DZAD: Diffusion-based Zero-shot Anomaly Detection

Tianrui Zhang[1, 2], **Liang Gao**[1], **Xinyu Li**[1*], **Yiping Gao**[1]

[1]Huazhong University of Science and Technology
[2]National University of Singapore
E1011227@u.nus.edu, gaoliang@hust.edu.cn, lixinyu@hust.edu.cn, gaoyiping@hust.edu.cn

Abstract

Zero-shot anomaly detection (ZSAD) aims to identify anomalies in new classes of images, and it's vital in industry and other fields. Most current methods are based on the multi-modal models CLIP and SAM, which have prior knowledge to assist model training, but they are highly dependent on the input of the prompts and their accuracy. We found that some diffusion model-based anomaly detection methods generate a large amount of semantic information and are very valuable for the ZSAD task. Therefore, we propose a diffusion model based zero-shot anomaly detection method, DZAD, and no additional prompt input is required. First, we propose the first diffusion-based zero-shot anomaly detection framework, which uses the proposed multi-timestep noise features extraction method to achieve anomaly detection in the denoising process of a latent space diffusion model with a semantic-guided (SG) network. Second, based on the detection results, we proposed a two-branch feature extractor for anomaly maps at different scales. Third, based on the difference between the anomaly detection task and other general image detection tasks, we propose a noise feature weight function for the diffusion model in the zero-shot anomaly detection task. Comparing with 7 recently state-of-the-art (SOTA) methods on MVTec AD and VisA datasets and analysis of the role of each component in ablation studies. The experiments demonstrate the validity of the method beyond the existing methods.

Introduction

In the field of computer vision, unsupervised anomaly detection (AD) tasks (Cao et al. 2024a) (Liang et al. 2023) (Wang et al. 2023b) aim to identify and localize anomalies in images using models trained only on normal images. Previous methods have focused on training dedicated models for each category, relying on vast collections of normal images as reference (Chen, Han, and Zhang 2023). However, in practical applications, it's challenging to collect extensive training images for each category of diverse industrial products. These models are typically evaluated only on data categories observed during training. Thus, the ability of a model to evaluate samples in an open-world environment, i.e., detecting anomalies in new classes not seen during training, is particularly meaningful. Consequently, we have initiated research on zero-shot anomaly detection (ZSAD) and propose a novel framework for this purpose.

ZSAD is a challenging task without real data from unseen classes. Prior knowledge about the expected behavior of normal data and the potential behavior of anomalies is easily available, even without visual examples (Wang et al. 2023a). With the emergence of visual language models such as CLIP (Radford et al. 2021) and SAM (Kirillov et al. 2023), some methods have demonstrated their capability to detect unknown features by encoding this prior knowledge in natural language and then aligning it with image features. Current zero-shot anomaly detection typically relies on the CLIP and SAM frameworks, describing product anomalies in natural language text (Cao et al. 2023). Building on this idea, some methods (Li et al. 2024) (Deng et al. 2023) (Zhang et al. 2023a) enhance these natural language text features to better suit zero-shot anomaly detection. These methods heavily depend on the accuracy of text features, often generating many text features for individual samples or enhancing the accuracy of prior information through extensive experimentation. To address this, we propose the zero-shot anomaly detection method that does not require text prompt input.

Diffusion models (Rombach et al. 2022) have recently demonstrated exceptional image generation capabilities, with some approaches adapting them for anomaly detection tasks. However, traditional denoising diffusion probabilistic models (DDPM) (Ho, Jain, and Abbeel 2020) and latent diffusion models (LDM) (Rombach et al. 2022) face challenges in reconstructing anomalies in input images while preserving semantic information through the addition of noise at different timesteps. By incorporating semantic features, some methods (He et al. 2024) (Zhang et al. 2023b) have addressed these reconstruction issues within the diffusion model framework. We find that semantic features during the diffusion process are particularly meaningful for zero-shot anomaly detection, yet these methods have only applied them in traditional anomaly detection tasks to influence the generation of latent features in the latent space. Therefore, we introduce a diffusion model-based approach for zero-shot anomaly detection.

To address the above problems, we propose the first diffusion-based framework DZAD for zero-shot anomaly detection and localization as shown in Figure 1, which

[*] Corresponding author: lixinyu@hust.edu.cn

consists of three parts: LDM diffusion model with linked SG networks, multi-timestep denoising feature anomaly detection, and two-branch anomaly map construction. Through our proposed multi-timestep noise feature extraction method, semantic features generated during the denoising process in the latent space of the diffusion model are collected and used for anomaly detection. Our proposed noise weight function for this diffusion zero-shot detection adjusts the relationship of diffusion process features. Then based on its results, two branching constructs multi-scale anomaly map. our contribution is summarized as follows:

- We propose the first diffusion-based zero-shot anomaly detection framework, DZAD, without the dependency on the prompt input.
- We propose a multi-timestep noise feature extraction method for an SD denoising network connected to an SG module to achieve zero-shot anomaly detection using diffusion noise features in the latent space. Based on the detection results obtained from the noise features, we propose a two-branch feature extraction method for anomaly maps of different scales to construct the anomaly maps.
- Based on the difference between the anomaly detection task and other general image detection tasks, the noise feature weight function of the diffusion model for the anomaly detection task is proposed.
- Abundant experiments demonstrate that DZAD not only has no prompt input requirement but also has sufficient advantages over the SOTA method. Compared with the current SOTA results, we have significant advantages in AUC results for three classical anomaly detection datasets.

Related Works

Zero-shot anomaly detection. Zero-shot anomaly detection tasks aim to develop a unified model capable of detecting anomalies across various domains without relying on reference normal samples (Jeong et al. 2023). Despite its potential for vast multifunctionality, this task faces significant challenges due to the absence of specific prior information related to the target domain. Current methods enhance anomaly detection capabilities by integrating external prior knowledge. CLIP acquires implicit knowledge to distinguish normal from abnormal by training on a vast dataset of visual-text pairs. The computed similarity can serve as an effective anomaly score (Cao et al. 2024a). WinCLIP (Jeong et al. 2023) achieves zero-shot anomaly detection by calculating the similarity between image patches and normal/abnormal textual captions. April-GAN (Chen, Han, and Zhang 2023) introduces a trainable linear layer to adapt CLIP for anomaly detection, thereby enhancing its applicability. AnomalyCLIP (Zhou et al. 2023) introduces the concept of learning object-agnostic text prompts to overcome the limitations of manually designed prompts. Similarly, based on SAM, SAA (Cao et al. 2023) introduces an integration framework that incorporates human expertise into the detection system. These methods require extensive text information features as prior knowledge to align with image features, typically generating hundreds of prompts per sample to aid training.

Due to the high dependence on prompt features, many methods are limited to breakthroughs in prompt generation. VQA-oriented AD (Zhang et al. 2023a) uses GPT-4V for granular region delineation followed by prompt generation. AnoVL (Deng et al. 2023) proposes a unified domain-aware contrastive state prompt template to reinforce prompts. PromptAD (Li et al. 2024) introduces a Cross-View Contrast Learning (CCL) strategy to optimize prompts. Since the quality of textual prompts critically determines the training effectiveness of the model (Radford et al. 2021; Jeong et al. 2023), these methods necessitate extensive text information to assist in image training.

Diffusion model. The diffusion model has gained widespread attention and research interest since its remarkable reconstruction ability (Zhang and Agrawala 2023) and LDM (Rombach et al. 2022) introduces conditions through cross-attention to control generation. Due to the diversity of anomaly features, image reconstruction for anomaly detection tasks often results in the loss of significant semantic features, which directly impacts the outcomes of reconstruction-based anomaly detection methods. The DiffAD (Zhang et al. 2023b) and DiAD (He et al. 2024) methods have recognized this issue and have proposed LDM-based anomaly detection reconstruction frameworks to integrate semantic features. DiffAD proposes noisy condition embedding and interpolation channels to diversify the reconstruction process and improve noise guidance mitigation. DiAD proposes the Semantic Feature Fusion (SFF) block to integrate multi-scale features, linking the Stable Diffusion (SD) and the Semantic-Guided (SG) network, thereby maintaining semantic consistency and reconstructing anomalies effectively. We find that the plethora of semantic features generated during the diffusion process plays a significant role in the novel task of zero-shot anomaly detection. However, these anomaly detection methods only utilize them to guide reconstruction without leveraging the noise features generated during the process.

To address the high dependence on the prompt input in zero-shot anomaly detection while maintaining the accuracy of the model, we propose a zero-shot anomaly detection method that utilizes noise features during the diffusion denoising process to enhance general detection capabilities.

Preliminaries

Denoising Diffusion Probabilistic Model. The Denoising Diffusion Probabilistic Model (DDPM) (Ho, Jain, and Abbeel 2020) consists of two main processes: the forward diffusion process and the reverse denoising process. During the forward diffusion process, a noisy sample x_t is generated using a Markov chain that incrementally adds Gaussian-distributed noise to an initial data sample x_0. The forward diffusion process can be characterized as follows:

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

The diffusion model defines $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i = \prod_{i=1}^{T}(1 - \beta_i)$, where β_i specifies the noise

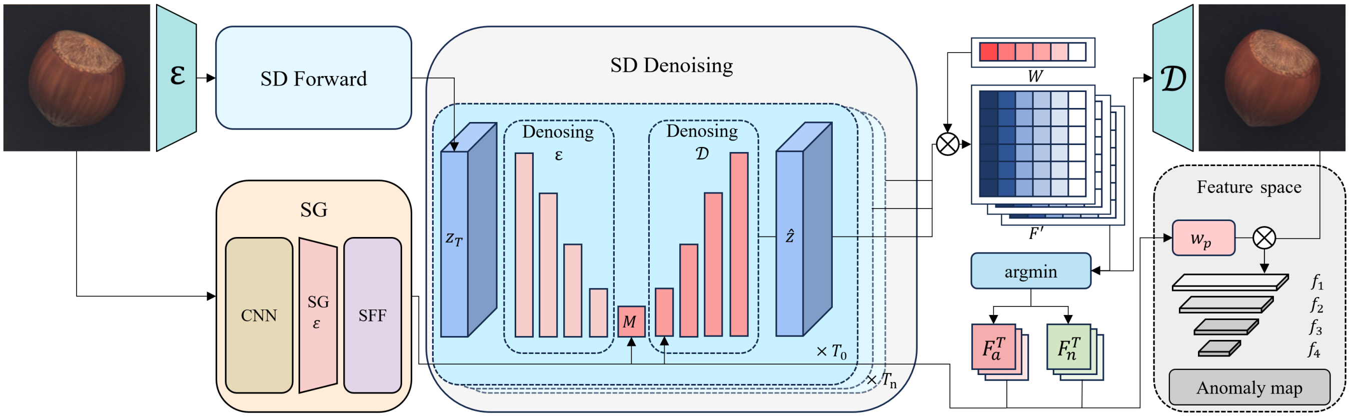

Figure 1: **Overall framework of DZAD.** ***1)*** The SG network encodes the input control feature c_i into latent space to guide the latent feature z_t generation. ***2)*** Noise and its semantic features are extracted by multi-timestep for anomaly detection in denoising the U-Net of SD. ***3)*** Construct multi-scale anomaly maps by pre-training feature extractor through two-branch feature extraction.

schedule controlling the amount of noise added at each timestep. In the reverse denoising stage, x_t is first sampled. Then, given the current sample x_t and the model prediction $\epsilon_\theta(x_t, t)$, reconstruct x_{t-1} using:

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

where $z \sim \mathcal{N}(0, I)$, σ_t is a fixed constant related to the variance schedule, $\epsilon_\theta(x_t, t)$ is a U-Net (Rombach et al. 2022) network used to predict the distribution, and the learnable parameters θ of the U-Net part are optimized by minimizing the following objective:

$$\min_\theta \mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0,I), t} \left\| \epsilon - \epsilon_\theta(x_t, t) \right\|_2^2 . \quad (3)$$

Latent Diffusion Model. The Latent Diffusion Model (LDM) (Rombach et al. 2022) consists of a pre-trained autoencoder model and a denoising U-Net. This network uses an encoder to compress images into a latent representation space, where diffusion is performed. Conditional features c are connected to the model through a cross-attention mechanism, thereby training the latent variables z_t. The training objective is defined as:

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

Diffusion Model ELBO. Diffusion models parameterize $p_\theta(x_{t-1} \mid x_t, c)$ as a Gaussian and train a neural network to map a noisy input x_t to a value used to compute the mean of $p_\theta(x_{t-1} \mid x_t, c)$ (Clark and Jaini 2024). Using this parameterization, the ELBO (Evidence Lower Bound) can be written as:

$$-\mathbb{E}_\epsilon \left[\sum_{t=2}^{T} w_t \left\| \epsilon - \epsilon_\theta(x_t, c) \right\|^2 - \log p_\theta(x_0 \mid x_1, c) \right] + C, \quad (5)$$

where C is a constant term that does not depend on c. Since T is over 1000 and $\log p_\theta(x_0 \mid x_1, c)$ is typically small, this term can be dropped and get:

$$-\mathbb{E}_{t,\epsilon} \left[\left\| \epsilon - \epsilon_\theta(x_t, c) \right\|^2 \right] + C. \quad (6)$$

Method

Our proposed method DZAD is shown in Figure 1. First, the pre-trained encoder downsamples the input image into a latent space representation. Then, the encoded image features are denoised using an SD denoising connected to an SG network. In the denoising process, the multi-timestep feature extraction method extracts the noise features in the latent space. Noise features at various t are extracted for the same input at different timesteps. The noise features at each stage n form a multi-timestep matrix corresponding to each input to realize anomaly detection. Based on the results of anomaly detection, a multi-scale anomaly map is constructed by controlling the distance between the features and the pre-trained decoder through a two-branch feature extraction.

Semantic-Guided Diffusion

DDPM and LDM are prone to the semantic loss problem when facing industrial image anomaly detection tasks, resulting in poor reconstruction quality (He et al. 2024; Li et al. 2024). The reason is that denoising based on Gaussian noise-like distribution after diffusion of many timesteps may produce samples belonging to different semantic classes. This is more serious for the zero-shot task since the model requires more adequate semantic features. The SG network can be used to solve the problem of LDM's inability to effectively reconstruct anomalies and retain semantic information about the input images (He et al. 2024). We found that this semantic information is beneficial for zero-shot anomaly detection. The pre-trained encoder $\mathcal{E}$ encodes $x_0 \in \mathbb{R}^{3 \times H \times W}$

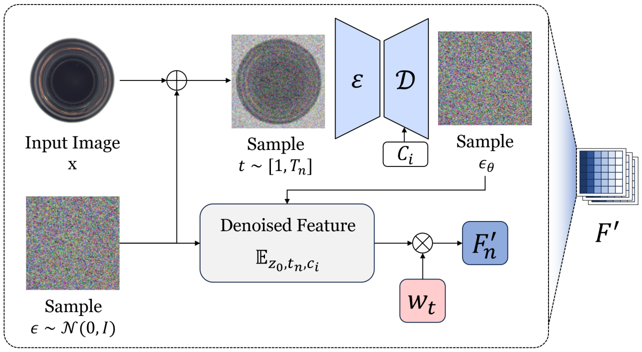

Figure 2: **Multi-timestep denoising anomaly detection.** Noise features on various t are extracted for the same input on several different timesteps, and the noise features at each stage n form a multi-timestep matrix corresponding to each input.

into a latent space representation $z = \mathcal{E}(x_0)$ where $z \in \mathbb{R}^{c\times h\times w}$. Get the forward diffusion process as:

$$z_t = \sqrt{\overline{\alpha}_t} z_0 + \sqrt{1-\overline{\alpha}_t}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I). \tag{7}$$

The latent representation feature z_t and the input x_0 are input to the SD denoising network and the SG network respectively. With T reverse denoising steps, the constraint of the LDM training is defined as:

$$\mathcal{L} = \mathbb{E}_{z_0,t,c_i,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_i)\|^2\right]. \tag{8}$$

The input image $x_0 \in \mathbb{R}^{3\times H\times W}$ undergoes transformation via SG network to maintain dimensionality compatible with the SD Encoder Block outputs. The combined outputs of x and z are processed through SG Encoder Blocks, with results continuously down sampled. The outputs from the SD and SG middle blocks (M_{SD} and M_{SG}) are summed up. To accommodate multi-class scenarios, outputs from the SG Decoder Blocks D_{SG} are merged with those from the SD decoder D_{SD} enhanced by an SFF block (He et al. 2024). The final output of the denoising network G is expressed as:

$$G = D_{\text{SD}}(M_{\text{SD}}(x_t)) + M_{\text{SG}}(x_0) + D_{\text{SG}}(M_{\text{SG}}(x_0)). \tag{9}$$

Anomaly Localization and Detection

Anomaly detection involves isolating images with characteristics that deviate significantly from the norm. In zero-shot anomaly detection tasks, classification typically uses a conditional generative model. This can be achieved by leveraging Bayesian inference to predict $p_\theta(\mathbf{x} \mid c_i)$ and prior knowledge $p(c)$ on the labels $\{c_i\}$:

$$p_\theta(c_i \mid x) = \frac{p(c_i)\, p_\theta(\mathbf{x} \mid c_i)}{\sum_j p(c_j)\, p_\theta(\mathbf{x} \mid c_j)}. \tag{10}$$

The prior knowledge $p(c)$ on the labels $\{c_i\}$ is typically extracted by training a model. Some effective methods expand this prior into large quantities of natural language to capture semantic information. However, these expansions require extensive validation with official linguistic data, and they may generate vast amounts of irrelevant features that impact feature extraction. To address this critical issue, we use SG networks to extract hidden features z_t during the training process as semantic features. For the expanded generative model, calculating $p_\theta(\mathbf{x} \mid c)$ is computationally intensive, so we use ELBO to handle this problem:

$$p_\theta(z_i \mid x) = \frac{\exp\{-\mathbb{E}_{t_i}[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, z_i^t)\|^2]\}}{\sum_j \exp\{-\mathbb{E}_{t_j}[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, z_j^t)\|^2]\}}. \tag{11}$$

The tensor data volume of the results obtained in this way for each input is relatively large due to the large differences in the features of the intra and inter between the different anomaly types and the small correlation. When used for classification and detection, it is often very ineffective due to its high degree of discretization. To address the issue of heterogeneity in anomaly types, we propose a multi-timestep feature matrix method to solve the problem of feature dispersion. For the input K class y_k, We calculate the unbiased Gaussian estimate at each expected timestep by sampling $N(t_i, \epsilon_i)$, where $t_i \sim [1, T]$ is divided into n equidistant timesteps $t_n \sim [1, T_n]$. These timesteps start from t_0 with an increment of d. For each corresponding timestep, $M = \frac{T}{d}$ times ELBO features of these noise images are calculated separately. These features combine to form a multi-timestep feature matrix $F \in \mathbb{R}^{N\times M}$. The multi-timestep feature matrix is then multiplied by a function w_t of the corresponding n in the weight matrix W and get F'. This method combines different diffusion features separately and avoids the problem of being too discrete when all noise features are stored in a very long tensor if the anomaly types are very different. The anomaly detection results y got from priori feature type y_k are expressed as:

$$y = \arg\min_{y_k} \sum_n \mathbb{E}_{z_0,t_n,c_i}\left[W_n \|(\epsilon - \epsilon_\theta(z_t, t_n, c_{y_k}))\|^2_{y_k}\right]. \tag{12}$$

For anomaly detection tasks, we found a significant difference between our method and others that utilize noise-augmented feature images for classification and segmentation tasks. Imagen (Kingma et al. 2021) and SD use "simple" noise augmentation training, where $w_t = \text{SNR}(t)$, representing the signal-to-noise ratio synchronized with time t. Google DeepMind (Clark and Jaini 2024) discovered that a single tuning function $w_t = \exp(-7t)$ is effective for cross-task image classification. However, these weighting functions are not suitable for the multi-timestep feature matrix we proposed for anomaly detection tasks. Based on their Learned Weighting Function, we then learn a model of 10 buckets $[b_0, \ldots, b_9]$ for class y_k with Monte Carlo sampling to perform standard maximum likelihood over the data:

$$p_v(y_k \mid x) = \frac{1 + \exp\left(\sum_{i=0}^{n} v_i \cdot (nb_i - x, y_k)\right)}{\sum_{y_j\in[y_k]}\left(1 + \exp\left(\sum_{i=0}^{n} v_i \cdot (nb_i - x, y_j)\right)\right)}. \tag{13}$$

We discovered that for anomaly detection tasks, the weighting function needs to slightly increase over time, unlike other tasks that require rapid exponential reduce. Finally, we found that the weighting function that approximates $w_t = 1 + \exp\left(-3 \cdot (10 - n)\right)$ can effectively be used for different timesteps of the multi-timestep feature matrix. The w_t of different stages n forms the weight matrix W.

The anomaly or normal features for each input are obtained from the anomaly detection results and are classified into sets F_a^T and F_n^T. With the results, the pixel-level weighting modules w_{pa} and w_{pn} are applied to the corresponding states, respectively. The features on the feature layer f_n at multiple scales n are later accumulated with the upsampling factor σ_n to form the final anomaly map. For F_n^T, w_{pn} moves F_n^T closer towards the input pre-trained model extractor features Φ^n. The anomaly map S_n is :

$$S_n = \sum_n \sigma_n \left(1 - w_{pn}(F_n^T)\frac{\Phi^n(x_0)^T \Phi^n(\hat{x}_0)}{\|\Phi^n(x_0)\| \|\Phi^n(\hat{x}_0)\|}\right). \quad (14)$$

For F_a^T, w_{pa} moves F_a^T away towards the extractor Φ^n output, and the most similar F_n^T is used to guide the features reconstruction, and the anomaly map S_a is:

$$S_a = \sum_n \sigma_n \left(1 - w_{pa}(F_n^T, F_a^T)\frac{\Phi^n(x_0)^T \Phi^n(\hat{x}_0)}{\|\Phi^n(x_0)\| \|\Phi^n(\hat{x}_0)\|}\right). \quad (15)$$

Experiments

Datasets and Evaluation Metrics

MVTec AD dataset. The MVTec AD (Bergmann et al. 2019) dataset simulates real-world industrial production scenarios. It consists of 5 types of textures and 10 types of objects in 3354 high-resolution images from different domains. The training set contains 3629 images with only normal samples and test set consists of 1725 images.

VisA dataset. The VisA (Zou et al. 2022) dataset consists of 12 subsets corresponding to 12 different objects. It includes 10,821 images, with 9,621 normal and 1,200 anomalous samples. Four subsets represent various types of PCB with complex structures like transistors, capacitors, and chips. Anomalous images feature various flaws, including surface defects like scratches, dents, color spots, or cracks, and structural defects such as misplacement.

BTAD dataset. The BTAD dataset (Mishra et al. 2021) specifically targets anomaly detection tasks within industrial texture quality control. It consists of 2377 high-resolution images meticulously organized into 3 main categories. These categories are further divided into specific types of material anomalies such as cracks, corrosion, and deformations.

Evaluation metrics. We utilize the widely recognized standard metric—Area Under the Receiver Operating Characteristic Curve (AUROC). Image-level AUROC is used to evaluate the model capability to identify anomalies at the overall image level, which is essential for determining whether an entire image is anomalous. On the other hand, pixel-level AUROC focuses on locating specific anomalous regions within an image, which provides a detailed localization of anomalies.

Experimental Setup

We tested the other datasets with diffusion models trained on the VisA dataset, and accordingly, tested the VisA dataset with diffusion models trained on the MVTec AD dataset. In this experiment, all images are resized to 256×256. We employ ResNet50 as the feature extraction network and select $n \in \{2, 3, 4\}$ as the feature layers for calculating anomaly localization. The training is conducted 1100 epochs on a single NVIDIA A100 40GB GPU, with a batch size of 32. We use the Adam optimizer (Loshchilov and Hutter 2019) with a learning rate of $1 \times e^{-5}$. For anomaly detection, the anomaly score of the image is derived from the maximum value of the anomaly localization score, which is processed through eight rounds of global average pooling, each with a size of 8×8. The initial denoising timestep T is set to 1,400 during inference. We employ DDIM (Song, Meng, and Ermon 2021) as the default sampler with 10 steps.

Comparison with SOTAs

We evaluate the proposed method and the other seven methods on the MVTec AD dataset to compare SOTA methods on the MVTec AD dataset. They are SAM (Kirillov et al. 2023), WinCLIP (Jeong et al. 2023), SAA (Cao et al. 2023), April-GAN (Chen, Han, and Zhang 2023), AdaCLIP (Cao et al. 2024b), AnomalyCLIP (Zhou et al. 2023). April-GAN and SAA took the first and second place respectively in 2023 CVPR Zero-Shot Visual Anomaly and Novelty Detection Challenge (VAND 2023). These methods require prompts as inputs into the model. We also compare the recently proposed MAEDAY (Schwartz et al. 2024) method that uses pre-trained models directly and does not require prompts as input.

Qualitative results. We show qualitative anomaly detection results for the MVTec AD datasets to visualize the effectiveness of the proposed method. As shown in Figure 3, the selected samples clearly demonstrate the accuracy of our method in anomaly detection, and it is intuitively clear that our method is more precise and accurate compared to other methods. The method demonstrates a robust detection capability, particularly when dealing with intricate background textures and anomalous region features closely resembling the background texture. For large complex shape detection, other results are usually discontinuous, our method solves this problem. In addition, for some complex shapes in normal samples, other methods can easily detect them as anomalies, while our method significantly outperforms other methods in this regard. These results demonstrate its advantages in practical applications.

Quantitative results. As shown in Table 1 and 2, our method achieves SOTA on most of the AUC score results for the MVTec AD, VisA, and BTAD datasets. Especially for the image level AUC results, our method is much higher than the current SOTA. Taking the classical dataset MVTec as an example, it is clear that our anomaly detection results achieve SOTA for most kinds of anomalies. For methods that

Category	Prompts required						No prompts	
	SAM *ICCV 2023*	WinCLIP *CVPR 2023*	SAA *VAND 2023*	April-GAN *VAND 2023*	AdaCLIP *ECCV 2024*	AnomalyCLIP *ICLR 2024*	MAEDAY *CVIU 2024*	Ours
Bottle	0.946	0.895	0.755	0.834	0.944	0.904	0.507	**1.000**
Cable	0.596	0.770	0.637	0.723	**0.906**	0.789	0.655	0.542
Capsule	0.545	0.869	0.420	0.920	0.915	0.958	0.481	**0.963**
Carpet	0.805	0.954	0.995	0.984	0.821	0.988	0.762	**1.000**
Grid	0.904	0.822	0.837	0.958	0.900	0.973	0.954	**0.974**
Hazelnut	0.580	0.943	0.832	0.961	0.802	**0.971**	0.941	0.916
Leather	0.904	0.967	0.993	0.991	0.998	0.986	0.946	**1.000**
Metal Nut	0.606	0.610	0.348	0.654	0.835	0.744	0.396	**0.870**
Pill	0.682	0.800	0.506	0.762	0.829	0.920	0.615	**0.989**
Screw	0.686	0.896	0.464	**0.978**	0.870	0.975	0.969	0.947
Tile	0.419	0.776	0.957	0.927	0.905	0.946	0.309	**1.000**
Toothbrush	0.683	0.869	0.222	0.958	0.936	0.919	0.723	**0.967**
Transistor	0.533	0.747	0.370	0.624	0.821	0.710	0.597	**0.896**
Wood	0.964	0.934	**0.998**	0.958	0.983	0.965	0.788	0.992
Zipper	0.760	0.916	0.194	0.911	0.915	0.914	0.762	**0.972**
Mean	0.707	0.851	0.635	0.876	0.892	0.911	0.694	**0.935**

Table 1: Zero-shot anomaly detection performances on MVTec AD measured in image level AUROC.

Dataset	Metric	Prompts Required					No Prompts	
		SAM *ICCV 2023*	WinCLIP *CVPR 2023*	April-GAN *VAND 2023*	SAA *VAND 2023*	AdaCLIP *ECCV 2024*	MAEDAY *CVIU 2024*	Ours
MVTec AD	Image AUC	0.708	0.918	0.823	0.635	0.892	0.745	**0.935**
	Pixel AUC	0.854	0.841	0.837	0.755	**0.942**	0.694	0.867
VisA	Image AUC	0.619	0.781	0.817	0.671	0.858	0.731	**0.902**
	Pixel AUC	0.926	0.796	0.952	0.765	**0.955**	0.706	0.920
BTAD	Image AUC	0.851	0.590	0.852	0.590	0.886	0.742	**0.928**
	Pixel AUC	0.898	0.726	0.895	0.658	0.921	0.733	**0.923**

Table 2: Evaluation metrics on different datasets and methods.

require prompt inputs, our method does not have these features to assist the training, but it still has a significant advantage in its anomaly detection capability. For methods that do not require prompt input, our method far outperforms existing methods in both metrics. On multiple datasets, the average image level AUC exceeds 0.182↑ , and the average pixel level AUC exceeds 0.199↑. These extensive experiments on different datasets illustrate that our method not only removes the high dependence on prompt input but can also efficiently perform zero-shot anomaly detection. In addition, it is obvious from ours and other methods that the Cable class is prone to failure in zero-shot anomaly detection. This is mainly due to the fact that the Cable class consists of a number of almost unrelated anomaly types, such as wire filaments, wrappers, and housings. This may require special semantic guidance to optimize the detection.

Ablation Studies

Effect of the architecture design. We investigate the importance of each module in DZAD on the MVTec AD dataset, as shown in Table 3. SD denotes only the diffusion model, which is purely based on the framework of LDM. SGSD denotes the intermediate block of the SG network added to the middle of SD to guide the subsequent generation. M_t denotes mlti-timestep extraction, F_b denotes two-branch feature extractor and w_t denotes weight function. Experimentally, it is proved that our proposed multi-timstep extraction can efficiently achieve the extraction and classification of anomalous features. Moreover, our proposed two-branch feature extractor for anomaly maps at different scales makes the reconstruction at the pixel level effectively guided.

Effect of weighting function. Designing an effective noise weighting function in diffusion anomaly detection models is essential, yet it has seen limited research in anomaly detection. Current studies are mainly based on traditional image classification and detection tasks, utilizing classic general image datasets. To bridge this gap, we design a particularly learned weighting function w_t and compare it with classic methods proposed in previous research. These methods include the Variational Diffusion Model (VDM) (Kingma et al. 2021), where w_t = SNR'(t) is optimized directly against the lower bound likelihood, the ”simple”

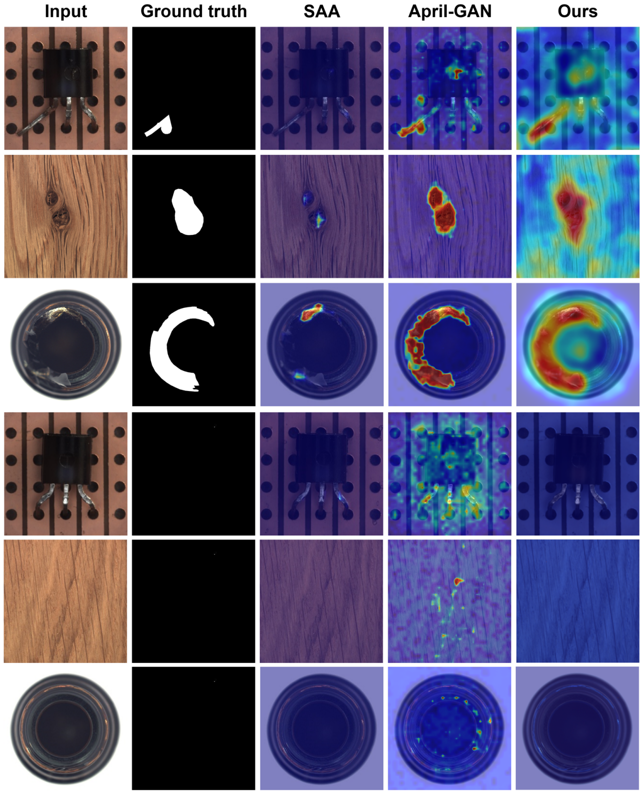

Figure 3: Qualitative illustration of anomaly location.

loss $w_t = \mathrm{SNR}(t)$ (Ho, Jain, and Abbeel 2020) based on human judgment and FID scores, and two strategies from DeepMind (Zhang and Agrawala 2023). One is a heuristic $w_t = \exp(-6t)$ that performs well on the CIFAR-100 dataset, and the other $w_t = \exp(-7t)$, a learned derived through clustering (Clark and Jaini 2024). As shown in the Table 4, these methods reduce rapidly with the timestep and are effective for general classification and detection tasks. However, they may not necessarily benefit anomaly detection.

To address this issue, we conducted research into the utilization of w_t in the context of zero-shot anomaly detection. From equation 13, we derived $w_t = 1+\exp(-3\cdot(10-n))$ in n stages timesteps. The outcomes presented in Table 4 illustrate that our proposed function has the potential to enhance the results of anomaly detection within the framework.

Effect of forward diffusion timesteps. We test the corresponding diffusion models in our framework with the MVTec AD and VisA datasets on the other dataset. Image and pixel-level AUC results are used as indicators. The experimental results are shown in Figure 4. As the number of forward diffusion timesteps increases, the anomaly reconstruction ability improves, but the degree of training is various. So we used 1225 epochs and 825 epochs respectively for the MVTec AD and VisA datasets.

SD	SGSD	M_t	F_b	w_t	*cls*	*seg*
✓					0.812	0.718
✓	✓				0.834	0.752
✓	✓	✓			0.930	0.823
✓	✓		✓		0.870	0.864
✓	✓	✓		✓	0.930	0.865
✓	✓	✓	✓	✓	**0.935**	**0.867**

Table 3: Ablation studies of the architecture design with AUROC metrics.

Weight Function w_t	MVTec AUC	VisA AUC
1	0.930	0.878
SNR	0.718	0.703
SNR'	0.709	0.707
$\exp(-6t)$	0.878	0.805
$\exp(-7t)$	0.877	0.784
$1+\exp(-1\cdot(10-n))$	0.922	0.890
$1+\exp(-5\cdot(10-n))$	0.933	0.901
$1+\exp(-3\cdot(10-n))$	**0.935**	**0.902**

Table 4: Ablation studies of the weight function w_t.

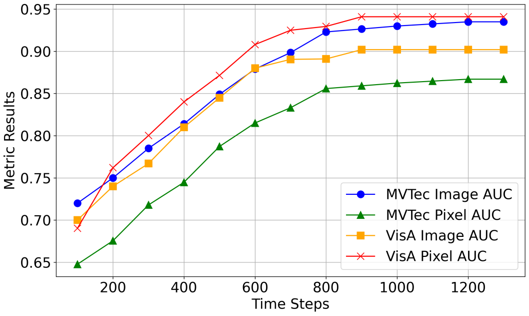

Figure 4: Ablation studies of timesteps.

Conclusion

We propose the first diffusion model-based zero-shot anomaly detection method, and our method solves the problem of high dependence on prompt input. Multi-timestep noise feature extraction method extracts noise features from the LDM model combined with the SG network. This differs from just using it to generate latent features in the latent space, as in the previous methods. In addition, we propose a two-branch feature extraction method to construct multi-scale anomaly maps. Numerous experiments show that our model outperforms existing zero-shot anomaly detection methods. Our method far outperforms other zero-shot anomaly detection methods that also do not require prompt inputs. Due to the relatively weak semantic embedding and alignment capabilities of zero-shot anomaly detection models, in future work, we will research the effect of diffusion noise on the accuracy of the pixel-level reconstruction results and improve the results further.

Acknowledgments

This work is supported by NSFC (No.52188102 and 52205523) and Fundamental Research Funds for the Central Universities under Grant No. 2024BRA004.

References

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD - A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Sections 1, 6.

Cao, Y.; Xu, X.; Sun, C.; Cheng, Y.; Du, Z.; Gao, L.; and Shen, W. 2023. Segment any anomaly without training via hybrid prompt regularization. arXiv:2305.10724.

Cao, Y.; Xu, X.; Zhang, J.; Cheng, Y.; Huang, X.; Pang, G.; and Shen, W. 2024a. A survey on visual anomaly detection: Challenge, approach, and prospect. arXiv:2401.16402.

Cao, Y.; Zhang, J.; Frittoli, L.; Cheng, Y.; Shen, W.; and Boracchi, G. 2024b. AdaCLIP: Adapting CLIP with Hybrid Learnable Prompts for Zero-Shot Anomaly Detection. arXiv:2407.15795.

Chen, X.; Han, Y.; and Zhang, J. 2023. A zero-/fewshot anomaly classification and segmentation method for CVPR 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2(4).

Clark, K.; and Jaini, P. 2024. Text-to-image diffusion models are zero shot classifiers. In *Advances in Neural Information Processing Systems*, volume 36.

Deng, H.; Zhang, Z.; Bao, J.; and Li, X. 2023. Anovl: Adapting vision-language models for unified zero-shot anomaly localization. arXiv:2308.15939.

He, H.; Zhang, J.; Chen, H.; Chen, X.; Li, Z.; Chen, X.; et al. 2024. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8472–8480.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; and Dabeer, O. 2023. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kingma, D.; Salimans, T.; Poole, B.; and Ho, J. 2021. Variational diffusion models. *Advances in Neural Information Processing Systems*, 34: 21696–21707.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Li, Y.; Goodge, A.; Liu, F.; and Foo, C. S. 2024. Promptad: Zero-shot anomaly detection using text prompts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1093–1102.

Liang, Y.; Zhang, J.; Zhao, S.; et al. 2023. Omni-frequency channel-selection representations for unsupervised anomaly detection. *IEEE Transactions on Image Processing*.

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

Mishra, P.; Verk, R.; Fornasier, D.; Piciarelli, C.; and Foresti, G. L. 2021. VT-ADL: A Vision Transformer Network for Image Anomaly Detection and Localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, 01–06. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763. PMLR.

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

Schwartz, E.; Arbelle, A.; Karlinsky, L.; Harary, S.; Scheidegger, F.; Doveh, S.; and Giryes, R. 2024. MAEDAY: MAE for few-and zero-shot AnomalY-Detection. *Computer Vision and Image Understanding*, 241: 103958.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net.

Wang, R.; Zheng, H.; Duan, X.; Liu, J.; Lu, Y.; Wang, T.; Xu, S.; and Zhang, B. 2023a. Few-shot learning with visual distribution calibration and cross-modal distribution alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23445–23454.

Wang, Y.; Peng, J.; Zhang, J.; Yi, R.; Wang, Y.; and Wang, C. 2023b. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8032–8041.

Zhang, J.; Chen, X.; Xue, Z.; Wang, Y.; Wang, C.; and Liu, Y. 2023a. Exploring grounding potential of VQA-oriented GPT-4V for zero-shot anomaly detection. arXiv:2311.02612.

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

Zhang, X.; Li, N.; Li, J.; Dai, T.; Jiang, Y.; and Xia, S. T. 2023b. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6782–6791.

Zhou, Q.; Pang, G.; Tian, Y.; He, S.; and Chen, J. 2023. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. arXiv:2310.18961.

Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; and Dabeer, O. 2022. Spot-the-difference self-supervised pretraining for anomaly detection and segmentation. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, 392–408. Tel Aviv, Israel: Springer.