arXiv:2403.05897v1 [cs.CV] 9 Mar 2024

RealNet: A Feature Selection Network with Realistic Synthetic Anomaly for Anomaly Detection

Ximiao Zhang[1] Min Xu[1]* Xiuzhuang Zhou[2]
[1]College of Information and Engineering, Capital Normal University
[2]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
{2211002048, xumin}@cnu.edu.cn[1], xiuzhuang.zhou@bupt.edu.cn[2]

Abstract

Self-supervised feature reconstruction methods have shown promising advances in industrial image anomaly detection and localization. Despite this progress, these methods still face challenges in synthesizing realistic and diverse anomaly samples, as well as addressing the feature redundancy and pre-training bias of pre-trained feature. In this work, we introduce RealNet, a feature reconstruction network with realistic synthetic anomaly and adaptive feature selection. It is incorporated with three key innovations: First, we propose Strength-controllable Diffusion Anomaly Synthesis (SDAS), a diffusion process-based synthesis strategy capable of generating samples with varying anomaly strengths that mimic the distribution of real anomalous samples. Second, we develop Anomaly-aware Features Selection (AFS), a method for selecting representative and discriminative pre-trained feature subsets to improve anomaly detection performance while controlling computational costs. Third, we introduce Reconstruction Residuals Selection (RRS), a strategy that adaptively selects discriminative residuals for comprehensive identification of anomalous regions across multiple levels of granularity. We assess RealNet on four benchmark datasets, and our results demonstrate significant improvements in both Image AUROC and Pixel AUROC compared to the current state-of-the-art methods. The code, data, and models are available at https://github.com/cnulab/RealNet.

1. Introduction

Image anomaly detection is a critical task in industrial production, with wide-ranging applications in quality control and safety monitoring. While self-supervised methods [20, 32, 48, 50, 53] have gained attention for training models using synthetic anomalies, they still face challenges in synthesizing realistic and diverse anomaly images, especially generating complex structural anomalies and unseen anomaly categories. Due to the lack of available anomaly images and prior knowledge about anomaly categories, existing methods rely on carefully crafted data augmentation strategies [20, 32] or external data [48] for anomaly synthesis, leading to significant distribution discrepancy between synthetic anomalies and real anomalies, thereby limiting the generalization ability of anomaly detection models to real-world applications. To address these issues, we introduce Strength-controllable Diffusion Anomaly Synthesis (SDAS), a novel synthesis strategy that generates diverse samples more closely aligned with natural distributions, and offers flexibility in controlling anomaly strength. SDAS employs DDPM [16] to model the distribution of normal samples and introduces perturbation terms during the sampling process to generate samples in low probability density regions. These samples simulate various natural anomaly patterns, such as aging, structural changes, abnormal textures, and color changes, as shown in Fig. 1.

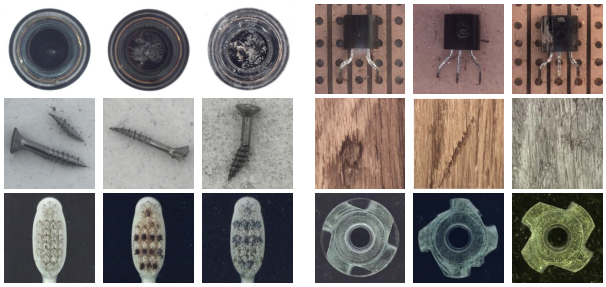


Figure 1. SDAS generates anomaly images using only normal images. The example images are sourced from the MVTec-AD dataset [3].

Parallel to this, feature reconstruction-based anomaly detection [8, 33, 44, 49, 53] is another promising research direction, which reconstructs the features of anomalous images as those of normal images and conducts anomaly detection and localization by reconstruction residuals. They have attracted considerable attention due to the simple

*Corresponding author.

RealNet：一种用于异常检测的具有真实合成异常的特征选择网络

张西淼[1] 徐敏[1]* 周秀壮[2] [1]首都师范大学信息工程学院 [2]北京邮电大学人工智能学院 {2211002048, xumin}@cnu.edu.cn[1], xiuzhuang.zhou@bupt.edu.cn[2]

4202raM9]VC.sc[1v79850.3042:viXra

摘要

Self-supervised feature reconstruction methods have shown promising advances in industrial image anomaly detection and localization. Despite this progress, these methods still face challenges in synthesizing realistic and diverse anomaly samples, as well as addressing the feature redundancy and pre-training bias of pre-trained feature. In this work, we introduce RealNet, a feature reconstruction network with realistic synthetic anomaly and adaptive feature selection. It is incorporated with three key innovations: First, we propose Strength-controllable Diffusion Anomaly Synthesis (SDAS), a diffusion process-based synthesis strategy capable of generating samples with varying anomaly strengths that mimic the distribution of real anomalous samples. Second, we develop Anomaly-aware Features Selection (AFS), a method for selecting representative and discriminative pre-trained feature subsets to improve anomaly detection performance while controlling computational costs. Third, we introduce Reconstruction Residuals Selection (RRS), a strategy that adaptively selects discriminative residuals for comprehensive identification of anomalous regions across multiple levels of granularity. We assess RealNet on four benchmark datasets, and our results demonstrate significant improvements in both Image AUROC and Pixel AUROC compared to the current state-of-the-art methods. The code, data, and models are available at https://github.com/cnulab/RealNet.

1. 引言

图像异常检测是工业生产中的关键任务，在质量控制与安全监测领域具有广泛应用。尽管自监督方法[20, 32, 48, 50, 53]通过合成异常训练模型受到关注，但在生成逼真且多样化的异常图像方面仍面临挑战，尤其——

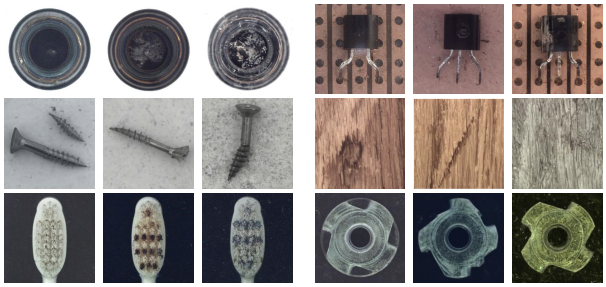


图1. SDAS仅使用正常图像生成异常图像。示例图像来源于MVTec-AD数据集[3]。

特别是在生成复杂的结构异常和未见过的异常类别方面。由于缺乏可用的异常图像以及关于异常类别的先验知识，现有方法依赖于精心设计的数据增强策略[20, 32]或外部数据[48]进行异常合成，这导致合成异常与真实异常之间存在显著的分布差异，从而限制了异常检测模型在实际应用中的泛化能力。为了解决这些问题，我们引入了强度可控扩散异常合成（SDAS），这是一种新颖的合成策略，能够生成更贴近自然分布的多样化样本，并提供了控制异常强度的灵活性。SDAS采用DDPM [16]对正常样本的分布进行建模，并在采样过程中引入扰动项，以在低概率密度区域生成样本。这些样本模拟了各种自然异常模式，如老化、结构变化、异常纹理和颜色变化，如图1所示。

与此同时，基于特征重构的异常检测[8, 33, 44, 49, 53]是另一个前景广阔的研究方向，该方法将异常图像的特征重构为正常图像的特征，并通过重构残差进行异常检测与定位。由于其方法简洁，这类研究已受到广泛关注。

*Corresponding author.

paradigm. However, due to the high computational demands of feature reconstruction and the lack of effective feature selection strategies, existing methods either employ small-scale pre-trained CNNs [33, 44, 49] for anomaly detection or handpick layer-specific features from pre-trained network [8, 53] for reconstruction. The latest work [14] highlights the importance of feature selection, indicating that existing anomaly detection methods [30, 46] are sensitive to feature selection. The optimal pre-trained feature subset for anomaly detection varies across different categories. Therefore, devising a unified feature selection approach has become a pressing need for advancing anomaly detection. In this paper, we propose RealNet, a feature reconstruction framework that incorporates Anomaly-aware Features Selection (AFS) and Reconstruction Residuals Selection (RRS). RealNet fully exploits the discriminative capabilities of large-scale pre-trained CNNs while reducing feature redundancy and pre-training bias, enhancing anomaly detection performance while effectively controlling computational demands. For different categories, RealNet selects different pre-trained feature subsets for anomaly detection, ensuring optimal anomaly detection performance while flexibly controlling the model size. Furthermore, RealNet effectively reduces missed detections by adaptively discarding reconstruction residuals lacking anomalous information, and significantly improves the recall of anomalous regions. In summary, our contributions are fourfold:

- We propose RealNet, a feature reconstruction network that effectively leverages multi-scale pre-trained features for anomaly detection by adaptively selecting pre-trained features and reconstruction residuals. RealNet achieves state-of-the-art performance while addressing the computational cost limitations suffered by previous methods.
- We introduce Strength-controllable Diffusion Anomaly Synthesis (SDAS), a novel anomaly synthesis strategy that generates realistic and diverse anomalous samples closely aligned with natural distributions.
- We evaluate RealNet on four datasets (MVTec-AD [3], MPDD [18], BTAD [24], and VisA [55]), surpassing existing state-of-the-art methods using the same set of network architectures and hyperparameters across datasets.
- We provide the Synthetic Industrial Anomaly Dataset (SIA). SIA is generated by SDAS and consists of a total of 360,000 anomalous images from 36 categories of industrial products. SIA can be conveniently utilized for anomaly synthesis to facilitate self-supervised anomaly detection methods.

2. Related work

Unsupervised anomaly detection and localization approaches use only normal images for model training, without any anomalous data. These methods can be roughly classified into four main categories: reconstruction-based methods [1, 2], self-supervised learning-based methods [20, 48], deep feature embedding-based methods [7, 30], and one-class classification-based methods [22, 43]. In this paper, we focus on the reconstruction-based and self-supervised learning-based methods, which are of particular relevance to our proposed RealNet framework.

Reconstruction-based methods follow a relatively consistent paradigm, which entails training a reconstruction model on normal images. The inability to effectively reconstruct anomalous regions in input images facilitates anomaly detection and localization through comparison of the original and reconstructed images. In this context, a variety of reconstruction techniques are explored, such as Autoencoder [2, 45], GAN [1, 31], Transformer [24, 28], and Diffusion model [23, 39, 52]. However, managing the reconstruction capability of the network remains challenging. In cases of complex image structures or textures, the network may produce a simplistic copy instead of selective reconstruction. Furthermore, inherent stylistic discrepancies between original and reconstructed images can lead to false positives or undetected anomalies.

Recent studies, as exemplified by [8, 33, 44, 49], have been primarily focused on anomaly detection through the reconstruction of pre-trained image features. In contrast to image-level reconstruction, multi-scale features pre-trained on ImageNet [9] demonstrate enhanced discriminative abilities to detect anomalies across a wide range of scales and diverse image patterns. However, due to the inherent feature redundancy in high-dimensional features and the pre-training bias introduced by classification tasks, the anomaly detection capability of large-scale pre-trained networks has not been fully utilized. Recent studies [33, 44, 49] use small-scale pre-trained networks to ensure controllable reconstruction costs, and other works [30, 38, 53] manually select partial layer features from pre-trained networks for anomaly detection. However, the optimal feature subset for anomaly detection varies across different categories [14], thus, these manually selecting methods often prove to be dataset-specific and suboptimal, resulting in a significant performance drop. Different from previous solutions, our RealNet presents a novel combination of efficient feature selection strategies and an optimized reconstruction process, effectively enhancing anomaly detection performance while maintaining computational efficiency.

Self-supervised learning-based methods aim to bypass the need for labels of anomalous images by setting a suitable proxy task. Notable works in this domain include CutPaste [20], which generates anomalies by transplanting image patches from one location to another, albeit with suboptimal continuity in the anomalous regions. NSA [32] uses Poisson image editing [26] for seamless image pasting to synthesize more natural anomaly regions. DRAEM [48] leverages the texture dataset DTD [5] to

范式。然而，由于特征重构的高计算需求以及缺乏有效的特征选择策略，现有方法要么采用小规模预训练CNN进行异常检测[33, 44, 49]，要么从预训练网络中手动挑选特定层特征进行重构[8, 53]。最新研究[14]强调了特征选择的重要性，指出现有异常检测方法[30, 46]对特征选择极为敏感。不同类别的最优预训练特征子集存在差异，因此设计统一的特征选择方法已成为推进异常检测发展的迫切需求。本文提出RealNet——一个融合异常感知特征选择与重构残差选择的特征重构框架。该框架充分挖掘大规模预训练CNN的判别能力，同时减少特征冗余与预训练偏差，在有效控制计算需求的同时提升异常检测性能。针对不同类别，RealNet选择不同的预训练特征子集进行异常检测，在灵活控制模型规模的同时确保最优检测性能。此外，通过自适应舍弃缺乏异常信息的重构残差，RealNet有效降低了漏检率，显著提升异常区域的召回率。综上所述，本文贡献包含四个方面：

- 我们提出了RealNet，一种特征重建网络，它通过自适应选择预训练特征和重建残差，有效利用多尺度预训练特征进行异常检测。RealNet在解决先前方法所面临的计算成本限制的同时，实现了最先进的性能。
- 我们提出了强度可控扩散异常合成（SDAS），这是一种新颖的异常合成策略，能够生成与自然分布高度吻合、真实且多样化的异常样本。
- 我们在四个数据集（MVTec-AD [3]、MPDD [18]、BTAD [24] 和 VisA [55]）上评估了RealNet，通过在各数据集上使用相同的网络架构和超参数，超越了现有的最先进方法。
- 我们提供了合成工业异常数据集（SIA）。SIA由SDAS生成，包含来自36类工业产品的总计360,000张异常图像。SIA可便捷地用于异常合成，以促进自监督异常检测方法的发展。

2. 相关工作

无监督异常检测与定位方法仅使用正常图像进行模型训练，无需任何异常数据。这些方法大致可分为四大类：基于重建的方法[1, 2]、基于自监督学习的方法[20, 48]、基于深度特征嵌入的方法[7, 30]以及基于单类分类的方法[22, 43]。本文重点关注基于重建和基于自监督学习的方法，这两类方法与我们所提出的RealNet框架尤为相关。

基于重构的方法遵循相对一致的范式，即在正常图像上训练重构模型。由于模型难以有效重建输入图像中的异常区域，通过对比原始图像与重构图像即可实现异常检测与定位。在此框架下，研究者探索了多种重构技术，如自编码器[2, 45]、生成对抗网络[1, 31]、Transformer[24, 28]以及扩散模型[23, 39, 52]。然而，如何有效控制网络的重构能力仍具挑战：当面对复杂图像结构或纹理时，网络可能仅生成简单复制而非选择性重建；此外，原始图像与重构图像之间固有的风格差异可能导致误报或漏检异常。

近期研究，如[8, 33, 44, 49]所示，主要集中于通过重建预训练图像特征进行异常检测。与图像级重建相比，在ImageNet[9]上预训练的多尺度特征展现出更强的判别能力，能够检测跨多种尺度和多样化图像模式的异常。然而，由于高维特征固有的冗余性以及分类任务带来的预训练偏差，大规模预训练网络的异常检测能力尚未得到充分利用。近期研究[33, 44, 49]采用小规模预训练网络以控制重建成本，另一些工作[30, 38, 53]则手动选取预训练网络中的部分层特征进行异常检测。但异常检测的最优特征子集因不同类别而异[14]，这些人工选择方法往往仅适用于特定数据集且非最优，导致性能显著下降。与先前方案不同，我们的RealNet提出了一种高效特征选择策略与优化重建过程的新颖组合，在保持计算效率的同时有效提升了异常检测性能。

基于自监督学习的方法旨在通过设定合适的代理任务来规避对异常图像标签的需求。该领域的显著工作包括CutPaste [20]，它通过将图像块从一个位置移植到另一个位置来生成异常，尽管异常区域的连续性欠佳。NSA [32] 使用泊松图像编辑 [26] 进行无缝图像粘贴，以合成更自然的异常区域。DRAEM [48] 则利用纹理数据集DTD [5]来

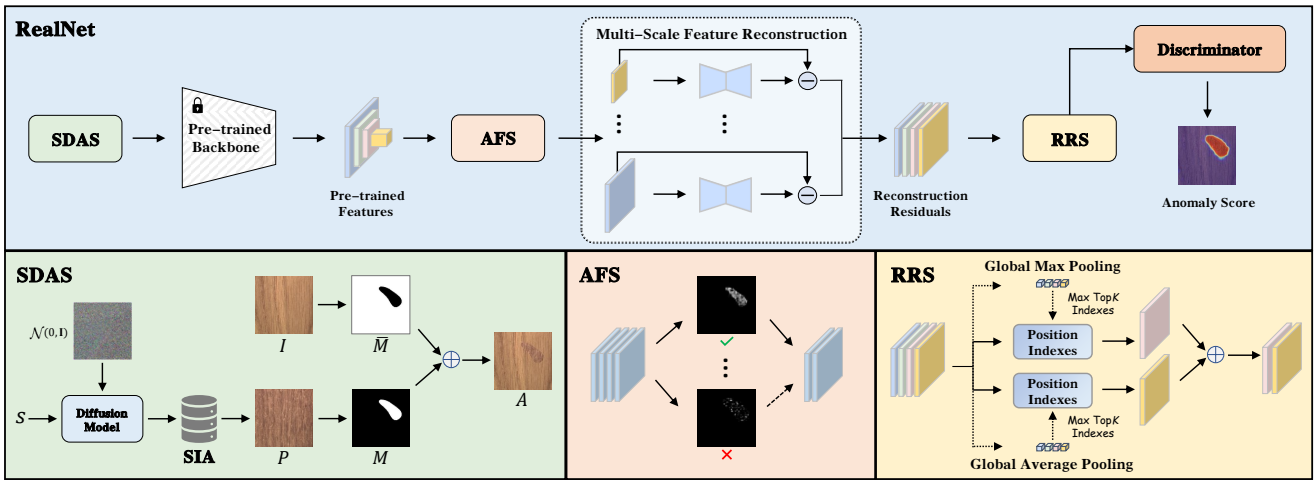


Figure 2. The pipeline of our RealNet consists of three core components: Strength-controllable Diffusion Anomaly Synthesis (SDAS), Anomaly-aware Features Selection (AFS), and Reconstruction Residuals Selection (RRS). 1) SDAS enables the synthesis of diverse, near-natural distribution anomalous images. 2) AFS refines features extracted by large-scale pre-trained CNN for dimensionality reduction. Refined features are reconstructed into corresponding normal image features by a set of reconstruction networks. 3) RRS selects reconstruction residuals most likely to identify anomalies, which are then fed into a discriminator for anomaly detection and localization.

synthesize various texture anomalies and achieve advanced self-supervised anomaly detection performance, however, it falls short when faced with specific structural anomalies, such as partial missing or misplaced elements.

The performance of self-supervised anomaly detection methods hinges on how closely the proxy task aligns with the real anomaly detection task. Anomaly synthesis, as a fundamental study in anomaly detection, has not yet received widespread exploration. Recent work [11] use StyleGAN2 [19] for image editing to generate anomalous images. However, the proposed method relies on real anomalous images and cannot generate unseen anomaly types. In contrast, SDAS operates in the probability space, free from constraints imposed by data augmentation rules or existing data, enabling effective control over anomaly strengths and the generation of realistic and diverse anomaly images using only normal images.

3. Method

In this section, we will introduce our proposed feature reconstruction framework, RealNet, which consists of three key components: Strength-controllable Diffusion Anomaly Synthesis (SDAS), Anomaly-aware Features Selection (AFS), and Reconstruction Residuals Selection (RRS). The pipeline of RealNet is illustrated in Fig. 2.

3.1. Strength-controllable Diffusion Anomaly Synthesis

Denoising Diffusion Probabilistic Models (DDPM) [16] employ a forward diffusion process to incrementally add noise $\mathcal{N}(0, \mathbf{I})$ to the original data distribution $q(x_0)$. At time t, the conditional probability distribution of the noisy data x_t is $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$, where $\{\beta_t\}_{t=1}^T$ is a fixed variance schedule, and $\{x_t\}_{t=1}^T$ are the latent variables. The diffusion process is defined as a Markov chain, with joint probability distribution $q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1})$. Following the sum rule of Gaussian random variables, the conditional probability distribution of x_t at time t is $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})$, where $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$.

The reverse process is described as another Markov chain, where the mean and variance of the reverse process are parameterized by θ, *i.e.*, $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$. There are various ways to model $\mu_\theta(x_t, t)$; typically, neural networks $\epsilon_\theta(x_t, t)$ are used to model the noise ϵ in the diffusion process, resulting in $\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t))$. In the training phase, our goal is to minimize the variational upper bound of the negative log-likelihood, which leads to the simplified objective:

$$\mathcal{L}_{simple} = \mathbb{E}_{t,x_0,\epsilon}[\|\epsilon - \epsilon_\theta(x_t, t)\|^2] \tag{1}$$

To generate realistic anomalous images, we first train a diffusion model to learn the distribution of normal images using Eq. (1). In reverse diffusion process characterized by $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$, x_{t-1} is the normal image obtained at time $t-1$. Due to the anomalous images being located in low-density regions near the normal images, we introduce an additional perturbation $s\Sigma$ to sample anomalous images, yielding $p(x'_{t-1}|x_{t-1}) = \mathcal{N}(x'_{t-1}; x_{t-1}, s\Sigma)$, where Σ is the additional introduced

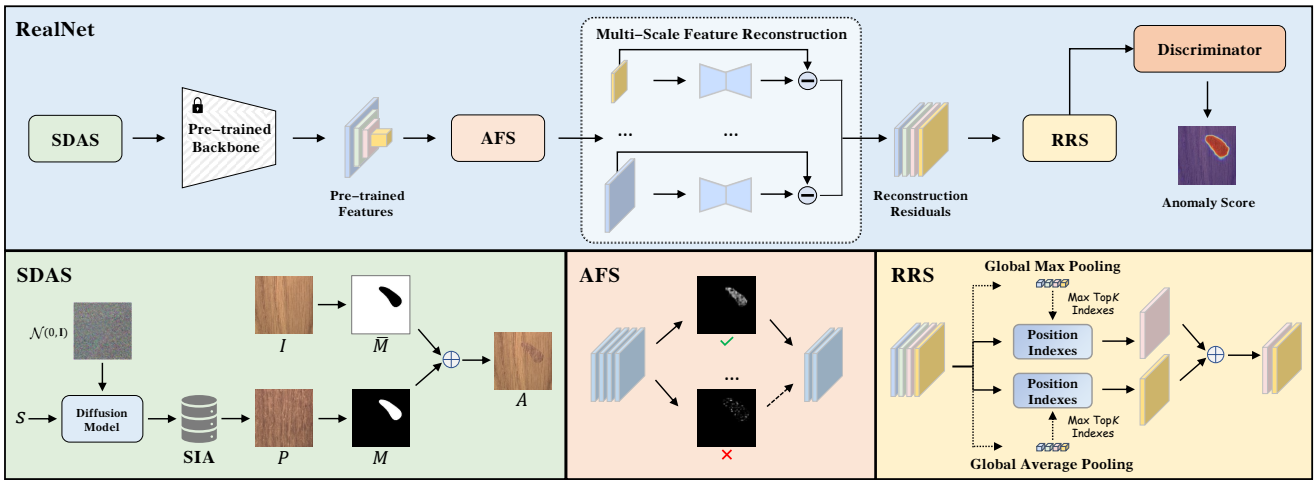


图2. 我们的RealNet流程包含三个核心组件：强度可控扩散异常合成（SDAS）、异常感知特征选择（AFS）与重建残差选择（RRS）。1）SDAS能够合成多样化、接近自然分布的异常图像。2）AFS对通过大规模预训练CNN提取的特征进行优化以实现降维，优化后的特征通过一组重建网络重构为对应的正常图像特征。3）RRS选择最可能识别异常的重建残差，随后将其输入判别器以进行异常检测与定位。

合成各种纹理异常并实现先进的自监督异常检测性能，但在面对特定结构异常时，例如部分缺失或错位元素，仍存在不足。

自监督异常检测方法的性能取决于代理任务与真实异常检测任务的契合程度。异常合成作为异常检测的基础研究，尚未得到广泛探索。近期研究[11]利用StyleGAN2[19]进行图像编辑以生成异常图像，但该方法依赖真实异常图像，无法生成未见过的异常类型。相比之下，SDAS在概率空间中运行，不受数据增强规则或现有数据的限制，能够有效控制异常强度，并仅使用正常图像生成逼真且多样化的异常图像。

3. 方法

在本节中，我们将介绍我们提出的特征重建框架RealNet，它包含三个关键组件：强度可控的扩散异常合成（SDAS）、异常感知特征选择（AFS）以及重建残差选择（RRS）。RealNet的流程如图2所示。

3.1. 强度可控的扩散异常合成

去噪扩散概率模型（DDPM）[16]采用前向扩散过程逐步添加噪声 $\mathcal{N}(0, \mathrm{I})$ 原始数据分布 $q(x_0)$。在时间 t，含噪声数据 x_t 的条件概率分布为 $q(x_t|x_{t-1}) = \mathcal{N}(x_t;\ \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathrm{I})$，其中 $\{\beta_t\}_{t=1}^T$ 是固定的方差调度，$\{x_t\}_{t=1}^T$ 是潜在变量。扩散过程被定义为马尔可夫链，其联合概率分布为 $q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1})$。根据高斯随机变量的求和规则，$x_t$ 在时间 t 的条件概率分布为 $q(x_t|x_0) = \mathcal{N}(x_t;\ \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathrm{I})$，其中 $\alpha_t = 1-\beta_t$，且 $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$。

逆过程被描述为另一个马尔可夫链，其中逆过程的均值和方差由θ、*i.e*、$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1};\ \mu_\theta(x_t,t), \Sigma_\theta(x_t,t))$参数化。有多种方法可以对$\mu_\theta(x_t,t)$进行建模；通常，神经网络$\epsilon_\theta(x_t,t)$被用来对扩散过程中的噪声$\epsilon$进行建模，从而得到 $\mu_\theta(x_t,t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t,t))$。在训练阶段，我们的目标是最小化负对数似然的变分上界，这导出了简化目标：

$$\mathcal{L}_{simple} = \mathbb{E}_{t,x_0,\epsilon}[\|\epsilon - \epsilon_\theta(x_t,t)\|^2] \quad (1)$$

为了生成逼真的异常图像，我们首先训练一个扩散模型，使用公式(1)学习正常图像的分布。在由 $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}、\ \mu_\theta(x_t,t), \Sigma_\theta(x_t,t))$描述的反向扩散过程中，$x_{t-1}$是在时间$t-1$处获得的正常图像。由于异常图像位于正常图像附近的低密度区域，我们引入额外的扰动$s\Sigma$来采样异常图像，得到 $p(x'_{t-1}|x_{t-1}) = \mathcal{N}(x'_{t-1};\ x_{t-1}, s\Sigma)$，其中$\Sigma$是额外引入的

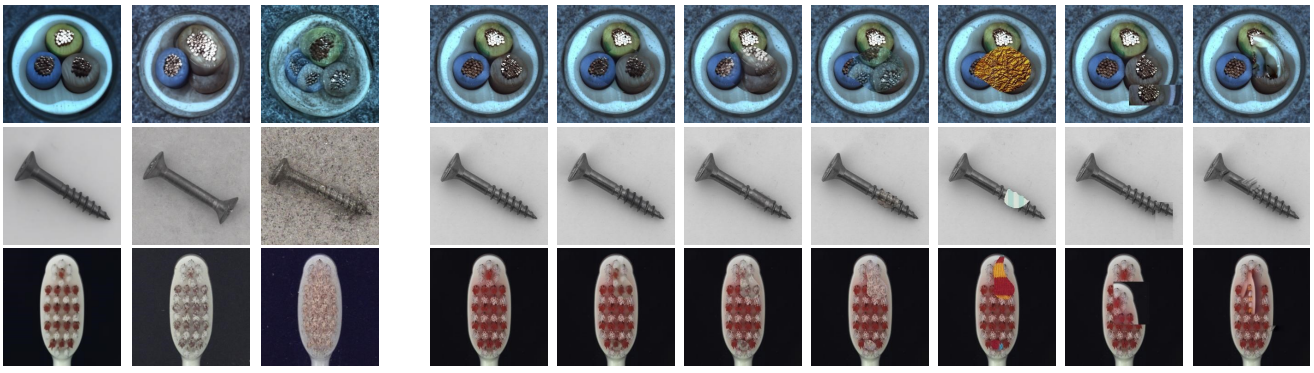


Figure 3. Anomaly image examples generated with different synthesis methods. (a) Examples generated using SDAS with different anomaly strengths s. (b) Examples featuring local anomaly regions generated by various anomaly synthesis methods.

variance, scalar s controls the anomaly strength ($s \geq 0$), and x'_{t-1} is the anomalous image obtained at time $t-1$. To simplify the anomalous synthesis process, we set $\Sigma = \Sigma_\theta(x_t, t)$, by which the conditional probability distribution of anomalous images x'_{t-1} can be written as follows:

$$p_\theta(x'_{t-1}|x_t) = \mathcal{N}(x'_{t-1}; \mu_\theta(x_t, t), (1+s)\Sigma_\theta(x_t, t)) \quad (2)$$

To ensure that the generated anomalous images are close to the distribution of normal images, we set $s \to 0$, resulting in $x'_{t-1} \approx x_{t-1}$; then we use x'_{t-1} for the next time step of the reverse diffusion process. The final form is $p_\theta(x'_{t-1}|x'_t) = \mathcal{N}(x'_{t-1}; \mu_\theta(x'_t, t), (1+s)\Sigma_\theta(x'_t, t))$. We term this process Strength-controllable Diffusion Anomaly Synthesis (SDAS), detailed in Algorithm 1. Specifically, SDAS will generate normal images if s is set to 0.

To incorporate these anomalous images during training of anomaly detection model, we follow the approach presented in [48], utilizing a Perlin noise generator [27] to capture various anomalous shapes and binarize them into an anomaly mask M. We denote the normal image as I, the anomalous image generated by SDAS as P, and the image with local anomalies synthesized by image blending as A:

$$A = \overline{M} \odot I + (1-\delta)(M \odot I) + \delta(M \odot P) \quad (3)$$

Algorithm 1 Strength-controllable Diffusion Anomaly Synthesis (SDAS)

Input: diffusion model $(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$
anomaly strength s
$x_T \sim \mathcal{N}(0, \mathbf{I})$
for all t from T to 1 **do**
  $\mu, \Sigma \leftarrow \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)$
  $x_{t-1} \sim \mathcal{N}(\mu, (1+s)\Sigma)$
end for
return x_0

where $\overline{M} = 1 - M$, $\odot$ denotes the element-wise multiplication operation, and δ is the opacity in the image blending. To ensure that the generated anomalous regions are located in the foreground, we use an adaptive threshold-based binarization method for foreground segmentation, similar to methods used in [32, 41, 42]. Fig. 3a shows the images generated by SDAS under different anomaly strengths, while Fig. 3b compares the images with local anomaly regions synthesized by different methods. The larger the value of s, the greater the distribution difference between the generated image and the normal image, and the more obvious the abnormal region obtained after image blending. When s is very small, imperceptible abnormal regions can be synthesized. Compared with alternative synthesis methods, the anomalies generated by SDAS are more continuous and can have very realistic structural anomalies.

3.2. Anomaly-aware Features Selection

In this section, we introduce the Anomaly-aware Features Selection (AFS) module within RealNet, a self-supervised method for pre-trained feature selection, reducing feature dimensionality and eliminating pre-training bias, as well as managing reconstruction costs. Firstly, we define a set of N triplets $\{A_n, I_n, M_n\}_{n=1}^N$, where $A_n, I_n \in R^{h \times w \times 3}$ represent anomaly images synthesized by SDAS and original normal images, and $M_n \in R^{h \times w}$ represents the corresponding anomaly mask. We denote the pre-trained network as ϕ_k, and $\phi_k(A_n) \in R^{h_k \times w_k \times c_k}$ represents the kth layer pre-trained feature extracted from A_n, where c_k represents the number of channels. For the ith feature map, $\phi_{k,i}(A_n) \in R^{h_k \times w_k}$, AFS selects m_k feature maps for reconstruction ($m_k \leq c_k$). Specifically, the feature maps indexed by k are from ResNet-like architectures, such as ResNet50 [13] or WideResNet50 [47], where $k \in \{1, 2, 3, 4\}$ represent the last layer outputs of blocks with different spatial resolutions.

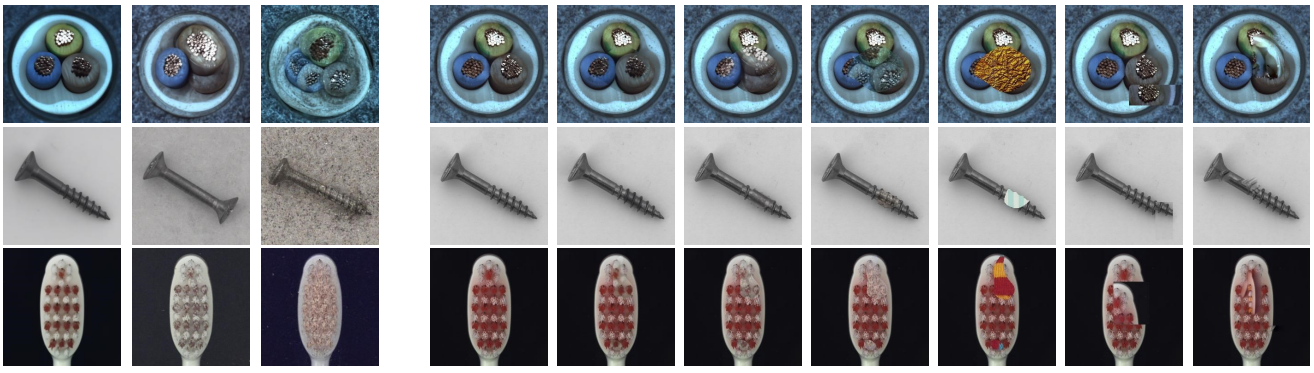


图3. 采用不同合成方法生成的异常图像示例。(a) 使用SDAS在不同异常强度s下生成的示例。(b) 展示多种异常合成方法生成的局部异常区域特征的示例。

方差，标量s控制异常强度（$s \geq 0$），而x'_{t-1}是在时间$t-1$处获得的异常图像。为了简化异常合成过程，我们设定$\Sigma = \Sigma_\theta(x_t, t)$，由此异常图像$x'_{t-1}$的条件概率分布可表示如下：

$$p_\theta(x'_{t-1}|x_t) = \mathcal{N}(x'_{t-1}; \mu_\theta(x_t, t), (1+s)\Sigma_\theta(x_t, t)) \quad (2)$$

为确保生成的异常图像接近正常图像的分布，我们设定$s \to$为0，从而得到$x'_{t-1} \approx x_{t-1}$；随后我们使用x'_{t-1}进行反向扩散过程的下一步。最终形式为$p_\theta(x'_{t-1}|x'_t) = \mathcal{N}(x'_{t-1};\ \mu_\theta(x'_t, t), (1+s)\Sigma_\theta(x'_t, t))$。我们将此过程称为强度可控扩散异常合成（SDAS），详见算法1。具体而言，若将s设为0，SDAS将生成正常图像。

为了在异常检测模型的训练中融入这些异常图像，我们遵循[48]中提出的方法，利用Perlin噪声生成器[27]来捕捉各种异常形状，并将其二值化为异常掩码M。我们将正常图像记为I，由SDAS生成的异常图像记为P，而通过图像混合合成的局部异常图像记为A：

$$A = \overline{M} \odot I + (1-\delta)(M \odot I) + \delta(M \odot P) \quad (3)$$

算法1 强度可控扩散异常合成（SDAS）

Input: diffusion model $(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$
anomaly strength s
$x_T \sim \mathcal{N}(0, \mathbf{I})$
for all t from T to 1 **do**
$\mu, \Sigma \leftarrow \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)$
$x_{t-1} \sim \mathcal{N}(\mu, (1+s)\Sigma)$
end for
return x_0

其中 $\overline{M} = 1 - M$, $\odot$ 表示逐元素乘法运算，δ 为图像融合中的不透明度。为确保生成的异常区域位于前景，我们采用基于自适应阈值的二值化方法进行前景分割，该方法与[32, 41, 42]中使用的技术类似。图3a展示了SDAS在不同异常强度下生成的图像，而图3b对比了不同方法合成的局部异常区域图像。s 值越大，生成图像与正常图像的分布差异越大，经图像融合后获得的异常区域也越明显。当 s 值极小时，可合成难以察觉的异常区域。与其他合成方法相比，SDAS生成的异常区域更具连续性，并能呈现高度真实的结构异常。

3.2. 异常感知特征选择

在本节中，我们介绍RealNet中的异常感知特征选择（AFS）模块，这是一种用于预训练特征选择的自监督方法，可降低特征维度、消除预训练偏差，并控制重建成本。首先，我们定义一组N三元组$\{A_n, I_n, M_n\}_{n=1}^N$，其中$A_n, I_n \in R^{h \times w \times 3}$表示由SDAS合成的异常图像和原始正常图像，$M_n \in R^{h \times w}$表示对应的异常掩码。我们将预训练网络表示为$\phi_k$，$\phi_k(A_n) \in R^{h_k \times w_k \times c_k}$表示从$A_n$中提取的第$k$层预训练特征，其中$c_k$表示通道数。对于第$i$个特征图$\phi_{k,i}(A_n) \in R^{h_k \times w_k}$，AFS选择$m_k$个特征图进行重建（$m_k \leq c_k$）。具体而言，由$k$索引的特征图来自类ResNet架构（如ResNet50 [13]或WideResNet50 [47]），其中$k \in \{1, 2, 3, 4\}$代表具有不同空间分辨率的块的最后层输出。

For the kth layer pre-trained features, we define the following AFS loss for evaluation of the ith feature map:

$$\mathcal{L}_{AFS}(\phi_{k,i}) = \frac{1}{N}\sum_{n=1}^{N} \|F([\phi_{k,i}(A_n) - \phi_{k,i}(I_n)]^2) - M_n\|_2^2 \tag{4}$$

where $F(\cdot)$ is a function that performs normalization operation and aligns the resolution of $[\phi_{k,i}(A_n) - \phi_{k,i}(I_n)]^2$ to M_n. Given the feature reconstruction process for anomalous images, we train a reconstruction network to infer $\phi_{k,i}(I_n)$ based on $\phi_{k,i}(A_n)$, which enables the detection and localization of anomalies through $[\phi_{k,i}(A_n) - \phi_{k,i}(I_n)]^2$. Ideally, $[\phi_{k,i}(A_n) - \phi_{k,i}(I_n)]^2$ should closely approximate M_n. The $\mathcal{L}_{AFS}(\phi_{k,i})$ represents the capability of $\phi_{k,i}$ in identifying anomalous regions. Due to the unavailability of real anomalous samples, we employ synthetic anomalies for feature selection. For the kth layer of pre-trained features, AFS selects m_k feature maps with the smallest $\mathcal{L}_{AFS}$ for reconstruction. We denote the AFS as $\varphi_k(\cdot)$, and $\varphi_k(A_n) \in R^{h_k \times w_k \times m_k}$, where $m_k \leq c_k$. We perform AFS on each layer of pre-trained features separately, and finally obtain selected multi-scale features $\{\varphi_1(A_n), ..., \varphi_K(A_n)\}$. In this process, each layer's feature dimension $\{m_1, ..., m_K\}$ serves as a set of hyperparameters. Specifically, for RealNet, AFS operation is performed only once on the pre-trained features of each layer, and the index of the selected feature maps is cached for subsequent training and inference.

AFS adaptively selects a subset of features from all available layers for anomaly detection, offering the following advantages compared to conventional methods [30, 38, 53] that select all features from partial layers: 1) AFS reduces feature redundancy within layers and mitigates pre-training bias, enhancing both feature representativeness and discriminability to improve anomaly detection performance. 2) AFS broadens the receptive field to enhance multi-scale anomaly detection capabilities. 3) AFS distinguishes the dimensions of pre-trained features from those employed for anomaly detection, ensuring efficient control over computational costs and flexible customization of the model size.

In RealNet, a set of reconstruction networks $\{G_1, ..., G_K\}$ are designed to reconstruct the selected synthetic anomalous features $\{\varphi_1(A_n), ..., \varphi_K(A_n)\}$ into the original image features $\{\varphi_1(I_n), ..., \varphi_K(I_n)\}$ at various resolutions. The loss function $\mathcal{L}_{recon}$ is defined as:

$$\mathcal{L}_{recon}(A, I) = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} \|G_k(\varphi_k(A_n)) - \varphi_k(I_n)\|_2^2 \tag{5}$$

During the reconstruction process, we intentionally forgo aligning multi-scale features [30, 33, 44] to preserve optimal performance. This choice is motivated by the potential drawbacks associated with aligning low-resolution features through down-sampling, which could compromise the network's detection resolution and increase the risk of misidentifying anomalies. On the other hand, aligning high-resolution features using up-sampling may result in unnecessary feature redundancy, leading to elevated reconstruction costs. A detailed discussion on the reconstruction network architectures can be found in Appendix C.

3.3. Reconstruction Residuals Selection

In this section, we present the Reconstruction Residuals Selection (RRS) module. Reconstruction residuals are denoted as $\{E_1(A_n), ..., E_K(A_n)\}$, where $E_k(A_n) = [\varphi_k(A_n) - G_k(\varphi_k(A_n))]^2$. To obtain the global reconstruction residual $E(A_n) \in R^{h' \times w' \times m'}$, we up-sample the low-resolution reconstruction residuals and concatenate them channel-wise, where $m' = \sum_{k=1}^{K} m_k$, $h' = max(h_1, ..., h_K)$, and $w' = max(w_1, ..., w_K)$.

The reconstruction residuals in $E(A_n)$ is obtained from the pre-trained features of reconstructing corresponding layer, and the features of the same resolution only have good ability to capture anomalies within a certain range. For instance, subtle low-level texture anomalies can be effectively captured exclusively by reconstruction residuals derived from low-level features. Therefore, RRS selects only a subset of reconstruction residuals that contain the most anomalous information for the anomaly score generation, to achieve the highest possible recall of anomalous regions.

Firstly, RRS performs GlobalMaxPooling (GMP) and GlobalAveragePooling (GAP) on $E(A_n)$ to obtain $E_{GMP}(A_n), E_{GAP}(A_n) \in R^{m'}$ respectively. The r largest elements in $E_{GMP}(A_n)$ and $E_{GAP}(A_n)$ are then used to index the positions of $E(A_n)$ and obtain $E_{max}(A_n, r), E_{avg}(A_n, r) \in R^{h' \times w' \times r}$, which respectively represent the Top$K$ reconstruction residuals with the highest maximum and average values. To avoid missed detections caused by inadequate resolution, reconstruction residuals with insufficient anomalous information are discarded in RRS.

As GMP and GAP respectively represent local and global properties spatially, E_{max} is more effective in capturing local anomalies in small areas, while E_{avg} focuses on selecting anomalies with large spans. Combining E_{max} and E_{avg} together can enhance the RRS's ability to capture anomalies of various scales. We define the RRS operator as $E_{RRS}(A_n, r) \in R^{h' \times w' \times r}$. $E_{RRS}(A_n, r)$ concatenates $E_{max}(A_n, r/2)$ and $E_{avg}(A_n, r/2)$. Finally, we feed the $E_{RRS}(A_n, r)$ into a discriminator, which maps the reconstruction residual to the image-level resolution, obtaining the final anomaly scores. The maximum value in anomaly scores is used as the image-level anomaly score. We use cross entropy loss $\mathcal{L}_{seg}(A, M)$ to supervise the training of

对于第k层预训练特征，我们定义以下AFS损失来评估第i个特征图：

$$\mathcal{L}_{AFS}(\phi_{k,i}) = \frac{1}{N}\sum_{n=1}^{N} \|F([\phi_{k,i}(A_n) - \phi_{k,i}(I_n)]^2) - M_n\|_2^2 \tag{4}$$

其中$F(\cdot)$是一个执行归一化操作并将[$\phi_{k,i}(A_n) - \phi_{k,i}(I_n)$]^{2}的分辨率对齐到$M_n$的函数。针对异常图像的特征重建过程，我们训练了一个重建网络以基于$\phi_{k,i}(A_n)$推断$\phi_{k,i}(I_n)$，从而通过[$\phi_{k,i}(A_n) - \phi_{k,i}(I_n)$]^{2}实现异常的检测与定位。理想情况下，[$\phi_{k,i}(A_n) - \phi_{k,i}(I_n)$]^{2}应紧密逼近$M_n$。$\mathcal{L}_{AFS}(\phi_{k,i})$代表了$\phi_{k,i}$识别异常区域的能力。由于真实异常样本的不可得性，我们采用合成异常进行特征选择。对于预训练特征的第k层，AFS选择具有最小$\mathcal{L}_{AFS}$的m_k个特征图进行重建。我们将AFS记为$\varphi_k(\cdot)$，且$\varphi_k(A_n) \in R^{h_k \times w_k \times m_k}$，其中$m_k \leq c_k$。我们对每一层预训练特征分别执行AFS，最终获得选定的多尺度特征$\{\varphi_1(A_n), ..., \varphi_K(A_n)\}$。在此过程中，每层特征维度$\{m_1, ..., m_K\}$作为一组超参数。具体而言，对于RealNet，AFS操作仅在每层预训练特征上执行一次，所选特征图的索引会被缓存以供后续训练和推理使用。

AFS自适应地从所有可用层中选择一个特征子集进行异常检测，相比传统方法[30, 38, 53]仅从部分层选取全部特征，具有以下优势：1) AFS减少了层内特征冗余并缓解预训练偏差，提升了特征的代表性与判别力，从而改善异常检测性能；2) AFS拓宽了感受野以增强多尺度异常检测能力；3) AFS将预训练特征的维度与异常检测所用维度解耦，确保在计算成本上实现高效控制，并能灵活定制模型规模。

在RealNet中，设计了一组重建网络$\{G_1, ..., G_K\}$，用于将选定的合成异常特征$\{\varphi_1(A_n), ..., \varphi_K(A_n)\}$在不同分辨率下重建为原始图像特征$\{\varphi_1(I_n), ..., \varphi_K(I_n)\}$。损失函数$\mathcal{L}_{recon}$定义为：

$$\mathcal{L}_{recon}(A, I) = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} \|G_k(\varphi_k(A_n)) - \varphi_k(I_n)\|_2^2 \tag{5}$$

在重建过程中，我们有意放弃对齐多尺度特征[30, 33, 44]以保持最佳性能。这一选择是基于以下考通过下采样对齐低分辨率特征可能带来的潜在缺点，是可能损害网络的检测分辨率，并增加误判异常的风险。另一方面，使用上采样对齐高分辨率特征可能导致不必要的特征冗余，从而增加重建成本。关于重建网络架构的详细讨论可参见附录C。

3.3. 重建残差选择

在本节中，我们介绍重构残差选择（RRS）模块。重构残差表示为$\{E_1(A_n), ..., E_K(A_n)\}$，其中$E_k(A_n) = [\varphi_k(A_n) - G_k(\varphi_k(A_n))]^2$。为获得全局重构残差$E(A_n) \in R^{h' \times w' \times m'}$，我们对低分辨率重构残差进行上采样并按通道维度拼接，其中$m' = \sum_{k=1}^{K} m_k$、$h' = max(h_1, ..., h_K)$，以及$w' = max(w_1, ..., w_K)$。

$E(A_n)$中的重建残差是从预训练的重建对应层特征中获得的，且相同分辨率的特征仅具备在一定范围内捕捉异常的良好能力。例如，细微的低级纹理异常只能通过源自低级特征的重建残差被有效捕获。因此，RRS仅选择包含最多异常信息的重建残差子集来生成异常分数，以实现异常区域尽可能高的召回率。

首先，RRS对$E(A_n)$分别执行全局最大池化（GMP）和全局平均池化（GAP）以获得$E_{GMP}(A_n), E_{GAP}(A_n) \in R^{m'}$。随后，利用$E_{GMP}(A_n)$和$E_{GAP}(A_n)$中前$r$个最大元素索引$E(A_n)$的位置并获取$E_{max}(A_n, r), E_{avg}(A_n, r) \in R^{h' \times w' \times r}$，其分别代表具有最高最大值和最高平均值的Top$K$个重建残差。为避免因分辨率不足导致的漏检，RRS会丢弃那些异常信息不足的重建残差。

由于GMP和GAP分别代表空间上的局部和全局特性，E_{max}能更有效地捕捉小范围内的局部异常，而E_{avg}则侧重于选择大跨度的异常。将E_{max}和E_{avg}结合可以增强RRS捕捉多尺度异常的能力。我们将RRS算子定义为$E_{RRS}(A_n, r) \in R^{h' \times w' \times r}$。$E_{RRS}(A_n, r)$将$E_{max}(A_n, r/2)$和$E_{avg}(A_n, r/2)$进行拼接。最后，我们将$E_{RRS}(A_n, r)$输入判别器，将重建残差映射到图像级分辨率，从而获得最终的异常分数。异常分数中的最大值被用作图像级异常分数。我们使用交叉熵损失$\mathcal{L}_{seg}(A, M)$来监督训练过程。

Table 1. Comparison of SIA with alternative anomaly synthesis approaches on the MVTec-AD dataset [3], employing Image AUROC (%), Pixel AUROC (%), and PRO (%) as evaluation metrics.

Category		SIA	DTD [5]	NSA [32]	CutPaste [20]
Texture	Carpet	(99.84, 99.19, 96.41)	(**100.0**, **99.27**, **96.96**)	(99.80, 98.60, 88.77)	(99.24, 98.42, 93.85)
	Grid	(**100.0**, 99.51, **97.28**)	(**100.0**, **99.57**, 97.14)	(**100.0**, 99.32, 91.31)	(**100.0**, 99.18, 92.53)
	Leather	(**100.0**, 99.76, 96.22)	(**100.0**, **99.77**, 96.41)	(**100.0**, 99.24, **96.85**)	(**100.0**, 99.41, 92.12)
	Tile	(99.96,**99.44**, **97.70**)	(**100.0**, 99.35, 95.27)	(**100.0**, 97.40, 86.45)	(99.86, 97.63, 84.39)
	Wood	(99.21, 98.22, 90.54)	(**99.65**, **98.28**, **91.23**)	(97.63, 93.30, 87.20)	(98.95, 95.29, 81.47)
	AVG	(99.80, 99.22, **95.63**)	(**99.93**, **99.25**,95.40)	(99.49, 97.57, 90.11)	(99.61, 97.99, 88.87)
Object	Bottle	(**100.0**, 99.30, **95.62**)	(**100.0**, 99.35, 95.57)	(**100.0**, **99.37**, 93.49)	(**100.0**, 99.14, 91.41)
	Cable	(99.19, **98.10**, **93.38**)	(98.95, 97.84, 90.36)	(**99.33**, 97.62, 93.26)	(96.35, 96.23, 86.05)
	Capsule	(**99.56**, **99.32**, 84.48)	(99.32, 99.19, 82.28)	(99.04, 99.27, **85.77**)	(98.48, 99.10, 79.55)
	hazelnut	(**100.0**, **99.68**, 93.14)	(**100.0**, 99.46, 93.46)	(**100.0**, 99.25, **94.41**)	(**100.0**, 99.03, 91.51)
	Metal Nut	(99.76, 98.58, 94.39)	(99.90, 98.58, **96.49**)	(**100.0**, **99.11**, 93.27)	(99.90, 98.03, 89.69)
	Pill	(**99.13**, **99.02**, 91.04)	(98.36, 98.88, 84.44)	(97.19, 98.28, **95.15**)	(97.22, 98.96, 86.48)
	Screw	(**98.83**, 99.45, 87.90)	(97.72, 99.36, 85.22)	(98.79, **99.62**, **93.74**)	(92.74, 98.53, 79.63)
	Toothbrush	(99.44, 98.71, **91.57**)	(99.44, 98.69, 90.87)	(**100.0**, **99.18**, 89.20)	(99.17, 98.85, 78.48)
	Transistor	(**100.0**, **98.00**, **92.92**)	(99.71, 97.15, 86.56)	(98.54, 95.67, 79.09)	(99.38, 96.32, 76.52)
	zipper	(99.82, **99.17**, **93.43**)	(99.68, 99.02, 88.77)	(**99.90**, 98.91, 93.05)	(99.61, 98.03, 92.26)
	AVG	(**99.57**, **98.93**, **91.79**)	(99.31, 98.75, 89.40)	(99.28, 98.63, 91.04)	(98.29, 98.22, 85.16)
AVG		(**99.65**, **99.03**, **93.07**)	(99.52, 98.92, 91.40)	(99.35, 98.28, 90.73)	(98.73, 98.14, 86.40)

Table 2. Comparison of RealNet with alternative anomaly detection methods on the MVTec-AD dataset [3].

Metric	*PatchCore* [30]	*SimpleNet* [21]	*FastFlow* [46]	DRAEM+SSPCAB [29]	DSR [49]	UniAD [44]	RD++ [38]	DeSTSeg [53]	DiffAD [52]	RealNet
Image AUROC	99.1	**99.6**	99.3	98.9	98.2	96.6	99.4	98.6	98.7	**99.6**
Pixel AUROC	98.1	98.1	98.1	97.2	-	96.6	98.3	97.9	98.3	**99.0**

discriminator. The overall loss of RealNet is:

$$\mathcal{L}(A, I, M) = \mathcal{L}_{recon}(A, I) + \mathcal{L}_{seg}(A, M) \qquad (6)$$

3.4. Synthetic Industrial Anomaly Dataset

To facilitate the reuse of generated anomaly images by SDAS, we constructed the Synthetic Industrial Anomaly Dataset (SIA). SIA comprises anomaly images for 36 categories from four industrial anomaly detection datasets, including MVTec-AD [3], MPDD [18], BTAD [24], and VisA [55]. We generated 10,000 anomaly images with a resolution of 256×256 for each category, with anomaly strength s uniformly sampled between 0.1 and 0.2. SIA can be conveniently used for synthesizing anomaly images through image blending, as described in Eq. (3), and can serve as an effective alternative to the widely used DTD dataset [5].

4. Experiment

4.1. Experimental setup

Datasets. We conduct extensive evaluations on four datasets, including MVTec-AD [3], MPDD [18], BTAD [24], and VisA [55]. MVTec-AD [3] contains 5,354 images from 15 categories for industrial anomaly detection tasks, including 10 object categories and 5 texture categories. MPDD [18] contains 1,346 images from 6 types of industrial metal products with varying lighting conditions, non-uniform backgrounds, and multiple products in each image. Furthermore, the placement orientation, shooting distance, and position of the products are also varied. BTAD [24] contains images of 3 industrial products from the real world. VisA [55] is comprised of 9,621 normal images and 1,200 anomaly images from 12 categories. Certain categories demonstrate intricate structures, as exemplified by PCBs, while others consist of multiple objects that require detection, such as Capsules, thus rendering detection and localization a challenging task.

Metrics. To evaluate the performance of image-level anomaly detection, we use the Area Under the Receiver Operator Curve (AUROC) metric, as in previous works [3, 18, 24, 55]. For pixel-level anomaly location, we use Pixel AUROC and Per Region Overlap (PRO) [4].

Implementation details. We evaluate RealNet on four datasets with consistent network architectures and hyperparameters, without specific tuning for individual categories. We use a WideResNet50 [47] pre-trained on ImageNet [9] as the backbone. In AFS, we set the dimension of pre-trained feature of each layer to $\{256, 512, 512, 256\}$ for reconstruction. For RRS, $1/3$ of the reconstruction residuals are reserved to generate the final anomaly scores. For SDAS, we train the diffusion model following [10] and use the SIA dataset for anomaly synthesis. Both SDAS and anomaly detection are performed at a resolution of 256×256 without center cropping, with a batch size of 16, and we use 64 batches of synthetic anomaly images for AFS. More details can be found in Appendix B.

表1. 在MVTec-AD数据集[3]上，采用图像AUROC（%）、像素AUROC（%）和PRO（%）作为评估指标，将SIA与替代异常合成方法的比较。

Category		SIA	DTD [5]	NSA [32]	CutPaste [20]
Texture	Carpet	(99.84, 99.19, 96.41)	(**100.0**, **99.27**, **96.96**)	(99.80, 98.60, 88.77)	(99.24, 98.42, 93.85)
	Grid	(**100.0**, 99.51, **97.28**)	(**100.0**, **99.57**, 97.14)	(**100.0**, 99.32, 91.31)	(**100.0**, 99.18, 92.53)
	Leather	(**100.0**, 99.76, 96.22)	(**100.0**, **99.77**, 96.41)	(**100.0**, 99.24, **96.85**)	(**100.0**, 99.41, 92.12)
	Tile	(99.96,**99.44**, **97.70**)	(**100.0**, 99.35, 95.27)	(**100.0**, 97.40, 86.45)	(99.86, 97.63, 84.39)
	Wood	(99.21, 98.22, 90.54)	(**99.65**, **98.28**, **91.23**)	(97.63, 93.30, 87.20)	(98.95, 95.29, 81.47)
	AVG	(99.80, 99.22, **95.63**)	(**99.93**, **99.25**,95.40)	(99.49, 97.57, 90.11)	(99.61, 97.99, 88.87)
Object	Bottle	(**100.0**, 99.30, **95.62**)	(**100.0**, 99.35, 95.57)	(**100.0**, **99.37**, 93.49)	(**100.0**, 99.14, 91.41)
	Cable	(99.19, **98.10**, **93.38**)	(98.95, 97.84, 90.36)	(**99.33**, 97.62, 93.26)	(96.35, 96.23, 86.05)
	Capsule	(**99.56**, **99.32**, 84.48)	(99.32, 99.19, 82.28)	(99.04, 99.27, **85.77**)	(98.48, 99.10, 79.55)
	hazelnut	(**100.0**, **99.68**, 93.14)	(**100.0**, 99.46, 93.46)	(**100.0**, 99.25, **94.41**)	(**100.0**, 99.03, 91.51)
	Metal Nut	(99.76, 98.58, 94.39)	(99.90, 98.58, **96.49**)	(**100.0**, **99.11**, 93.27)	(99.90, 98.03, 89.69)
	Pill	(**99.13**, **99.02**, 91.04)	(98.36, 98.88, 84.44)	(97.19, 98.28, **95.15**)	(97.22, 98.96, 86.48)
	Screw	(**98.83**, 99.45, 87.90)	(97.72, 99.36, 85.22)	(98.79, **99.62**, **93.74**)	(92.74, 98.53, 79.63)
	Toothbrush	(99.44, 98.71, **91.57**)	(99.44, 98.69, 90.87)	(**100.0**, **99.18**, 89.20)	(99.17, 98.85, 78.48)
	Transistor	(**100.0**, **98.00**, **92.92**)	(99.71, 97.15, 86.56)	(98.54, 95.67, 79.09)	(99.38, 96.32, 76.52)
	zipper	(99.82, **99.17**, **93.43**)	(99.68, 99.02, 88.77)	(**99.90**, 98.91, 93.05)	(99.61, 98.03, 92.26)
	AVG	(**99.57**, **98.93**, **91.79**)	(99.31, 98.75, 89.40)	(99.28, 98.63, 91.04)	(98.29, 98.22, 85.16)
AVG		(**99.65**, **99.03**, **93.07**)	(99.52, 98.92, 91.40)	(99.35, 98.28, 90.73)	(98.73, 98.14, 86.40)

表2. 在MVTec-AD数据集[3]上RealNet与替代异常检测方法的比较。

Metric	*PatchCore* [30]	*SimpleNet* [21]	*FastFlow* [46]	DRAEM+SSPCAB [29]	DSR [49]	UniAD [44]	RD++ [38]	DeSTSeg [53]	DiffAD [52]	RealNet
Image AUROC	99.1	**99.6**	99.3	98.9	98.2	96.6	99.4	98.6	98.7	**99.6**
Pixel AUROC	98.1	98.1	98.1	97.2	-	96.6	98.3	97.9	98.3	**99.0**

判别器。RealNet的整体损失为：

$$\mathcal{L}(A, I, M) = \mathcal{L}_{recon}(A, I) + \mathcal{L}_{seg}(A, M) \qquad (6)$$

3.4. 合成工业异常数据集

为促进SDAS生成的异常图像复用，我们构建了合成工业异常数据集（SIA）。SIA包含来自四个工业异常检测数据集的36个类别的异常图像，包括MVTec-AD [3]、MPDD [18]、BTAD [24]和VisA [55]。我们为每个类别生成了10,000张分辨率为256×256的异常图像，异常强度s在0.1至0.2间均匀采样。如公式(3)所述，SIA可通过图像混合便捷地用于合成异常图像，并能作为广泛使用的DTD数据集[5]的有效替代方案。

4. 实验

4.1. 实验设置

数据集。我们在四个数据集上进行了广泛的评估，包括MVTec-AD [3]、MPDD [18]、BTAD [24]和VisA [55]。MVTec-AD [3]包含来自15个类别的5,354张图像，用于工业异常检测任务，其中包括10个物体类别和5个纹理类别。MPDD [18]包含来自6种工业金属产品的1,346张图像，这些图像具有不同的光照条件、非均匀背景以及多个产品在每张图像。此外，产品的摆放方向、拍摄距离和位置也各不相同。BTAD [24] 包含来自现实世界的3种工业产品的图像。VisA [55] 由来自12个类别的9,621张正常图像和1,200张异常图像组成。某些类别展现出复杂的结构，例如PCB，而其他类别则包含多个需要检测的对象，例如胶囊，这使得检测和定位成为一项具有挑战性的任务。

指标。为了评估图像级异常检测的性能，我们采用接收者操作特征曲线下面积（AUROC）指标，与先前的研究一致[3, 18, 24, 55]。对于像素级异常定位，我们使用像素AUROC和区域重叠度（PRO）[4]。

实现细节。我们在四个数据集上评估RealNet，采用一致的网络架构和超参数，未针对个别类别进行特定调优。我们使用在ImageNet上预训练的WideResNet50作为骨干网络。在AFS中，我们将每层预训练特征的维度设置为{256, 512, 512, 256}用于重建。对于RRS，我们保留1/3的重建残差以生成最终异常分数。对于SDAS，我们遵循[10]的方法训练扩散模型，并使用SIA数据集进行异常合成。SDAS和异常检测均在256×256分辨率下执行（不进行中心裁剪），批大小为16，并在AFS中使用64批合成异常图像。更多细节可参见附录B。

Table 3. Comparison of SIA with DTD [5] and CutPaste [20] on the MPDD dataset [18], employing Image AUROC (%), Pixel AUROC (%), and PRO (%) as evaluation metrics.

Category	SIA	DTD [5]	CutPaste [20]
Bracket Black	(**94.95**, **99.27**, 87.10)	(89.49, 98.90, **88.57**)	(66.42, 96.67, 56.53)
Bracket Brown	(**96.83**, **97.81**, **94.36**)	(92.99, 97.35, 92.64)	(95.48, 97.54, 55.17)
Bracket White	(**88.78**, 97.44, **84.00**)	(86.67, **98.59**, 77.08)	(88.44, 96.51, 64.32)
Connector	(**100.0**, 97.46, **84.79**)	(99.05, 97.76, 65.91)	(99.05, **98.47**, 74.05)
Metal Plate	(**100.0**, 99.28, **94.44**)	(**100.0**, **99.35**, 93.78)	(99.95, 98.83, 92.69)
Tubes	(**97.51**, 97.94, 93.29)	(92.62, **99.01**, **96.49**)	(91.49, 98.09, 92.99)
AVG	(**96.35**, 98.20, **89.66**)	(93.47, **98.49**, 85.75)	(90.14, 97.69, 72.63)

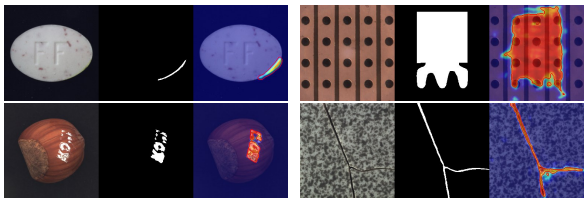


Figure 4. Qualitative results of RealNet on the MVTec-AD dataset [3]. Within each group, from left to right, are the anomaly image, ground-truth, and predicted anomaly score.

Table 4. Comparison of RealNet with alternative anomaly detection methods on the MPDD dataset [18].

Metric	*PatchCore* [30]	*CFlow* [12]	*PaDiM* [7]	*SPADE* [6]	DAGAN [36]	Skip-GANomaly [1]	RealNet
Image AUROC	82.1	86.1	74.8	77.1	72.5	64.8	**96.3**
Pixel AUROC	95.7	97.7	96.7	95.9	83.3	82.2	**98.2**

4.2. Anomaly detection on MVTec-AD

We train RealNet using SIA and alternative anomaly synthetic methods on the MVTec-AD dataset [3], to evaluate the model's performance in anomaly detection and localization. These methods include: 1) **DTD** [5]: This method utilizes the DTD dataset [5] to blend images with generated anomalous textures, and the data augmentation strategy in [48] is employed during the blending process. 2) **NSA** [32]: This method employs Poisson image editing [26] to seamless image editing, following parameter setting in [32]. 3) **CutPaste** [20]: This method involves random image cropping and pasting to synthesize anomaly regions.

The experimental results are shown in Tab. 1. SDAS demonstrates flexibility in controlling the anomaly strength and generates synthetic anomalies with multiple anomaly patterns, especially for the object categories, where it achieves the best detection and localization performance. Compared to other methods, SDAS is not constrained by data augmentation rules or external data, enabling the synthesis of more natural and rich functional anomalies, as shown in Fig. 3. RealNet trained using SIA achieves remarkable performance on the MVTec-AD dataset [3], with an Image AUROC of 99.65%, a Pixel AUROC of 99.03%, and a PRO score of 93.07%. Fig. 4 presents the qualitative anomaly localization results of RealNet on the MVTec-AD dataset [3]. The method exhibits remarkable pixel-level anomaly localization, proficiently identifying diverse anomaly patterns at various scales. Furthermore, RealNet can achieve a rapid inference speed of 31.93 FPS when using a single Nvidia GeForce RTX 3090, and it can perform inference using only 4GB of GPU memory. A detailed computational efficiency analysis can be found in Appendix C.

We also compare RealNet with several state-of-the-art anomaly detection methods, and the results are shown in Tab. 2. Built on the same pre-trained network, RealNet outperforms the state-of-the-art alternatives, including Deep Feature Embedding-Based method (PatchCore [30] and SimpleNet [21]) and the NF-Based method (FastFlow [46]). When compared to previous reconstruction-based methods, RealNet achieves significant performance improvement.

4.3. Anomaly detection on MPDD

We evaluate RealNet on MPDD dataset [18] with SIA, DTD [5], and CutPaste [20], and the results are shown in Tab. 3. Notably, RealNet trained with SIA achieves a significant improvement of 2.88% in Image AUROC over DTD [5]. Tab. 4 shows the Image AUROC and the Pixel AUROC of RealNet and other methods on the MPDD dataset [18]. RealNet achieves an Image AUROC of 96.3%, surpassing the current best performance (CFlow [12]) by 10.2%, even without any dataset-specific tuning.

4.4. Anomaly detection on other benchmarks

To comprehensively assess the effectiveness of RealNet, we conduct experiments on the BTAD [24] and VisA [55] datasets. On the VisA dataset [55], characterized by complex structures and multiple detection objects, RealNet employing SIA yields a significant performance improvement, achieving an Image AUROC of 97.8% and a Pixel AUROC of 98.8%. In the case of the BTAD dataset [24], RealNet with SIA achieves competitive results, securing an Image AUROC of 96.1% and a Pixel AUROC of 97.9%. Detailed results can be found in Appendix C.

4.5. Ablation studies

To evaluate the effectiveness of each module of RealNet, we conduct comprehensive ablation studies on MVTec-AD dataset [3]. First, we evaluate the impact of AFS and RRS on the performance of RealNet.

W/O AFS: We replace AFS with two alternative dimensionality reduction methods, namely Random Dimensionality Reduction (RDR) [7] and Random Linear Projections

表3. 在MPDD数据集[18]上，采用图像AUROC（%）、像素AUROC（%）和PRO（%）作为评估指标，对比SIA与DTD[5]及CutPaste[20]的性能。

Category	SIA	DTD [5]	CutPaste [20]
Bracket Black	(**94.95**, **99.27**, 87.10)	(89.49, 98.90, **88.57**)	(66.42, 96.67, 56.53)
Bracket Brown	(**96.83**, **97.81**, **94.36**)	(92.99, 97.35, 92.64)	(95.48, 97.54, 55.17)
Bracket White	(**88.78**, 97.44, **84.00**)	(86.67, **98.59**, 77.08)	(88.44, 96.51, 64.32)
Connector	(**100.0**, 97.46, **84.79**)	(99.05, 97.76, 65.91)	(99.05, **98.47**, 74.05)
Metal Plate	(**100.0**, 99.28, **94.44**)	(**100.0**, **99.35**, 93.78)	(99.95, 98.83, 92.69)
Tubes	(**97.51**, 97.94, 93.29)	(92.62, **99.01**, **96.49**)	(91.49, 98.09, 92.99)
AVG	(**96.35**, 98.20, **89.66**)	(93.47, **98.49**, 85.75)	(90.14, 97.69, 72.63)

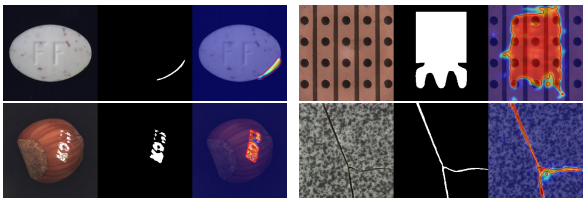


图4. RealNet在MVTec-AD数据集[3]上的定性结果。每组中从左至右分别为异常图像、真实标注及预测的异常得分。

表4. MPDD数据集[18]上RealNet与替代异常检测方法的对比。

Metric	*PatchCore* [30]	*CFlow* [12]	*PaDiM* [7]	*SPADE* [6]	DAGAN [36]	Skip-GANomaly [1]	RealNet
Image AUROC	82.1	86.1	74.8	77.1	72.5	64.8	**96.3**
Pixel AUROC	95.7	97.7	96.7	95.9	83.3	82.2	**98.2**

4.2. MVTec-AD数据集上的异常检测

我们在MVTec-AD数据集[3]上使用SIA及替代性异常合成方法训练RealNet，以评估模型在异常检测与定位中的性能。这些方法包括：1）DTD[5]：该方法利用DTD数据集[5]将图像与生成的异常纹理混合，并在混合过程中采用[48]提出的数据增强策略；2）NSA[32]：该方法采用泊松图像编辑[26]进行无缝图像合成，参数设置遵循[32]；3）CutPaste[20]：该方法通过随机裁剪并粘贴图像区域来合成异常区域。

实验结果如表1所示。SDAS在控制异常强度方面展现出灵活性，并能生成具有多种异常模式的合成异常，特别是在物体类别上实现了最佳的检测与定位性能。相较于其他方法，SDAS不受数据增强规则或外部数据的限制，能够合成更自然且丰富的功能异常，如图3所示。使用SIA训练的RealNet在MVTec-AD数据集[3]上取得了显著性能，图像AUROC达99.65%，像素AUROC达99.03%，PRO分数为93.07%。图4展示了RealNet在MVTec-AD数据集[3]上的定性异常定位结果。该方法在像素级异常定位方面表现突出，能够有效识别不同尺度的多种异常模式。此外，RealNet在使用单张Nvidia GeForce RTX 3090显卡时推理速度可达31.93 FPS，且仅需4GB显存即可完成推理。详细的计算效率分析见附录C。

我们还比较了RealNet与几种最先进的异常检测方法，结果如表2所示。基于相同的预训练网络，RealNet的表现优于包括Deep在内的最先进替代方案。基于特征嵌入的方法（PatchCore [30] 和 SimpleNet [21]）以及基于归一化流的方法（FastFlow [46]）。与先前的基于重建的方法相比，RealNet 实现了显著的性能提升。

4.3. MPDD上的异常检测

我们在MPDD数据集[18]上使用SIA、DTD[5]和CutPaste[20]对RealNet进行评估，结果如表3所示。值得注意的是，采用SIA训练的RealNet在图像AUROC上比DTD[5]显著提升了2.88%。表4展示了RealNet及其他方法在MPDD数据集[18]上的图像AUROC和像素AUROC。RealNet实现了96.3%的图像AUROC，即使未进行任何数据集特定调优，仍超越当前最佳性能（CFlow[12]）10.2%。

4.4. 其他基准测试上的异常检测

为了全面评估RealNet的有效性，我们在BTAD [24]和VisA [55]数据集上进行了实验。在具有复杂结构和多检测对象特点的VisA数据集[55]上，采用SIA的RealNet实现了显著的性能提升，获得了97.8%的图像AUROC和98.8%的像素AUROC。在BTAD数据集[24]上，配备SIA的RealNet取得了具有竞争力的结果，获得了96.1%的图像AUROC和97.9%的像素AUROC。详细结果可参见附录C。

4.5. 消融研究

为了评估RealNet各模块的有效性，我们在MVTec-AD数据集[3]上进行了全面的消融实验。首先，我们评估了AFS和RRS对RealNet性能的影响。

无AFS：我们将AFS替换为两种替代的降维方法，即随机降维（RDR）[7]和随机线性投影。

Table 5. Ablation studies of RealNet on the MVTec-AD dataset [3].

(a) The impact of AFS and RRS on RealNet.

	AFS	RRS	Image AUROC	Pixel AUROC	PRO
1	-	-	94.46 / 95.67	93.38 / 95.84	79.81 / 82.26
2	✓	-	96.86	96.32	84.13
3	-	✓	99.39 / 99.09	98.66 / 98.22	92.01 / 88.38
4	✓	✓	**99.65**	**99.03**	**93.07**

(b) The impact of anomaly strengths on RealNet.

Metric	$s = 0$	$s = 0.1$	$s = 0.2$	$s \in [0.1, 0.2]$
Image AUROC	99.35	**99.65**	99.61	**99.65**
Pixel AUROC	98.85	98.96	98.95	**99.03**
PRO	91.80	92.16	89.36	**93.07**

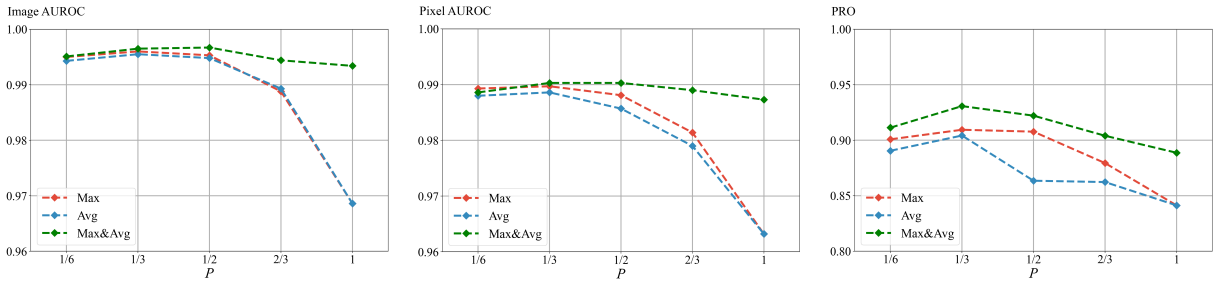


Figure 5. Performance of RealNet on MVTec-AD dataset [3] under various modes of reconstruction residuals selection (`Max`, `Avg`, and `Max&Avg`) and varying retention ratio P of reconstruction residuals.

Reduction (RLPR) [30, 40]. RDR randomly selects some dimensional features from high-dimensional features, while RLPR employs an untrained linear transformation layer for linear projection. We report the results of our RealNet with RDR and RLPR, respectively, as shown in the experiments 1 and 3 in Tab. 5a. **W/O RRS:** We feed the global reconstruction residual $E(A_n)$ into the discriminator to generate anomaly scores, and the results are shown in the experiments 1 and 2 in Tab. 5a.

As indicated by the ablation results in Tab. 5a, RRS contributes significantly to performance improvement. Using all the reconstruction residuals to generate anomaly scores, reconstruction residuals lacking of anomaly information can lead to missed anomaly regions, resulting in a significant decrease in anomaly detection performance. Furthermore, AFS yields better anomaly detection results compared to RDR and RLPR. A straightforward visualization result about AFS is provided in Appendix D.

We further investigate the impact of anomaly strength s in SDAS, and the results are shown in Tab. 5b. When s equals 0, SDAS generates normal images in high probability density regions. Blending images may introduce false positive anomaly regions, which lowers the reconstruction difficulty and confuses the discriminator, leading to suboptimal performance. Conversely, when s is too large, the synthetic anomaly images deviate from the true distribution of anomalous images, causing RealNet's performance to deteriorate. Our findings indicate that uniformly sampling s within a specific range serves as a robust approach for generating anomalous images. This method enables RealNet to cover a wider range of anomalous patterns, ultimately improving the overall anomaly detection performance.

In Fig. 5, we report the impact of different RRS modes and retention ratios on the performance of RealNet. Since the setting of r is related to m_k, we introduce the retention ratio P, defined as: $P = \frac{r}{\sum_{k=1}^{K} m_k}$. Compared to `Max` and `Avg` modes, `Max&Avg` mode demonstrates superior robustness in detecting anomalies across various scales. At equal retention rates, the `Max&Avg` mode discards more reconstruction residuals lacking anomalous information than the `Max` and `Avg` modes, mitigating performance degradation and further emphasizing the effectiveness of the `Max&Avg` mode in enhancing RealNet's anomaly detection capabilities. More ablation experiments and analysis can be found in Appendix C.

5. Conclusion

In this work, we introduce RealNet, an innovative self-supervised anomaly detection framework. Our approach integrates three core components: Strength-controllable Diffusion Anomaly Synthesis (SDAS), Anomaly-aware Features Selection (AFS), and Reconstruction Residuals Selection (RRS). These components synergistically contribute to RealNet, enabling effective exploitation of large-scale pre-trained models in anomaly detection while keeping the computational overhead within a reasonably low and acceptable range. RealNet provides a flexible foundation for future research in anomaly detection utilizing pre-trained feature reconstruction techniques. Through extensive experiments, we illustrate RealNet's proficiency in addressing diverse real-world anomaly detection challenges.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 62177034 and Grant 61972046.

Table 5. Ablation studies of RealNet on the MVTec-AD dataset [3].

(AFS和RRS对RealNet的影响。

	AFS	RRS	Image AUROC	Pixel AUROC	PRO
1	-	-	94.46 / 95.67	93.38 / 95.84	79.81 / 82.26
2	✓	-	96.86	96.32	84.13
3	-	✓	99.39 / 99.09	98.66 / 98.22	92.01 / 88.38
4	✓	✓	**99.65**	**99.03**	**93.07**

(异常强度对RealNet的影响。

Metric	$s = 0$	$s = 0.1$	$s = 0.2$	$s \in [0.1, 0.2]$
Image AUROC	99.35	**99.65**	99.61	**99.65**
Pixel AUROC	98.85	98.96	98.95	**99.03**
PRO	91.80	92.16	89.36	**93.07**

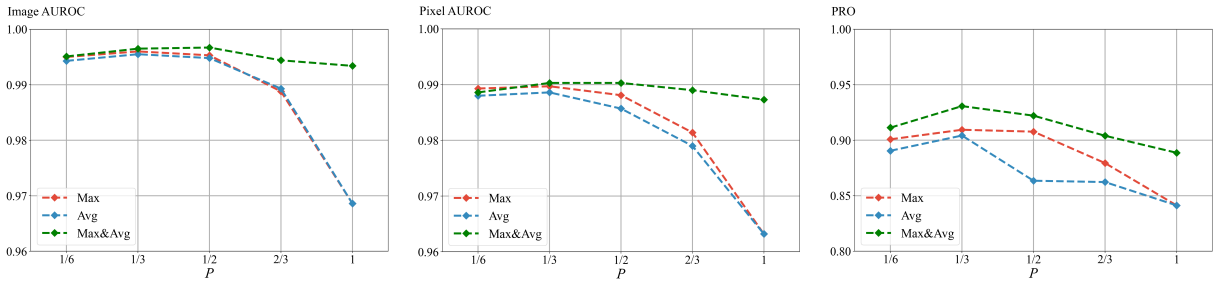


图5. RealNet在MVTec-AD数据集[3]上，采用不同重建残差选择模式（Max、Avg及Max&Avg）及不同重建残差保留比例{v*}时的性能表现。

降维（RLPR）[30, 40]。RDR从高维特征中随机选取部分维度特征，而RLPR则采用未经训练的线性变换层进行线性投影。我们分别报告了RealNet采用RDR和RLPR的结果，如Tab. 5a中的实验1和3所示。W/O RRS：我们将全局重建残差$E(A_n)$输入判别器以生成异常分数，结果如Tab. 5a中的实验1和2所示。

如Tab. 5a中的消融实验结果所示，RRS对性能提升有显著贡献。若使用所有重建残差生成异常分数，缺乏异常信息的重建残差可能导致异常区域漏检，从而导致异常检测性能显著下降。此外，与RDR和RLPR相比，AFS能获得更好的异常检测结果。关于AFS的直观可视化结果见附录D。

我们进一步研究了SDAS中异常强度s的影响，结果如表5b所示。当s等于0时，SDAS大概率在高概率密度区域生成正常图像。混合图像可能会引入误报的异常区域，这会降低重建难度并干扰判别器，导致性能欠佳。相反，当s过大时，合成异常图像会偏离真实异常图像的分布，导致RealNet的性能下降。我们的研究结果表明，在特定范围内均匀采样s是生成异常图像的稳健方法。这种方法使RealNet能够覆盖更广泛的异常模式，最终提升整体异常检测性能。

在图5中，我们报告了不同RRS模式的影响以及保留率对RealNet性能的影响。由于r的设置与m_k相关，我们引入了保留率P，其定义为：$P = \frac{r}{\sum^K m_k}$。与Max和Avg模式相比，Max&Avg模式在检测不同尺度异常时表现出更优的鲁棒性。在相同保留率下，Max&Avg模式比Max和Avg模式丢弃了更多缺乏异常信息的重建残差，从而减轻了性能下降，并进一步凸显了Max&Avg模式在增强RealNet异常检测能力方面的有效性。更多消融实验与分析可参见附录C。

5. 结论

在本工作中，我们提出了RealNet——一种创新的自监督异常检测框架。该方法整合了三个核心组件：强度可控的扩散异常合成（SDAS）、异常感知特征选择（AFS）以及重建残差选择（RRS）。这些组件协同作用于RealNet，使得在异常检测中能够有效利用大规模预训练模型，同时将计算开销保持在合理较低且可接受的范围内。RealNet为未来基于预训练特征重建技术的异常检测研究提供了灵活的基础。通过大量实验，我们展示了RealNet在处理多样化现实世界异常检测挑战方面的卓越能力。

致谢。本研究部分得到国家自然科学基金（项目编号：62177034 和 61972046）的资助。

References

[1] Samet Akçay, Amir Atapour-Abarghouei, and Toby P Breckon. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 2, 7

[2] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, pages 161–169. Springer, 2019. 2

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec-ad: A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 1, 2, 6, 7, 8, 13, 14, 15, 16, 17, 18, 19, 20

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020. 6

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 2, 6, 7, 12, 13, 15, 16

[6] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020. 7, 13

[7] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*, pages 475–489. Springer, 2021. 2, 7, 15

[8] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022. 1, 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 6, 12, 15

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 6

[11] Yuxuan Duan, Yan Hong, Li Niu, and Liqing Zhang. Few-shot defect image generation via defect-aware feature manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 571–578, 2023. 3, 16

[12] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 98–107, 2022. 7

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 13, 14, 15

[14] Lars Heckler, Rebecca König, and Paul Bergmann. Exploring the importance of pretrained feature extractors for unsupervised anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2916–2925, 2023. 2

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 15

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 3

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 12

[18] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pages 66–71. IEEE, 2021. 2, 6, 7, 17, 18, 19, 20

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[20] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021. 1, 2, 6, 7, 15, 16

[21] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, 2023. 6, 7

[22] Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus Robert Muller. Explainable deep one-class classification. In *International Conference on Learning Representations*, 2021. 2

[23] Fanbin Lu, Xufeng Yao, Chi-Wing Fu, and Jiaya Jia. Removing anomalies as noises for industrial defect localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16166–16175, 2023. 2

[24] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localiza-

参考文献

[1] Samet Akcay, Amir Atapour-Abarghouei, Toby P Breckon。Skip-GANomaly：用于异常检测的跳跃连接对抗训练编码器-解码器。发表于 *2019 International Joint Conference on Neural Networks (IJCNN)*，第1–8页。IEEE，2019年。2, 7[2] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, Nassir Navab。用于脑部MR图像无监督异常分割的深度自编码模型。发表于 *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*，第161–169页。Springer，2019年。2[3] Paul Bergmann, Michael Fauser, David Sattlegger, Carsten Steger。MVTec-AD：一个用于无监督异常检测的综合真实世界数据集。发表于 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*，第9592–9600页，2019年。1, 2, 6, 7, 8, 13, 14, 15, 16, 17, 18, 19, 20[4] Paul Bergmann, Michael Fauser, David Sattlegger, Carsten Steger。无先验知识的学生：基于判别性潜在嵌入的师生异常检测。发表于 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*，第4183–4192页，2020年。6[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, Andrea Vedaldi。描述自然场景中的纹理。发表于 *Proceedings of the IEEE conference on computer vision and pattern recognition*，第3606–3613页，2014年。2, 6, 7, 12, 13, 15, 16[6] Niv Cohen, Yedid Hoshen。基于深度金字塔对应关系的子图像异常检测。*arXiv preprint arXiv:2005.02357*，2020年。7, 13[7] Thomas Defard, Aleksandr Setkov, Angelique Loesch, Romaric Audigier。PaDiM：一种用于异常检测与定位的块分布建模框架。发表于 *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*，第475–489页。Springer，2021年。2, 7, 15[8] Hanqiu Deng, Xingyu Li。通过从单类嵌入进行反向蒸馏的异常检测。发表于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第9737–9746页，2022年。1, 2[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei。ImageNet：一个大规模分层图像数据库。发表于 *2009 IEEE conference on computer vision and pattern recognition*，第248–255页。IEEE，2009年。2, 6, 12, 15[10] Prafulla Dhariwal, Alexander Nichol。扩散模型在图像合成上击败GAN。*Advances in Neural Information Processing Systems*，34:8780–8794，2021年。6[11] Yuxuan Duan, Yan Hong, Li Niu, Liqing Zhang。通过缺陷感知特征操作的少样本缺陷图像生成。发表于 *Proceedings of the AAAI Conference on Artificial Intelligence*，第571–578页，2023年。3, 16[12] Denis Gudovskiy, Shun Ishizaka, Kazuki Kozuka。CFlow-AD：具有基于条件归一化流的定位。于*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*，第98–107页，2022年。7 [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, 与 Jian Sun。用于图像识别的深度残差学习。于*Proceedings of the IEEE conference on computer vision and pattern recognition*，第770–778页，2016年。4, 13, 14, 15 [14] Lars Heckler, Rebecca König, 与 Paul Bergmann。探索预训练特征提取器对于无监督异常检测与定位的重要性。于*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第2916–2925页，2023年。2 [15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, 与 Sepp Hochreiter。通过两时间尺度更新规则训练的GAN收敛至局部纳什均衡。*Advances in neural information processing systems*, 30, 2017年。15 [16] Jonathan Ho, Ajay Jain, 与 Pieter Abbeel。去噪扩散概率模型。*Advances in Neural InformationProcessing Systems*, 33:6840–6851, 2020年。1, 3 [17] Sergey Ioffe 与 Christian Szegedy。批量归一化：通过减少内部协变量偏移加速深度网络训练。于*International conference on machine learn-ing*，第448–456页。pmlr, 2015年。12 [18] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, 与 Milos Skotak。基于深度学习的金属零件缺陷检测：在复杂条件下评估现有方法。于*2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*，第66–71页。IEEE, 2021年。2, 6, 7, 17, 18, 19, 20 [19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, 与 Timo Aila。分析与改进StyleGAN的图像质量。于*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*，第8110–8119页，2020年。3 [20] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, 与 Tomas Pfister。CutPaste：用于异常检测与定位的自监督学习。于*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第9664–9674页，2021年。1, 2, 6, 7, 15, 16 [21] Zhikang Liu, Yiming Zhou, Yuansheng Xu, 与 Zilei Wang。Simplenet：用于图像异常检测与定位的简单网络。于*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第20402–20411页，2023年。6, 7 [22] Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Marius Kloft, 与 Klaus Robert Muller。可解释的深度单类分类。于*International Conference on Learning Representations*, 2021年。2 [23] Fanbin Lu, Xufeng Yao, Chi-Wing Fu, 与 Jiaya Jia。将异常作为噪声去除以进行工业缺陷定位。于*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*，第16166–16175页，2023年。2 [24] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, 与 Gian Luca Foresti。VT-ADL：一种用于图像异常检测与定位的视觉Transformer网络。

tion. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021. 2, 6, 7, 12, 13, 17, 18, 19, 20

[25] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 12

[26] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003. 2, 7

[27] Ken Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3):287–296, 1985. 4

[28] Jonathan Pirnay and Keng Chai. Inpainting transformer for anomaly detection. In *Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part II*, pages 394–406. Springer, 2022. 2

[29] Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised predictive convolutional attentive block for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13576–13586, 2022. 6

[30] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022. 2, 5, 6, 7, 8, 13

[31] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*, pages 146–157. Springer, 2017. 2

[32] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 474–489. Springer, 2022. 1, 2, 4, 6, 7, 15, 16

[33] Yong Shi, Jie Yang, and Zhiquan Qi. Unsupervised anomaly segmentation via deep feature reconstruction. *Neurocomputing*, 424:9–22, 2021. 1, 2, 5, 14

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 12, 13

[35] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 13, 14

[36] Ta-Wei Tang, Wei-Han Kuo, Jauh-Hsiang Lan, Chien-Fang Ding, Hakiem Hsu, and Hong-Tsu Young. Anomaly detection neural network with dual auto-encoders gan and its industrial inspection applications. *Sensors*, 20(12):3336, 2020. 7

[37] Xian Tao, Dapeng Zhang, Wenzhi Ma, Zhanxin Hou, Zhenfeng Lu, and Chandranath Adak. Unsupervised anomaly detection for surface defects with dual-siamese network. *IEEE Transactions on Industrial Informatics*, 18(11):7707–7717, 2022. 14

[38] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D Tr Nguyen, and Steven QH Truong. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24511–24520, 2023. 2, 5, 6, 13

[39] Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 650–656, 2022. 2

[40] Jiang Xi, Jianlin Liu, Jinbao Wang, Qiang Nie, WU Kai, Yong Liu, Chengjie Wang, and Feng Zheng. Softpatch: Unsupervised anomaly detection with noisy data. In *Advances in Neural Information Processing Systems*. 8

[41] Minghui Yang, Peng Wu, and Hui Feng. Memseg: A semi-supervised method for image surface defect detection using differences and commonalities. *Engineering Applications of Artificial Intelligence*, 119:105835, 2023. 4

[42] Xincheng Yao, Ruoqi Li, Jing Zhang, Jun Sun, and Chongyang Zhang. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24490–24499, 2023. 4

[43] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2, 13

[44] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 5, 6, 14, 15

[45] Sanyapong Youkachen, Miti Ruchanurucks, Teera Phatrapomnant, and Hirohiko Kaneko. Defect segmentation of hot-rolled steel strip surface by using convolutional auto-encoder and conventional image processing. In *2019 10th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*, pages 1–5. IEEE, 2019. 2

[46] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021. 2, 6, 7, 13

[47] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, 2016. 4, 6, 13, 14, 21

[48] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem: A discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021. 1, 2, 4, 7, 13, 15, 16

在 *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*，第01–06页。IEEE，2021年。2, 6, 7, 12, 13, 17, 18, 19, 20 [25] Alexander Quinn Nichol 和 Prafulla Dhariwal。改进的去噪扩散概率模型。在 *International Conference on Machine Learning*，第8162–8171页。PMLR，2021年。12 [26] Patrick Pérez、Michel Gangnet 和 Andrew Blake。泊松图像编辑。在 *ACM SIGGRAPH 2003 Papers*，第313–318页。2003年。2, 7 [27] Ken Perlin。一种图像合成器。*ACM Siggraph Computer Graphics*，19(3):287–296，1985年。4 [28] Jonathan Pirnay 和 Keng Chai。用于异常检测的修复变换器。在 *Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part II*，第394–406页。Springer，2022年。2 [29] Nicolae-Cătălin Ristea、Neelu Madan、Radu Tudor Ionescu、Kamal Nasrollahi、Fahad Shahbaz Khan、Thomas B Moes- lund 和 Mubarak Shah。用于异常检测的自监督预测卷积注意力块。在 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第13576–13586页，2022年。6 [30] Karsten Roth、Latha Pemula、Joaquin Zepeda、Bernhard Schölkopf、Thomas Brox 和 Peter Gehler。迈向工业异常检测的完全召回。在 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第14318–14328页，2022年。2, 5, 6, 7, 8, 13 [31] Thomas Schlegl、Philipp Seeböck、Sebastian M Waldstein、Ursula Schmidt-Erfurth 和 Georg Langs。使用生成对抗网络进行无监督异常检测以指导标记发现。在 *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*，第146–157页。Springer，2017年。2 [32] Hannah M Schlüter、Jeremy Tan、Benjamin Hou 和 Bern- hard Kainz。用于自监督异常检测和定位的自然合成异常。在 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*，第474–489页。Springer，2022年。1, 2, 4, 6, 7, 15, 16 [33] Yong Shi、Jie Yang 和 Zhiquan Qi。通过深度特征重建的无监督异常分割。*Neurocomput- ing*，424:9–22，2021年。1, 2, 5, 14 [34] Jiaming Song、Chenlin Meng 和 Stefano Ermon。去噪扩散隐式模型。在 *International Conference on Learning Representations*，2021年。12, 13 [35] Mingxing Tan 和 Quoc Le。EfficientNet：重新思考卷积神经网络的模型缩放。在 *International conference on machine learning*，第6105–6114页。PMLR，2019年。13, 14 [36] Ta-Wei Tang、Wei-Han Kuo、Jauh-Hsiang Lan、Chien-Fang Ding、Hakiem Hsu 和 Hong-Tsu Young。具有双自编码器GAN的异常检测神经网络及其工业检测应用。*Sensors*，20(12):3336，2020年。7 [37] 陶贤，张大鹏，马文志，侯占鑫，卢振峰，Chandranath Adak。基于双孪生网络的表面缺陷无监督异常检测。*IEEE Transactions on Industrial Informatics*，18(11):7707–7717，2022。14[38] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D Tr Nguyen, Steven QH Truong。重访用于异常检测的反向蒸馏。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第24511–24520页，2023。2, 5, 6, 13[39] Julian Wyatt, Adam Leach, Sebastian M Schmon, Chris G Willcocks。Anoddpm：使用单纯形噪声的降噪扩散概率模型进行异常检测。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第650–656页，2022。2[40] 郑玺，刘建林，王金宝，聂强，吴凯，刘勇，王成杰，郑锋。Softpatch：含噪声数据的无监督异常检测。收录于 *Advances in Neural Information Processing Systems*。8[41] 杨明辉，吴鹏，冯辉。Memseg：利用差异性和共性进行图像表面缺陷检测的半监督方法。*Engineering Applications of Artificial Intelligence*，119:105835，2023。4[42] 姚新成，李若琪，张静，孙俊，张崇阳。显式边界引导的半推-拉对比学习用于有监督异常检测。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第24490–24499页，2023。4[43] Jihun Yi, Sungroh Yoon。Patch svdd：用于异常检测与分割的块级支持向量数据描述。收录于 *Proceedings of the Asian Conference on Computer Vision*，2020。2, 13[44] 游志远，崔磊，沈宇骏，杨凯，卢昕，郑宇，乐心怡。一种统一的多类别异常检测模型。收录于 *Advances in Neural Information Pro- cessing Systems*，2022。1, 2, 5, 6, 14, 15[45] Sanyapong Youkachen, Miti Ruchanurucks, Teera Phatrapomnant, Hirohiko Kaneko。基于卷积自编码器和传统图像处理的热轧钢带表面缺陷分割。收录于 *2019 10th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*，第1–5页。IEEE，2019。2[46] 余家伟，郑晔，王翔，李伟，吴雨霜，赵瑞，吴立伟。Fastflow：基于二维归一化流的无监督异常检测与定位。*arXiv preprint arXiv:2111.07677*，2021。2, 6, 7, 13[47] Sergey Zagoruyko, Nikos Komodakis。宽残差网络。收录于 *Proceedings of the British Machine Vision Conference (BMVC)*，第87.1–87.12页。BMVA Press，2016。4, 6, 13, 14, 21[48] Vitjan Zavrtanik, Matej Kristan, Danijel Skočaj。Draem：一种判别性训练的重构嵌入用于表面异常检测。收录于 *Proceedings of the IEEE/CVF International Conference on Computer Vision*，第8330–8339页，2021。1, 2, 4, 7, 13, 15, 16

[49] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dsr: A dual subspace re-projection network for surface anomaly detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 539–554. Springer, 2022. 1, 2, 6

[50] Hui Zhang, Zuxuan Wu, Zheng Wang, Zhineng Chen, and Yu-Gang Jiang. Prototypical residual networks for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16281–16291, 2023. 1

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 15, 16

[52] Xinyi Zhang, Naiqi Li, Jiawei Li, Tao Dai, Yong Jiang, and Shu-Tao Xia. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6782–6791, 2023. 2, 6

[53] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3914–3923, 2023. 1, 2, 5, 6

[54] Ying Zhao. Omnial: A unified cnn framework for unsupervised anomaly localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3924–3933, 2023. 15

[55] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 392–408. Springer, 2022. 2, 6, 7, 12, 13, 17, 18, 19, 20

[49] Vitjan Zavrtanik、Matej Kristan 和 Danijel Skočaj。DSR：一种用于表面异常检测的双子空间重投影网络。收录于 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*，第 539–554 页。Springer，2022 年。1, 2, 6[50] Hui Zhang、Zuxuan Wu、Zheng Wang、Zhineng Chen 和 Yu-Gang Jiang。用于异常检测与定位的原型残差网络。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第 16281–16291 页，2023 年。1[51] Richard Zhang、Phillip Isola、Alexei A Efros、Eli Shechtman 和 Oliver Wang。深度特征作为感知度量的不合理有效性。收录于 *Proceedings of the IEEE conference on computer vision and pattern recogni- tion*，第 586–595 页，2018 年。15, 16[52] Xinyi Zhang、Naiqi Li、Jiawei Li、Tao Dai、Yong Jiang 和 Shu-Tao Xia。基于扩散概率模型的无监督表面异常检测。收录于 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*，第 6782–6791 页，2023 年。2, 6[53] Xuan Zhang、Shiyu Li、Xi Li、Ping Huang、Jiulong Shan 和 Ting Chen。DestSeg：用于异常检测的分割引导去噪师生模型。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第 3914–3923 页，2023 年。1, 2, 5, 6[54] Ying Zhao。OmniAL：一种用于无监督异常定位的统一 CNN 框架。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第 3924–3933 页，2023 年。15[55] Yang Zou、Jongheon Jeong、Latha Pemula、Dongqing Zhang 和 Onkar Dabeer。用于异常检测与分割的“找不同”自监督预训练。收录于 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*，第 392–408 页。Springer，2022 年。2, 6, 7, 12, 13, 17, 18, 19, 20

RealNet: A Feature Selection Network with Realistic Synthetic Anomaly for Anomaly Detection

Supplementary Material

A. Overview

We organize this supplementary material into the following sections: Appendix B provides additional implementation details for RealNet. Appendix C provides detailed results on the BTAD [24] and VisA [55] datasets, supplementary ablation study results, an analysis of RealNet's computational efficiency, anomaly detection results in multi-class setting, as well as synthetic anomaly image quality assessment results. Appendix D offers additional visualization results, including qualitative results of RealNet in anomaly localization, images generated by SDAS, and a straightforward visualization result of AFS. Appendix E discusses the limitations of our method.

B. More details

In SDAS, we use the learnable reverse diffusion variance [25] as $\Sigma_\theta(x_t, t)$, given by:

$$\Sigma_\theta(x_t, t) = \exp(v \log \beta_t + (1 - v) \log \tilde{\beta}_t) \tag{S1}$$

Here, β_t represents the variance of the diffusion process, while $\tilde{\beta}_t$ represents the variance of the conditional posterior distribution $q(x_{t-1}|x_t, x_0)$, and $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. The vector v is predicted by the model and weighted with β_t and $\tilde{\beta}_t$ in the log space. We optimize $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ with the loss $\mathcal{L}_{hybrid}$:

$$\mathcal{L}_{hybrid} = \mathcal{L}_{simple} + \gamma \mathcal{L}_{vlb} \tag{S2}$$

where

$$\begin{aligned} \mathcal{L}_{vlb} &= \mathcal{L}_0 + \mathcal{L}_1 + ... + \mathcal{L}_{T-1} + \mathcal{L}_T \\ \mathcal{L}_0 &= -\log p_\theta(x_0|x_1) \\ \mathcal{L}_{t-1} &= D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) \\ \mathcal{L}_T &= D_{KL}(q(x_T|x_0)||p(x_T)) \end{aligned} \tag{S3}$$

We set γ to 0.001 in Eq. (S2), and stop the gradient of $\mu_\theta(x_t, t)$ in $\mathcal{L}_{vlb}$ during the training phase. To accelerate the convergence of the diffusion model, we initialize it with weights pre-trained on ImageNet [9]. We set the reverse diffusion step T of 20, and generating 10,000 images at a resolution of 256×256 takes 6 hours using a single NVIDIA GeForce RTX 3090.

The SDAS with DDIM [34] is described in Algorithm S1, which provides three options for applying perturbation variance in the deterministic reverse diffusion process: $\Sigma = \beta_t$, $\Sigma = \tilde{\beta}_t$, and $\Sigma = \Sigma_\theta(x_t, t)$. Experimental observations show that the anomaly images obtained by IDDPM [25] are slightly better than those obtained by DDIM [34], and therefore, we use IDDPM [25] for SDAS. Some examples can be found in Appendix D.

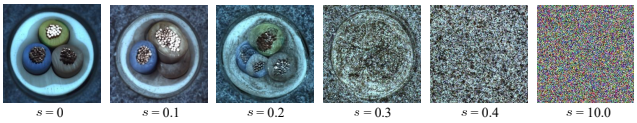


Figure S1. Sample anomaly images generated by SDAS with different anomaly strengths s.

Fig. S1 presents examples of images generated by SDAS with a broader range of anomaly strengths. As the anomaly strength increases, the generated anomalous images contain more noise, reducing their authenticity. In the experiments, we set the anomaly strength between 0.1 and 0.2, allowing SDAS to encompass a wider range of real-world anomalies.

In RRS, the global reconstruction residual $E(A_n)$ originates from distinct reconstruction networks, leading to disparate distributions across its dimensions. We apply a `BatchNorm` [17] layer (without Affine) to $E(A_n)$ and then perform reconstruction residuals selection to ensure a consistent distribution across the dimensions of $E(A_n)$.

The discriminator is implemented using a basic MLP with upsampling layers to map anomaly scores from feature resolution to image resolution. During the training phase of RealNet, we do not use any data augmentation for the synthesis of anomalous images, and maintain an equal ratio between normal images and synthetic anomalous images. In the process of image blending, we uniformly sample the opacity δ from 0.5 to 1.0 in Eq. (3). The training of RealNet is performed on a single NVIDIA GeForce RTX 3090, with an approximate average training time of 2 hours.

C. More results

C.1. Experimental results on BTAD

We evaluate the anomaly detection and localization performance of RealNet and alternative methods on the BTAD dataset [24], with the results shown in Tab. S1. Although SIA does not show a significant performance improvement compared to DTD [5] due to the absence of complex structural anomalies in the three industrial products of the BTAD dataset [24], RealNet demonstrates state-of-the-art performance in anomaly detection and localization when compared to other methods, without any structural or hyperparameter tuning.

RealNet：一种用于异常检测的具有真实合成异常的特征选择网络

补充材料

A. 概述

我们将本补充材料组织为以下部分：附录B提供了RealNet的额外实现细节。附录C详细展示了在BTAD[24]和VisA[55]数据集上的结果、补充消融研究结果、RealNet计算效率分析、多类别设置下的异常检测结果，以及合成异常图像质量评估结果。附录D提供了额外的可视化结果，包括RealNet在异常定位中的定性结果、SDAS生成的图像，以及AFS的直观可视化结果。附录E讨论了我们方法的局限性。

B. 更多细节

在SDAS中，我们采用可学习的反向扩散方差[25]作为$\Sigma_\theta(x_t,t)$，其表达式为：

$$\Sigma_\theta(x_t,t)=\exp(v\log\beta_t+(1-v)\log\tilde{\beta}_t) \tag{S1}$$

此处，β_t代表扩散过程的方差，而$\tilde{\beta}_t$代表条件后验分布$q(x_{t-1}|x_t,x_0)$的方差，以及$\tilde{\beta}_t=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$。向量$v$由模型预测，并在$log$空间中以$\beta_t$和$\tilde{\beta}_t$加权。我们通过损失函数$\mathcal{L}_{hybrid}$优化$\mu_\theta(x_t,t)$和$\Sigma_\theta(x_t,t)$：

$$\mathcal{L}_{hybrid}=\mathcal{L}_{simple}+\gamma\mathcal{L}_{vlb} \tag{S2}$$

其中

$$\begin{aligned}\mathcal{L}_{vlb}&=\mathcal{L}_0+\mathcal{L}_1+...+\mathcal{L}_{T-1}+\mathcal{L}_T\\ \mathcal{L}_0&=-\log p_\theta(x_0|x_1)\\ \mathcal{L}_{t-1}&=D_{KL}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t))\\ \mathcal{L}_T&=D_{KL}(q(x_T|x_0)||p(x_T))\end{aligned} \tag{S3}$$

在方程(S2)中，我们将γ设为0.001，并在训练阶段对$\mathcal{L}_{vlb}$中的$\mu_\theta(x_t,t)$进行梯度截断。为加速扩散模型的收敛，我们使用在ImageNet[9]上预训练的权重进行初始化。设置反向扩散步数T为20，使用单张NVIDIA GeForce RTX 3090生成10,000张分辨率为256×256的图像需耗时6小时。

SDAS与DDIM[34]在算法S1中描述，该算法为确定性反向扩散过程中的扰动方差应用提供了三种选项：$\Sigma=\beta_t$、$\Sigma=\tilde{\beta}_t$和$\Sigma=\Sigma_\theta(x_t,t)$。实验

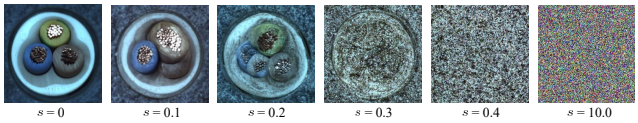


图 S1. 由SDAS生成的不同异常强度s下的样本异常图像。

观察表明，ID-DPM [25] 获得的异常图像略优于DDIM [34] 所获结果，因此我们在SDAS中采用IDDPM [25]。部分示例如附录D所示。

图S1展示了SDAS生成的具有更广泛异常强度范围的图像示例。随着异常强度的增加，生成的异常图像包含更多噪声，降低了其真实性。在实验中，我们将异常强度设置在0.1至0.2之间，使SDAS能够涵盖更广泛的现实世界异常情况。

在RRS中，全局重建残差$E(A_n)$源自不同的重建网络，导致其各维度分布存在差异。我们对$E(A_n)$应用BatchNorm[17]层（不含仿射变换），随后执行重建残差选择，以确保$E(A_n)$各维度的分布保持一致性。

判别器采用基础MLP配合上采样层实现，将异常分数从特征分辨率映射至图像分辨率。在RealNet的训练阶段，我们未使用任何数据增强技术来合成异常图像，并保持正常图像与合成异常图像的数量比例为1:1。在图像融合过程中，我们根据公式(3)从0.5到1.0范围内均匀采样不透明度参数δ。RealNet的训练在单张NVIDIA GeForce RTX 3090显卡上完成，平均训练时间约为2小时。

C. 更多结果

C.1. BTAD上的实验结果

我们在BTAD数据集[24]上评估了RealNet及其他替代方法的异常检测与定位性能，结果如表S1所示。尽管由于BTAD数据集[24]中的三种工业产品未包含复杂结构异常，SIA相比DTD[5]未展现出显著的性能提升，但RealNet在与其他方法比较时，无需任何结构或超参数调整，仍展现出最先进的异常检测与定位性能。

Algorithm S1 SDAS with DDIM [34]

Input: diffusion model $\epsilon_\theta(x_t, t)$, perturbation variance Σ, anomaly strength s
$x_T \sim \mathcal{N}(0, \mathbf{I})$
for all t from T to 1 **do**
$x_{t-1} \sim \mathcal{N}(\sqrt{\bar{\alpha}_{t-1}}(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(x_t, t), s\Sigma)$
end for
return x_0

Table S1. Comparison of RealNet with alternative anomaly detection methods on the BTAD dataset [24], employing Image AUROC (%) and Pixel AUROC (%) as evaluation metrics.

Category	VT-ADL [24]	P-SVDD [43]	FastFlow [46]	SPADE [6]	RD++ [38]	RealNet (SIA)	RealNet (DTD [5])
01	(-, **99**)	(95.7, 91.6)	(-, 95)	(91.4, 97.3)	(96.8, 96.2)	(**100.0**, 98.2)	(**100.0**, 98.1)
02	(-, 94)	(72.1, 93.6)	(-, 96)	(71.4, 94.4)	(**90.1**, **96.4**)	(88.6, 96.3)	(87.5, 96.3)
03	(-, 77)	(82.1, 91.0)	(-, 99)	(99.9, 99.1)	(**100.0**, **99.7**)	(99.6, 99.4)	(99.4, 99.6)
AVG	(-, 90.0)	(83.3, 92.1)	(-, 96.7)	(87.6, 96.9)	(95.6,97.4)	(**96.1**, 97.9)	(95.7, **98.0**)

Table S2. Comparison of RealNet with alternative anomaly detection methods on the VisA dataset [55], employing Image AUROC (%) and Pixel AUROC (%) as evaluation metrics.

Category	SPADE [6]	FastFlow [46]	DRAEM [48]	PatchCore [30]	RealNet (SIA)	RealNet (DTD [5])
Candle	(91.0, 97.9)	(92.8, 94.9)	(91.8, 96.6)	(**98.6**, **99.5**)	(96.1, 99.1)	(95.0, 99.0)
Capsules	(61.4, 60.7)	(71.2, 75.3)	(74.7, 98.5)	(81.6, **99.5**)	(**93.2**, 98.7)	(88.1, 97.6)
Cashew	(**97.8**, 86.4)	(91.0, 91.4)	(95.1, 83.5)	(97.3, **98.9**)	(**97.8**, 98.3)	(95.9, 97.6)
Chewing gum	(85.8, 98.6)	(91.4, 98.6)	(94.8, 96.8)	(99.1, 99.1)	(99.9, **99.8**)	(**100.0**, **99.8**)
Fryum	(88.6, 96.7)	(88.6, **97.3**)	(**97.4**, 87.2)	(96.2, 93.8)	(97.1, 96.2)	(95.3, 95.2)
Macaroni1	(95.2, 96.2)	(98.3, 97.3)	(97.2, **99.9**)	(97.5, 99.8)	(**99.8**, **99.9**)	(98.2, 99.7)
Macaroni2	(87.9, 87.5)	(86.3, 89.2)	(85.0, 99.2)	(78.1, 99.1)	(**95.2**, **99.6**)	(91.8, 99.3)
PCB1	(72.1, 66.9)	(77.4, 75.2)	(47.6, 88.7)	(**98.5**, **99.9**)	(**98.5**, 99.7)	(97.1, 99.4)
PCB2	(50.7, 71.1)	(61.9, 67.3)	(89.8, 91.3)	(97.3, **99.0**)	(**97.6**, 98.0)	(97.5, 97.8)
PCB3	(90.5, 95.1)	(74.3, 94.8)	(92.0, 98.0)	(97.9, **99.2**)	(**99.1**, 98.8)	(97.6, 98.4)
PCB4	(83.1, 89.0)	(80.9, 89.9)	(98.6, 96.8)	(99.6, **98.6**)	(**99.7**, **98.6**)	(99.2, **98.6**)
Pipe fryum	(81.1, 81.8)	(72.0, 87.3)	(**100.0**, 85.8)	(99.8, 99.1)	(99.9, **99.2**)	(99.9, 98.6)
AVG	(82.1, 85.6)	(82.2, 88.2)	(88.7, 93.5)	(95.1, **98.8**)	(**97.8**, **98.8**)	(96.3, 98.4)

C.2. Experimental results on VisA

We present the performance of RealNet and alternative methods on the VisA dataset under the one-class protocol [55] in Tab. S2. RealNet achieves the best performance in both anomaly detection and localization. Compared to DTD [5], the RealNet trained using SIA shows an improvement of 1.5% in Image AUROC and 0.4% in Pixel AUROC.

C.3. Supplementary ablation studies

To further investigate RealNet's anomaly detection performance on the MVTec-AD dataset [3], we examine various backbones and reconstruction feature dimension settings. As shown in Tab. S3, when WideResNet50 [47] is employed as the backbone and the reconstruction feature dimensions $\{m_1, ..., m_K\}$ are reduced from {256, 512, 512, 256} to {128, 256, 256, 128}, there is a slight decrease of 0.16% in Image AUROC. Despite this reduction, RealNet maintains its competitive performance compared to other methods. Additionally, the adoption of EfficientNetB4 [35] and ResNet34 [13] as backbones also results in competitive performance, demonstrating the effectiveness of RealNet across various settings.

C.4. Computational efficiency analysis

We investigate the computational efficiency and detection performance of three different multi-scale feature reconstruction architectures on the MVTec-AD dataset [3], as illustrated in Fig. S2. To provide a comprehensive analysis, Tab. S4 presents the inference speed, model size (including backbone), and anomaly detection performance of these architectures. The inference is performed using a single Nvidia GeForce RTX 3090, with all other settings adhering to the specifications detailed in Sec. 4.1.

We utilize a consistent reconstruction network based on the U-Net model with skip connections across three distinct architectures. The employed U-Net model initiates with a

Algorithm S1 SDAS with DDIM [34]

Input: diffusion model $\epsilon_\theta(x_t, t)$, perturbation variance Σ, anomaly strength s
$x_T \sim \mathcal{N}(0, \mathbf{I})$
for all t from T to 1 **do**
$x_{t-1} \sim \mathcal{N}(\sqrt{\bar{\alpha}_{t-1}}(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(x_t, t), s\Sigma)$
end for
return x_0

表S1. 在BTAD数据集[24]上，使用图像AUROC（%）和像素AUROC（%）作为评估指标，RealNet与替代异常检测方法的比较。

Category	VT-ADL [24]	P-SVDD [43]	FastFlow [46]	SPADE [6]	RD++ [38]	RealNet (SIA)	RealNet (DTD [5])
01	(-, **99**)	(95.7, 91.6)	(-, 95)	(91.4, 97.3)	(96.8, 96.2)	(**100.0**, 98.2)	(**100.0**, 98.1)
02	(-, 94)	(72.1, 93.6)	(-, 96)	(71.4, 94.4)	(**90.1**, **96.4**)	(88.6, 96.3)	(87.5, 96.3)
03	(-, 77)	(82.1, 91.0)	(-, 99)	(99.9, 99.1)	(**100.0**, **99.7**)	(99.6, 99.4)	(99.4, 99.6)
AVG	(-, 90.0)	(83.3, 92.1)	(-, 96.7)	(87.6, 96.9)	(95.6,97.4)	(**96.1**, 97.9)	(95.7, **98.0**)

表S2。在VisA数据集[55]上，使用图像AUROC（%）和像素AUROC（%）作为评估指标，比较RealNet与替代异常检测方法的表现。

Category	SPADE [6]	FastFlow [46]	DRAEM [48]	PatchCore [30]	RealNet (SIA)	RealNet (DTD [5])
Candle	(91.0, 97.9)	(92.8, 94.9)	(91.8, 96.6)	(**98.6**, **99.5**)	(96.1, 99.1)	(95.0, 99.0)
Capsules	(61.4, 60.7)	(71.2, 75.3)	(74.7, 98.5)	(81.6, **99.5**)	(**93.2**, 98.7)	(88.1, 97.6)
Cashew	(**97.8**, 86.4)	(91.0, 91.4)	(95.1, 83.5)	(97.3, **98.9**)	(**97.8**, 98.3)	(95.9, 97.6)
Chewing gum	(85.8, 98.6)	(91.4, 98.6)	(94.8, 96.8)	(99.1, 99.1)	(99.9, **99.8**)	(**100.0**, **99.8**)
Fryum	(88.6, 96.7)	(88.6, **97.3**)	(**97.4**, 87.2)	(96.2, 93.8)	(97.1, 96.2)	(95.3, 95.2)
Macaroni1	(95.2, 96.2)	(98.3, 97.3)	(97.2, **99.9**)	(97.5, 99.8)	(**99.8**, **99.9**)	(98.2, 99.7)
Macaroni2	(87.9, 87.5)	(86.3, 89.2)	(85.0, 99.2)	(78.1, 99.1)	(**95.2**, **99.6**)	(91.8, 99.3)
PCB1	(72.1, 66.9)	(77.4, 75.2)	(47.6, 88.7)	(**98.5**, **99.9**)	(**98.5**, 99.7)	(97.1, 99.4)
PCB2	(50.7, 71.1)	(61.9, 67.3)	(89.8, 91.3)	(97.3, **99.0**)	(**97.6**, 98.0)	(97.5, 97.8)
PCB3	(90.5, 95.1)	(74.3, 94.8)	(92.0, 98.0)	(97.9, **99.2**)	(**99.1**, 98.8)	(97.6, 98.4)
PCB4	(83.1, 89.0)	(80.9, 89.9)	(98.6, 96.8)	(99.6, **98.6**)	(**99.7**, **98.6**)	(99.2, **98.6**)
Pipe fryum	(81.1, 81.8)	(72.0, 87.3)	(**100.0**, 85.8)	(99.8, 99.1)	(99.9, **99.2**)	(99.9, 98.6)
AVG	(82.1, 85.6)	(82.2, 88.2)	(88.7, 93.5)	(95.1, **98.8**)	(**97.8**, **98.8**)	(96.3, 98.4)

C.2. VisA 上的实验结果

我们在Tab. S2中展示了RealNet及其他方法在VisA数据集上采用单类协议[55]的性能表现。RealNet在异常检测和定位方面均取得最佳性能。与DTD[5]相比，采用SIA训练的RealNet在图像AUROC上提升了1.5%，在像素AUROC上提升了0.4%。

C.3. 补充消融研究

为了进一步研究RealNet在MVTec-AD数据集[3]上的异常检测性能，我们考察了不同骨干网络与重建特征维度设置。如表S3所示，当采用WideResNet50[47]作为骨干网络，并将重建特征维度$\{m_1, ..., m_K\}$从{256, 512, 512, 256}降低至{128, 256, 256, 128}时，图像AUROC指标轻微下降了0.16%。尽管存在这一降幅，RealNet相较于其他方法仍保持着竞争优势。

方法。此外，采用EfficientNetB4 [35]和ResNet34 [13]作为主干网络也取得了有竞争力的性能，这证明了Real-Net在各种设置下的有效性。

C.4. 计算效率分析

我们在MVTec-AD数据集[3]上研究了三种不同多尺度特征重建架构的计算效率与检测性能，如图S2所示。为提供全面分析，表S4展示了这些架构的推理速度、模型大小（包含主干网络）及异常检测性能。推理过程使用单张Nvidia GeForce RTX 3090完成，其余设置均遵循第4.1节详述的规范。

我们采用基于U-Net模型的一致性重建网络，该网络在三种不同架构中均包含跳跃连接。所使用的U-Net模型初始结构为

Table S3. Performance evaluation of RealNet with varying backbones and reconstruction feature dimension settings on the MVTec-AD dataset [3], employing Image AUROC (%), Pixel AUROC (%), and PRO (%) as evaluation metrics.

Backbone	EfficientNetB4 [35]	ResNet34 [13]	WideResNet50 [47]	
$\{m_1,...,m_K\}$	{24, 32, 56, 160}	{64, 128, 256, 128}	{128, 256, 256, 128}	{256, 512, 512, 256}
Bottle	(**100.0**, 98.83, **95.96**)	(**100.0**, 98.56, 95.91)	(**100.0**, **99.41**, 94.37)	(**100.0**, 99.30, 95.62)
Cable	(96.36, 96.33, 88.61)	(96.31, 96.32, 88.68)	(98.35, 98.01, 92.99)	(**99.19**, **98.10**, **93.38**)
Capsule	(97.97, 99.16, **91.46**)	(96.81, 98.78, 87.87)	(99.44, **99.39**, 79.76)	(**99.56**, 99.32, 84.48)
Carpet	(**100.0**, 98.27, 96.35)	(99.76, 98.37, 94.45)	(99.80, 98.91, 96.32)	(99.84, **99.19**, **96.41**)
Grid	(99.92, 99.31, 97.35)	(**100.0**, 99.26, **97.39**)	(**100.0**, **99.55**, 96.38)	(**100.0**, 99.51, 97.28)
Hazelnut	(99.89, 98.45, **94.98**)	(99.93, 99.35, 94.36)	(**100.0**, 99.67, 93.06)	(**100.0**, **99.68**, 93.14)
Leather	(**100.0**, 99.34, 97.75)	(99.97, 99.40, **98.28**)	(**100.0**, **99.81**, 96.99)	(**100.0**, 99.76, 96.22)
Metal Nut	(99.07, 96.90, 92.65)	(99.17, 96.68, 93.34)	(**99.90**, **98.75**, **95.10**)	(99.76, 98.58, 94.39)
Pill	(96.10, 94.86, 86.60)	(97.55, 98.23, **93.17**)	(97.85, **99.19**, 80.73)	(**99.13**, 99.02, 91.04)
Screw	(92.95, 99.05, **92.68**)	(96.99, 99.09, 89.57)	(97.99, 99.28, 88.60)	(**98.83**, **99.45**, 87.90)
Tile	(99.49, 95.69, 92.10)	(99.93, 97.40, 91.65)	(**100.0**, 99.27, 97.20)	(99.96, **99.44**, **97.70**)
Toothbrush	(99.44, 98.90, **92.39**)	(**100.0**, 98.26, 91.74)	(**100.0**, **99.26**, 91.22)	(99.44, 98.71, 91.57)
Transistor	(99.58, **98.57**, **93.63**)	(99.33, 97.70, 88.53)	(99.79, 98.26, 83.34)	(**100.0**, 98.00, 92.92)
Wood	(98.77, 94.47,**92.67**)	(98.16, 96.35, 91.46)	(**99.56**, **98.22**, 90.76)	(99.21, **98.22**, 90.54)
Zipper	(99.71, 98.01, 91.68)	(**99.90**, 98.55, **93.91**)	(99.74, **99.20**, 90.73)	(99.82, 99.17, 93.43)
AVG	(98.62, 97.74, **93.12**)	(98.92, 98.15, 92.69)	(99.49, **99.07**, 91.17)	(**99.65**, 99.03, 93.07)

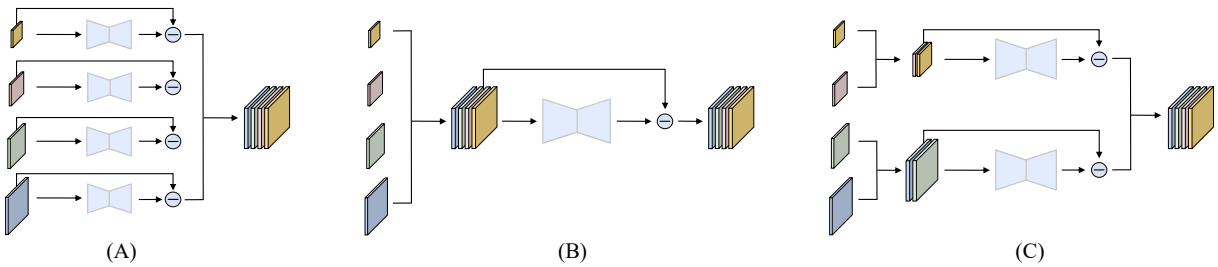


Figure S2. Various architectures of multi-scale feature reconstruction for anomaly detection. (A) Independent Reconstruction Architecture uses separate networks for multi-scale feature reconstruction. (B) Fully Aligned Feature Reconstruction Architecture aligns all features for reconstruction. (C) Neighboring Aligned Feature Reconstruction Architecture aligns and reconstructs neighboring resolution features.

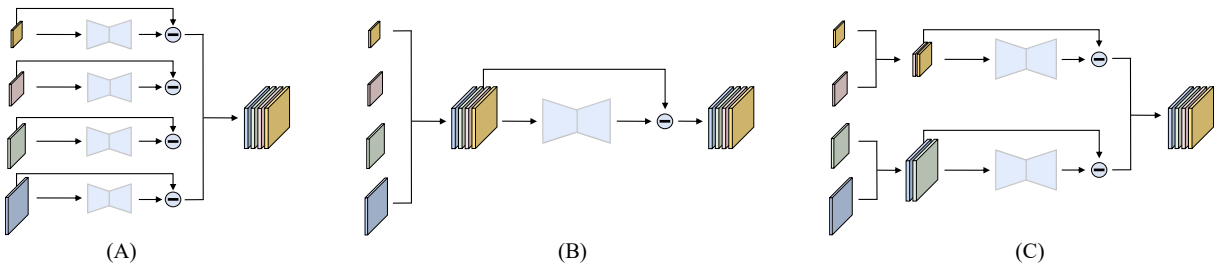
stack of residual layers and down-sampling layers, gradually decreasing the spatial dimensions while increasing the number of channels. Subsequently, the model utilizes a stack of residual layers and up-sampling layers to inversely reconstruct features. Throughout this process, skip connections are incorporated at equivalent spatial resolutions to ensure a smooth and logical flow.

Specifically, architecture **A** adopts separate reconstruction networks to reconstruct multi-scale features without the need for feature interpolation or alignment. This method ensures outstanding anomaly detection performance while maintaining high computational efficiency. With a resolution of 256×256 and reconstruction feature dimensions of $\{256, 512, 512, 256\}$, architecture **A** with model size of 2.2 GB achieves a rapid inference speed of 31.93 FPS. And it can perform inference using only 4GB of GPU memory. Concurrently, it attains an Image AUROC of 99.65% and a Pixel AUROC of 99.03%. By decreasing the reconstruction feature dimensions to $\{128, 256, 256, 128\}$, architecture **A** reduces the model size to 0.74 GB and achieves a higher inference speed of 40.42 FPS, while preserving an Image AUROC of 99.49% and a Pixel AUROC of 99.07%. Furthermore, at a high resolution of 512×512, it delivers an inference speed of 13.53 FPS, along with an Image AUROC of 99.40% and a Pixel AUROC of 98.71%. These inference speeds indicate that architecture **A** satisfies the real-time requirements for industrial inspection applications.

Regarding architecture **B**, as referenced in [33, 37, 44], it is used to align the multi-scale features of a small pre-trained network. As aligning down-sampled features will reduce the resolution of model detection and cause predictable performance loss, the experiment only discusses

表S3。在MVTec-AD数据集[3]上，采用不同骨干网络和重建特征维度设置的RealNet性能评估，使用图像AUROC（%）、像素AUROC（%）和PRO（%）作为评估指标。

Backbone	EfficientNetB4 [35]	ResNet34 [13]	WideResNet50 [47]	
$\{m_1,...,m_K\}$	{24, 32, 56, 160}	{64, 128, 256, 128}	{128, 256, 256, 128}	{256, 512, 512, 256}
Bottle	(**100.0**, 98.83, **95.96**)	(**100.0**, 98.56, 95.91)	(**100.0**, **99.41**, 94.37)	(**100.0**, 99.30, 95.62)
Cable	(96.36, 96.33, 88.61)	(96.31, 96.32, 88.68)	(98.35, 98.01, 92.99)	(**99.19**, **98.10**, **93.38**)
Capsule	(97.97, 99.16, **91.46**)	(96.81, 98.78, 87.87)	(99.44, **99.39**, 79.76)	(**99.56**, 99.32, 84.48)
Carpet	(**100.0**, 98.27, 96.35)	(99.76, 98.37, 94.45)	(99.80, 98.91, 96.32)	(99.84, **99.19**, **96.41**)
Grid	(99.92, 99.31, 97.35)	(**100.0**, 99.26, **97.39**)	(**100.0**, **99.55**, 96.38)	(**100.0**, 99.51, 97.28)
Hazelnut	(99.89, 98.45, **94.98**)	(99.93, 99.35, 94.36)	(**100.0**, 99.67, 93.06)	(**100.0**, **99.68**, 93.14)
Leather	(**100.0**, 99.34, 97.75)	(99.97, 99.40, **98.28**)	(**100.0**, **99.81**, 96.99)	(**100.0**, 99.76, 96.22)
Metal Nut	(99.07, 96.90, 92.65)	(99.17, 96.68, 93.34)	(**99.90**, **98.75**, **95.10**)	(99.76, 98.58, 94.39)
Pill	(96.10, 94.86, 86.60)	(97.55, 98.23, **93.17**)	(97.85, **99.19**, 80.73)	(**99.13**, 99.02, 91.04)
Screw	(92.95, 99.05, **92.68**)	(96.99, 99.09, 89.57)	(97.99, 99.28, 88.60)	(**98.83**, **99.45**, 87.90)
Tile	(99.49, 95.69, 92.10)	(99.93, 97.40, 91.65)	(**100.0**, 99.27, 97.20)	(99.96, **99.44**, **97.70**)
Toothbrush	(99.44, 98.90, **92.39**)	(**100.0**, 98.26, 91.74)	(**100.0**, **99.26**, 91.22)	(99.44, 98.71, 91.57)
Transistor	(99.58, **98.57**, **93.63**)	(99.33, 97.70, 88.53)	(99.79, 98.26, 83.34)	(**100.0**, 98.00, 92.92)
Wood	(98.77, 94.47,**92.67**)	(98.16, 96.35, 91.46)	(**99.56**, **98.22**, 90.76)	(99.21, **98.22**, 90.54)
Zipper	(99.71, 98.01, 91.68)	(**99.90**, 98.55, **93.91**)	(99.74, **99.20**, 90.73)	(99.82, 99.17, 93.43)
AVG	(98.62, 97.74, **93.12**)	(98.92, 98.15, 92.69)	(99.49, **99.07**, 91.17)	(**99.65**, 99.03, 93.07)



图S2. 用于异常检测的多尺度特征重建的不同架构。(A) 独立重建架构使用独立的网络进行多尺度特征重建。(B) 完全对齐特征重建架构将所有特征对齐以进行重建。(C) 相邻对齐特征重建架构对齐并重建相邻分辨率特征。

残差层和下采样层的堆叠，逐渐减小空间维度同时增加通道数。随后，模型利用残差层和上采样层的堆叠进行反向特征重建。在此过程中，在等效空间分辨率处引入了跳跃连接，以确保流程的平滑与合理。

具体而言，架构A采用独立的重建网络来重建多尺度特征，无需进行特征插值或对齐。该方法在保持高计算效率的同时，确保了出色的异常检测性能。在分辨率为256×256、重建特征维度为{256, 512, 512, 256}的条件下，模型大小为2.2GB的架构A实现了31.93 FPS的快速推理速度，且仅需4GB GPU显存即可完成推理。

同时，它实现了99.65%的图像AUROC和99.03%的像素AUROC。通过将重建特征维度降低至{128、256、256、128}，架构A将模型大小缩减至0.74 GB，并将推理速度提升至40.42 FPS，同时保持了99.49%的图像AUROC和99.07%的像素AUROC。此外，在512×512的高分辨率下，其推理速度达到13.53 FPS，并取得99.40%的图像AUROC和98.71%的像素AUROC。这些推理速度表明，架构A满足了工业检测应用对实时性的要求。

关于架构B，如[33, 37, 44]中所述，它被用于对齐小型预训练网络的多尺度特征。由于对齐下采样特征会降低模型检测的分辨率并导致可预见的性能损失，实验仅讨论了

Table S4. Performance evaluation of various reconstruction architectures on the MVTec-AD dataset [3]. The metrics include Image AUROC (%), Pixel AUROC (%), and PRO (%).

	Speed (FPS) ↑	Model Size (GB) ↓	Metrics ↑
$\{m_1,...,m_K\}$ is {128, 256, 256, 128} and image size is 256×256			
A	**40.42**	**0.74**	(99.49, **99.07**, 91.17)
$\{m_1,...,m_K\}$ is {256, 512, 512, 256} and image size is 256×256			
A	31.93	2.20	(**99.65**, 99.03, 93.07)
B	10.83	7.22	(98.44, 98.17, 94.27)
C	22.39	3.75	(99.62, 98.90, **94.71**)
$\{m_1,...,m_K\}$ is {256, 512, 512, 256} and image size is 512×512			
A	13.53	2.20	(99.40, 98.71, 94.01)

up-sampling alignment. Compared to architecture **A**, architecture **B** reconstructs the interpolated features, significantly reducing computational efficiency and increasing model size. Moreover, due to the limited number of normal images, the overly large reconstruction network in architecture **B** is prone to overfitting, resulting in reduced detection performance. Consequently, for large-scale pre-trained networks with high-dimensional features, aligning and reconstructing all features is suboptimal.

Moreover, we observe that utilizing multiple reconstruction networks for feature reconstruction in architecture **A** causes minor deviations in localizing small-area anomalies, resulting in a reduced PRO. To address this, we propose architecture **C**, which aligns and reconstructs features from two neighboring resolution, thereby reducing the number of reconstruction networks, controlling the model size, and striking a balance between computational efficiency and localization accuracy. At a 256×256 resolution, with reconstruction feature dimensions of {256, 512, 512, 256}, architecture **C** has a 3.75 GB model size and achieves an inference speed of 22.39 FPS, while attaining an Image AUROC of 99.62%, a Pixel AUROC of 98.90%, and a PRO of 94.71%.

In summary, the design of RealNet balances both anomaly detection performance and computational efficiency. The introduction of AFS allows us to flexibly customize models of various sizes to accommodate different usage scenarios. Furthermore, among our three key innovations, both AFS and RRS introduce no additional learnable parameters, ensuring strong interpretability. As for SDAS, it only introduces perturbation during the reverse diffusion process, without requiring any prior knowledge about the distribution of real anomaly images.

C.5. Anomaly detection in multi-class setting

In the multi-class setting [44, 54], anomaly detection is performed across multiple target classes concurrently, without access to sample class labels during both training and inference phases. Learning the data distributions of multiple classes jointly makes the reconstruction more complex.

Table S5. Comparison of RealNet with alternative methods in multi-class anomaly detection on the MVTec-AD dataset [3].

Methods	Image AUROC	Pixel AUROC
DRAEM [48]	88.1	87.2
PaDiM [7]	84.2	89.5
UniAD [44]	96.5	96.8
OmniAL [54]	97.2	98.3
RealNet	**97.3**	**98.4**

Table S6. Image quality comparison of SIA with alternative anomaly synthesis approaches on the MVTec-AD dataset [3].

Methods	FID [15] ↓	LPIPS [51] ↑
DTD [5]	120.52±0.63	0.16±0.00
CutPaste [20]	77.34±0.09	0.11±0.00
NSA [32]	68.76±0.16	0.09±0.01
SIA	**60.39±1.26**	**0.18±0.01**

In such settings, previous reconstruction methods tend to output copies of the input images instead of performing selective reconstruction, which leads to a significant decrease in performance. We evaluate the performance of RealNet in multi-class anomaly detection on the MVTec-AD dataset [3] and compare it with alternative state-of-the-art methods. We use DTD [5] for anomaly synthesis as class labels are unavailable during training. The remaining settings are consistent with Sec. 4.1.

The results are shown in Tab. S5. When detecting anomalies across 15 categories of the MVTec-AD dataset [3] concurrently, RealNet achieves an Image AUROC of 97.3% and a Pixel AUROC of 98.4% using a ResNet50 [13] pre-trained on ImageNet [9], surpassing state-of-the-art multi-class anomaly detection methods [44, 54]. To ensure that normal regions can be reconstructed correctly, we do not explicitly constrain the generalization ability of the reconstructed network in RealNet. Instead, we implicitly constrain the reconstruction network to ensure that anomalous regions can be correctly detected by discarding a part of the reconstruction residuals.

C.6. Synthetic anomaly image quality assessment

In this section, we evaluate the quality of anomaly images generated by various anomaly synthesis methods on the MVTec-AD dataset [3]. Specifically, we use the following evaluation metrics:

- FID (Fréchet Inception Distance) [15]: FID measures the distance between the distribution of synthetic anomaly images and real anomaly images, evaluating both the realism and diversity of the synthetic anomaly images. A lower value indicates better performance.

表 S4. 在 MVTec-AD 数据集 [3] 上对各种重建架构的性能评估。评估指标包括图像 AUROC (%)、像素 AUROC (%) 和 PRO (%)。

	Speed (FPS) ↑	Model Size (GB) ↓	Metrics ↑
$\{m_1,...,m_K\}$ is {128, 256, 256, 128} and image size is 256×256			
A	**40.42**	**0.74**	(99.49, **99.07**, 91.17)
$\{m_1,...,m_K\}$ is {256, 512, 512, 256} and image size is 256×256			
A	31.93	2.20	(**99.65**, 99.03, 93.07)
B	10.83	7.22	(98.44, 98.17, 94.27)
C	22.39	3.75	(99.62, 98.90, **94.71**)
$\{m_1,...,m_K\}$ is {256, 512, 512, 256} and image size is 512×512			
A	13.53	2.20	(99.40, 98.71, 94.01)

上采样对齐。与架构A相比，架构B重建了插值特征，显著降低了计算效率并增加了模型规模。此外，由于正常图像数量有限，架构B中过大的重建网络容易过拟合，导致检测性能下降。因此，对于具有高维特征的大规模预训练网络，对齐并重建所有特征并非最优选择。

此外，我们观察到，在架构A中使用多个重建网络进行特征重建会导致小面积异常定位出现微小偏差，从而降低PRO。为解决这一问题，我们提出了架构C，该架构对齐并重建来自两个相邻分辨率的特征，从而减少了重建网络的数量，控制了模型规模，并在计算效率与定位精度之间取得了平衡。在256 × 256分辨率下，重建特征维度为{256、512、512、256}，架构C的模型大小为3.75 GB，推理速度达到22.39 FPS，同时实现了99.62%的图像AUROC、98.90%的像素AUROC以及94.71%的PRO。

总而言之，RealNet的设计在异常检测性能和计算效率之间取得了平衡。AFS的引入使我们能够灵活定制不同尺寸的模型，以适应多样化的使用场景。此外，在我们的三项关键创新中，AFS和RRS均未引入额外的可学习参数，确保了强大的可解释性。至于SDAS，它仅在反向扩散过程中引入扰动，无需任何关于真实异常图像分布的先验知识。

C.5. 多类别设置中的异常检测

在多类别设置[44, 54]中，异常检测同时针对多个目标类别进行，在训练和推理阶段均无法获取样本类别标签。联合学习多个类别的数据分布使得重构过程更为复杂。

表 S5. 在 MVTec-AD 数据集 [3] 上多类别异常检测中 RealNet 与替代方法的比较。

Methods	Image AUROC	Pixel AUROC
DRAEM [48]	88.1	87.2
PaDiM [7]	84.2	89.5
UniAD [44]	96.5	96.8
OmniAL [54]	97.2	98.3
RealNet	**97.3**	**98.4**

表S6。在MVTec-AD数据集[3]上，SIA与替代异常合成方法的图像质量比较。

Methods	FID [15] ↓	LPIPS [51] ↑
DTD [5]	120.52±0.63	0.16±0.00
CutPaste [20]	77.34±0.09	0.11±0.00
NSA [32]	68.76±0.16	0.09±0.01
SIA	**60.39±1.26**	**0.18±0.01**

在此类设定下，先前的重建方法倾向于输出输入图像的复制品，而非执行选择性重建，这导致性能显著下降。我们在MVTec-AD数据集[3]上评估RealNet在多类别异常检测中的性能，并将其与当前最先进的替代方法进行比较。由于训练期间缺乏类别标签，我们使用DTD[5]进行异常合成。其余设定与第4.1节保持一致。

结果如表S5所示。在同时检测MVTec-AD数据集[3]的15个类别异常时，RealNet使用ImageNet[9]预训练的ResNet50[13]实现了97.3%的图像AUROC和98.4%的像素AUROC，超越了当前最先进的多类别异常检测方法[44, 54]。为确保正常区域能够被正确重建，我们在RealNet中并未显式约束重建网络的泛化能力，而是通过隐式约束重建网络，使其通过舍弃部分重建残差来确保异常区域能被准确检测。

C.6. 合成异常图像质量评估

在本节中，我们评估了在MVTec-AD数据集[3]上各种异常合成方法生成的异常图像质量。具体而言，我们采用以下评估指标：

- FID（Fréchet Inception Distance）[15]：FID通过衡量合成异常图像与真实异常图像分布之间的距离，评估合成异常图像的真实性和多样性。数值越低，表示性能越好。

- LPIPS (Learned Perceptual Image Patch Similarity) [51]: We employ cluster-based LPIPS [11] to evaluate the diversity of synthetic anomaly images. Supposing a category contains N real anomaly images, we partition the synthesized anomaly images into N groups by finding the lowest LPIPS, then we compute the mean pairwise LPIPS within each group and compute the average of all groups. A higher cluster LPIPS indicates greater diversity.

We employ various anomaly synthesis methods to generate 1,000 anomaly images for evaluation, with each method independently assessed three times. The experimental results are shown in Tab. S6. In comparison to other anomaly synthesis methods, SIA achieves the best FID and LPIPS metrics, highlighting the outstanding performance of SDAS in generating both realistic and diverse anomaly images, and demonstrating the effectiveness of SDAS in improving anomaly detection performance.

D. Visualization

We conduct a comprehensive visual analysis of RealNet on the four datasets. Fig. S3 shows the qualitative results of RealNet in anomaly localization, showcasing its outstanding performance in pixel-level anomaly localization. Figs. S4 and S5 display the anomaly images and normal images generated by SDAS, respectively. Fig. S6 illustrates images synthesized using SIA with localized anomalous regions. Fig. S7 provides an intuitive explanation of pre-training bias, indicating that not all feature maps contribute equally to anomaly detection and localization, which validates the efficacy of AFS.

E. Limitations

In some categories with more texture anomalies, such as the texture categories in MVTec-AD dataset [3], SIA's performance may slightly underperform when compared to DTD [5]. Given that DTD dataset [5] includes a diverse range of real-world texture images, it effectively simulates common anomaly types in the textural category, such as color, oil, and glue. Nonetheless, SIA excels in the majority of scenarios, outperforming DTD [5] and offering superior capability in synthesizing anomalies in images with intricate structures.

Compared to anomaly synthesis methods based on data augmentation [20, 32] or external data [48], SDAS increases additional offline training time. For instance, we generate 10,000 anomaly images at a resolution of 256×256 for each category, and it will take 6 hours using a single NVIDIA GeForce RTX 3090. However, it is pivotal to clarify that RealNet omits SDAS without any additional computational cost during inference and real-world applications. Therefore, we believe that the slight increase in training time to enhance performance is necessary and worthwhile.

In order to achieve higher computational efficiency, we do not upsample multi-scale features. Instead, we employ multiple reconstruction networks for feature reconstruction, which reduce the resolution of anomaly detection. The lower feature reconstruction resolution may introduce minor deviations in localizing small anomalous areas, leading to a decrease in PRO. However, we found that increasing the resolution of anomaly detection by reducing the number of reconstruction networks can improve PRO. For instance, architecture **C** in Fig. S2 achieved a higher PRO score of 94.71%. Furthermore, increasing the resolution of images can also lead to an improvement in PRO, as detailed in Tab. S4.

- LPIPS（学习感知图像块相似度）[51]：我们采用基于聚类的LPIPS [11]来评估合成异常图像的多样性。假设某类别包含 N 张真实异常图像，我们通过寻找最低LPIPS值将合成异常图像划分为 N 组，随后计算每组内的平均成对LPIPS值，并计算所有组的平均值。更高的聚类LPIPS值表示更强的多样性。

我们采用多种异常合成方法生成了1000张异常图像进行评估，每种方法均独立进行三次测试。实验结果如表S6所示。与其他异常合成方法相比，SIA在FID和LPIPS指标上均取得最优结果，凸显了SDAS在生成既逼真又多样的异常图像方面的卓越性能，同时证明了SDAS在提升异常检测性能方面的有效性。

D. 可视化

我们对RealNet在四个数据集上进行了全面的可视化分析。图S3展示了RealNet在异常定位方面的定性结果，凸显了其在像素级异常定位中的卓越性能。图S4和图S5分别展示了由SDAS生成的异常图像和正常图像。图S6展示了使用SIA合成带有局部异常区域的图像。图S7直观地解释了预训练偏差，表明并非所有特征图都对异常检测和定位有同等贡献，从而验证了AFS的有效性。

E. 局限性

在一些纹理异常较多的类别中，例如MVTec-AD数据集[3]中的纹理类别，与DTD[5]相比，SIA的性能可能略有不足。鉴于DTD数据集[5]包含了多样化的真实世界纹理图像，它能有效模拟纹理类别中的常见异常类型，如颜色、油污和胶渍。尽管如此，SIA在大多数场景中表现卓越，不仅超越了DTD[5]，还在合成具有复杂结构图像中的异常方面展现出更优异的能力。

与基于数据增强[20, 32]或外部数据[48]的异常合成方法相比，SDAS增加了额外的离线训练时间。例如，我们为每个类别生成10,000张分辨率为256×256的异常图像，使用单张NVIDIA GeForce RTX 3090将耗时6小时。然而，关键需要澄清的是，RealNet在推理和实际应用过程中无需任何额外计算成本即可省略SDAS。因此，我们认为为提升性能而略微增加训练时间是必要且值得的。

为了获得更高的计算效率，我们不对多尺度特征进行上采样，而是采用多个重建网络进行特征重建，这降低了异常检测的分辨率。较低的特征重建分辨率可能会在定位小异常区域时引入微小偏差，导致PRO指标下降。然而，我们发现通过减少重建网络数量来提高异常检测分辨率可以改善PRO。例如，图S2中的架构C实现了94.71%的更高PRO分数。此外，提高图像分辨率也能带来PRO的提升，具体细节见表S4。

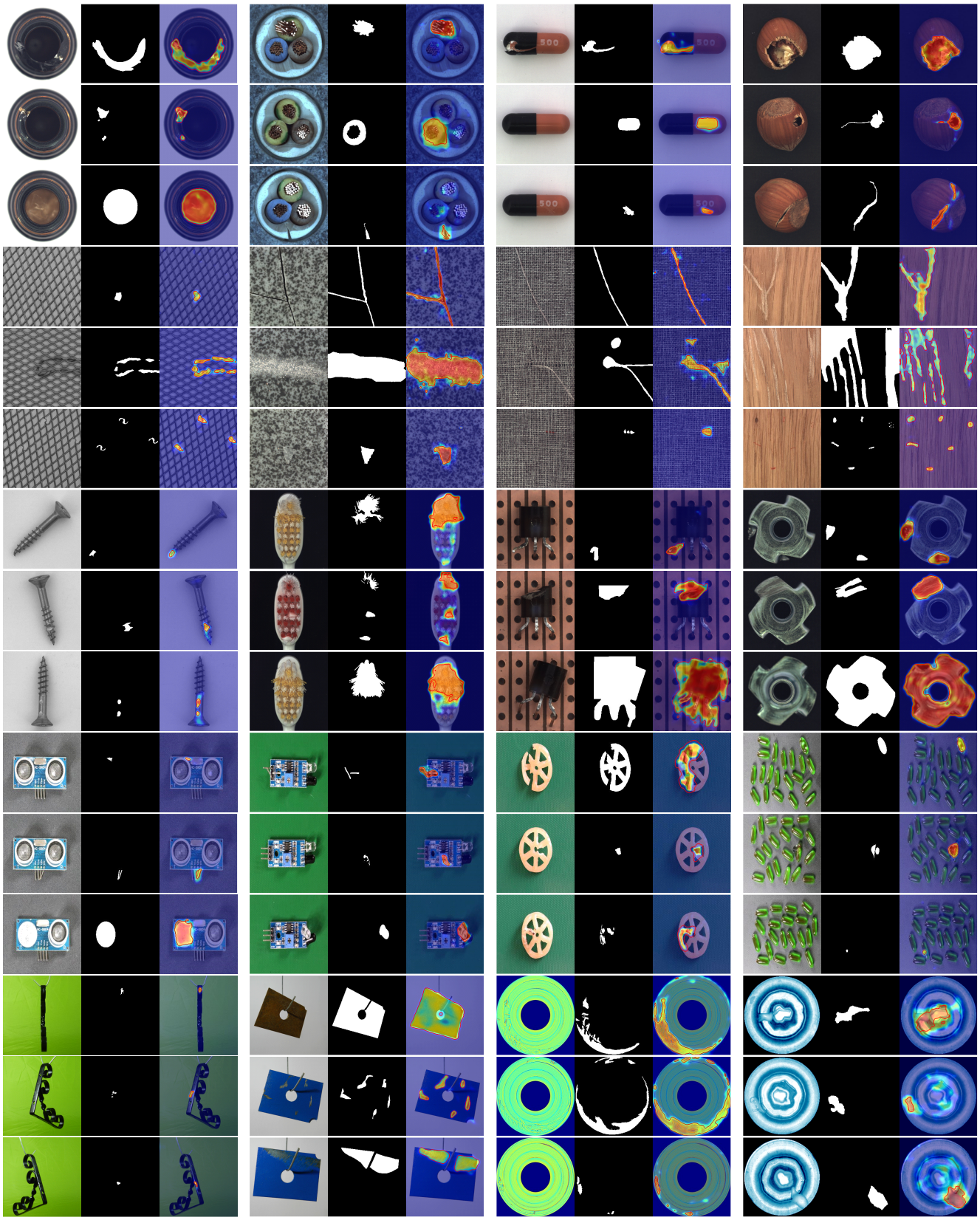


Figure S3. Qualitative results of RealNet. Within each group, from left to right, are the anomaly image, ground-truth, and predicted anomaly score. The examples are from the MVTec-AD [3], MPDD [18], BTAD [24], and VisA [55] datasets.

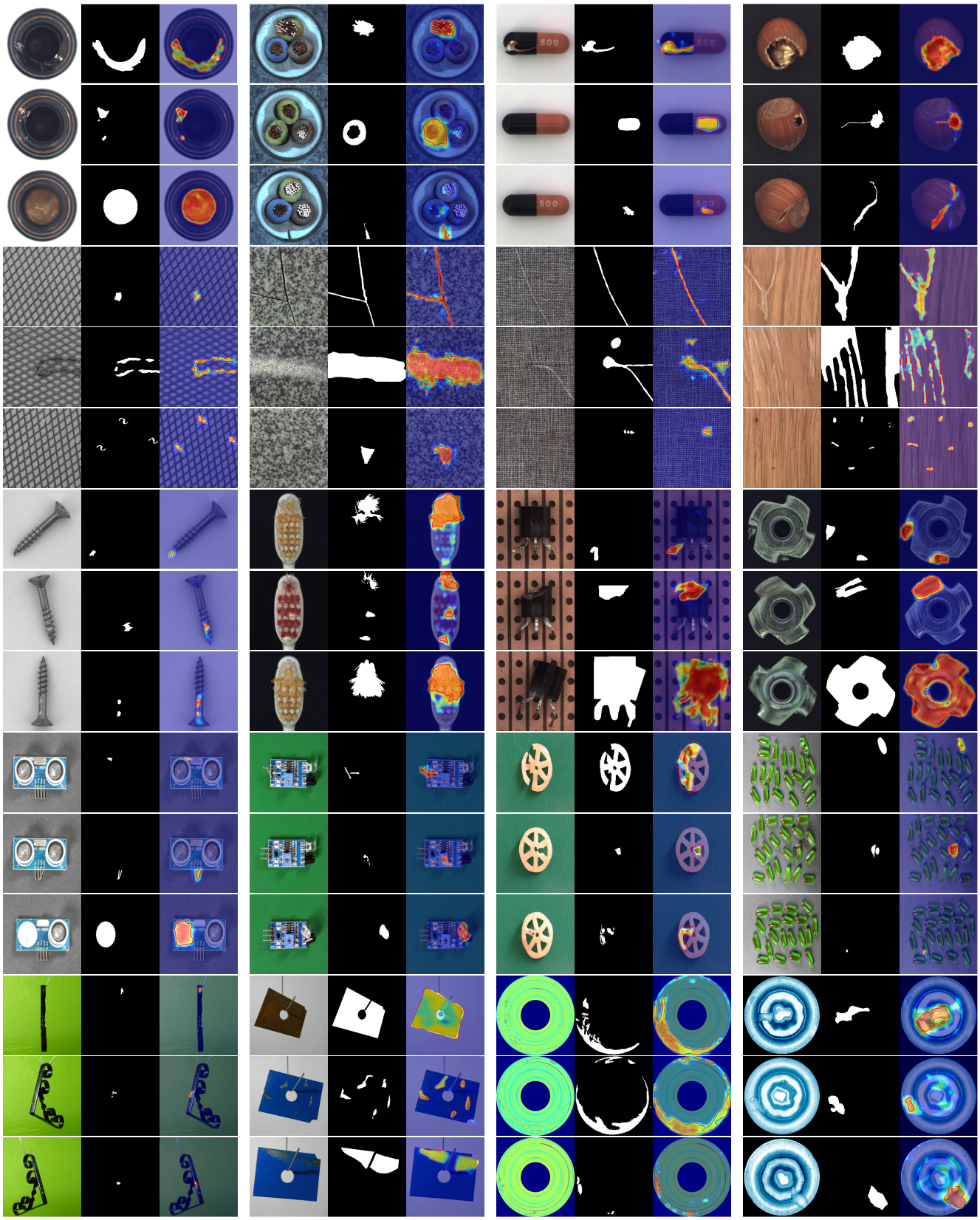


图 S3. RealNet 的定性结果。每组内从左至右分别为异常图像、真实标注及预测异常分数。示例选自 MVTec-AD [3]、MPDD [18]、BTAD [24] 和 VisA [55] 数据集。

Figure S4. Anomaly images generated by SDAS. The examples are from the MVTec-AD [3], MPDD [18], BTAD [24], and VisA [55] datasets. Within each group, from top to bottom, the anomaly strength gradually increases.

图 S4. SDAS 生成的异常图像。示例选自 MVTec-AD [3]、MPDD [18]、BTAD [24] 和 VisA [55] 数据集。每组图像中，从上至下异常强度{v*}逐渐增强。

Figure S5. Normal images generated by SDAS (when $s = 0$). The examples are from the MVTec-AD [3], MPDD [18], BTAD [24], and VisA [55] datasets.

图 S5. SDAS 生成的正样本图像（当 $s = 0$ 时）。示例选自 MVTec-AD [3]、MPDD [18]、BTAD [24] 和 VisA [55] 数据集。

Figure S6. Local anomaly images synthesized by SIA. The examples are from the MVTec-AD [3], MPDD [18], BTAD [24], and VisA [55] datasets. Within each group, from top to bottom, the anomaly strength gradually increases.

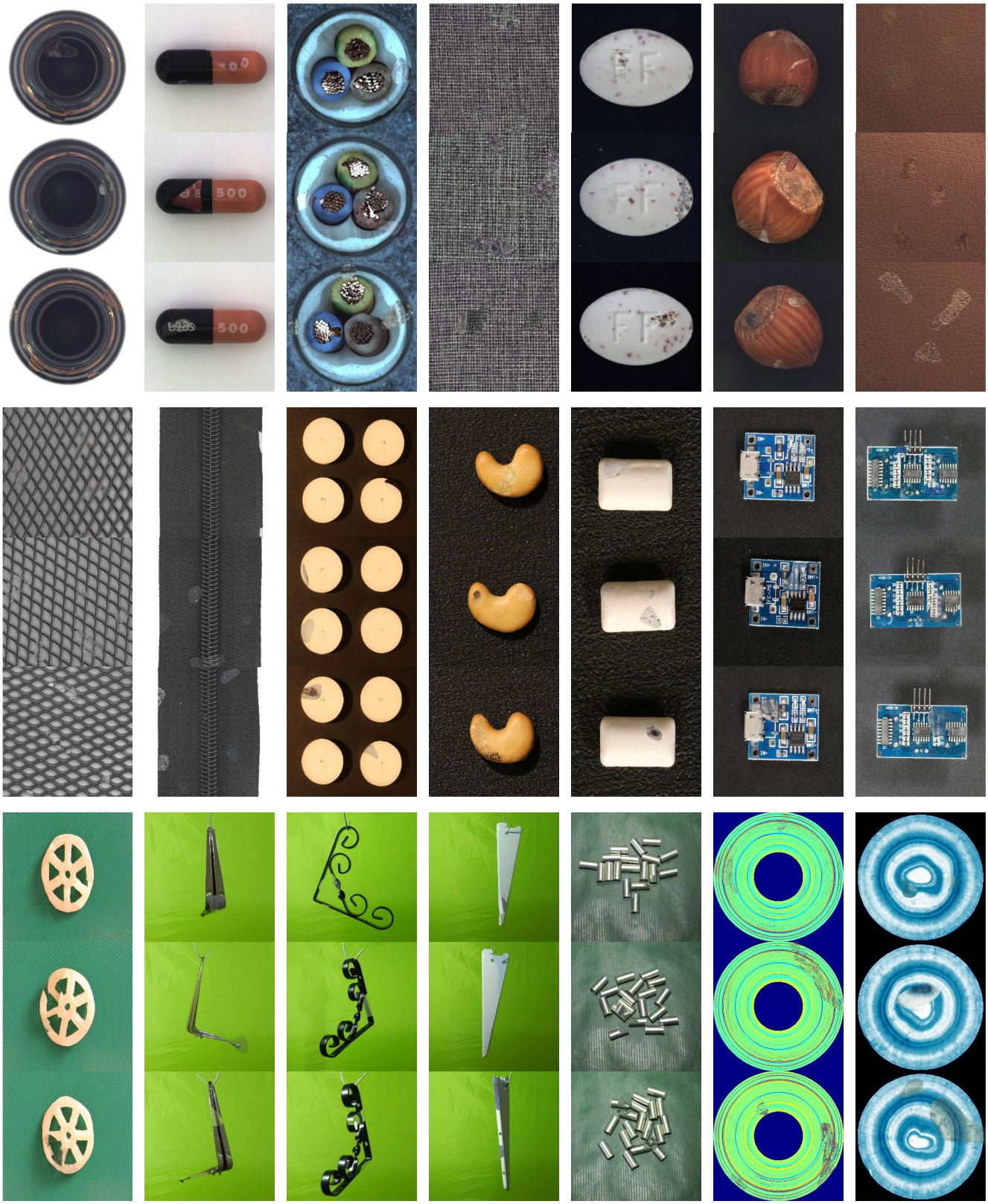


图 S6. SIA 合成的局部异常图像。示例来自 MVTec-AD [3]、MPDD [18]、BTAD [24] 和 VisA [55] 数据集。每组图像中，从上至下异常强度逐渐增强。

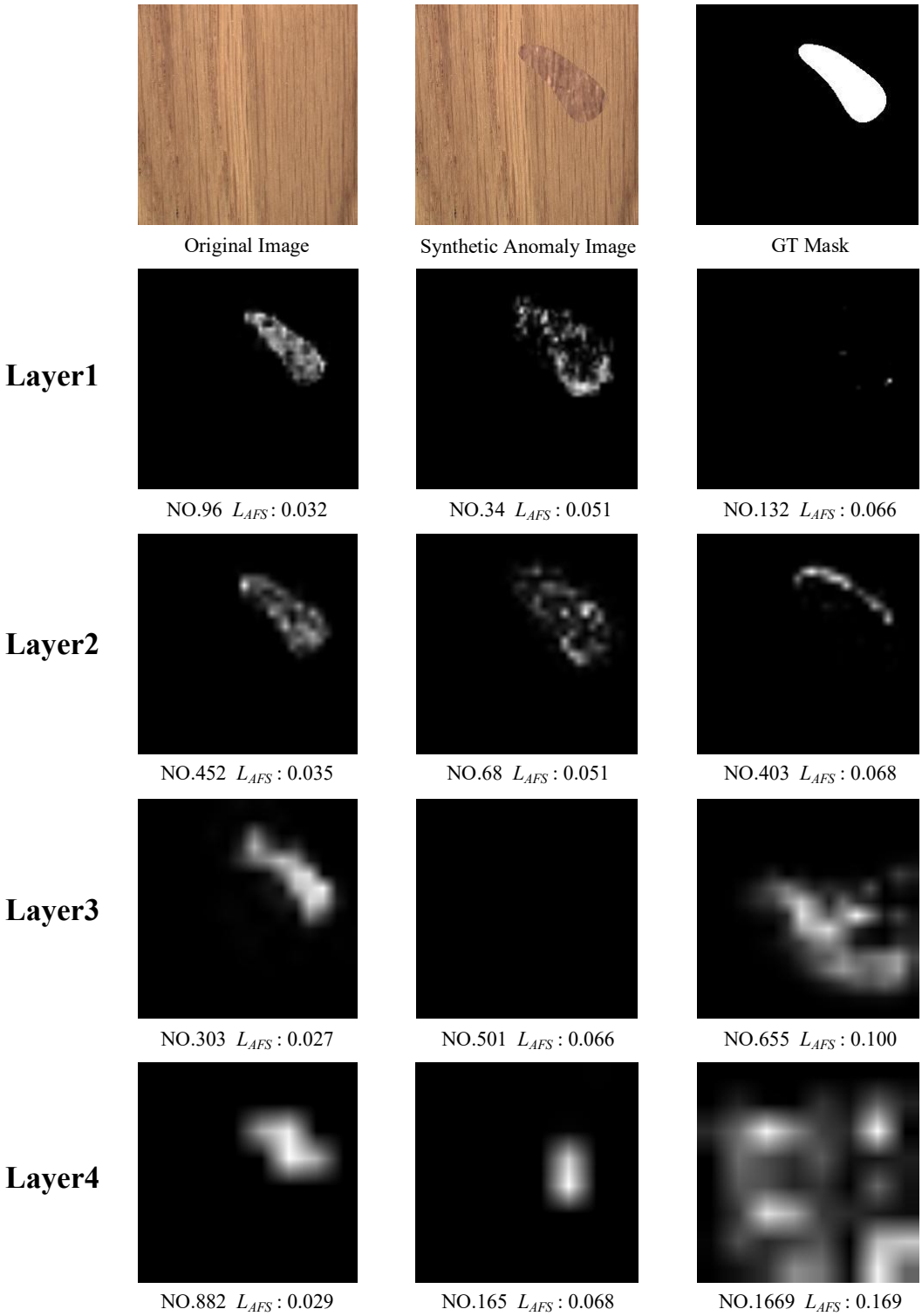


Figure S7. Visualization of AFS. For an original image and a synthetic anomaly image, we visualize the normalized difference between their corresponding feature maps across different layers of a pre-trained WideResNet50 [47]. From top to bottom, the feature map respectively come from the first layer to the fourth layer. Each feature map is labelled with its index in the layer and the corresponding AFS loss. From left to right, the localization performance of the feature maps gradually decreases. Our visualization intuitively demonstrates the localization bias caused by pre-training, indicating that not all feature maps contribute equally to anomaly detection and localization, as well as emphasizing the effectiveness of AFS.

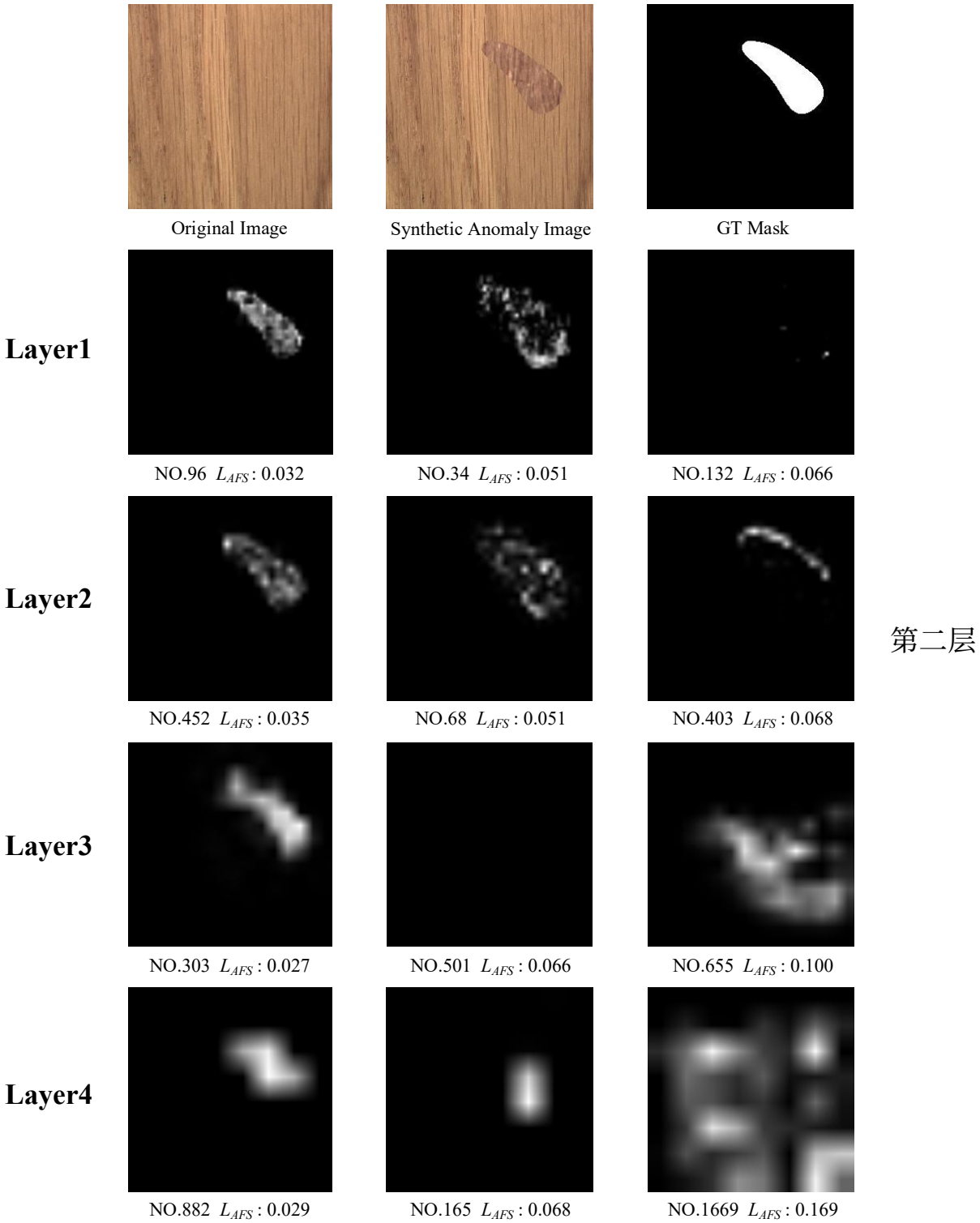


图 S7. AFS 的可视化。对于原始图像和合成异常图像，我们可视化了它们在预训练 WideResNet50 [47] 不同层对应特征图之间的归一化差异。从上到下，特征图分别来自第一层至第四层。每个特征图均标注了其所在层的索引及对应的 AFS 损失值。从左至右，特征图的定位性能逐渐下降。我们的可视化直观展示了预训练带来的定位偏差，表明并非所有特征图对异常检测与定位的贡献度相同，同时也印证了 AFS 的有效性。