a single NVIDIA Tesla V100 32GB with a batch size of 12. Adam optimiser (Loshchilov and Hutter 2019) with a learning rate of $1e^{-5}$ is set. A Gaussian filter with $\sigma = 5$ is used to smooth the anomaly localization score. For anomaly detection, the anomaly score of the image is the maximum value of the averagely pooled anomaly localization score which undergoes 8 rounds of global average pooling operations with a size of 8×8. During inference, the initial denoising timestep T is set from 1,000. We use DDIM (Song, Meng, and Ermon 2021) as the sampler with 10 steps by default.

Comparison with SOTAs

We conduct and analyze a range of qualitative and quantitative comparison experiments on MVTec-AD, VisA, MVTec-3D and *Medical* datasets. We choose a synthesizing-based method DRAEM (Zavrtanik, Kristan, and Skočaj 2021a), three embedding-based methods MKD (Salehi et al. 2021), PaDiM (Defard et al. 2021) and RD4AD (Deng and Li 2022), a reconstruction-based method EdgRec (Liu et al. 2022), a unified SOTA UniAD (You et al. 2022) method and diffusion-based DDPM and LDM methods. Specifically, we categorize the aforementioned methods into two types: non-diffusion and diffusion-based methods. For the experiments on *Medical* dataset, we follow the BMAD (Bao et al. 2023) benchmark and add two methods STFPM (Yamada and Hotta 2021) and CFLOW (Gudovskiy, Ishizaka, and Kozuka 2022) for comparison.

Qualitative Results. We conducted substantial qualitative experiments on MVTec-AD and VisA datasets to visually demonstrate the superiority of our method in image reconstruction and the accuracy of anomaly localization. As shown in Figure 4, our method exhibits better reconstruction capabilities for anomalous regions compared to the EdgRec on MVTec-AD dataset. In comparison to UniAD shown in Figure 5, our method exhibits more accurate anomaly localization abilities on VisA dataset. More qualitative results will be presented in *Appendix*.

Quantitative Results. As shown in Table 1 and in Ta-

Category		Non-Diffusion Method					Diffusion-based Method		
		PaDiM	MKD	DRAEM	RD4AD	UniAD	DDPM	LDM	Ours
Objects	Bottle	96.1/-	91.8/-	87.6/62.5/56.9	97.8/**68.2**/67.6	98.1/66.0/**69.2**	59.9/ 4.9/11.7	86.9/49.1/50.0	**98.4**/52.2/54.8
	Cable	81.0/-	89.3/-	71.3/14.7/17.8	85.1/26.3/33.6	97.3/39.9/45.2	66.5/ 6.7/10.6	89.3/18.5/26.2	**96.8/50.1/57.8**
	Capsule	96.9/-	88.3/-	50.5/ 6.0/10.0	**98.8/43.4/50.0**	98.5/42.7/46.5	63.1/ 6.2/ 9.7	90.0/ 7.9/27.3	97.1/42.0/45.3
	Hazelnut	96.3/-	91.2/-	96.9/70.0/60.5	97.9/36.2/51.6	98.1/55.2/56.8	91.2/24.1/28.3	95.1/51.2/53.5	**98.3/79.2/80.4**
	Metal Nut	84.8/-	64.2/-	62.2/31.1/21.0	93.8/**62.3**/65.4	94.8/55.5/**66.4**	62.7/14.6/29.2	70.5/19.3/30.7	**97.3**/30.0/38.3
	Pill	87.7/-	69.7/-	94.4/59.1/44.1	**97.5/63.4/65.2**	95.0/44.0/53.9	55.3/ 4.0/ 8.4	74.9/10.2/15.0	95.7/46.0/51.4
	Screw	94.1/-	92.1/-	95.5/33.8/40.6	**99.4**/40.2/44.6	98.3/28.7/37.6	91.1/ 1.8/ 3.8	91.7/ 2.2/ 4.6	97.9/**60.6/59.6**
	Toothbrush	95.6/-	88.9/-	97.7/55.2/55.8	**99.0**/53.6/58.8	98.4/34.9/45.7	76.9/ 4.0/ 7.7	93.7/20.4/ 9.8	**99.0/78.7/72.8**
	Transistor	92.3/-	71.7/-	64.5/23.6/15.1	85.9/42.3/45.2	**97.9/59.5/64.6**	53.2/ 5.8/11.4	85.5/25.0/30.7	95.1/15.6/31.7
	Zipper	94.8/-	86.1/-	98.3/**74.3/69.3**	**98.5**/53.9/60.3	96.8/40.1/49.9	67.4/ 3.5/ 7.6	66.9/ 5.3/ 7.4	96.2/60.7/60.0
Textures	Carpet	97.6/-	95.5/-	98.6/**78.7/73.1**	99.0/58.5/60.4	98.5/49.9/51.1	89.2/18.8/44.3	**99.1**/70.6/66.0	98.6/42.2/46.4
	Grid	71.0/-	82.3/-	98.7/44.5/46.2	**99.2**/46.0/47.4	96.5/23.0/28.4	63.1/ 0.7/ 1.9	52.4/ 1.1/ 1.9	96.6/**66.0/64.1**
	Leather	84.8/-	96.7/-	97.3/**60.3**/57.4	**99.3**/38.0/45.1	98.8/32.9/34.4	97.3/38.9/43.2	99.0/45.9/44.0	98.8/56.1/**62.3**
	Tile	80.5/-	85.3/-	**98.0/93.6/86.0**	95.3/48.5/60.5	91.8/42.1/50.6	87.0/35.2/36.6	90.1/43.9/51.6	92.4/65.7/64.1
	Wood	89.1/-	80.5/-	**96.0/81.4/74.6**	95.3/47.8/51.0	93.2/37.2/41.5	84.7/30.9/37.3	92.3/44.1/46.6	93.3/43.3/43.5
Mean		89.5/-	84.9/-	87.2/52.5/48.6	96.1/48.6/53.8	**96.8**/43.4/49.5	75.6/13.3/19.5	85.1/27.6/31.0	**96.8/52.6/55.5**

表3：在MVTec-AD数据集上使用$AUROC_{seg}/AP_{seg}/F1max_{seg}$指标进行多类别异常定位的SOTA方法对比。

Method	Non-Diffusion		Diffusion-based		
	DRAEM	UniAD	DDPM	LDM	Ours
PRO	71.1	90.4	49.0	66.3	**90.7**

表4：在MVTec-AD数据集上使用PRO指标的多类别异常定位结果。

分别使用RGB图像和3D点云。训练集包含2,656张仅含正常样本的图像。测试集包含1,197张图像，涵盖正常与异常样本。本实验仅使用RGB图像。

医学数据集。我们还将三种医学数据集BraTS2021（Baid等人，2021年）、BTCV（Landman等人，2015年）和LiTs（Bilic等人，2023年）合并为一个*Medical*数据集，用于多类别异常检测。训练集包含9,042个切片，测试集包含5,208个切片。评估指标。遵循先前工作，异常检测和异常定位均采用受试者工作特征曲线下面积（AUROC）、平均精度（AP）和最大F1分数（F1max），其中cls表示图像级异常检测，seg表示像素级异常定位。此外，异常定位中还使用了每区域重叠度（PRO）。DICE分数在医学领域被广泛使用。

实现细节

MVTec-AD和VisA中的所有图像均被调整为256 × 256的尺寸。对于去噪网络，我们采用SGDB的第四模块与SDDB连接。在本实验中，我们采用ResNet50作为特征提取网络，并选择$n \in \{2, 3, 4\}$作为用于计算异常定位的特征层。我们使用KL方法作为自编码器，并在训练去噪网络前对模型进行微调。我们在单张NVIDIA Tesla V100 32GB显卡上以批次大小为12训练了1000轮，优化器采用Adam。优化器（Loshchilov和Hutter，2019）的学习率设置为1e^{-5}。使用σ =为5的高斯滤波器对异常定位分数进行平滑处理。对于异常检测，图像的异常分数是经过8轮大小为8×8的全局平均池化操作后的平均池化异常定位分数的最大值。在推理过程中，初始去噪时间步长T从1,000开始设置。默认情况下，我们使用DDIM（Song、Meng和Ermon，2021）作为采样器，共10步。

与SOTAs的比较

我们在MVTec-AD、VisA、MVTec-3D和*Medical*数据集上进行了一系列定性与定量对比实验。我们选择了基于合成的方法DRAEM（Zavrtanik, Kristan和Skočaj 2021a）、三种基于嵌入的方法MKD（Salehi等人2021）、PaDiM（Defard等人2021）和RD4AD（Deng和Li 2022）、基于重建的方法EdgRec（Liu等人2022）、统一的SOTA方法UniAD（You等人2022）以及基于扩散的DDPM和LDM方法。具体而言，我们将上述方法分为两类：非扩散方法与基于扩散的方法。在*Medical*数据集的实验中，我们遵循BMAD（Bao等人2023）基准，并额外加入STFPM（Yamada和Hotta 2021）与CFLOW（Gudovskiy, Ishizaka和Kozuka 2022）两种方法进行比较。

定性结果。我们在MVTec-AD和VisA数据集上进行了大量定性实验，以直观展示本方法在图像重建方面的优越性及异常定位的准确性。如图4所示，在MVTec-AD数据集上，相较于EdgRec，我们的方法对异常区域展现出更优的重建能力。与图5所示的UniAD相比，我们的方法在VisA数据集上表现出更精确的异常定位能力。更多定性结果将在*Appendix*中呈现。

定量结果。如表1和表

Metrics	Non-Diffusion		Diffusion-based		
	DRAEM	UniAD	DDPM	LDM	**Ours**
$AUROC_{cls}$	63.2	78.9	66.3	68.5	**84.6**
AP_{cls}	86.1	93.4	78.0	90.6	**94.8**
$F1max_{cls}$	89.2	91.4	86.6	91.6	**95.5**
$AUROC_{seg}$	93.2	**96.5**	90.7	92.2	96.4
AP_{seg}	16.8	21.2	6.0	9.3	**25.3**
$F1max_{seg}$	20.2	28.0	10.7	13.5	**32.2**
PRO	55.0	**88.1**	69.7	73.8	87.8

Table 5: Quantitative comparisons on MVTec-3D dataset.

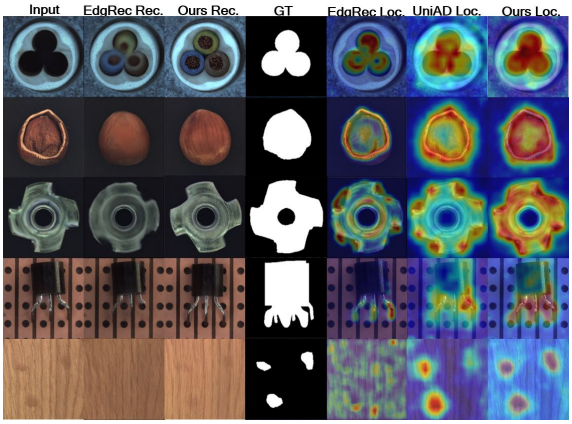

Figure 4: Qualitative illustration on MVTec-AD dataset.

ble 3, our method achieves SOTA AUROC/AP/F1max metrics of 97.2/99.0/96.5 and 96.8/52.6/55.5 for image-wise and pixel-wise respectively for multi-class setting on MVTec-AD dataset. For the diffusion-based methods, our approach significantly outperforms existing DDPM and LDM methods in terms of 11.7↑ in AUROC and 25↑ in AP for anomaly localization. For non-diffusion methods, our approach surpasses existing methods in both metrics, especially at the pixel level, where our method exceeds UniAD by 9.2↑/6.0↑ in AP/F1max. Our method has also demonstrated its superiority on VisA dataset, as shown in Table 2. Our approach exhibits significant improvements compared to diffusion-based methods of 30.1↑/9.4↑ than the LDM method in image/pixel AUROC. It also performs well compared to UniAD by 4.9↑/6.0↑ in pixel AP/F1max metrics. Detailed experiments for each category are provided in *Appendix*. We have extended the method to 3D datasets and medical domain datasets. Table 5 and Table 6 show the effectiveness and scalability of our method on MVtec-3D and *Medical* datasets, with results surpassing the state of the art (SOTA).

Metrics	MKD	CFLOW	RD4AD	PaDiM	PatchCore	STFPM	UniAD	Ours
$AUROC_{cls}$	70.9	62.0	74.7	64.6	76.0	72.2	76.4	**77.2**
$AUROC_{seg}$	92.8	93.2	96.2	93.0	96.8	93.4	96.7	**96.9**
PRO	79.3	79.0	88.0	79.2	86.6	86.0	87.4	**87.7**
$DICE$	21.9	13.5	19.5	15.2	21.7	17.1	28.7	**32.3**

Table 6: Quantitative comparisons on *Medical* dataset.

SD	MSG	SGEB3	SGEB4	BN+ReLU	IN+SiLU	cls	seg
✓						79.3	89.5
✓	✓					95.1	91.1
✓	✓	✓				95.3	89.1
✓	✓	✓	✓			93.8	91.2
✓	✓	✓		✓		96.7	96.7
✓	✓	✓			✓	**97.2**	**96.8**

Table 7: Ablation studies on the design of DiAD with AUROC metrics.

Ablation Studies

The architecture design of DiAD. We investigate the importance of each module in DiAD as shown in Table 7. SD indicates only the diffusion model without connecting to the SG network which is the LDM's architecture. MSG indicates only the middle block of the SG network adding to the middle of SD. SGEB3 and SGEB4 indicate directly skip-connecting to the corresponding SDDB. When connecting SGDB3 and SGDB4 at the same time, more details of the original images are preserved in terms of texture, but the reconstruction ability for large anomaly areas decreases. Using the combination of IN+SiLU in the SFF block yields better results compared to using BN+ReLU.

Effect of pre-trained feature extractors. Table 8 shows the quantitative comparison of using different pre-trained backbones as feature extraction networks. ResNet50 achieved the best performance in anomaly classification metrics, while WideResNet101 excelled in anomaly segmentation.

Backbone		$AUROC_{cls}$	AP_{cls}	$F1max_{cls}$	$AUROC_{seg}$	AP_{seg}	$F1max_{seg}$	PRO
VGG	16	91.8	97.2	93.9	92.1	47.2	50.5	80.1
	19	91.3	96.9	93.7	92.3	47.5	50.6	80.4
ResNet	18	94.7	98.1	96.0	96.0	49.9	53.3	89.1
	34	95.2	98.3	95.7	96.2	51.2	54.5	89.6
	50	**97.2**	**99**	**96.5**	96.8	52.6	55.5	90.7
	101	96.2	98.4	**96.5**	**96.9**	52.9	56.4	91.2
WideResNet	50	95.9	98.6	**96.5**	96.4	51.8	55.1	89.3
	101	95.6	98.3	95.8	**96.9**	**54.6**	**56.5**	**91.4**
EfficientNet	b0	93.5	97.7	94.7	94.0	50.0	52.4	84.0
	b2	94.2	98.0	95.1	94.1	48.6	52.1	84.2
	b4	92.8	97.5	94.8	93.6	47.2	50.7	83.5

Table 8: Ablation studies on different feature extractors.

Effect of feature layers used in anomaly score calculating. After extracting feature maps of 5 different scales using a pre-trained backbone, the anomaly scores are calculated by computing the cosine similarity between feature maps from different layers. The experimental results, as shown in Table 9, indicate that using feature maps from layers f_2, f_3, and f_4 (with corresponding sizes of 64×64, 32×32, and 16×16) yields the best performance.

f_1	f_2	f_3	f_4	f_5	$AUROC_{cls}$	AP_{cls}	$F1max_{cls}$	$AUROC_{seg}$	AP_{seg}	$F1max_{seg}$
✓	✓	✓	✓	✓	93.8	97.8	95.0	94.0	42.0	45.9
✓	✓	✓	✓		96.7	98.7	96.1	96.7	52.5	55.2
	✓	✓	✓	✓	93.4	97.1	93.6	95.2	48.5	51.3
✓	✓	✓			97.1	**99.0**	**96.8**	96.4	49.4	53.1
	✓	✓	✓		**97.2**	**99.0**	96.5	**96.8**	**52.6**	**55.5**
	✓	✓			94	97.4	94.2	95.3	48.5	51.7
		✓	✓		97.1	**99.0**	**96.8**	96.4	49.4	53.1

Table 9: Ablation studies on the feature layers used in calculating the anomaly localization score based on ResNet50.

Metrics	Non-Diffusion		Diffusion-based		
	DRAEM	UniAD	DDPM	LDM	**Ours**
$AUROC_{cls}$	63.2	78.9	66.3	68.5	**84.6**
AP_{cls}	86.1	93.4	78.0	90.6	**94.8**
$F1max_{cls}$	89.2	91.4	86.6	91.6	**95.5**
$AUROC_{seg}$	93.2	**96.5**	90.7	92.2	96.4
AP_{seg}	16.8	21.2	6.0	9.3	**25.3**
$F1max_{seg}$	20.2	28.0	10.7	13.5	**32.2**
PRO	55.0	**88.1**	69.7	73.8	87.8

表5：MVTec-3D数据集上的定量比较。

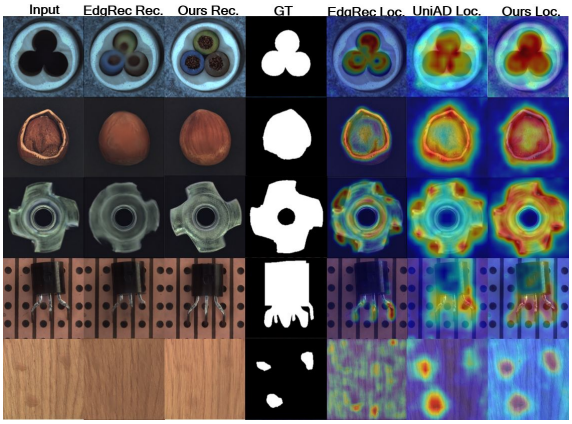

图4：MVTec-AD数据集上的定性说明。

如表3所示，在多类别MVTec-AD数据集上，我们的方法在图像级和像素级分别取得了97.2/99.0/96.5和96.8/52.6/55.5的SOTA性能（AUROC/AP/F1max指标）。对于基于扩散的方法，我们的方法在异常定位任务上显著优于现有DDPM和LDM方法，AUROC提升11.7↑，AP提升25↑。在非扩散方法中，我们的方法在各项指标上均超越现有方法，尤其在像素级别表现突出，AP/F1max指标较UniAD分别提升9.2↑/6.0↑。如表2所示，我们的方法在VisA数据集上也展现出优越性：相比基于扩散的LDM方法，图像/像素级AUROC提升30.1↑/9.4↑；在像素级AP/F1max指标上较UniAD提升4.9↑/6.0↑。各类别的详细实验数据见*Appendix*。我们还将方法扩展至3D数据集和医学领域数据集，表5和表6展示了本方法在MVtec-3D和*Medical*数据集上的有效性与可扩展性，其结果均超越当前最优水平（SOTA）。

Metrics	MKD	CFLOW	RD4AD	PaDiM	PatchCore	STFPM	UniAD	Ours
$AUROC_{cls}$	70.9	62.0	74.7	64.6	76.0	72.2	76.4	**77.2**
$AUROC_{seg}$	92.8	93.2	96.2	93.0	96.8	93.4	96.7	**96.9**
PRO	79.3	79.0	88.0	79.2	86.6	86.0	87.4	**87.7**
$DICE$	21.9	13.5	19.5	15.2	21.7	17.1	28.7	**32.3**

表6：在*Medical*数据集上的定量比较。

SD	MSG	SGEB3	SGEB4	BN+ReLU	IN+SiLU	cls	seg
✓						79.3	89.5
✓	✓					95.1	91.1
✓	✓	✓				95.3	89.1
✓	✓	✓	✓			93.8	91.2
✓	✓	✓		✓		96.7	96.7
✓	✓	✓			✓	**97.2**	**96.8**

表7：基于AU-ROC指标的DiAD设计消融研究。

消融研究

DiAD的架构设计。如表7所示，我们研究了DiAD中每个模块的重要性。SD表示仅使用扩散模型而未连接SG网络，即LDM的架构。MSG表示仅在SD的中间层添加SG网络的中间块。SGEB3和SGEB4表示直接跳跃连接到对应的SDDB。当同时连接SGDB3和SGDB4时，原始图像的纹理细节保留得更好，但对大异常区域的重建能力会下降。在SFF块中使用IN+SiLU组合相比使用BN+ReLU能获得更好的结果。

预训练特征提取器的影响。表8展示了使用不同预训练骨干网络作为特征提取网络的定量比较。ResNet50在异常分类指标上取得了最佳性能，而WideResNet101在异常分割方面表现优异。

Backbone		$AUROC_{cls}$	AP_{cls}	$F1max_{cls}$	$AUROC_{seg}$	AP_{seg}	$F1max_{seg}$	PRO
VGG	16	91.8	97.2	93.9	92.1	47.2	50.5	80.1
	19	91.3	96.9	93.7	92.3	47.5	50.6	80.4
ResNet	18	94.7	98.1	96.0	96.0	49.9	53.3	89.1
	34	95.2	98.3	95.7	96.2	51.2	54.5	89.6
	50	**97.2**	**99**	**96.5**	96.8	52.6	55.5	90.7
	101	96.2	98.4	**96.5**	**96.9**	52.9	56.4	91.2
WideResNet	50	95.9	98.6	**96.5**	96.4	51.8	55.1	89.3
	101	95.6	98.3	95.8	**96.9**	**54.6**	**56.5**	**91.4**
EfficientNet	b0	93.5	97.7	94.7	94.0	50.0	52.4	84.0
	b2	94.2	98.0	95.1	94.1	48.6	52.1	84.2
	b4	92.8	97.5	94.8	93.6	47.2	50.7	83.5

表8：不同特征提取器的消融研究。

异常分数计算中使用的特征层效果。通过使用预训练骨干网络提取5个不同尺度的特征图后，通过计算不同层特征图之间的余弦相似度来计算异常分数。如表9所示的实验结果表明，使用来自层f_2、f_3和f_4 (的特征图（对应尺寸分别为64×64、32×32和16×16)）可获得最佳性能。

f_1	f_2	f_3	f_4	f_5	$AUROC_{cls}$	AP_{cls}	$F1max_{cls}$	$AUROC_{seg}$	AP_{seg}	$F1max_{seg}$
✓	✓	✓	✓	✓	93.8	97.8	95.0	94.0	42.0	45.9
✓	✓	✓	✓		96.7	98.7	96.1	96.7	52.5	55.2
	✓	✓	✓	✓	93.4	97.1	93.6	95.2	48.5	51.3
✓	✓	✓			97.1	**99.0**	**96.8**	96.4	49.4	53.1
	✓	✓	✓		**97.2**	**99.0**	96.5	**96.8**	**52.6**	**55.5**
	✓	✓			94	97.4	94.2	95.3	48.5	51.7
		✓	✓		97.1	**99.0**	**96.8**	96.4	49.4	53.1

表9：基于ResNet50计算异常定位分数所用特征层的消融研究。

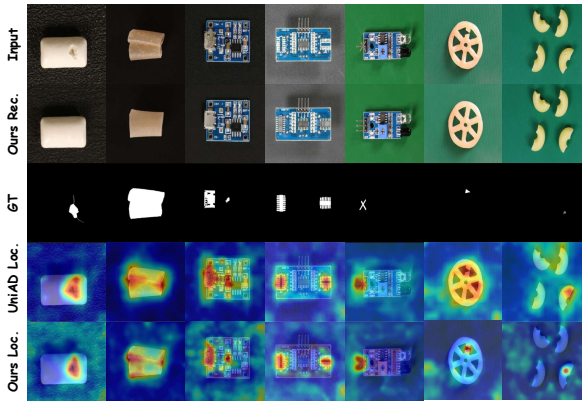

Figure 5: Qualitative results on VisA dataset.

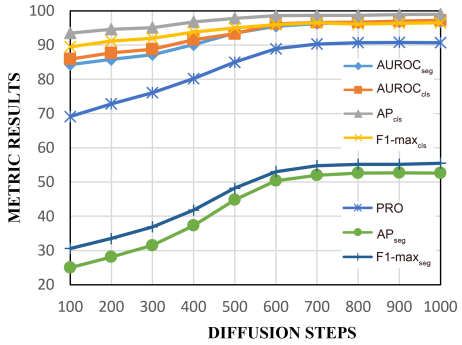

Figure 6: Ablation studies on different diffusion timesteps.

Effect of forward diffusion timesteps. Increasing the number of diffusion steps in the forward process impacts the performance of image reconstruction. The experimental results, depicted in Figure 6, indicate that with an increasing number of forward diffusion steps, the image approaches pure Gaussian noise, while the anomaly reconstruction ability improves as well. Nevertheless, when the number of forward diffusion steps is less than 600, a significant decline in performance occurs because the number of steps is insufficient for anomaly reconstruction.

Conclusion

This paper proposes a diffusion-based DiAD framework to address the issue of category and semantic loss in the stable diffusion model for multi-class anomaly detection. We propose the Semantic-Guided network and Spatial-aware Feature Fusion block to better reconstruct the abnormal regions while maintaining the same semantic information as the input image. Our approach achieves state-of-the-art performance on MVTec-AD and VisA datasets, significantly outperforming the non-diffusion and diffusion-based methods.
Limitation. Although our method has demonstrated exceptional performance in reconstructing anomalies, it can be susceptible to the influence of background impurities, resulting in errors in localization and classification. In the future, we will further explore diffusion models and enhance the background's anti-interference capability for multi-class anomaly detection. Additionally, we will incorporate multi-modal assistance in our anomaly detection. Lastly, we will utilize larger models to enhance reconstruction performance.

References

Amit, T.; Shaharbany, T.; Nachmani, E.; and Wolf, L. 2022. SegDiff: Image Segmentation with Diffusion Probabilistic Models. arXiv:2112.00390.

Baid, U.; Ghodasara, S.; Mohan, S.; Bilello, M.; Calabrese, E.; Colak, E.; Farahani, K.; Kalpathy-Cramer, J.; Kitamura, F. C.; Pati, S.; et al. 2021. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*.

Bao, J.; Sun, H.; Deng, H.; He, Y.; Zhang, Z.; and Li, X. 2023. BMAD: Benchmarks for Medical Anomaly Detection. *arXiv preprint arXiv:2306.11876*.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 9592–9600.

Bergmann, P.; Jin, X.; Sattlegger, D.; and Steger, C. 2022. The MVTec 3D-AD Dataset for Unsupervised 3D Anomaly Detection and Localization. In *VISGRAPP*. SCITEPRESS - Science and Technology Publications.

Bilic, P.; Christ, P.; Li, H. B.; Vorontsov, E.; Ben-Cohen, A.; Kaissis, G.; Szeskin, A.; Jacobs, C.; Mamani, G. E. H.; Chartrand, G.; et al. 2023. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84: 102680.

Cao, Y.; Wan, Q.; Shen, W.; and Gao, L. 2022. Informative knowledge distillation for image anomaly segmentation. *Knowledge-Based Systems*, 248: 108846.

Cao, Y.; Xu, X.; Sun, C.; Cheng, Y.; Du, Z.; Gao, L.; and Shen, W. 2023. Segment Any Anomaly without Training via Hybrid Prompt Regularization. *arXiv preprint arXiv:2305.10724*.

Chen, R.; Xie, G.; Liu, J.; Wang, J.; Luo, Z.; Wang, J.; and Zheng, F. 2023a. Easynet: An easy network for 3d industrial anomaly detection. In *ACM MM*, 7038–7046.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2022. DiffusionDet: Diffusion Model for Object Detection. arXiv:2211.09788.

Chen, X.; Han, Y.; and Zhang, J. 2023. A Zero-/Few-Shot Anomaly Classification and Segmentation Method for CVPR 2023 VAND Workshop Challenge Tracks 1&2: 1st Place on Zero-shot AD and 4th Place on Few-shot AD. *arXiv preprint arXiv:2305.17382*.

Chen, X.; Zhang, J.; Tian, G.; He, H.; Zhang, W.; Wang, Y.; Wang, C.; Wu, Y.; and Liu, Y. 2023b. CLIP-AD: A Language-Guided Staged Dual-Path Model for Zero-shot Anomaly Detection. *arXiv preprint arXiv:2311.00453*.

Defard, T.; Setkov, A.; Loesch, A.; and Audigier, R. 2021. Padim: a patch distribution modeling framework for anomaly detection and localization. In *ICPR*, 475–489. Springer.

Deng, H.; and Li, X. 2022. Anomaly detection via reverse distillation from one-class embedding. In *CVPR*, 9737–9746.

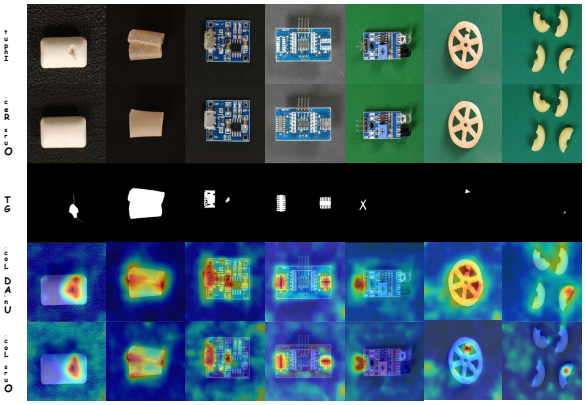

图5：VisA数据集上的定性结果。

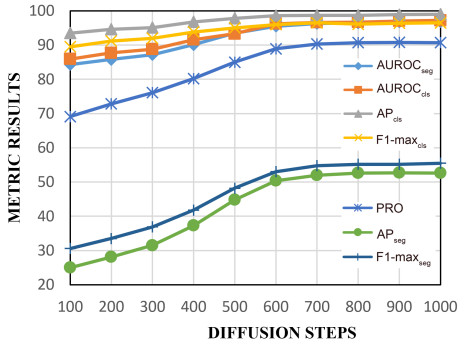

图6：不同扩散时间步的消融研究。

前向扩散时间步数的影响。增加前向过程中的扩散步数会影响图像重建的性能。实验结果如图6所示，表明随着前向扩散步数的增加，图像逐渐接近纯高斯噪声，同时异常重建能力也有所提升。然而，当正向扩散步数少于600步时，由于步数不足以进行异常重建，性能会出现显著下降。

结论

本文提出了一种基于扩散的DiAD框架，以解决多类别异常检测中稳定扩散模型存在的类别和语义丢失问题。我们提出了语义引导网络和空间感知特征融合模块，以更好地重建异常区域，同时保持与输入图像相同的语义信息。我们的方法在MVTec-AD和VisA数据集上实现了最先进的性能，显著超越了非扩散方法和基于扩散的方法。

局限性。尽管我们的方法在异常重建方面表现出卓越性能，但它容易受到背景杂质的影响，导致定位和分类出现误差。未来，我们将进一步探索扩散模型并加以改进。背景对多类异常检测的抗干扰能力。此外，我们将在异常检测中融入多模态辅助。最后，我们将利用更大的模型来提升重建性能。

参考文献

阿米特，T.；沙哈尔巴尼，T.；纳赫马尼，E.；和沃尔夫，L. 2022。SegDiff：基于扩散概率模型的图像分割。arXiv:2112.00390。

Baid, U.; Ghodasara, S.; Mohan, S.; Bilello, M.; Calabrese, E.; Colak, E.; Farahani, K.; Kalpathy-Cramer, J.; Kitamura, F. C.; Pati, S.; 等. 2021. rsna-asnr-miccai brats 2021脑肿瘤分割与放射基因组分类基准.

arXiv preprint arXiv:2107.02314. Bao, J.; Sun, H.; Deng, H.; He, Y.; Zhang, Z.; Li, X. 2023. BMAD：医学异常检测基准. *arXiv preprint arXiv:2306.11876*. Bergmann, P.; Fauser, M.; Sattlegger, D.; Steger, C. 2019. MVTec AD——一个用于无监督异常检测的综合真实世界数据集. 于 *CVPR*, 9592–9600. Bergmann, P.; Jin, X.; Sattlegger, D.; Steger, C. 2022. 用于无监督3D异常检测与定位的MVTec 3D-AD数据集. 于 *VISGRAPP*. SCITEPRESS - Science and Technology Publications. Bilic, P.; Christ, P.; Li, H. B.; Vorontsov, E.; Ben-Cohen, A.; Kaissis, G.; Szeskin, A.; Jacobs, C.; Mamani, G. E. H.; Chartrand, G.; 等. 2023. 肝脏肿瘤分割基准（lits）. *Medical Image Analysis*, 84: 102680. Cao, Y.; Wan, Q.; Shen, W.; Gao, L. 2022. 用于图像异常分割的信息化知识蒸馏. *Knowledge-Based Systems*, 248: 108846. Cao, Y.; Xu, X.; Sun, C.; Cheng, Y.; Du, Z.; Gao, L.; Shen, W. 2023. 通过混合提示正则化实现无需训练的任意异常分割. *arXiv preprint arXiv:2305.10724*. Chen, R.; Xie, G.; Liu, J.; Wang, J.; Luo, Z.; Wang, J.; Zheng, F. 2023a. Easynet：一个用于3D工业异常检测的简易网络. 于 *ACM MM*, 7038–7046. Chen, S.; Sun, P.; Song, Y.; Luo, P. 2022. DiffusionDet：用于目标检测的扩散模型. arXiv:2211.09788. Chen, X.; Han, Y.; Zhang, J. 2023. 一种用于CVPR 2023 VAND研讨会挑战赛道1&2的零样本/少样本异常分类与分割方法：零样本异常检测第一名，少样本异常检测第四名. *arXiv preprint arXiv:2305.17382*. Chen, X.; Zhang, J.; Tian, G.; He, H.; Zhang, W.; Wang, Y.; Wang, C.; Wu, Y.; Liu, Y. 2023b. CLIP-AD：一种语言引导的分阶段双路径模型，用于零样本异常检测.

arXiv preprint arXiv:2311.00453. Defard, T.; Setkov, A.; Loesch, A.; Audigier, R. 2021. Padim：一种用于异常检测与定位的块分布建模框架. 于 *ICPR*, 475–489. Springer. Deng, H.; Li, X. 2022. 通过从单类嵌入的反向蒸馏进行异常检测. 于 *CVPR*, 9737–9746.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. Ieee.

Ding, C.; Pang, G.; and Shen, C. 2022. Catching both gray and black swans: Open-set supervised anomaly detection. In *CVPR*, 7388–7398.

Elfwing, S.; Uchibe, E.; and Doya, K. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107: 3–11.

Gu, Z.; Liu, L.; Chen, X.; Yi, R.; Zhang, J.; Wang, Y.; Wang, C.; Shu, A.; Jiang, G.; and Ma, L. 2023. Remembering Normality: Memory-guided Knowledge Distillation for Unsupervised Anomaly Detection. In *ICCV*, 16401–16409.

Gudovskiy, D.; Ishizaka, S.; and Kozuka, K. 2022. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *WACV*, 98–107.

Hahnloser, R. H.; Sarpeshkar, R.; Mahowald, M. A.; Douglas, R. J.; and Seung, H. S. 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *nature*, 405(6789): 947–951.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; and Salimans, T. 2022. Imagen Video: High Definition Video Generation with Diffusion Models. arXiv:2210.02303.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*, volume 33, 6840–6851.

Huang, C.; Guan, H.; Jiang, A.; Zhang, Y.; Spratling, M.; and Wang, Y.-F. 2022. Registration based few-shot anomaly detection. In *ECCV*, 303–319. Springer.

Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Bach, F. R.; and Blei, D. M., eds., *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, 448–456. JMLR.org.

Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; and Dabeer, O. 2023. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, 19606–19616.

Kingma, D. P.; and Welling, M. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114.

Landman, B.; Xu, Z.; Igelsias, J.; Styner, M.; Langerak, T.; and Klein, A. 2015. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, 12.

Li, C.-L.; Sohn, K.; Yoon, J.; and Pfister, T. 2021. Cutpaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, 9664–9674.

Liang, Y.; Zhang, J.; Zhao, S.; Wu, R.; Liu, Y.; and Pan, S. 2023. Omni-frequency channel-selection representations for unsupervised anomaly detection. *IEEE Transactions on Image Processing*.

Liu, J.; Xie, G.; Wang, J.; Li, S.; Wang, C.; Zheng, F.; and Jin, Y. 2023. Deep Industrial Image Anomaly Detection: A Survey. *arXiv preprint arXiv:2301.11514*, 2.

Liu, T.; Li, B.; Zhao, Z.; Du, X.; Jiang, B.; and Geng, L. 2022. Reconstruction from edge image combined with color and gradient difference for industrial surface anomaly detection. arXiv:2210.14485.

Liznerski, P.; Ruff, L.; Vandermeulen, R. A.; Franks, B. J.; Kloft, M.; and Müller, K. 2021. Explainable Deep One-Class Classification. In *ICLR*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

Mousakhan, A.; Brox, T.; and Tayyub, J. 2023. Anomaly Detection with Conditioned Denoising Diffusion Models. arXiv:2305.15956.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards total recall in industrial anomaly detection. In *CVPR*, 14318–14328.

Salehi, M.; Mirzaei, H.; Hendrycks, D.; Li, Y.; Rohban, M. H.; and Sabokrou, M. 2022. A Unified Survey on Anomaly, Novelty, Open-Set, and Out-of-Distribution Detection: Solutions and Future Challenges. arXiv:2110.14051.

Salehi, M.; Sadjadi, N.; Baselizadeh, S.; Rohban, M. H.; and Rabiee, H. R. 2021. Multiresolution knowledge distillation for anomaly detection. In *CVPR*, 14902–14912.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *ICLR*. OpenReview.net.

Tan, D. S.; Chen, Y.-C.; Chen, T. P.-C.; and Chen, W.-C. 2021. Trustmae: A noise-resilient defect classification framework using memory-augmented auto-encoders with trust regions. In *WACV*, 276–285.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 6105–6114. PMLR.

Tao, X.; Gong, X.; Zhang, X.; Yan, S.; and Adak, C. 2022. Deep Learning for Unsupervised Anomaly Localization in Industrial Images: A Survey. *IEEE Transactions on Instrumentation and Measurement*, 71: 1–21.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2017. Instance Normalization: The Missing Ingredient for Fast Stylization. arXiv:1607.08022.

Wang, Y.; Peng, J.; Zhang, J.; Yi, R.; Wang, Y.; and Wang, C. 2023. Multimodal Industrial Anomaly Detection via Hybrid Fusion. In *CVPR*, 8032–8041.

Wu, J.; Li, J.; Zhang, J.; Zhang, B.; Chi, M.; Wang, Y.; and Wang, C. 2023. PVG: Progressive Vision Graph for Vision Recognition. *arXiv preprint arXiv:2308.00574*.

邓，J.；董，W.；Socher，R.；李，L.-J.；李，K.；和 Fei-Fei，L. 2009。Imagenet：一个大规模分层图像数据库。于 *CVPR*，248–255。Ieee。 丁C.；庞，G.；和沈，C. 2022。捕捉灰天鹅与黑天鹅：开放集监督异常检测。于 *CVPR*，7388–7398。Elfwing，S.；Uchibe，E.；和 Doya，K. 2018。用于强化学习中神经网络函数逼近的 Sigmoid 加权线性单元。*Neural networks*，107：3–11。 顾Z.；刘，L.；陈，X.；易，R.；张，J.；王，Y.；王，C.；舒，A.；江，G.；和马，L. 2023。记住正常性：用于无监督异常检测的记忆引导知识蒸馏。于 *ICCV*，16401–16409。Gudovskiy，D.；Ishizaka，S.；和 Kozuka，K. 2022。Cflow-ad：通过条件归一化流实现具有定位功能的实时无监督异常检测。于 *WACV*，98–107。Hahnloser，R. H.；Sarpeshkar，R.；Mahowald，M. A.；Douglas，R. J.；和 Seung，H. S. 2000。数字选择与模拟放大共存于受皮层启发的硅电路中。*nature*，405(6789)：947–951。 何K.；张，X.；任，S.；和孙，J. 2016。用于图像识别的深度残差学习。于 *CVPR*，770–778。Ho，J.；Chan，W.；Saharia，C.；Whang，J.；Gao，R.；Gritsenko，A.；Kingma，D. P.；Poole，B.；Norouzi，M.；Fleet，D. J.；和 Salimans，T. 2022。Imagen Video：使用扩散模型生成高清视频。arXiv:2210.02303。Ho，J.；Jain，A.；和 Abbeel，P. 2020。去噪扩散概率模型。于 *NeurIPS*，第 33 卷，6840–6851。 黄，C.；管，H.；江，A.；张，Y.；Spratling，M.；和王，Y.-F. 2022。基于配准的小样本异常检测。于 *ECCV*，303–319。Springer。Ioffe，S.；和 Szegedy，C. 2015。批量归一化：通过减少内部协变量偏移加速深度网络训练。载于 Bach，F. R.；和 Blei，D. M. 编，*ICML*，*JMLR Workshop and Conference Proceedings* 第 37 卷，448–456。JMLR.org。

Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; 和 Dabeer, O. 2023. Winclip: 零样本/少样本异常分类与分割。于 *CVPR*, 19606–19616。

Kingma, D. P.; 和 Welling, M. 2022. 自动编码变分贝叶斯. arXiv:1312.6114. Landman, B.; Xu, Z.; Igelsias, J.; Styner, M.; Langerak, T.; 和 Klein, A. 2015. MICCAI 颅腔外多图谱标记–研讨会与挑战赛. 于 *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 第5卷, 12. Li, C.-L.; Sohn, K.; Yoon, J.; 和 Pfister, T. 2021. CutPaste: 用于异常检测与定位的自监督学习. 于 *CVPR*, 9664–9674. Liang, Y.; Zhang, J.; Zhao, S.; Wu, R.; Liu, Y.; 和 Pan, S. 2023. 用于无监督异常检测的全频通道选择表示. *IEEE Transactions on Image Processing*.

刘, J.; 谢, G.; 王, J.; 李, S.; 王, C.; 郑, F.; 和 金, Y. 2023. 深度工业图像异常检测综述.

arXiv preprint arXiv:2301.11514, 2. 刘T.; 李, B.; 赵, Z.; 杜, X.; 姜, B.; 和 耿, L. 2022. 结合颜色与梯度差异的边缘图像重建用于工业表面异常检测. arXiv:2210.14485.Liznerski, P.; Ruff, L.; Vandermeulen, R. A.; Franks, B. J.; Kloft, M.; 和 Müller, K. 2021. 可解释的深度单类分类. 于 *ICLR*.Loshchilov, I.; 和 Hutter, F. 2019. 解耦权重衰减正则化. arXiv:1711.05101.Mousakhan, A.; Brox, T.; 和 Tayyub, J. 2023. 基于条件去噪扩散模型的异常检测. arXiv:2305.15956.Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; 和 Ommer, B. 2022. 基于潜在扩散模型的高分辨率图像合成. arXiv:2112.10752.Ronneberger, O.; Fischer, P.; 和 Brox, T. 2015. U-Net: 用于生物医学图像分割的卷积网络. 于 *MICCAI*, 234–241. Springer.Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; 和 Gehler, P. 2022. 迈向工业异常检测的完全召回. 于 *CVPR*, 14318–14328.Salehi, M.; Mirzaei, H.; Hendrycks, D.; Li, Y.; Rohban, M. H.; 和 Sabokrou, M. 2022. 关于异常、新颖性、开放集和分布外检测的统一综述：解决方案与未来挑战. arXiv:2110.14051.Salehi, M.; Sadjadi, N.; Baselizadeh, S.; Rohban, M. H.; 和 Rabiee, H. R. 2021. 用于异常检测的多分辨率知识蒸馏. 于 *CVPR*, 14902–14912.Song, J.; Meng, C.; 和 Ermon, S. 2021. 去噪扩散隐式模型. 于 *ICLR*. OpenReview.net.Tan, D. S.; Chen, Y.-C.; Chen, T. P.-C.; 和 Chen, W.-C. 2021. TrustMAE: 一种使用带信任区域的记忆增强自编码器的抗噪缺陷分类框架. 于 *WACV*, 276–285.Tan, M.; 和 Le, Q. 2019. EfficientNet: 重新思考卷积神经网络的模型缩放. 于 *ICML*, 6105–6114. PMLR.Tao, X.; Gong, X.; Zhang, X.; Yan, S.; 和 Adak, C. 2022. 工业图像无监督异常定位的深度学习：综述.

IEEE Transactions on Instru- mentation and Measurement, 71: 1–21.Ulyanov, D.; Vedaldi, A.; 和 Lempitsky, V. 2017. 实例归一化：快速风格化的缺失成分. arXiv:1607.08022. 王Y.; 彭, J.; 张, J.; 易, R.; 王, Y.; 和 王, C. 2023. 通过混合融合的多模态工业异常检测. 于 *CVPR*, 8032–8041. 吴J.; 李, J.; 张, J.; 张, B.; 迟, M.; 王, Y.; 和 王, C. 2023. PVG: 用于视觉识别的渐进式视觉图.

arXiv preprint arXiv:2308.00574.

Wyatt, J.; Leach, A.; Schmon, S. M.; and Willcocks, C. G. 2022. AnoDDPM: Anomaly Detection with Denoising Diffusion Probabilistic Models using Simplex Noise. In *CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, 649–655. IEEE.

Xie, G.; Wang, J.; Liu, J.; Jin, Y.; and Zheng, F. 2023. Pushing the Limits of Fewshot Anomaly Detection in Industry Vision: Graphcore. In *ICLR*.

Yamada, S.; and Hotta, K. 2021. Reconstruction student with attention for student-teacher pyramid matching. *arXiv preprint arXiv:2111.15376*.

Yan, X.; Zhang, H.; Xu, X.; Hu, X.; and Heng, P. 2021. Learning Semantic Context from Normal Samples for Unsupervised Anomaly Detection. In *AAAI*, 3110–3118.

Yi, J.; and Yoon, S. 2020. Patch SVDD: Patch-level SVDD for Anomaly Detection and Segmentation. In *ACCV*.

Yoon, J.; Sohn, K.; Li, C.-L.; Arik, S. O.; Lee, C.-Y.; and Pfister, T. 2022. Self-supervise, Refine, Repeat: Improving Unsupervised Anomaly Detection. *Transactions on Machine Learning Research*.

You, Z.; Cui, L.; Shen, Y.; Yang, K.; Lu, X.; Zheng, Y.; and Le, X. 2022. A Unified Model for Multi-class Anomaly Detection. In *NeurIPS*, volume 35, 4571–4584.

Yu, J.; Zheng, Y.; Wang, X.; Li, W.; Wu, Y.; Zhao, R.; and Wu, L. 2021. FastFlow: Unsupervised Anomaly Detection and Localization via 2D Normalizing Flows. arXiv:2111.07677.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. In *BMVC*. BMVA Press.

Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2021a. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, 8330–8339.

Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2021b. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112: 107706.

Zhang, H.; Wang, Z.; Wu, Z.; and Jiang, Y.-G. 2023a. DiffusionAD: Denoising Diffusion for Anomaly Detection. arXiv:2303.08730.

Zhang, J.; Chen, X.; Xue, Z.; Wang, Y.; Wang, C.; and Liu, Y. 2023b. Exploring Grounding Potential of VQA-oriented GPT-4V for Zero-shot Anomaly Detection. *arXiv preprint arXiv:2311.02612*.

Zhang, J.; Li, X.; Li, J.; Liu, L.; Xue, Z.; Zhang, B.; Jiang, Z.; Huang, T.; Wang, Y.; and Wang, C. 2023c. Rethinking Mobile Block for Efficient Attention-based Models. In *ICCV*, 1389–1400.

Zhang, J.; Li, X.; Wang, Y.; Wang, C.; Yang, Y.; Liu, Y.; and Tao, D. 2022. Eatformer: Improving vision transformer inspired by evolutionary algorithm. *arXiv preprint arXiv:2206.09325*.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543.

Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; and Dabeer, O. 2022. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *ECCV*, 392–408. Springer.

Appendices

Effect of DDIM sampler steps

In order to accelerate the sampling speed in the denoising process, UiAD adopts the DDIM sampling strategy. We investigated the impact of different DDIM sampler steps on the results, as shown in Table 10. The results indicate that increasing the number of sampling steps does not significantly affect the results. Therefore, using a 10-step sampling process can achieve the best performance while greatly accelerating the sampling speed.

Steps	1	5	10	20	50	100	200
seg	72.5	96.5	**96.8**	**96.8**	96.7	96.7	**96.8**
cls	66.1	96.4	**97.2**	97.1	97.0	96.8	96.9

Table 10: Ablation studies on DDIM sampler steps.

Effect of Global average pooling

Global average pooling is used to reduce the potential occurrence of false positives. For *m-n* in the table below, *m* represents the iterations and *n* represents the kernel size. Through quantitative analysis, the most effective approach is employing an 8×8 size global average pooling with *8* iterations. Also, the best-performing combinations exhibit the same feature map size.

Global Average Pooling	1-16	4-16	5-12	6-10	8-8	10-7	15-5	20-4
AUROC-cls	96.0	96.7	96.9	97.1	**97.2**	**97.2**	97.0	96.8

Limitations of the datasets

We found that there are several categories of image-level anomaly detection results that are significantly lower than others, such as capsules and screws. As shown in Fig 7, we discovered some false positives in input good images during the test. Our method performs well in reconstructing the objects in the objects' main bodies, but the background region of the original image contains impurities, causing the pre-trained feature extraction network to extract features that perceive the background impurities as anomalies. As anomaly detection is expected to identify anomalies within the object rather than the background region, there are certain deficiencies in the Mvtec-AD as well as the VisA datasets that lead to false positives. In response to this issue, we increase the number of global average pooling operations to alleviate the problem of high anomaly scores caused by impurities in the background.

Hyperparameters of DiAD

We provided a comprehensive set of hyperparameters for the three models in DiAD as shown in Table 11.

Wyatt, J.; Leach, A.; Schmon, S. M.; 与 Willcocks, C. G. 2022. AnoDDPM：使用单纯形噪声的去噪扩散概率模型进行异常检测。于 *CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*，649–655。IEEE。 Xie, G.; Wang, J.; Liu, J.; Jin, Y.; 与 Zheng, F. 2023. 挑战工业视觉中少样本异常检测的极限：Graphcore。于 *ICLR*。 Yamada, S.; 与 Hotta, K. 2021. 基于注意力重构学生网络用于师生金字塔匹配。*arXiv preprint arXiv:2111.15376*。 Yan, X.; Zhang, H.; Xu, X.; Hu, X.; 与 Heng, P. 2021. 从正常样本中学习语义上下文以进行无监督异常检测。于 *AAAI*，3110–3118。 Yi, J.; 与 Yoon, S. 2020. Patch SVDD：用于异常检测与分割的块级支持向量数据描述。于 *ACCV*。 Yoon, J.; Sohn, K.; Li, C.-L.; Arik, S. O.; Lee, C.-Y.; 与 Pfister, T. 2022. 自监督、精炼、重复：改进无监督异常检测。*Transactions on Ma- chine Learning Research*。

游, Z.; 崔, L.; 沈, Y.; 杨, K.; 陆, X.; 郑, Y.; 和乐, X. 2022. 多类别异常检测的统一模型. 于 *NeurIPS*, 第35卷, 4571–4584.

Yu, J.; Zheng, Y.; Wang, X.; Li, W.; Wu, Y.; Zhao, R.; and Wu, L. 2021. FastFlow：基于二维归一化流的无监督异常检测与定位。arXiv:2111.07677。Zagoruyko, S.; and Komodakis, N. 2016. 宽残差网络。于 *BMVC*。BMVA Press。Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2021a. Draem——一种用于表面异常检测的判别性训练重建嵌入方法。于 *ICCV*，8330–8339。Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2021b. 通过修复重建进行视觉异常检测。*Pattern Recognition*，112：107706。Zhang, H.; Wang, Z.; Wu, Z.; and Jiang, Y.-G. 2023a. DiffusionAD：用于异常检测的去噪扩散模型。arXiv:2303.08730。Zhang, J.; Chen, X.; Xue, Z.; Wang, Y.; Wang, C.; and Liu, Y. 2023b. 探索面向VQA的GPT-4V在零样本异常检测中的基础潜力。*arXiv preprint arXiv:2311.02612*。Zhang, J.; Li, X.; Li, J.; Liu, L.; Xue, Z.; Zhang, B.; Jiang, Z.; Huang, T.; Wang, Y.; and Wang, C. 2023c. 重新思考高效基于注意力模型中的移动模块。于 *ICCV*，1389–1400。Zhang, J.; Li, X.; Wang, Y.; Wang, C.; Yang, Y.; Liu, Y.; and Tao, D. 2022. Eatformer：受进化算法启发改进视觉Transformer。*arXiv preprint arXiv:2206.09325*。Zhang, L.; and Agrawala, M. 2023. 为文本到图像扩散模型添加条件控制。arXiv:2302.05543。Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; and Dabeer, O. 2022. 用于……的找差异自监督预训练异常检测与分割。于*ECCV*，第392–408页。施普林格出版社。

附录

DDIM采样器步数的影响

为了在去噪过程中加速采样速度，UiAD采用了DDIM采样策略。我们研究了不同DDIM采样步数对结果的影响，如表10所示。结果表明，增加采样步数不会显著影响结果。因此，采用10步采样过程可以在大幅提升采样速度的同时实现最佳性能。

Steps	1	5	10	20	50	100	200
seg	72.5	96.5	**96.8**	**96.8**	96.7	96.7	**96.8**
cls	66.1	96.4	**97.2**	97.1	97.0	96.8	96.9

表10：DDIM采样器步数的消融研究。

全局平均池化的效果

全局平均池化用于减少潜在误报的发生。对于下表中的*m-n*，*m*代表迭代次数，*n*代表卷积核大小。通过定量分析，最有效的方法是采用8×8大小的全局平均池化，并进行*8*次迭代。此外，表现最佳的组合展现出相同的特征图尺寸。

Global Average Pooling	1-16	4-16	5-12	6-10	8-8	10-7	15-5	20-4
AUROC-cls	96.0	96.7	96.9	97.1	**97.2**	**97.2**	97.0	96.8

数据集的局限性

我们发现，有几类图像级异常检测结果明显低于其他类别，例如胶囊和螺丝。如图7所示，我们在测试过程中发现输入的正常图像中存在一些误报。我们的方法在重建物体主体部分表现良好，但原始图像的背景区域含有杂质，导致预训练的特征提取网络提取的特征将背景杂质感知为异常。由于异常检测旨在识别物体内部的异常而非背景区域，Mvtec-AD及VisA数据集本身存在一定缺陷，导致了误报的产生。针对这一问题，我们增加了全局平均池化操作的数量，以缓解因背景杂质导致异常分数偏高的问题。

DiAD的超参数

我们在DiAD中为三个模型提供了一套全面的超参数，如表11所示。

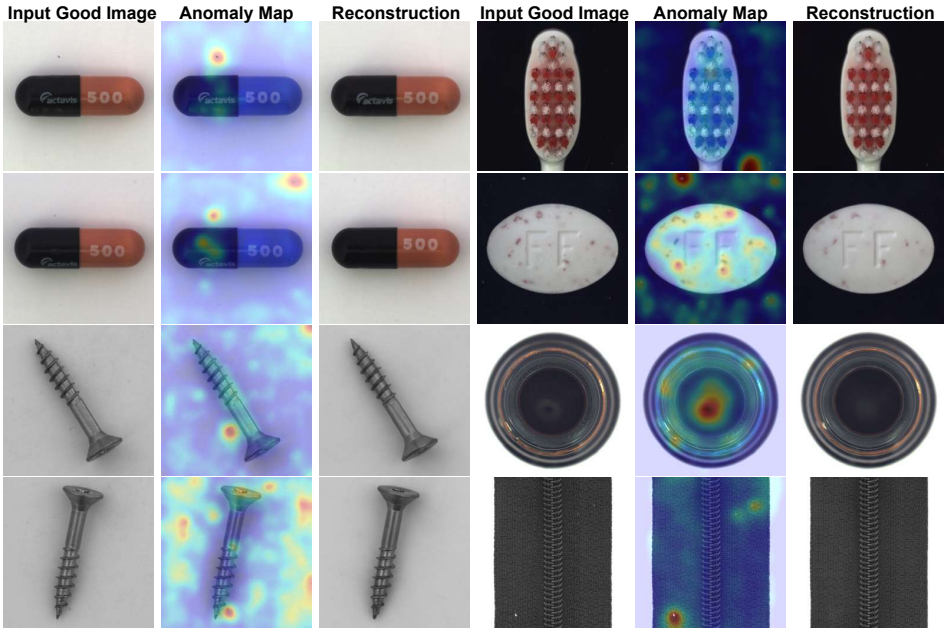

Figure 7: Visualization of false positive classifications and localizations.

Parameters Name	Model Name		
	SD Denoising Network	SG Network	Autoencoder
z shape	$32 \times 32 \times 4$		
$\|z\|$	4096		
Diffusion steps T	1000		
DDIM sampling steps T	10		
Noise Schedule	linear		
Model input shape	$32 \times 32 \times 4$	$256 \times 256 \times 3$	$256 \times 256 \times 3$
N params	859M	471M	83.7M
Embed dim	-	-	4
Channels	320	320	128
Num res blocks	2	2	2
Channel Multiplier	1,2,4,4	1,2,4,4	1,2,4,4
Attention resolutions	4,2,1	4,2,1	-
Num Heads	8	8	-
Batch Size	12		
Accumulate_grad_batches	4		
Epochs	1000		
Learning Rate	1.0e-5		

Table 11: Hyperparameters for the DiAD. All models trained on a single NVIDIA Tesla V100 32GB.

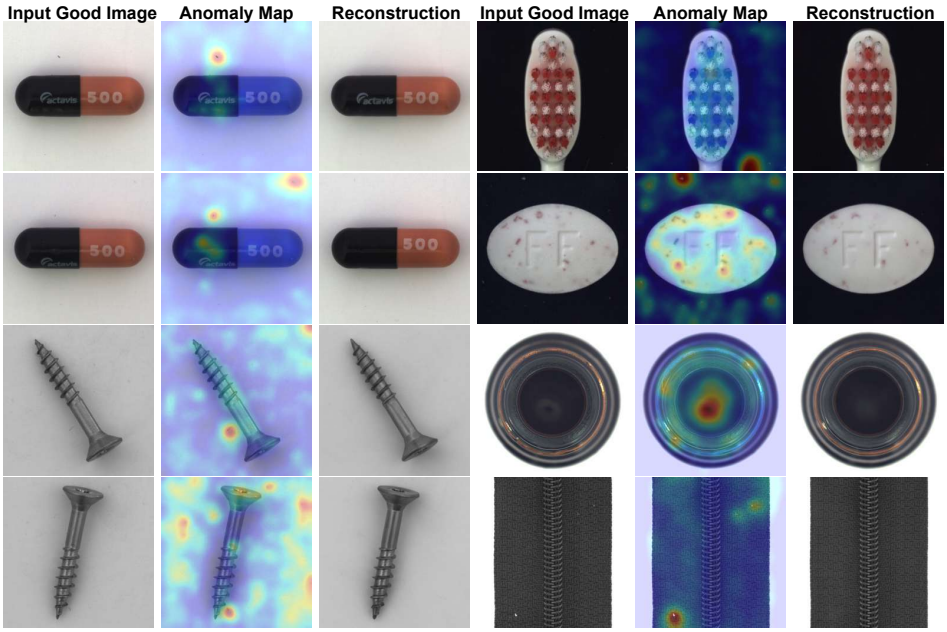

图7：误报分类与定位的可视化。

Parameters Name	Model Name		
	SD Denoising Network	SG Network	Autoencoder
z shape	$32 \times 32 \times 4$		
$\|z\|$	4096		
Diffusion steps T	1000		
DDIM sampling steps T	10		
Noise Schedule	linear		
Model input shape	$32 \times 32 \times 4$	$256 \times 256 \times 3$	$256 \times 256 \times 3$
N params	859M	471M	83.7M
Embed dim	-	-	4
Channels	320	320	128
Num res blocks	2	2	2
Channel Multiplier	1,2,4,4	1,2,4,4	1,2,4,4
Attention resolutions	4,2,1	4,2,1	-
Num Heads	8	8	-
Batch Size	12		
Accumulate_grad_batches	4		
Epochs	1000		
Learning Rate	1.0e-5		

表11：DiAD的超参数。所有模型均在单张NVIDIA Tesla V100 32GB显卡上训练。

Category	Non-Diffusion Method		Diffusion-based Method		
	DRAEM	UniAD	DDPM	LDM	Ours
pcb1	71.9/72.2/70.0	92.8/92.7/87.8	54.1/47.7/67.1	51.2/46.9/66.8	88.1/88.7/80.7
pcb2	78.4/78.2/76.2	87.8/87.7/83.1	50.8/48.5/66.6	57.0/63.4/67.5	91.4/91.4/84.7
pcb3	76.6/77.4/74.7	78.6/78.6/76.1	53.4/51.2/66.8	62.7/69.6/72.0	86.2/87.6/77.6
pcb4	97.3/97.5/93.5	98.8/98.8/94.3	56.0/48.4/66.4	54.4/47.1/66.8	99.6/99.5/97.0
macaroni1	69.8/68.5/70.9	79.9/79.8/72.7	50.9/55.1/68.0	56.2/49.6/68.4	85.7/85.2/78.8
macaroni2	59.4/60.7/68.0	71.6/71.6/69.9	54.4/51.8/67.1	56.8/52.7/66.6	62.5/57.4/69.6
capsules	83.4/91.1/82.1	55.6/55.6/76.9	58.9/62.7/78.2	57.7/71.4/77.3	58.2/69.0/78.5
candle	69.3/73.9/68.0	94.1/94.0/86.1	52.7/48.3/66.6	50.4/52.2/68.2	92.8/92.0/87.6
cashew	81.7/89.7/87.3	92.8/92.8/91.4	63.5/78.9/80.6	61.1/71.0/80.0	91.5/95.7/89.7
chewinggum	93.7/97.1/91.0	96.3/96.2/95.2	50.9/65.6/80.0	53.9/65.8/81.3	99.1/99.5/95.9
fryum	89.1/95.0/86.6	83.0/83.0/85.0	51.0/62.4/80.0	63.5/71.6/81.6	89.8/95.0/87.2
pipe_fryum	82.8/91.2/83.9	94.7/94.7/93.9	56.9/74.9/80.0	56.1/75.5/80.3	96.2/98.1/93.7
Mean	79.1/81.9/78.9	85.5/85.5/84.4	54.5/57.9/72.3	56.7/61.4/73.1	86.8/88.3/85.1

Table 12: Comparison with SOTA methods on VisA dataset for multi-class anomaly detection with $AUROC_{cls}$/AP_{cls}/$F1max_{cls}$ metrics.

Category	Non-Diffusion Method		Diffusion-based Method		
	DRAEM	UniAD	DDPM	LDM	Ours
pcb1	94.6/31.8/37.2/52.8	93.3/ 3.9/ 8.3/64.1	75.7/ 1.1/ 2.8/36.1	84.5/ 2.1/ 4.9/54.3	98.7/49.6/52.8/80.2
pcb2	92.3/10.0/18.6/66.2	93.9/ 4.2/ 9.2/66.9	76.2/ 0.7/ 1.6/30.8	89.5/ 2.5/ 6.7/52.7	95.2/ 7.5/16.7/67.0
pcb3	90.8/14.1/24.4/42.9	97.3/13.8/21.9/70.6	83.3/ 1.0/ 2.5/56.1	94.4/ 9.2/17.4/67.8	96.7/ 8.0/18.8/68.9
pcb4	94.4/31.0/37.6/75.7	94.9/14.7/22.9/72.3	73.0/ 1.4/ 3.5/29.9	80.4/ 2.1/ 4.2/40.3	97.0/17.6/27.2/85.0
macaroni1	95.0/19.1/24.1/67.0	97.4/ 3.7/ 9.7/84.0	87.4/ 0.4/ 1.0/61.2	81.6/ 0.3/ 1.3/47.3	94.1/10.2/16.7/68.5
macaroni2	94.6/ 3.9/12.4/65.2	95.2/ 0.9/ 4.3/76.6	84.8/ 0.2/ 0.6/54.1	87.2/ 0.3/ 0.6/57.2	93.6/ 0.9/ 2.8/73.1
capsules	97.1/27.8/33.7/62.8	88.7/ 3.0/ 7.4/43.7	77.1/ 1.1/ 2.8/34.6	75.5/ 1.1/ 2.7/34.8	97.3/10.0/21.0/77.9
candle	82.2/10.1/19.0/65.6	98.5/17.6/27.9/91.6	76.4/ 0.4/ 1.4/34.1	85.3/ 0.9/ 1.9/46.8	97.3/12.8/22.8/89.4
cashew	80.7/ 9.9/15.7/38.5	98.6/51.7/58.3/87.9	74.5/ 2.7/ 5.2/58.7	90.5/ 5.1/10.1/68.3	90.9/53.1/60.9/61.8
chewinggum	91.0/62.3/63.3/40.9	98.8/54.9/56.1/81.3	74.7/ 1.4/ 2.8/37.9	84.1/ 3.1/ 6.9/52.9	94.7/11.9/25.8/59.5
fryum	92.4/38.8/38.5/69.5	95.9/34.0/40.6/76.2	85.7/ 9.4/17.2/58.4	89.9/14.8/24.8/60.1	97.6/58.6/60.1/81.3
pipe_fryum	91.1/38.1/39.6/61.8	98.9/50.2/57.7/91.5	87.0/ 6.9/12.9/69.6	96.4/31.0/37.2/77.6	99.4/72.7/69.9/89.9
Mean	91.3/23.5/29.5/58.8	95.9/21.0/27.0/75.6	79.7/ 2.2/ 4.5/46.8	86.6/ 6.0/ 9.9/55.0	96.0/26.1/33.0/75.2

Table 13: Comparison with SOTA methods on VisA dataset for multi-class anomaly localization with $AUROC_{seg}$/AP_{seg}/$F1max_{seg}$/PRO metrics.

Category	Non-Diffusion Method		Diffusion-based Method		
	DRAEM	UniAD	DDPM	LDM	Ours
pcb1	71.9/72.2/70.0	92.8/92.7/87.8	54.1/47.7/67.1	51.2/46.9/66.8	88.1/88.7/80.7
pcb2	78.4/78.2/76.2	87.8/87.7/83.1	50.8/48.5/66.6	57.0/63.4/67.5	91.4/91.4/84.7
pcb3	76.6/77.4/74.7	78.6/78.6/76.1	53.4/51.2/66.8	62.7/69.6/72.0	86.2/87.6/77.6
pcb4	97.3/97.5/93.5	98.8/98.8/94.3	56.0/48.4/66.4	54.4/47.1/66.8	99.6/99.5/97.0
macaroni1	69.8/68.5/70.9	79.9/79.8/72.7	50.9/55.1/68.0	56.2/49.6/68.4	85.7/85.2/78.8
macaroni2	59.4/60.7/68.0	71.6/71.6/69.9	54.4/51.8/67.1	56.8/52.7/66.6	62.5/57.4/69.6
capsules	83.4/91.1/82.1	55.6/55.6/76.9	58.9/62.7/78.2	57.7/71.4/77.3	58.2/69.0/78.5
candle	69.3/73.9/68.0	94.1/94.0/86.1	52.7/48.3/66.6	50.4/52.2/68.2	92.8/92.0/87.6
cashew	81.7/89.7/87.3	92.8/92.8/91.4	63.5/78.9/80.6	61.1/71.0/80.0	91.5/95.7/89.7
chewinggum	93.7/97.1/91.0	96.3/96.2/95.2	50.9/65.6/80.0	53.9/65.8/81.3	99.1/99.5/95.9
fryum	89.1/95.0/86.6	83.0/83.0/85.0	51.0/62.4/80.0	63.5/71.6/81.6	89.8/95.0/87.2
pipe_fryum	82.8/91.2/83.9	94.7/94.7/93.9	56.9/74.9/80.0	56.1/75.5/80.3	96.2/98.1/93.7
Mean	79.1/81.9/78.9	85.5/85.5/84.4	54.5/57.9/72.3	56.7/61.4/73.1	86.8/88.3/85.1

表12：在VisA数据集上使用$AUROC_{cls}/AP_{cls}/F1max_{cls}$指标进行多类别异常检测与SOTA方法的对比。

Category	Non-Diffusion Method		Diffusion-based Method		
	DRAEM	UniAD	DDPM	LDM	Ours
pcb1	94.6/31.8/37.2/52.8	93.3/ 3.9/ 8.3/64.1	75.7/ 1.1/ 2.8/36.1	84.5/ 2.1/ 4.9/54.3	98.7/49.6/52.8/80.2
pcb2	92.3/10.0/18.6/66.2	93.9/ 4.2/ 9.2/66.9	76.2/ 0.7/ 1.6/30.8	89.5/ 2.5/ 6.7/52.7	95.2/ 7.5/16.7/67.0
pcb3	90.8/14.1/24.4/42.9	97.3/13.8/21.9/70.6	83.3/ 1.0/ 2.5/56.1	94.4/ 9.2/17.4/67.8	96.7/ 8.0/18.8/68.9
pcb4	94.4/31.0/37.6/75.7	94.9/14.7/22.9/72.3	73.0/ 1.4/ 3.5/29.9	80.4/ 2.1/ 4.2/40.3	97.0/17.6/27.2/85.0
macaroni1	95.0/19.1/24.1/67.0	97.4/ 3.7/ 9.7/84.0	87.4/ 0.4/ 1.0/61.2	81.6/ 0.3/ 1.3/47.3	94.1/10.2/16.7/68.5
macaroni2	94.6/ 3.9/12.4/65.2	95.2/ 0.9/ 4.3/76.6	84.8/ 0.2/ 0.6/54.1	87.2/ 0.3/ 0.6/57.2	93.6/ 0.9/ 2.8/73.1
capsules	97.1/27.8/33.7/62.8	88.7/ 3.0/ 7.4/43.7	77.1/ 1.1/ 2.8/34.6	75.5/ 1.1/ 2.7/34.8	97.3/10.0/21.0/77.9
candle	82.2/10.1/19.0/65.6	98.5/17.6/27.9/91.6	76.4/ 0.4/ 1.4/34.1	85.3/ 0.9/ 1.9/46.8	97.3/12.8/22.8/89.4
cashew	80.7/ 9.9/15.7/38.5	98.6/51.7/58.3/87.9	74.5/ 2.7/ 5.2/58.7	90.5/ 5.1/10.1/68.3	90.9/53.1/60.9/61.8
chewinggum	91.0/62.3/63.3/40.9	98.8/54.9/56.1/81.3	74.7/ 1.4/ 2.8/37.9	84.1/ 3.1/ 6.9/52.9	94.7/11.9/25.8/59.5
fryum	92.4/38.8/38.5/69.5	95.9/34.0/40.6/76.2	85.7/ 9.4/17.2/58.4	89.9/14.8/24.8/60.1	97.6/58.6/60.1/81.3
pipe_fryum	91.1/38.1/39.6/61.8	98.9/50.2/57.7/91.5	87.0/ 6.9/12.9/69.6	96.4/31.0/37.2/77.6	99.4/72.7/69.9/89.9
Mean	91.3/23.5/29.5/58.8	95.9/21.0/27.0/75.6	79.7/ 2.2/ 4.5/46.8	86.6/ 6.0/ 9.9/55.0	96.0/26.1/33.0/75.2

表13：在VisA数据集上使用$AUROC_{seg}/AP_{seg}/F1max_{seg}/PRO$指标进行多类别异常定位与SOTA方法的比较。

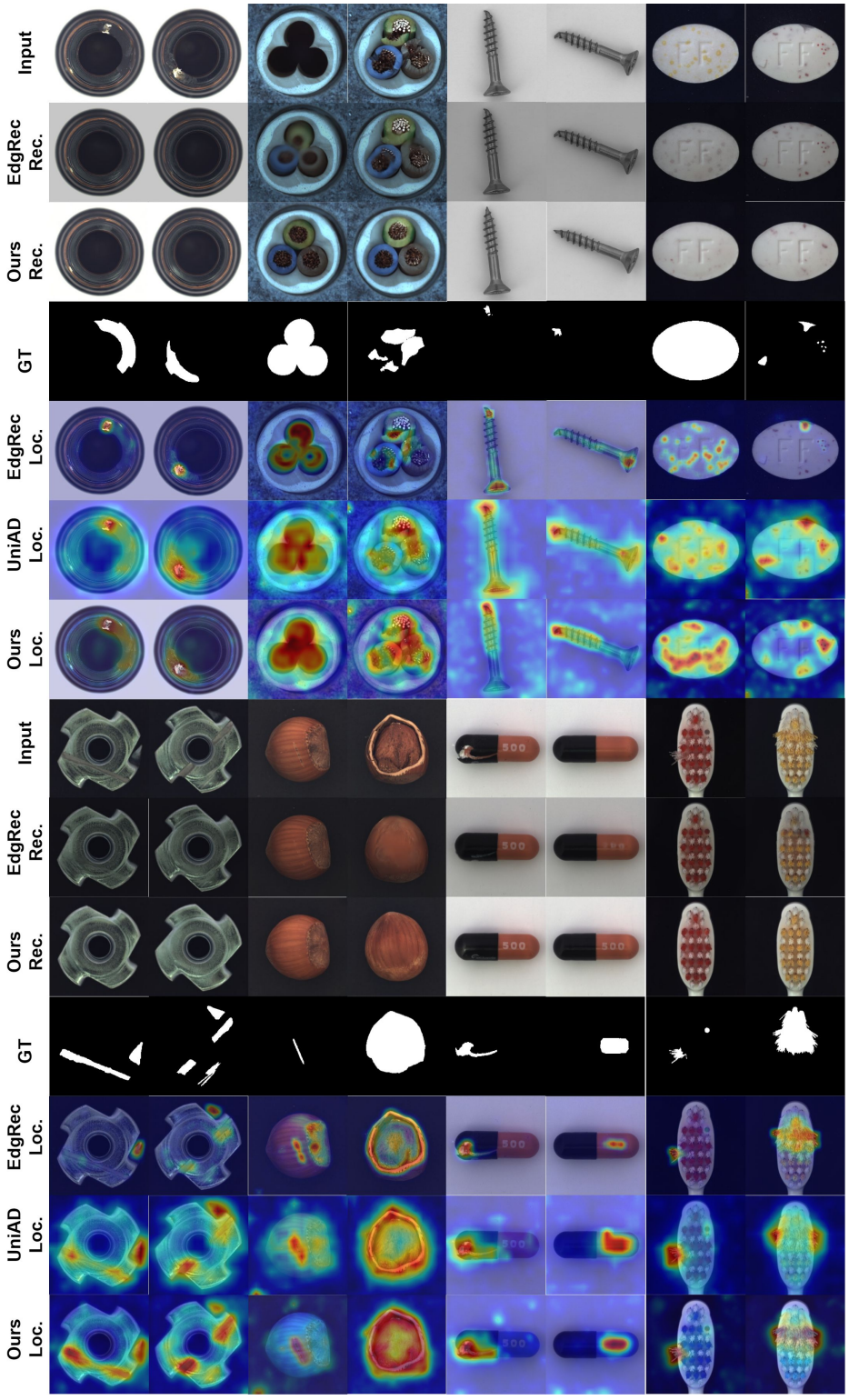

Figure 8: Qualitative comparison results for anomaly localization on MVTec-AD dataset.

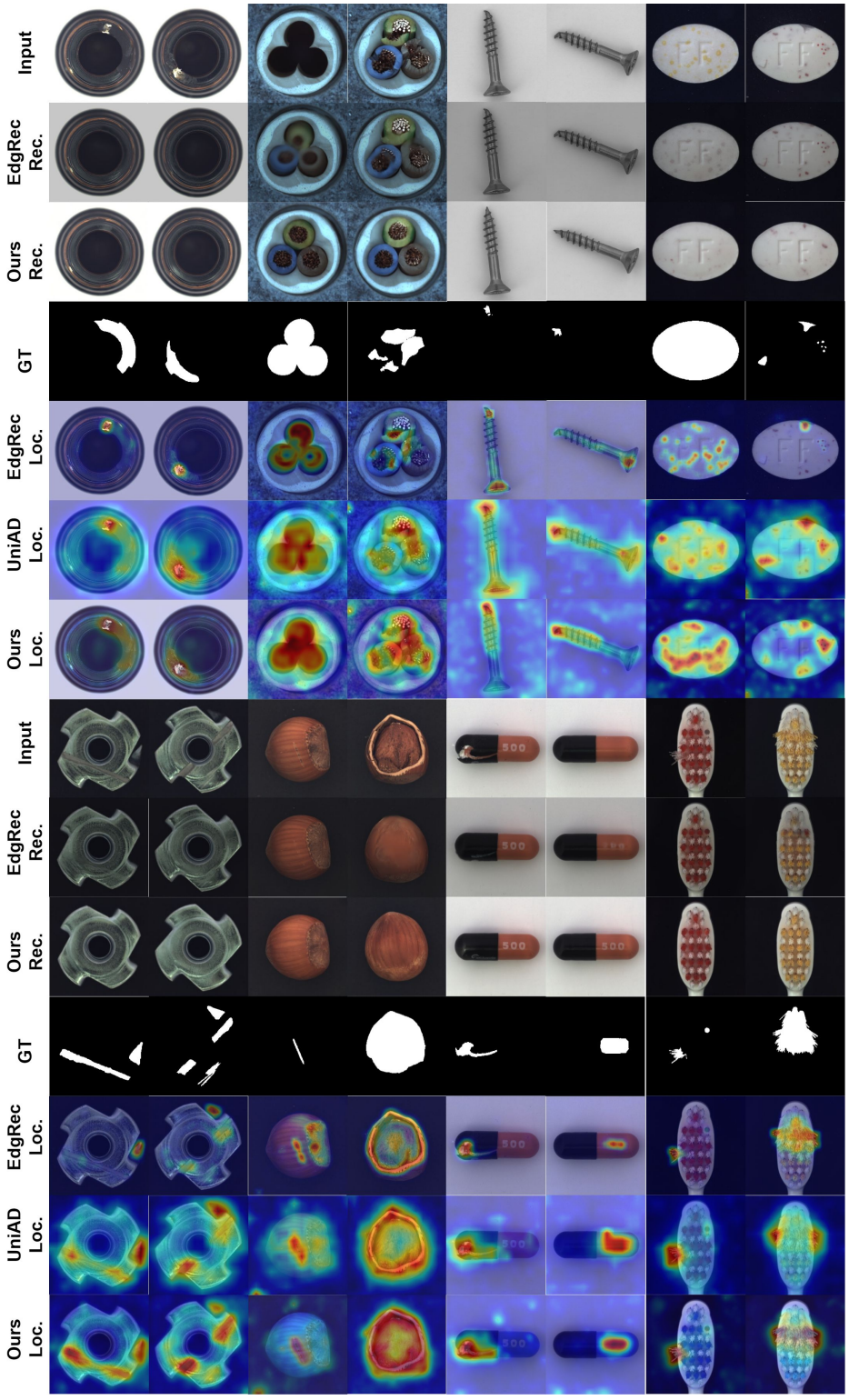

图8：MVTec-AD数据集上异常定位的定性比较结果。

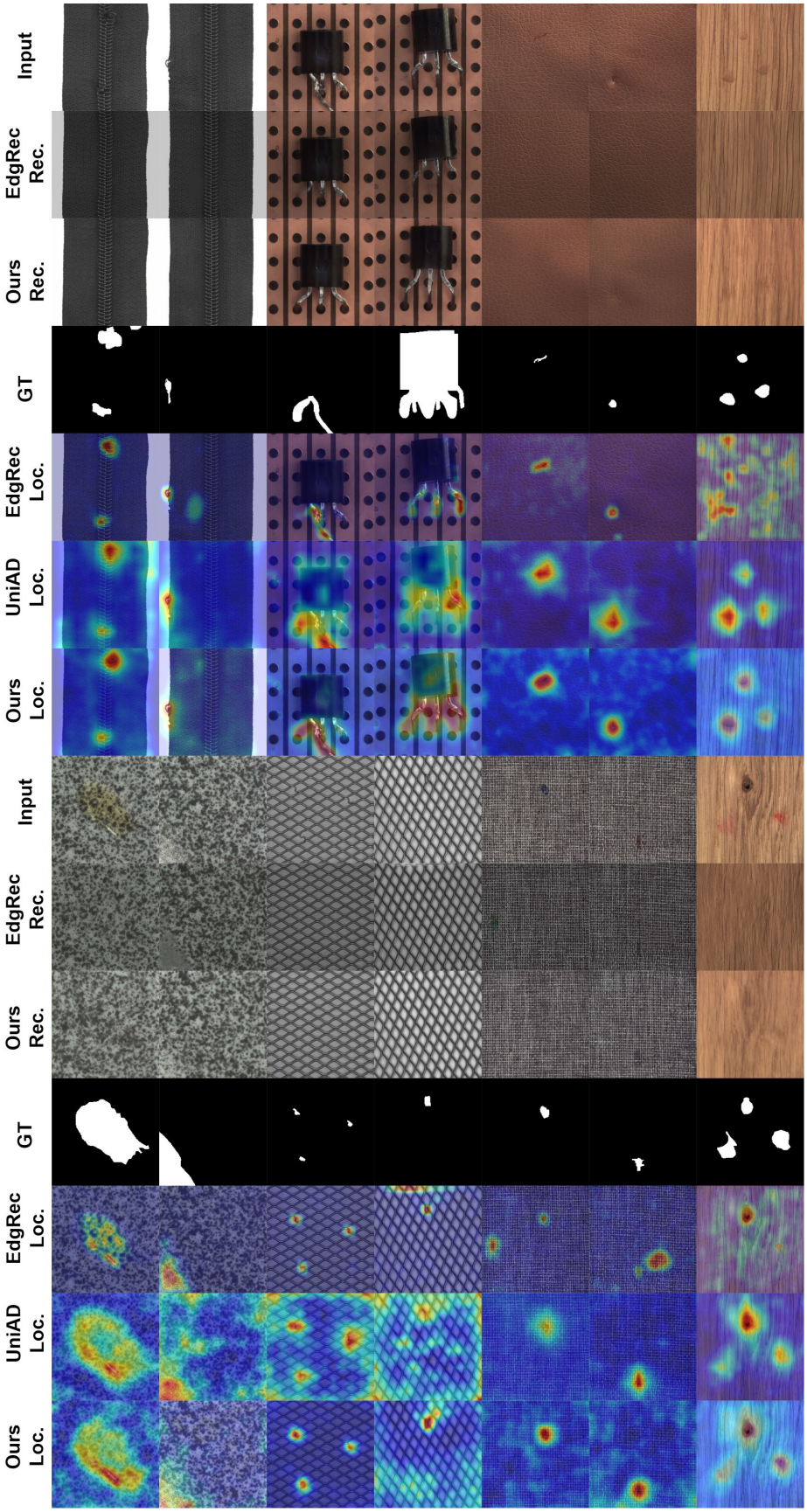

Figure 9: Qualitative comparison results for anomaly localization on MVTec-AD dataset.

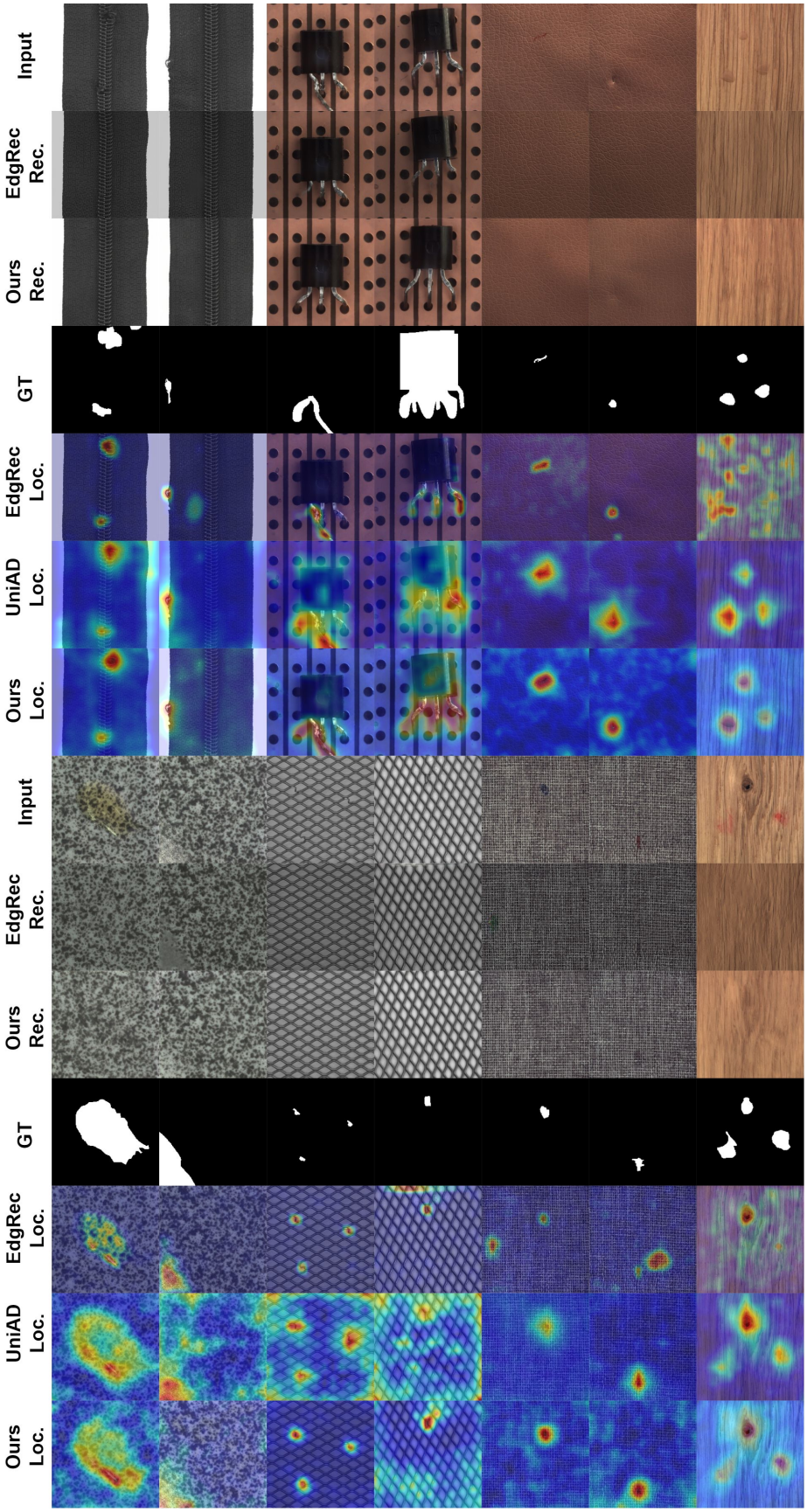

图9：MVTec-AD数据集上异常定位的定性比较结果。

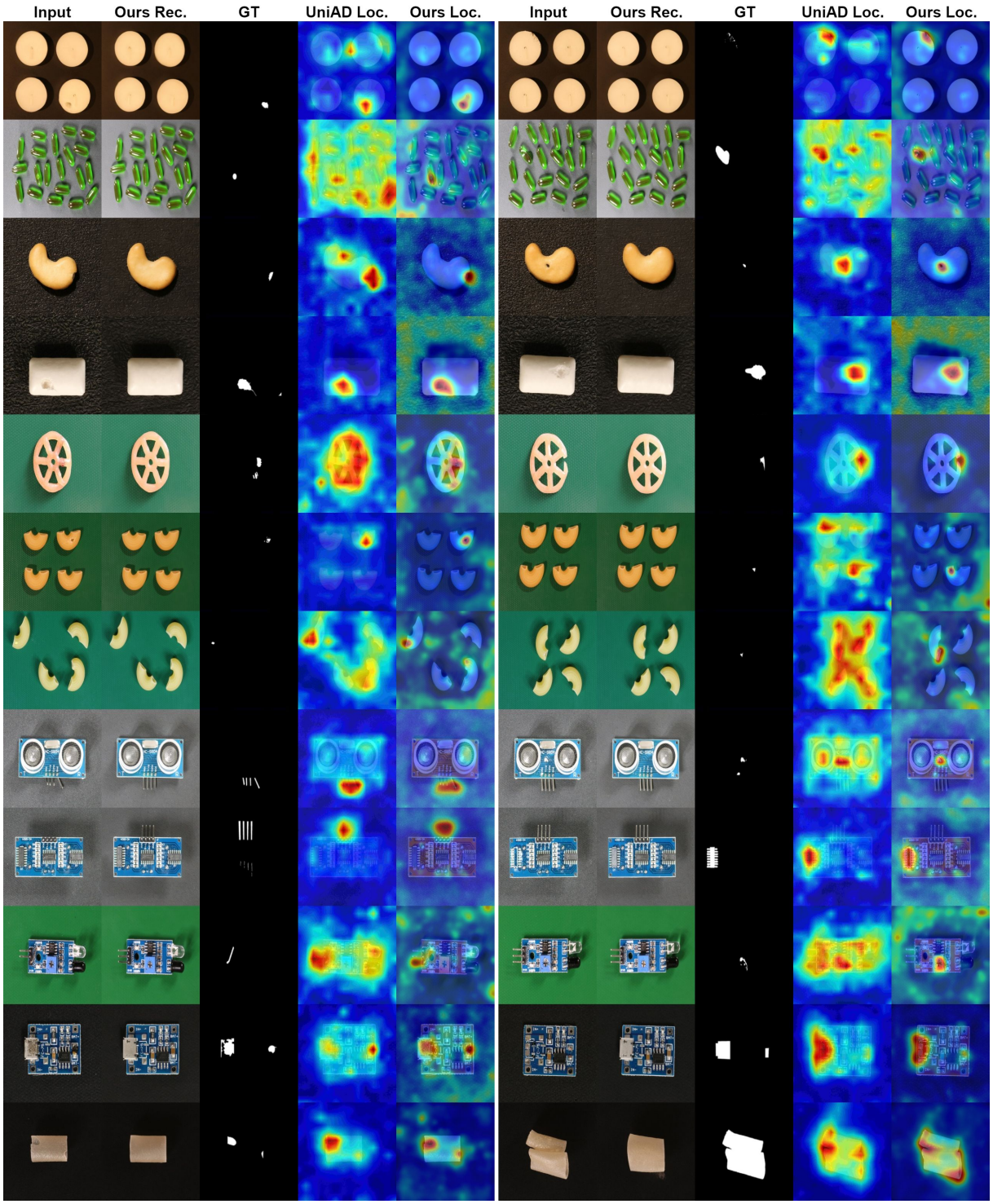

Figure 10: Qualitative comparison results for anomaly localization on VisA dataset.

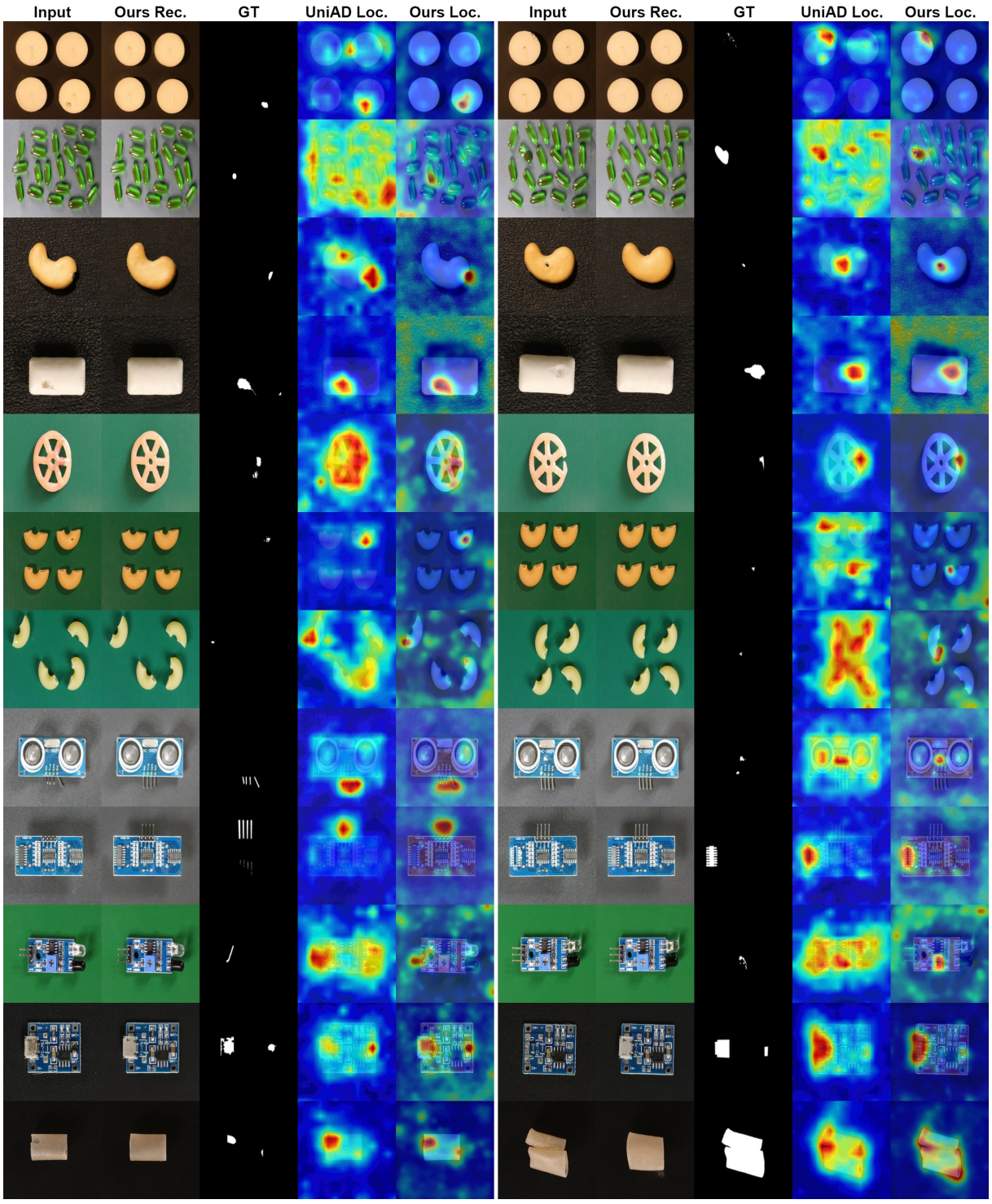

图 10: Q　　VisA异常定位的定性比较结果　　数据集。