CADiff: Context-Aware Diffusion for Controllable Anomaly Generation in Anomaly Detection

Xuan Tong[1], Yuxuan Lin[2], Junxiong Lin[1], Xinji Mai[1], Haoran Wang[1], Zeng Tao[1], Yang Yao[3], Ruofan Wang[2], Wenqiang Zhang[1,2*]

[1]College of Intelligent Robotics and Advanced Manufacturing, Fudan University
[2]College of Computer Science and Artificial Intelligence, Fudan University
[3]University of Hong Kong
{xtong23, yuxuanlin24, linjx23, xjmai23, hrwang23, ztao19, rfwang23, wqzhang}@m.fudan.edu.cn,
yaoyangacademia@outlook.com

Abstract

Generating anomalies is a crucial method to enhance detection and classification performance by expanding anomalous data repository. However, existing anomaly generation methods overlook the intrinsic entanglement between diverse anomaly types and product structures, leading to semantic ambiguity. We propose CADiff, a context-aware generation framework that reframes anomalies as compositional perturbations. Firstly, we propose Context-aware Text Prompt (CTP), a mechanism which contains multiple tokens that characterize anomalies and products separately to enhance the contextual consistency of generated images and refine the local variability of anomalies. Secondly, we develop Self-adaptive Spatial Control (SSC), a self-adaptive interaction design that mitigates anomaly leakage or missing phenomena. Thirdly, we introduce Intensity-controllable Attention Re-weighting (IAR), an inference scheduling scheme with the ability to amplify or attenuate abnormal semantic effects to improve generation diversity. Extensive experiments on MVTec AD and VisA datasets demonstrate the superiority of our proposed method over state-of-the-art methods in both realism and diversity of the generated results, and significantly improve the performance of downstream tasks, including anomaly detection, anomaly localization, and anomaly classification tasks.

Introduction

Visual anomaly detection is essential for maintaining quality control and plays a crucial role in advancing industrial automation. However, due to the scarcity and difficulty of obtaining anomalous samples, the detection process faces significant challenges (Yu et al. 2024; Mei, Yang, and Yin 2018). Although existing methods (Chen et al. 2020; Li et al. 2021; Roth et al. 2022; Zavrtanik, Kristan, and Skočaj 2021a) rely on unsupervised learning, it remains difficult for them to build a robust detection system given the unpredictability of anomalies, and they can not deal with the task of anomaly classification (Hu et al. 2024). To address these challenges, recent studies (Zhang et al. 2023a, 2021; Duan et al. 2023; Hu et al. 2024) have demonstrated that generating highly realistic anomalous samples can significantly enhance the performance of detection algorithms.

Various studies have focused on high-quality anomaly generation as illustrated in Fig. 1. First, traditional methods randomly crop and paste random patterns from external datasets or itself to produce anomalies having different visual structures with the normal samples, which are unnatural and conspicuous (Zavrtanik, Kristan, and Skočaj 2021a; Li et al. 2021; Yang, Wu, and Feng 2023). Second, among methods utilizing Generative Adversarial Networks (GANs) (Goodfellow et al. 2020), DFMGAN employs StyleGANv2 (Duan et al. 2023) to generate anomalous images using anomaly-aware residual blocks, but the authenticity of the generated anomalies is lacking. Most recently, several models leverage prior information of pretrained Latent Diffusion Models (LDMs) like Stable Diffusion (SD) (Rombach et al. 2022). CAGen(Jiang et al. 2024), guided by natural texts, suffers from color and shape distortions due to semantic priors of pretrained U-Net (Ronneberger, Fischer, and Brox 2015) which is inherently biased toward natural scenes rather than industrial domains. AnomalyDiffusion(AnoDiff) (Hu et al. 2024) generates anomalies by learning anomalous appearance and positions separately. However, these diffusion-based methods all focus on modeling anomalous regions, neglecting their contextual coherence with the entire products. As shown by the attention maps, they tend to capture entire objects rather than distinguish anomalies from the product, leading to low-quality results such as weak responses to inconspicuous texture changes or boundary misalignment with backgrounds. Furthermore, they lack the flexibility to compose novel anomalies beyond test sets.

Motivated by this, we regard anomaly generation as the sparse and diverse perturbations over structurally consistent product surfaces, and propose CADiff to break their intrinsic entanglement and uncover the generative mechanism of anomalies. First, Context-aware Text Prompt (CTP) uses multiple learnable tokens to learn the features of products and anomalies separately, allowing the anomaly tokens to capture diverse anomalous semantics, while contextual tokens to learn overall normal product surface. We restrict localizing cross-attention masks to concentrate on learning coherent image regions and use contrastive loss to facilitate disentanglement of different tokens, which contributes

*Corresponding author.

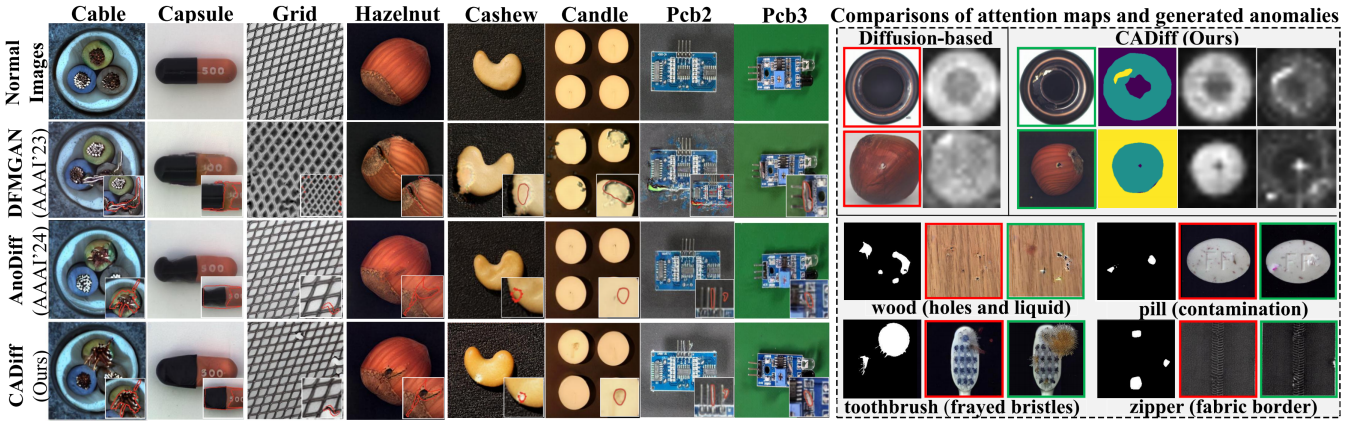

Figure 1: **Left:** Generated anomalies of CADiff and others on MVTec AD and VisA. Red lines indicate anomalous masks. **Right:** Comparisons of attention maps and generated combined anomalies between others *(Red boxes)* and ours *(green boxes)*.

to seamless blending of anomalies with the products. Second, we empirically found that relying on one-way guidance from ControlNet(Zhang, Rao, and Agrawala 2023), there are anomaly leakage beyond the masks to contaminate background or anomaly missing phenomenon, which leave the training process fragile. To strengthen spatial control of anomaly masks, we introduce Self-adaptive Spatial Control (SSC) based on cross-attention adaptive interaction. It involves extracting the key and value from the output of the zero-convolution layer and applying an adaptive fusion MLP to better interact with the global features in the main branch. Third, leveraging the disentangled anomalous entities, we propose Intensity-controllable Attention Re-weighting (IAR) to modulate the intensity of generated anomalies. It incorporates a delayed focus scheduler that dynamically determines the time to re-weight cross-attention maps, significantly enhancing generated diversity and authenticity. Our contributions are as follows:

- We propose CADiff, an anomaly generation framework that reveals the intrinsic anomaly–background entanglement underlying semantic ambiguity in generation. A Context-aware Text Prompt (CTP) is introduced to disentangle local anomalies from contextual normal products.
- We introduce Self-adaptive Spatial Control (SSC) and Intensity-controllable Attention Re-weighting (IAR). They involve a self-adaptive interaction module that mitigates anomaly leakage and missing problems, improving the realism of anomalies and training stability. IAR constructs anomalies with varying intensities, enhancing the diversity of anomalies to mimic natural distributions.
- Experiments on MVTec AD and VisA datasets demonstrate the superiority of our method over existing methods in terms of both generation quality and performance of downstream anomaly inspection tasks.

Related Work

Anomaly Detection. Detecting approaches are divided into reconstruction-based, embedding-based and generation-based. Firstly, reconstruction-based methods (Zavrtanik, Kristan, and Skočaj 2021a,b) reconstruct normal images in the training stage, assuming that the model would reconstruct anomalous images with a large error in the test stage to detect anomalies. However, these methods may also reconstruct anomalies well due to the strong ability of neural networks, thereby violating the underlying assumption. Secondly, embedding-based methods (Roth et al. 2022; Zhang et al. 2023b; Deng and Li 2022) usually use a pretrained network on ImageNet (Deng et al. 2009) to capture the high-level features of images. The anomaly score is calculated by measuring the distance between the test sample and normal samples in the feature space. But industrial image features are different from natural images, so that directly using pretrained features may cause a mismatch problem. Thirdly, generation-based methods (Zhang et al. 2021; Duan et al. 2023; Hu et al. 2024) generate anomalous images to simulate potential deviations from the normal distribution as negative samples, which help the network learn to recognize and differentiate anomalous patterns more effectively, demonstrating strong potential in downstream tasks.

Anomaly Generation. Traditional methods DRAEM (Zavrtanik, Kristan, and Skočaj 2021a) extract textures from DTD dataset (Sharan, Rosenholtz, and Adelson 2014), while CutPaste (Li et al. 2021) and NSA (Schlüter et al. 2022) augment samples by pasting unnatural patches from other training images. PRN (Zhang et al. 2023a) enhances detection performance by incorporating real anomalies into the data augmentation process. GAN-based models like SDGAN (Niu et al. 2020), Defect-GAN (Zhang et al. 2021), and DFMGAN (Duan et al. 2023) suffer from generation problems such as color inconsistencies and mode collapse. Recently, text-guided approaches have emerged with the advancement of LDMs. AnomalyDiffusion (Hu et al. 2024) fine-tunes LDMs, decoupling anomaly appearance and location through Textual Inversion (Gal et al. 2022) and a spatial encoder. RealNet (Zhang, Xu, and Zhou 2024) trains a denoising diffusion model (Ho, Jain, and Abbeel 2020) on normal samples, disrupts the normal denoising process during inference to generate anomalous images. CAGEN (Jiang et al. 2024) uses ControlNet (Zhang, Rao, and Agrawala

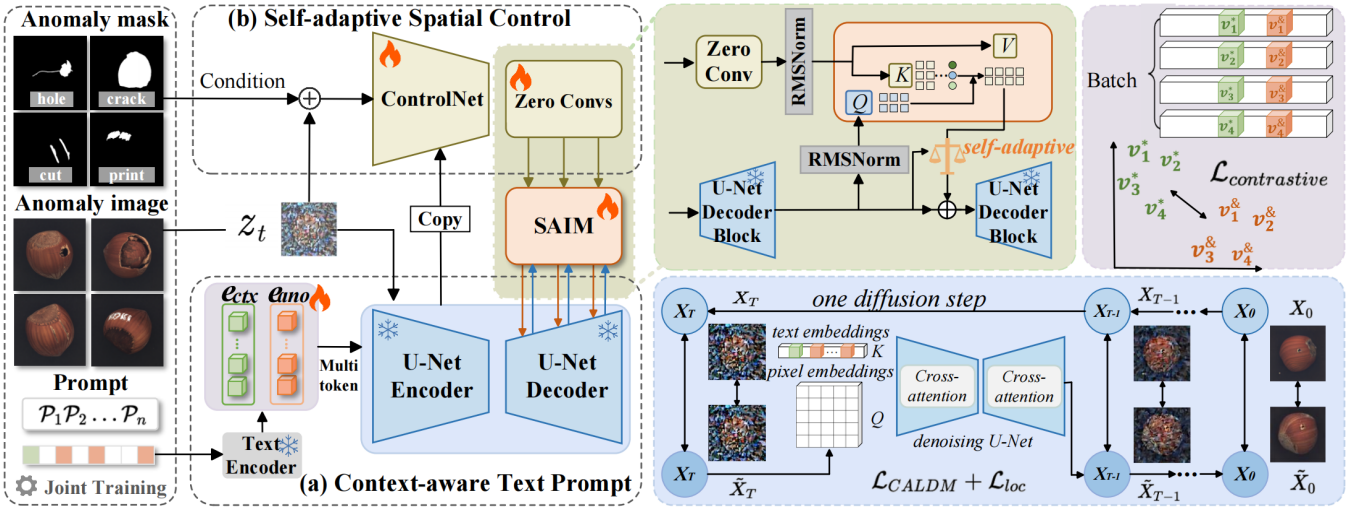

Figure 2: Overall framework of CADiff. (a) Context-aware Text Prompt: an adapted diffusion process for multi-token learning with two regularization terms; (b) Self-adaptive Spatial Control: integrating ControlNet with a Self-adaptive Interaction Module (SAIM) to modulate spatial features; (c) Intensity-controllable Attention Re-weighting: changing anomalous intensities in the inference stage as shown in Fig. 4.

2023) to control the regions where anomaly features are injected into normal images. Although they leverage the strong prior information of diffusion models, relying solely on a natural text prompt or a single learnable prompt does not fully harness the potential of LDMs (Rombach et al. 2022).

Method

As shown in Fig. 2, first, our method CADiff employs a multi-token learning scheme to disentangle conditioned text into contextual and anomaly tokens, incorporating two regularization terms to constrain and facilitate the learning process of a modified diffusion framework. Second, to precisely control the locations of generated anomalies, we introduce a self-adaptive interaction module within ControlNet, which integrates learned textual embeddings with spatial masks to dynamically determine the flow of these two information streams. Finally, leveraging the decoupled tokens, we propose strategies of masked focus attention and delayed focus scheduler to re-weight the cross-attention maps of a designated token, enabling interpretable modulation of anomalous intensity to enhance generation diversity.

Preliminaries

Stable Diffusion. Stable Diffusion consists of a variational auto-encoder (VAE), U-Net, and text encoder. The VAE encoder ε compresses the image x to a latent representation z, which is perturbed by Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ in the forward diffusion process. The U-Net, parameterized by θ, denoises the noisy latent representation by predicting the noise. This denoising process can be conditioned on text prompts y encoded by text encoder τ_θ. The training process is to minimize the loss function below:

$$L_{\text{LDM}} = \mathbb{E}_{\varepsilon(x),\epsilon,y,t}\left[||\epsilon - \epsilon_\theta\left(\varepsilon\left(x\right), t, \tau_\theta\left(y\right)\right)||_2^2\right] \quad (1)$$

Textual inversion in anomaly generation. In few-shot anomaly generation task, given a textual embedding v learned by Textual Inversion (Gal et al. 2022) from q real-world anomalous images I_q, which is then injected into diffusion model as anomalous knowledge.

$$v^* = \arg\min_v \mathbb{E}_{\varepsilon(x),\epsilon,y,t}\left[||\epsilon - \epsilon_\theta\left(\varepsilon\left(I_q^t\right), t, \tau_\theta\left(y\right)\right)||_2^2\right] \quad (2)$$

Text condition in cross-attention mechanism. The encoder converts text y into d-dimensional embeddings $\tau_\theta\left(y\right) = c \in \mathbb{R}^{n\times d}$. The cross-attention layers take image embeddings $z \in \mathbb{R}^{(h\times w)\times f}$ and text embeddings c as inputs. Here, $Q = W^q z, K = W^k c,$ and $V = W^v c$ are acquired using learned linear layers $W^q \in \mathbb{R}^{f\times d'}, W^k, W^v \in \mathbb{R}^{d\times d'}$. Each cross-attention layer is represented as follows. $A\left[i,j,k\right] \in [0,1]^{(h\times w)\times n}$ represents the amount of information flow from the k-th text token to the (i,j) latent pixel.

$$\text{attn}\left(Q,K,V\right) = A\cdot V = \text{soft}\max\left(\frac{QK^\top}{\sqrt{d'}}\right)\cdot V \quad (3)$$

Context-aware Text Prompt

Multiple learnable text tokens. Our framework CADiff learns a set of tokens $\mathcal{P} = \left[p^*, p^{\&}, ..., p^{@}\right]$, which are initialized by the text encoder τ_θ as embedding vectors $\mathcal{V} = \left[v^*, v^{\&}, ..., v^{@}\right]$. In specific, we disentangle learnable tokens into contextual tokens $e_{cxt} = \left[v^*\right]$ and anomaly tokens $e_{ano} = \left[v^{\&}, ..., v^{@}\right]$, which represent a single product and diverse anomaly types. The optimization process is still guided by the image-level LDM, but now updating while keeping τ_θ and ϵ_θ frozen (Gal et al. 2022). Context-aware learning loss of multiple tokens can be denoted as follows:

$$\mathcal{L}_{\text{CALDM}} = \mathbb{E}_{\varepsilon(x),\epsilon,y,t,\mathcal{V}}\left[||\epsilon - \epsilon_\theta\left(\varepsilon\left(I_q^t\right), t, \mathcal{V}\right)||_2^2\right] \quad (4)$$

CADiff enables the simultaneous learning of multiple tokens representing different concepts within a prompt.

Localizing masked loss. Cross-attention maps have the potential to delineate the location and outline of a target entity (Hertz et al. 2022). However, a single latent pixel can attend to all text tokens, resulting in cluttered attention maps. To address this, we strive to generate attention maps with the following constraints. Let $\mathcal{M}^l = \{M^{*,l}, M^{\&,l}, ..., M^{@,l}\}$ represent the binary masks of products and anomalies in the l-th layer of U-Net by downsampling them to different scales, $\mathcal{I} = \{i_1, i_2, ..., i_n\}$ be the index list indicating which anomaly type corresponds to a token in the text prompt, and $A_t^{i,l} = A_t^l[:,:,i] \in [0,1]^{(h \times w)}$ be the cross-attention map of the i-th token at time step t with $h \times w$ resolution. We compute the average attention maps for each learnable token $p \in \mathcal{P}$ over all time steps $A^{i,l} = 1/T \sum_{t=1}^{T} A_t^{i,l}$. A balanced L2 loss is employed to minimize the distance between the cross-attention map and the segmentation mask:

$$\mathcal{L}_{\text{loc}} = \sum_{l=1}^{L} \| \frac{1}{n} \sum_{i=1}^{n} \left(A^{i,l} - M^{i,l} \right) \|^2 \tag{5}$$

where L is the total number of U-Net layers, n is number of learnable tokens, and M^* is a reverse mask of the sum of the other masks in $\mathcal{M}$.

Contrastive disentanglement loss. To accelerate the separation of different tokens, we use the most widely adopted training objective NT-Xent loss (Chen et al. 2020). At each learning step, we randomly sample a mini-batch of B augmented example images and suppose there are C learnable embedding vectors for every example, resulting in BC data points. We represent the contrastive learning task as $\{v_b^c\}_{b=1}^{B}{}_{,c=1}^{C}$. Our goal is to disentangle each token to represent distinct concepts, encompassing both the separation of contextual information from anomalies and the finer granularity among different anomaly tokens. Therefore, we regard v_i^c and v_j^c as positives sharing a same concept $c \in C$ and other different concepts in the same mini-batch as negatives. Then NT-Xent loss between a pair can be formulated as follows:

$$l_{i,j}^c = -\ln\left(\frac{\exp\left(\cos\left(v_i^c, v_j^c\right)/\tau\right)}{\sum_{c=1}^{C} \sum_{j=1}^{B} 1_{[j \neq i]} \exp\left(\cos\left(v_i^c, v_j^c\right)/\tau\right)} \right) \tag{6}$$

For all positive pairs in a mini-batch of size N, the contrastive loss is defined as:

$$\mathcal{L}_{contrastive} = \frac{1}{C} \cdot \frac{1}{B} \sum_{c=1}^{C} \sum_{i=1}^{B} \sum_{j=1}^{B} l_{i,j}^c \tag{7}$$

where $\cos(\cdot,\cdot)$ is the cosine similarity, $1_{[j \neq i]} \in \{0,1\}$ is an indicator function evaluating to 1 if $j \neq i$ and τ denotes a temperature parameter.

Joint training. In the training process, rather than training separate U-Net models for each anomaly type, we employ joint training for each product, using a unified U-Net model across all anomaly types. We found that such a joint training strategy not only mitigates the overfitting caused by the limited number of each anomaly type, but also increases the diversity of the generated anomaly images. To integrate overall losses, we introduce factors γ and β to regulate the training of context-aware diffusion process:

$$\mathcal{L} = \mathcal{L}_{\text{CALDM}} + \gamma \mathcal{L}_{loc} + \beta \mathcal{L}_{contrastive} \tag{8}$$

Self-adaptive Spatial Control

With the multiple text tokens learning the image semantics and binary masks indicating the spatial locations of anomalies, it becomes intuitive to introduce ControlNet (Zhang, Rao, and Agrawala 2023) for improved control of diffusion process. But experiments showed that this straightforward application encountered problems: 1) Over-training the learnable text tokens leads to anomaly leakage into the background or missing anomalies due to contextual intrusion into masked anomaly regions; 2) while under-training generates blurred and unclear edges of anomalies. The results in Fig. 4a indicate that the interaction between textual prompts and spatial localization control is insufficient under the guidance of ControlNet.

To address this issue, we design a Self-Adaptive Interactive Module (SAIM). In the ControlNet, each locked block corresponds to a trainable copy block with a zero-convolution layer. We respectively denote the outputs of a locked block and a trainable copy block after zero-convolution as y and y_c. Different from the original ControlNet, which directly concatenates them, SAIM injects an additional cross-attention layer, to better integrate the textual and spatial information across two branches. RMSNorm (Zhang and Sennrich 2019) is applied to enable rapid convergence.

$$y_a = \text{CrossAttn}(\text{RMSNorm}(y), \text{RMSNorm}(y_c)) \tag{9}$$

Furthermore, we apply a multi-layer perceptron (MLP) to adaptively weight how much information from the cross-attention output y_a should be used. After applying the weighting to integrate y_a and y, we concatenate the summation with y as the output of this whole module y_f.

$$y_f = \text{concat}\left(y,\ y + \text{MLP}\left(y_a\right)\right) \tag{10}$$

Intensity-controllable Attention Re-weighting

Given that our multi-token learning has disentangled text embeddings into anomalous parts and contextual background information. It is available to generate anomalous images with varying anomaly intensities, in order to control the difficulty of network training in downstream tasks. For example, we consider the product category "hazelnut" and anomaly type "print". Fig. 4b presents a real anomaly and our generated image, and we want to make the anomalous semantics "print" more concealed or clearer. We scale the cross-attention map values of the corresponding token $v^{\&}$ of "print" with parameter $w \in [-2, 2]$ (Hertz et al. 2022).

However, our empirical findings indicate that such naive iterative re-weighting results in uneven enhancement effects. This occurs because the attention maps are distributed and unrefined in the early phases of the denoising process (Fig. 4c), gradually becoming more localized and focused in the later stages as noise diminishes. Premature re-weighting disrupts this process by unintentionally amplifying regions outside the conditional mask while failing to adequately enhance less noticeable areas within the mask. To address this problem, first, we propose Masked Focus Attention (MFA), providing explicit masks to restrict the attention re-weighting regions. Second, we develop a simple Delayed

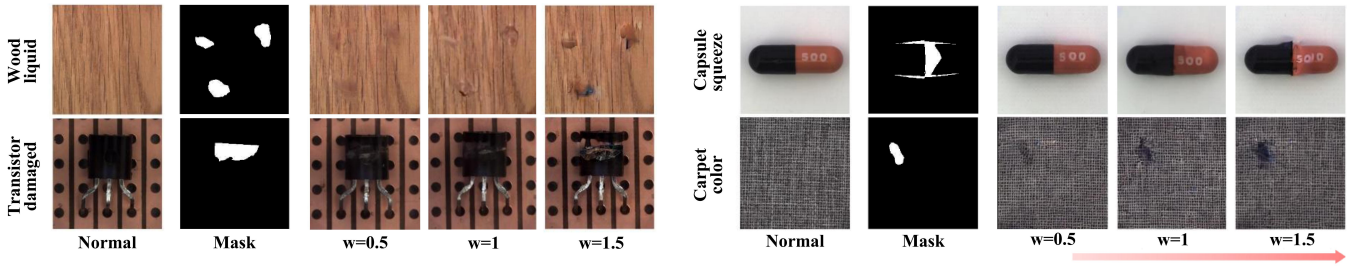

Figure 3: Anomalous images generated with different intensities, demonstrating how the adjustable factor w influences the prominence of anomalies, with higher w amplifying the anomalies.

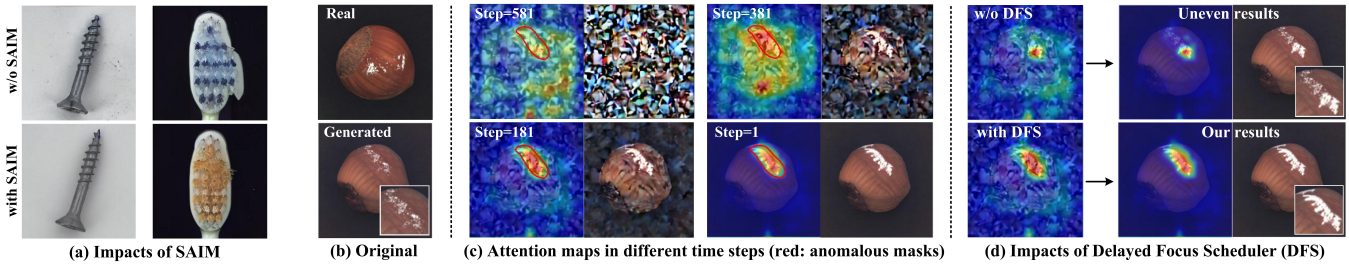

Figure 4: **Left:** Comparison of generated anomalies with and without SAIM; **Right:** Examples of amplified "print" anomalies, with and without the applications of Masked Focus Attention (MFA) and Delayed Focus Scheduler (DFS).

Focus Scheduler (DFS). Given the attention map $A_t^{\&}$ at time step t and the mask $M^{\&}$ with 16×16 resolution (Chefer et al. 2023), we calculate the number of activated pixels n_t above the average value in the attention map within an anomaly mask, and the pixel number of the mask is n_{start}. When the degree of attention focus $\alpha_t = n_t/n_{start}$ reaches a threshold α_{thr}, we perform attention re-weighting as follows:

$$\left(A_t^{\&}\right)' = \begin{cases} w \cdot \left\|M^{\&}\right\|_1 \cdot A_t^{\&} & \text{if } \alpha_t > \alpha_{thr} \\ A_t^{\&} & \text{otherwise} \end{cases} \tag{11}$$

Fig. 4d demonstrates that applying all strategies yields the best results. Additional results are shown in Fig. 3.

Experiment

Datasets. We conduct experiments on the widely used MVTec AD (Bergmann et al. 2019) and VisA (Zou et al. 2022) datasets. During training, we follow the settings of AnomalyDiffusion (Hu et al. 2024), using only the first third of the anomaly samples.

Evaluation Metrics. We classify indicators into generation and performance metrics. Generation metrics evaluate the quality and diversity of generated images. Inception Score (IS) measures image quality, while Intra-cluster Pairwise LPIPS (IC-LPIPS) assesses diversity. Performance metrics, including Area Under the Receiver Operating Characteristic curve (AUROC), Average Precision (AP) and the F1-max score (Duan et al. 2023), measure the effectiveness of generated images in anomaly detection and localization tasks.

Implementation Details. We fine-tuned the pretrained Stable Diffusion v1.5 model using ControlNet model and retain the original hyper-parameter choices (Zhang, Rao, and Agrawala 2023). Firstly for training, we train our models for each product over a total of 1000 epochs on a single V100 GPU, with a constant learning rate of 1e-5 and a batch size of 16. The dimension of learnable embeddings are 768 and they are initialized as "anomaly" to avoid the intrinsic error of tokenizer. When calculating overall loss, we apply $\alpha_{thr} = 0.75$, the coefficients of loss terms $\gamma = 0.001$ and $\beta = 0.0005$. We perform random cropping, translation, and rotation on both the image and its corresponding mask to get augmented views. Secondly during inference, we adopt DDIM (Nichol and Dhariwal 2021) sampling method with the denoising step $T = 50$ to generate 1,000 images per anomaly type and train corresponding U-Net (Ronneberger, Fischer, and Brox 2015) for comparative evaluation.

Comparison with State-of-the-art

Among the generation-based methods, we mainly compare CADiff with representative DRAEM (Zavrtanik, Kristan, and Skočaj 2021a), PRN (Zhang et al. 2023a), DFM-GAN (Duan et al. 2023) and AnomalyDiffusion (Hu et al. 2024). Their detection and localization results on MVTec AD dataset are shown in Tab. 3. It can be observed that the U-Net trained in our generated data reaches the highest Image AUROC of **99.5%**, Pixel AUROC of **99.3%**, Pixel AP of **84.6%**, surpassing the second-ranked by a margin of **3.2%** (AP). For VisA dataset, we outperform AnoDiff across all metrics. We visualize the pixel-wise predictions in Fig. 5, which exhibits our remarkable anomaly localization.

Generation metrics on MVTec AD dataset are demonstrated in Tab. 1. CADiff achieves **1.84** on IS and **0.35** on IC-LPIPS, showing that our method generates anomalies with the highest fidelity and diversity. We present generated anomalous images on two datasets in Fig. 5 and Fig. 1. It is

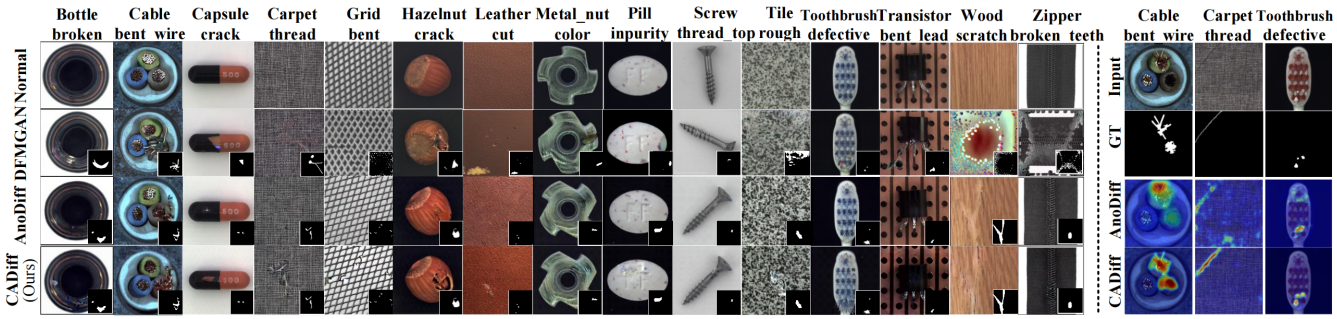

Figure 5: **Left:** Generation results on MVTec AD. The sub-image in the lower right corner is the generated mask, among which AnoDiff and our method share identical masks for comparisons. **Right:** Qualitative results of anomaly localization.

Metric	Method	bottle	cable	capsule	carpet	grid	hazelnut	leather	metal nut	pill	screw	tile	toothbrush	transistor	wood	zipper	**Average**
IS↑	DFMGAN	1.62	1.96	1.59	1.23	1.97	1.93	**2.06**	1.49	1.63	1.12	2.39	**1.82**	**1.64**	2.12	1.29	1.72
	AnoDiff	1.58	2.13	1.59	1.16	2.04	**2.13**	1.94	**1.96**	1.61	1.28	**2.54**	1.68	1.57	2.33	1.39	1.80
	Ours	**1.91**	**2.15**	**1.60**	**1.25**	**2.05**	1.92	1.87	1.65	**1.63**	**1.29**	2.50	1.72	1.62	**2.38**	**2.05**	**1.84**
IC-L↑	DFMGAN	0.12	0.25	0.11	0.13	0.13	0.24	0.17	**0.32**	0.16	0.14	0.22	0.18	0.25	0.35	0.27	0.20
	AnoDiff	0.19	0.41	0.21	0.24	**0.44**	0.31	**0.41**	0.30	0.26	0.30	**0.55**	0.21	**0.34**	0.37	0.25	0.32
	Ours	**0.29**	**0.45**	**0.29**	**0.25**	0.40	**0.34**	0.34	0.31	**0.28**	**0.32**	0.52	**0.32**	0.33	**0.42**	**0.34**	**0.35**

Table 1: Generation Metrics of IS and IC-LPIPS (IC-L) indicators on MVTec AD. Bold represents the best results.

observed that DFMGAN (Duan et al. 2023) collapses into identical masks (e.g., *wood_scratch*) and introduces noise to the images (e.g., *zipper_broken_teeth*). For AnoDiff (Hu et al. 2024), under the control of identical anomaly masks, it shows worse alignment with the masks than ours, especially when encountered with irregular mask shapes (e.g., *cable_bent_wire*) and subtle texture anomalies (e.g., *grid_bent*).

Performance metrics on MVTec AD dataset are shown in Tab. 2, we compare it with the advanced reconstruction-based and embedding-based anomaly detection methods, including RD4AD (Deng and Li 2022), PatchCore (Roth et al. 2022), CFA (Lee, Lee, and Song 2022), DRAEM (Zavrtanik, Kristan, and Skočaj 2021a), RealNet (Zhang, Xu, and Zhou 2024), SSPCAB (Ristea et al. 2022), DeSTSeg (Zhang et al. 2023b), DevNet (Pang et al. 2021), DRA (Ding, Pang, and Shen 2022), and PRN (Zhang et al. 2023a). Furthermore, we train an anomaly classification model following (Hu et al. 2024). CADiff achieves an average accuracy improvement of **7.38%** compared to the second-ranked, outperforming others in the most categories.

Ablation Study

We evaluate our key components in Tab. 5. For CTP, we compare the differences between single-token learning of Textual Inversion (Gal et al. 2022) on the masked anomalous regions and our multi-token learning (MTL) scheme. In Fig. 6, we investigate the disentanglement of learned embeddings with t-SNE. A single learned token produces scattered distributions, indicating its inability to differentiate between anomalous semantics across products, as reflected in attention maps that capture entire objects. In contrast, multiple tokens form distinct clusters revealing an effective decoupling of heterogeneous anomalous concepts. Tab. 5 shows that our MTL combined with all regularization (Reg) terms

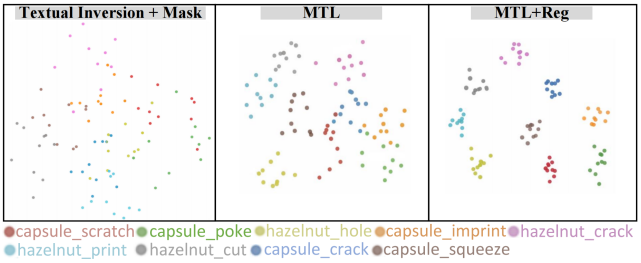

Figure 6: The t-SNE projection of learned embeddings of different categories (e.g., hazelnut_print, capsule_crack).

improves the Pixel AUROC from 92.8% to 98.6%, illustrating this strategy can most effectively distinguish all learned concepts. For SSC, removing the self-adaptive interaction module (SAIM) leads to a sharp decrease in all metrics.

We further investigate the impact of anomaly intensity factor w, Masked Focus Attention (MFA) and Delayed Focus Scheduler (DFS) in Tab. 4. Specifically, values of $w < 1$ attenuate anomalous effects, posing greater challenges for the detection network in identifying faint anomalies. However, overly small values of w make the anomaly regions nearly imperceptible, causing potential misalignment between generated anomalies and masks, thereby compromising the accuracy of supervision signals. Tab. 4 shows that using only the generated images of $w = 0.5$ results in a 1.2% reduction in Pixel AP compared to $w = 1$. Conversely, $w > 1$ amplifies anomalous effects, but excessive amplification sharpens edges, leading the detection network to rely on superficial shortcuts rather than capturing the true underlying anomaly patterns, which is demonstrated unfavorable. Our experiments indicate that uniformly sampling $w \in [0.5, 1.5]$ is effective and largely improves generation

Category	Unsupervised							Supervised			
	RD4AD	PatchCore	CFA	DRAEM	RealNet	SSPCAB	DeSTSeg	DevNet	DRA	PRN	**Ours**
bottle	98.8/51.0	97.6/75.0	98.9/50.9	99.1/88.5	99.2/86.9	98.9/88.6	98.2/89.3	96.7/67.9	91.7/41.5	**99.4**/92.3	99.2/**95.5**
cable	98.8/77.0	96.8/65.9	98.4/79.8	94.8/61.4	97.8/55.6	93.1/52.1	98.1/61.4	97.9/67.6	86.1/34.8	98.8/78.9	**99.5/91.0**
capsule	99.0/60.5	98.6/46.6	98.9/**71.1**	97.6/47.9	**99.3**/59.2	90.4/48.7	99.2/57.3	91.1/46.6	88.5/11.0	98.5/62.2	98.8/62.9
carpet	**99.4**/46.0	98.7/65.0	99.1/47.7	96.3/62.5	98.8/59.1	92.3/49.1	96.4/72.9	94.6/19.6	98.2/54.0	99.0/82.0	99.3/**82.1**
grid	98.0/75.4	97.2/23.6	98.6/**82.9**	99.5/53.2	99.6/62.0	**99.6**/58.2	98.8/60.3	90.2/44.9	86.2/28.6	98.4/45.7	98.9/57.2
hazelnut	94.2/57.2	97.6/55.2	98.5/80.2	99.5/88.1	99.5/76.9	99.6/**94.5**	99.5/88.4	76.9/46.8	88.8/20.3	99.7/93.8	**99.9**/92.3
leather	96.6/53.5	98.9/43.4	96.2/60.9	98.8/68.5	99.7/71.3	97.2/60.3	99.4/75.5	94.3/66.2	97.2/ 5.1	99.7/69.7	**99.8/85.7**
metal_nut	97.3/53.8	97.5/86.8	98.6/74.6	98.7/91.6	98.1/70.2	99.3/95.1	98.7/93.6	93.3/57.4	80.3/30.6	99.7/98.0	**99.7/98.0**
pill	98.4/58.1	97.0/75.9	98.8/67.9	97.7/44.8	98.8/78.1	96.5/48.1	98.6/82.9	98.9/79.9	79.6/22.1	99.5/91.3	**99.8/96.3**
screw	99.1/51.8	98.7/34.2	98.7/61.4	**99.7/72.9**	99.5/44.9	99.1/62.0	98.1/58.5	66.5/21.1	51.0/ 5.1	97.5/44.9	98.4/49.4
tile	97.4/78.2	94.9/56.0	98.6/92.6	99.4/96.4	99.3/92.4	99.2/96.3	98.2/90.4	88.7/63.9	91.0/54.4	**99.6**/96.5	99.4/**98.0**
toothbrush	99.0/63.1	97.6/37.1	98.4/61.7	97.3/49.2	98.8/64.6	97.5/38.9	99.2/75.1	96.3/52.4	74.5/ 4.8	**99.6**/78.1	99.5/**85.4**
transistor	**99.6**/50.3	91.8/66.7	98.6/82.9	92.2/56.0	98.0/68.8	85.3/36.5	89.2/64.8	55.2/ 4.4	79.3/11.2	98.4/85.6	99.5/**95.5**
wood	**99.3**/39.1	95.7/54.3	97.6/25.6	97.6/81.6	98.1/71.1	97.2/77.1	97.7/81.8	93.1/47.9	82.9/21.0	97.8/82.6	99.0/**91.3**
zipper	**99.7**/52.7	98.5/63.1	95.9/53.9	98.6/73.6	99.0/65.9	98.1/78.2	98.9/85.1	92.4/53.1	96.8/42.3	98.8/77.6	99.5/**87.8**
Average	98.3/57.8	97.1/56.6	98.3/66.3	97.7/69.0	98.9/68.5	96.2/65.5	97.9/75.8	86.4/49.3	84.8/25.7	99.0/78.6	**99.3/84.6**

Table 2: Comparison on pixel-level anomaly localization (AUROC / AP) between the simple U-Net trained on our generated dataset and the existing anomaly detection methods with their official codes or pretrained models.

Method	AUC-P	AP-P	F1-P	AUC-I	AP-I
DREAM	92.2/89.8	54.1/32.9	53.1/32.3	94.6/89.9	97.0/87.4
PRN	96.9/91.5	66.2/40.3	64.7/42.3	91.6/88.5	96.6/86.7
DFMGAN	90.0/88.7	62.7/38.1	62.1/40.6	87.2/86.8	94.8/86.4
AnoDiff	99.1/97.0	81.4/42.9	76.3/45.8	99.2/90.1	99.7/88.9
Ours	**99.3/99.0**	**84.6/65.0**	**80.5/62.8**	**99.5/95.7**	**99.8/96.1**

Table 3: Comparison of pixel-level (-P) and image-level (-I) AUROC, AP, and F1-max on MVTec AD (left) / VisA (right) datasets using a U-Net trained on generated data.

Metric	$w = 0.5$	$w = 1$	$w = 1.5$	$w \in [0.5, 1.5]$	w/o MFA	w/o DFS
AP-P	79.9	81.1	80.2	**84.6**	81.4	82.7
AUC-P	97.2	97.9	97.4	**99.3**	97.3	98.2
AUC-I	98.0	98.2	98.1	**99.5**	98.2	98.7
IC-L	0.22	0.28	0.26	**0.35**	0.31	0.29

Table 4: Impacts of anomaly intensity factor w, MFA and DFS.

MTL	Reg	SAIM	AUC-P	AP-P	AUC-I	IS	IC-L
			92.8	72.9	90.5	1.68	0.19
✓			97.9	78.1	96.5	1.76	0.25
✓	✓		98.6	80.9	99.2	1.80	0.28
✓		✓	98.8	82.4	98.9	1.79	0.31
✓	✓	✓	**99.3**	**84.6**	**99.5**	**1.84**	**0.35**

Table 5: The impact of CTP and SSC on CADiff.

diversity in IC-LPIPS, as further decrease or increase in w yields changes in visual quality and detection performance.

Applications

By revealing the decoupled nature between anomalies and products during generation, such a characteristic can be extended to logical anomalies (Bergmann et al. 2022) and medical domains (Johnson et al. 2019). As shown in Fig. 7, logical anomalies inherently involve compositional semantics. For instance, by disentangling the concept of fruits from a meal box and amplifying it, we simulate a quantity-related anomaly with wrong number of fruits. In medical imagery where anatomical structures are naturally delineated, we disentangle smoky consolidation from the lung and attenuate its effect to restore a healthy appearance. Such a disentangled formulation highlights how concept-level control enables bidirectional editing of normal and anomalous features, supporting broader applications in structured domains that demand semantic manipulation or segmentation.

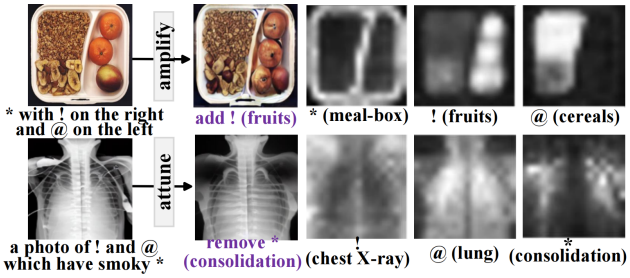

Figure 7: Applications on logical and medical anomalies.

Conclusion

We propose Context-aware Diffusion (CADiff), an algorithm to generate high-quality anomalies, tailored for downstream multiple anomaly inspection tasks. CADiff transcends the limitation of solely focus on anomalous regions in the previous works, with a disentanglement of contextual and anomaly tokens to explore intrinsic relations of global uniform product surfaces and localized variable anomalies. It further allows self-adaptive spatial control of anomalous locations to achieve seamless integration with products. Leveraging the decoupled concepts, CADiff enables explainable adjustment of anomalous effects to enhance generation diversity. Extensive experiments demonstrate CADiff's effectiveness in tackling a wide range of real-world anomaly detection scenarios.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No.62576109, 62072112)

References

Bergmann, P.; Batzner, K.; Fauser, M.; Sattlegger, D.; and Steger, C. 2022. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130(4): 947–969.

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/CVF conference on computer vision and pattern recognition*, 9592–9600.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4): 1–10.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Deng, H.; and Li, X. 2022. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9737–9746.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Ding, C.; Pang, G.; and Shen, C. 2022. Catching both gray and black swans: Open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7388–7398.

Duan, Y.; Hong, Y.; Niu, L.; and Zhang, L. 2023. Few-shot defect image generation via defect-aware feature manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 571–578.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, T.; Zhang, J.; Yi, R.; Du, Y.; Chen, X.; Liu, L.; Wang, Y.; and Wang, C. 2024. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8526–8534.

Jiang, B.; Xie, Y.; Li, J.; Li, N.; Jiang, Y.; and Xia, S.-T. 2024. CAGEN: Controllable Anomaly Generator using Diffusion Model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3110–3114. IEEE.

Johnson, A. E.; Pollard, T. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Peng, Y.; Lu, Z.; Mark, R. G.; Berkowitz, S. J.; and Horng, S. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Lee, S.; Lee, S.; and Song, B. C. 2022. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10: 78446–78454.

Li, C.-L.; Sohn, K.; Yoon, J.; and Pfister, T. 2021. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9664–9674.

Mei, S.; Yang, H.; and Yin, Z. 2018. An unsupervised-learning-based approach for automated defect inspection on textured surfaces. *IEEE transactions on instrumentation and measurement*, 67(6): 1266–1277.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, 8162–8171. PMLR.

Niu, S.; Li, B.; Wang, X.; and Lin, H. 2020. Defect image sample generation with GAN for improving defect recognition. *IEEE Transactions on Automation Science and Engineering*, 17(3): 1611–1622.

Pang, G.; Ding, C.; Shen, C.; and Hengel, A. v. d. 2021. Explainable deep few-shot anomaly detection with deviation networks. *arXiv preprint arXiv:2108.00462*.

Ristea, N.-C.; Madan, N.; Ionescu, R. T.; Nasrollahi, K.; Khan, F. S.; Moeslund, T. B.; and Shah, M. 2022. Self-supervised predictive convolutional attentive block for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13576–13586.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14318–14328.

Schlüter, H. M.; Tan, J.; Hou, B.; and Kainz, B. 2022. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *European Conference on Computer Vision*, 474–489. Springer.

Sharan, L.; Rosenholtz, R.; and Adelson, E. H. 2014. Accuracy and speed of material categorization in real-world images. *Journal of vision*, 14(9): 12–12.

Yang, M.; Wu, P.; and Feng, H. 2023. MemSeg: A semi-supervised method for image surface defect detection using differences and commonalities. *Engineering Applications of Artificial Intelligence*, 119: 105835.

Yu, J.; Chi, C. B.; Fichera, S.; Paoletti, P.; Mehta, D.; and Luo, S. 2024. Multi-class Road Defect Detection and Segmentation using Spatial and Channel-wise Attention for Autonomous Road Repairing. *arXiv preprint arXiv:2402.04064*.

Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2021a. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8330–8339.

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

Zhang, B.; and Sennrich, R. 2019. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32.

Zhang, G.; Cui, K.; Hung, T.-Y.; and Lu, S. 2021. Defect-GAN: High-fidelity defect synthesis for automated defect inspection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2524–2534.

Zhang, H.; Wu, Z.; Wang, Z.; Chen, Z.; and Jiang, Y.-G. 2023a. Prototypical residual networks for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16281–16291.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, X.; Li, S.; Li, X.; Huang, P.; Shan, J.; and Chen, T. 2023b. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3914–3923.

Zhang, X.; Xu, M.; and Zhou, X. 2024. RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16699–16708.

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