This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Unseen Visual Anomaly Generation

Han Sun
EPFL, Switzerland

Yunkang Cao
HUST, China

Hao Dong
ETH Zurich, Switzerland

Olga Fink
EPFL, Switzerland

{han.sun, olga.fink}@epfl.ch, cyk_hust@hust.edu.cn, hao.dong@ibk.baug.ethz.ch

Abstract

*Visual anomaly detection (AD) presents significant challenges due to the scarcity of anomalous data samples. While numerous works have been proposed to synthesize anomalous samples, these synthetic anomalies often lack authenticity or require extensive training data, limiting their applicability in real-world scenarios. In this work, we propose **Anomaly Anything** (AnomalyAny), a novel framework that leverages Stable Diffusion (SD)'s image generation capabilities to generate diverse and realistic unseen anomalies. By conditioning on a single normal sample during test time, AnomalyAny is able to generate unseen anomalies for arbitrary object types with text descriptions. Within AnomalyAny, we propose attention-guided anomaly optimization to direct SD's attention on generating hard anomaly concepts. Additionally, we introduce prompt-guided anomaly refinement, incorporating detailed descriptions to further improve the generation quality. Extensive experiments on MVTec AD and VisA datasets demonstrate AnomalyAny's ability in generating high-quality unseen anomalies and its effectiveness in enhancing downstream AD performance. Our demo and code are available at https://hansunhayden.github.io/CUT.github.io/.*

1. Introduction

Visual anomaly detection (AD) [5] aims to identify unusual or unexpected patterns in image data that deviate from a normal distribution of features, and is essential in fields such as industrial inspection and quality control [3]. Since anomalies are rare and difficult to collect, most existing AD methods rely on unsupervised learning [12, 22, 32] using solely normal samples. Despite recent advancements in the AD field, the scarcity of anomalous samples for training remains a persistent challenge. To address this, various studies have explored visual anomaly generation, as illustrated in Fig. 1. Some approaches [33, 35] augment normal samples by cropping and pasting random patterns, either from natural patterns in other datasets or from the images themselves. While this generates diverse anomalous samples and is applicable to unseen objects without training, these samples often appear unrealistic. Another approach involves generating visual anomalies using generative models [17], which produce more realistic images but require sufficient and representative normal and abnormal samples for training, which is typically challenging to obtain for AD. Given the rarity and variability of anomalies, collecting a representative set of anomalous samples is difficult. Additionally, the diversity of product variants and configurations in industrial applications often results in a lack of representative normal samples [18], presenting an additional challenge. These limitations lead to a shortage of both abnormal and normal samples. Consequently, generative models are often not suitable for real-world applications with limited data and may be biased towards the few available training observations.

Given the limitations of existing visual anomaly generation methods, we aim to generate diverse and realistic unseen anomalies using minimal normal data *without* any anomalous samples. The goal motivates us to explore Stable Diffusion (SD) [27], a latent text-to-image diffusion model known for generating diverse images across various domains. Although SD demonstrates impressive image generation capabilities, it is not specifically designed for visual anomaly generation. Consequently, when applied directly for anomaly generation, SD may produce images that deviate from the desired normal distribution and fail to accurately represent realistic anomaly patterns (Fig. 3 (b)). Prior approaches [17, 35] suggest fine-tuning SD on available normal or anomalous samples, but this approach is constrained in data scarcity scenarios and can compromise SD's ability to generalize to unseen data and anomaly types.

In this work, we propose **Anomaly Anything** (AnomalyAny) for realistic and diverse unseen visual anomaly generation. We leverage SD for unseen visual anomaly generation with anomaly text descriptions, which can be defined manually or automatically acquired from GPT-4 [24] with object category as input. Instead of fine-tuning SD on normal data, we introduce ***test-time normal sample condi-***

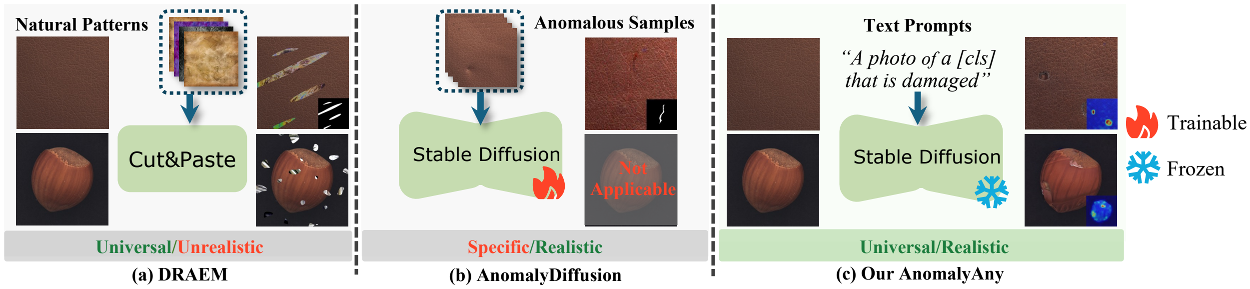


Figure 1. **Comparison between different visual anomaly generation methods.** In comparison to existing methods, AnomalyAny can generate diverse and realistic unseen anomalies without training.

tioning, guiding the generation process on a single normal sample. This approach preserves the diversity and generalization ability of SD, enabling anomaly generation on novel data and new anomaly types. With a single model, we can generate realistic and diverse anomalous samples for any new objects and from arbitrary anomaly descriptions. We observe that the original SD struggles to generate realistic anomaly samples due to two inherent challenges: first, anomalies are comparatively rare in SD's training data; second, unlike common objects and patterns, anomalous patterns typically occupy only a small region of the image, making them easily overlooked in generation. To address this, we propose ***attention-guided anomaly optimization*** to enforce the model's attention on generating anomaly concepts by maximizing the attention value correlated with the anomaly token. To further guide and refine the generation results, we propose ***prompt-guided anomaly refinement***, which leverages more detailed anomaly descriptions as additional semantic guidance. Our main contributions are:

- We propose AnomalyAny, a visual anomaly generation framework for unseen anomalies. Users can provide arbitrary normal images of objects and anomaly descriptions to generate realistic and diverse anomalous samples.
- We introduce attention-guided anomaly optimization along with prompt-guided anomaly refinement to overcome the limitations of SD in anomaly generation. Our proposed method achieves more authentic generation results than existing methods, without requiring additional training or abundant normal/abnormal data samples.
- We demonstrate the effectiveness of our proposed AnomalyAny in both generation quality and in facilitating the training of the downstream anomaly detection models.

2. Related Works

2.1. Anomaly Generation

The scarcity of anomalous data has motivated numerous studies to synthesize anomaly samples [10]. One approach augments normal training samples by cropping and pasting abnormal patterns from few-shot abnormal samples in the test set [21, 30] , natural patterns of external datasets [7, 33], or directly from the normal image itself [19, 29]. Despite being simple and effective, these methods lack authenticity and diversity in the generated samples. Another approach employs generative models, such as generative adversarial networks (GANs) to synthesize anomalies [11, 34]. More recently, advancements in diffusion models [14, 17, 31, 35] have enabled methods that fine-tune these models on anomaly data to produce more diverse and realistic anomalies. However, these methods typically require substantial amounts of normal and/or abnormal data samples to learn the dataset-specific distribution, which is often impractical in data-limited scenarios. Also, these methods can only generate samples similar to the training set, *i.e.*, seen anomalies, but fail in generating unseen anomalies. In real-world applications, however, it is often impractical to assume prior knowledge of all possible anomalies, highlighting the need for methods that can generalize to unseen types. To generate unseen yet realistic anomalies, AnomalyAny directly deploys the pretrained SD model for anomaly synthesis without additional training, offering a more flexible solution for high-quality anomaly generation across arbitrary objects.

2.2. Anomaly Detection

The scarcity of available anomalous samples has made **unsupervised AD** [4, 9, 28] the dominant paradigm in the AD field, where the goal is to model the normal distribution and then identify anomalies as outliers [4, 9, 28]. However, these methods rely on the availability of a sufficient amount of normal samples that adequately represent the underlying distribution. Due to the lack of representative normal samples across the wide range of product variations and industrial configurations, **few-shot AD** has gained increasing attention [22, 32]. Existing few-shot methods typically leverage additional knowledge from vision-language models pretrained on large-scale datasets, such as Contrastive Language-Image Pre-training (CLIP) [25], to compute similarities between data samples and the normal or abnormal text prompts [13, 18, 20]. Nevertheless, training AD models solely with normal samples remains challenging due to a lack of awareness of anomaly distributions, underscoring the need for realistic anomalous samples to enhance AD performance.

3. Preliminary: Stable Diffusion

Diffusion models. Denoising diffusion probabilistic models (DDPMs) [15] learn the desired data distribution by defining a Markov chain of length T. In the forward pass, this chain gradually adds noise to a given data sample x_0 to obtain a sequence of noisy samples $x_t, t \in T$. In the reverse process, a model ϵ_θ parameterized by θ is learned to predict the noise added for each step t. Our method is based on SD [27], a latent diffusion model (LDM) that applies the denoising diffusion process to the latent representation z of x in the latent space of a variational auto-encoder (VAE) with the learning objective of predicting the added noise at each time step t as:

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

where z_t represents the noised latent representation at time step t. During inference, the reverse process starts with a random noise $x_T \sim \mathcal{N}(0, \mathbf{I})$ and gradually generates an image sample from the noise from step T to 0.

Text condition. SD introduces text guidance via a cross-attention mechanism. The denoising UNet network in the latent space of SD consists of self-attention layers followed by cross-attention layers at resolutions $P \in (64, 32, 16, 8)$. Given a text prompt c composed of N tokens, a guidance vector $\tau(c)$ is then obtained via the CLIP text-encoder τ [26]. $\tau(c)$ is then mapped to intermediate feature maps of the DDPM model ϵ_θ through each cross-attention layer as follows:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \quad A \in [P, P, N], \quad (2)$$

$$Q = W_Q^{(i)} \cdot \varphi_i\left(z_t\right), K = W_K^{(i)} \cdot \tau(c), V = W_V^{(i)} \cdot \tau(c), \quad (3)$$

where $\varphi_i(z_t)$ is the intermediate feature of the UNet. A, Q, K, V denote the attention, query, key, and value matrices. $W^{(i)}$ represents learnable projection matrices. Based on image-captioning pairs $\{x, c\}$, the text-conditioned diffusion model can be learned with the objective:

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

4. Methodology

In this work, we propose AnomalyAny, a novel framework for generating realistic, promptable anomalies on unseen objects and anomaly types, as illustrated in Fig. 2. AnomalyAny comprises three core modules. The framework begins with Test-time Normal Sample Conditioning (Sec. 4.1) to guide the generation process on a single normal sample, enabling anomaly generation on novel objects and anomaly types. During generation, we introduce a two-stage optimization process: Attention-Guided Anomaly Optimization (Sec. 4.2) first focuses generation on hard anomaly concepts, followed by Prompt-guided Anomaly Refinement (Sec. 4.3) that leverages detailed text descriptions to further enhance the generation quality.

4.1. Test-time Normal Sample Conditioning

In our work, we leverage SD to generate unseen visual anomalies. While SD is effective for image generation, the inherent diversity of its underlying distribution causes the generated images to significantly differ from the normal images in a specific AD dataset, as shown in Fig. 3 (b). To generate images that align with the target normal distribution, instead of fine-tuning SD, we incorporate information from normal samples directly during the generation process. Given a normal sample x^{normal} and its latent representation $z^{normal} = \mathcal{E}(x^{normal})$ encoded by the VAE encoder, we get a sequence of samples $z_1^{normal}, z_2^{normal}, ..., z_T^{normal}$ with progressively added noise controlled by the DDPM noise scheduler. Instead of starting the inference from $z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we begin from step $t_{start} = T \cdot (1 - \gamma)$ with the noisy latent representation $z_{t_{start}}^{normal}$, which conveys corrupted features from the guidance normal image x^{normal}. This enables the generation of anomalies for novel object types during test time without requiring additional training. The parameter γ controls the starting step, i.e., the noise scale added to the guidance data sample. In our experiments, we set $\gamma = 0.25$ to balance similarity to the original distribution with inference steps that promote diverse outcomes.

For more precise control conditioned on a given normal sample, we use a mask input as an optional constraint to specify the location of the generated anomalies. This mask can function either as a foreground mask to restrict anomalies to the intended object or as a detailed, pre-defined mask that accurately defines the anomaly region. For the predicted latent representation z_t at time step t, we apply the following operation:

$$z_t = mask \odot z_t + (1 - mask) \odot z_t^{normal}. \quad (5)$$

This formulation preserves the distribution of the normal guidance sample outside the masked region, ensuring that modifications are restricted solely to the specified area.

4.2. Attention-Guided Anomaly Optimization

We generate anomalous samples of a specific object type, denoted as $[cls]$ (e.g., bottle), by providing anomaly text descriptions, such as *"A photo of a* $[cls]$ *that is damaged"*. Due to the challenging nature of anomaly generation as introduced, the desired anomaly semantics are often neglected in the generated images, as shown in Fig. 3 (d). To overcome this challenge, we introduce ***attention-guided***

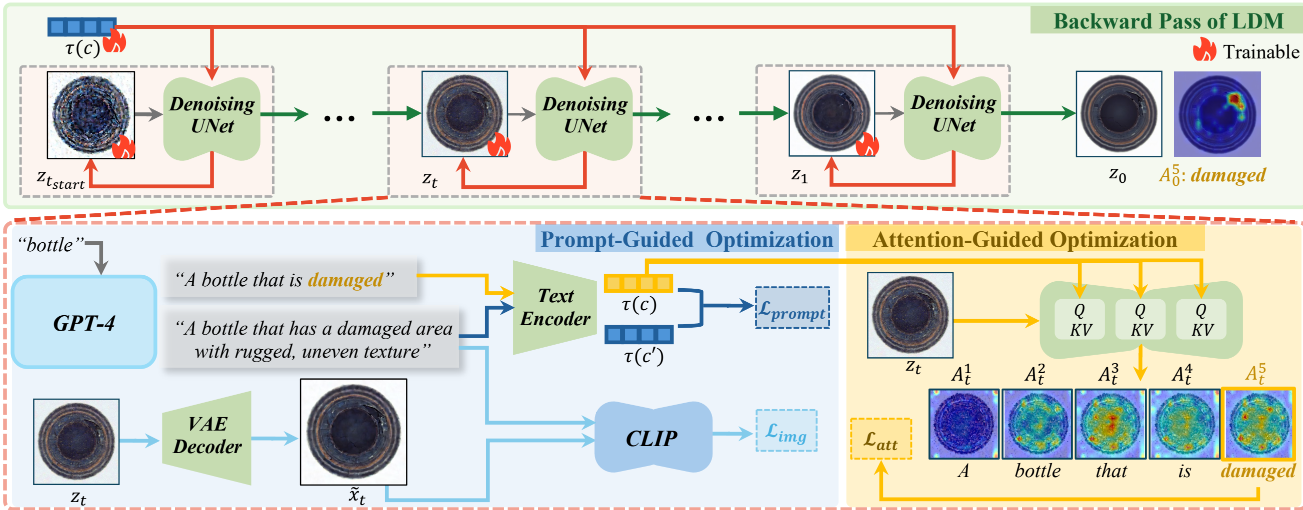
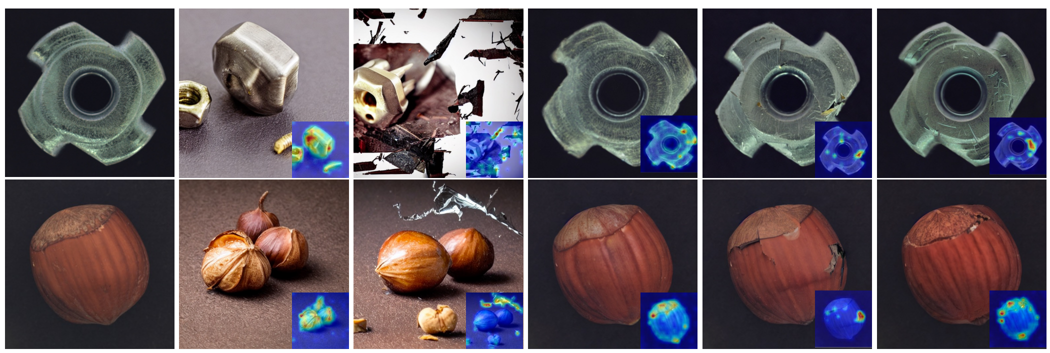


Figure 2. **Illustration of AnomalyAny** with details of the attention-guided & prompt-guided optimization process at time step t.



(a) (b) (c) (d) (e) (f)

Figure 3. **Examples of generated anomaly samples and attention map of *damage*.** We present (a) Normal guidance image, and the results generated by (b) Stable Diffusion, (c) Ours w/o normal sample conditioning, (d) Ours w/o attention-guided optimization, (e) Ours w/o prompt-guided optimization, and (f) Our proposed AnomalyAny.

anomaly optimization to enforce the generation of these critical yet difficult-to-capture anomaly concepts. First, we aggregate the attention maps from SD for subsequent optimization. As introduced in Sec. 3, SD introduces text guidance using a cross-attention mechanism. Given a text prompt c composed of N tokens, at each inference step t, we obtain a collection of attention maps $A_t \in P \times P \times N$ corresponding to c at different resolutions. These attention maps capture the cross-correlation with each text token, where $A[:,:,i]$ represents the probability assigned to token $c_i, i \in (1, N)$. Following the methodology of [6], we average all the attention maps at a resolution of 16×16, as these are identified to be the most semantically informative. The resulting attention map $\bar{A}_t$ is then normalized across the token dimension and smoothed using a Gaussian$(\cdot)$ function:

$$\bar{A}_t = \text{Gaussian}(\text{softmax}(\bar{A}_t)), \quad \bar{A}_t \in 16 \times 16 \times N. \tag{6}$$

We enforce the generated image to convey the semantic meaning of the anomaly type token c_j from the given c, e.g., *"damaged"* in *"A photo of a* $[cls]$ *that is damaged"* via optimization guided by $\bar{A}_t$. During the generation process, at each time step t, we optimize the intermediate latent representation z_t by maximizing the attention associated with the anomaly description. The optimized latent representation is then denoised to obtain z_{t-1}, which is used in the subsequent optimization and denoising steps. as shown in Fig. 2. Specifically, during optimization at step t, we extract the attention map $\bar{A}_t^j = \bar{A}_t[:,:,j]$ correlated with the anomaly token c_j. We then compute the gradient update for z_t by applying the loss function $\mathcal{L}_{att}$ to $\bar{A}_t^j$. Furthermore, we increase controllability by allowing masks as optional inputs and restricting the optimization of z_t to within the specified mask. The optimization at step t is formulated as follows:

$$\mathcal{L}_{att} = 1 - \max\left(\bar{A}_t^j \odot \text{mask}\right), \tag{7}$$

$$z_t \leftarrow z_t - \alpha_t \cdot \nabla_{z_t} \mathcal{L}_{att} \odot \text{mask}, \tag{8}$$

where α_t is a scalar defining the step size. This objective encourages a high maximum value of $\bar{A}_t^j$ to strengthen the activations of the anomaly token. Through iterative opti-

Figure 4. Visualization of **intermediate generation results** and **the attention maps** of anomaly tokens at different denoising steps.

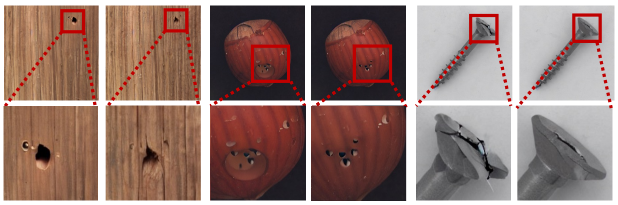


Figure 5. Examples of **anomalies** generated after attention-based optimization on anomaly tokens (a) w/o and (b) w/ localization-aware scheduler.

mization, we progressively integrate the semantic characteristics of anomalies into the generated image, as illustrated in Fig. 4.

Our empirical results demonstrate that the iterative updates described above can cause redundant attention on specific pixels, resulting in image artifacts. To mitigate over-optimization, we propose a ***localization-aware scheduler***. Starting with the initial generation result $z_{t_{start}}$ and its anomaly attention map $\bar{A}^j_{t_{start}}$, we determine the number of activated pixels $n_{t_{start}}$ by counting pixels with attention values exceeding the mean in the smoothed attention map. At each optimization step t, we calculate the activated pixel count n_t and compute the scalar α_t as follows:

$$\alpha_t = \lambda(1 + \Delta_t \cdot t) \cdot \frac{n_t}{n_{t_{start}}}, \tag{9}$$

where λ is a scaling factor that controls the strength of the optimization, and Δt adjusts the step size to gradually decrease the update rate. As the activated pixels become increasingly localized, we reduce α_t to mitigate overfitting. As illustrated in Fig. 5, this strategy significantly mitigates unrealistic artifacts in the generated samples.

4.3. Refined Anomaly Generation via Prompt-Guided Optimization

Attention-based optimization enforces the generation of anomaly keywords consisting of one to two tokens. However, the limited token length often results in ambiguous descriptions, thereby reducing the authenticity and richness of the generated context. To enhance semantic guidance and improve the generation quality, we propose a ***prompt-guided optimization for anomaly refinement*** that incorporates detailed anomaly descriptions. To enrich the generated anomaly distribution, we leverage GPT-4 [1] to generate potential anomaly types c_j for a given object, along with their corresponding detailed descriptions c'. For example, given the object as *bottle*, we prompt GPT-4 to identify possible anomaly types and their descriptions, such as *"damaged"* $\rightarrow$ *"close-up of a bottle that has a damaged area with rugged, uneven texture"*, thereby providing more nuanced guidance for generating a diverse range of damage types or anomalies.

To incorporate the semantic guidance from these long-text descriptions during the generation process, we introduce a CLIP-based image generation loss in the final denoising steps to ensure semantic consistency between the generated image and text guidance. At time step t, given the latent representation z_t, we obtain the decoded generation result $\tilde{x}_t = \mathcal{D}(z_t)$. We then minimize its distance from c' in the CLIP [26] embedding space as follows:

$$\mathcal{L}_{img} = 1.0 - \text{cosine}(\Phi^T(c'), \Phi^V(\tilde{x}_t)), \tag{10}$$

where $\Phi^T(\cdot)$ and $\Phi^V(\cdot)$ denotes the CLIP text and visual encoders, respectively. Minimizing this similarity loss aligns the generated anomalies more closely with the semantic attributes of c', refining the anomaly type with detailed descriptions. We then jointly optimize z_t for attention and semantic alignment as:

$$\mathcal{L} = \mathcal{L}_{img} + \alpha_t \cdot \mathcal{L}_{att}, \quad z_t \leftarrow z_t - \nabla_{z_t}\mathcal{L} \odot mask. \tag{11}$$

To further incorporate semantic guidance, we adapt c to the finer guidance c' by computing their similarity and optimize the prompt embedding as:

$$\mathcal{L}_{prompt} = 1.0 - \text{cosine}(\tau(c), \tau(c')), \tag{12}$$

$$\tau(c) \leftarrow \tau(c) - \nabla_{\tau(c)}(\mathcal{L}_{prompt} + \mathcal{L}_{img}). \tag{13}$$

This process enriches the semantic information within the original prompt embedding $\tau(c)$ to guide the generation process effectively. In our framework, we optimize this joint loss in the final 30 of the total inference steps. This integrated guidance not only strengthens semantic consistency but also mitigates unrealistic artifacts often encountered with standalone attention-based optimization, enabling more precise, contextually rich anomaly generation aligned with detailed prompt specifications.

With our proposed AnomalyAny framework, we achieve promptable unseen anomaly generation. Notably, since AnomalyAny is not trained on specific datasets, it is unrestricted by the distribution of available normal samples, making it applicable across a broad range of unseen object categories and anomaly types. Additionally, the attention maps define a probability distribution over each text token, allowing us to use the final smoothed attention map $\bar{A}^j_0$ at time step 0 as pixel-level annotation to localize the anomaly described by c_j.

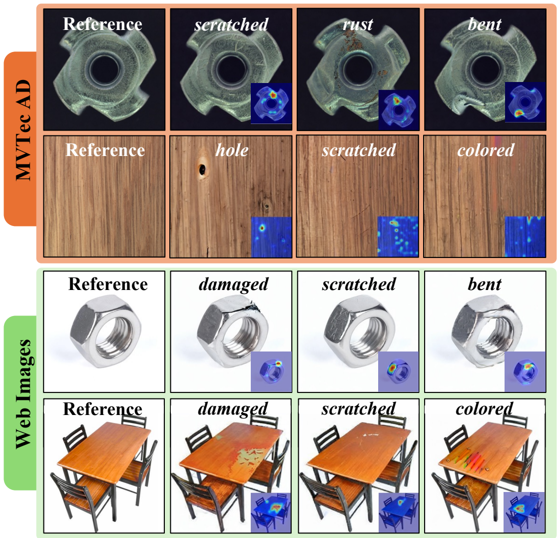


Figure 6. **Anomaly generation results for arbitrary objects and anomaly descriptions.** The bottom right presents the attention maps of the anomaly tokens.

Category	NSA [29]		RealNet [35]		AnoDiff [17]		Ours	
	IS ↑	IC-L ↑	IS ↑	IC-L ↑	IS ↑	IC-L ↑	IS ↑	IC-L ↑
bottle	1.35	0.11	1.58	0.16	1.58	0.19	1.73	0.17
cable	1.55	0.38	1.65	0.41	2.13	0.41	2.06	0.41
capsule	1.38	0.17	1.62	0.18	1.59	0.21	2.16	0.23
carpet	1.03	0.22	1.03	0.25	1.16	0.24	1.10	0.34
grid	2.26	0.36	2.24	0.37	2.04	0.44	2.31	0.38
hazelnut	2.04	0.31	2.21	0.31	2.13	0.31	2.55	0.32
leather	1.07	0.26	1.67	0.36	1.94	0.41	2.26	0.41
metal nut	1.66	0.23	1.67	0.24	1.96	0.30	1.82	0.27
pill	1.34	0.24	1.45	0.26	1.61	0.26	2.91	0.30
screw	1.20	0.31	1.20	0.31	1.28	0.30	1.33	0.32
tile	1.31	0.37	1.57	0.48	2.54	0.55	2.66	0.53
toothbrush	1.18	0.17	1.19	0.18	1.68	0.21	1.64	0.22
transistor	1.34	0.22	1.49	0.30	1.57	0.34	1.66	0.28
wood	1.30	0.32	2.22	0.41	2.33	0.37	1.93	0.41
zipper	1.52	0.22	1.88	0.24	1.39	0.25	2.14	0.33
average	1.44	0.26	1.64	0.30	1.80	0.32	**2.02**	**0.33**

Table 1. **Comparisons between anomaly generation quality and diversity with different anomaly generation methods.** AnoDiff: AnomalyDiffusion.

5. Experiment

5.1. Experimental Setup

Datasets. We conduct extensive experiments on MVTec AD [2] and VisA [36], two benchmarks for industrial anomaly detection. MVTec AD includes 5,000 images across 15 object and texture categories with pixel-level annotations for defects, while VisA contains 10,821 images from 12 categories with annotations.

Evaluation Metrics. To evaluate the anomaly generation results quantitatively, we employ Inception Score (IS) to measure the generation quality and Intra-cluster pairwise LPIPS distance (IC-LPIPS) [23] to evaluate the generation diversity. We also validate the effectiveness of our generated image samples by training anomaly detection frameworks using our generated data. For the anomaly detection task, we employ five metrics to thoroughly evaluate the detection performance: image-level and pixel-level Area Under the Receiver Operating Characteristic (AUROC), denoted as I-AUC and P-AUC, respectively, the image-level and pixel-level max-F1 scores [18], denoted as I-F1 and P-F1, and Per-Region-Overlap [3] denoted as PRO.

Implementation Details. For anomaly generation, we deploy the pretrained SD model with $T = 100$ and $\gamma = 0.25$. Following [35], we apply binary thresholding to extract a foreground mask for each normal guidance image, which acts as mask guidance for generation. More implementation details are provided in the supplementary material.

5.2. Anomaly Generation Results

Unseen Anomaly Generation. In Fig. 6, we present the generation results of AnomalyAny on MVTec AD and web images. Compared to existing data generation methods like [16, 17, 34], our AnomalyAny requires no training on normal or anomalous samples. Thus, AnomalyAny is not constrained by the limited available training data distribution, presenting universal anomaly generation capabilities on unseen data and arbitrary anomaly types. This shows

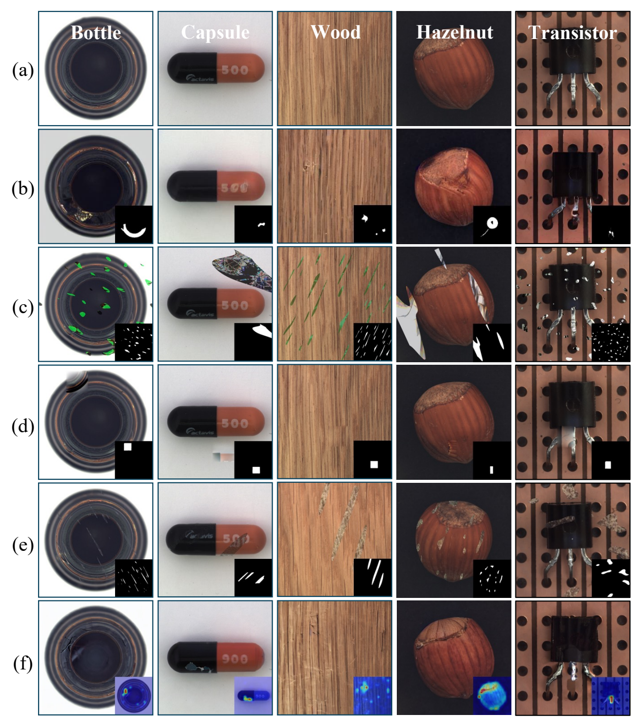


Figure 7. **Qualitative comparisons between existing anomaly generation methods.** (a) Reference, (b) AnomalyDiffusion, (c) DREAM, (d) NSA, (e) RealNet, (f) Our AnomalyAny.

Metrics →	MVTec AD					VisA				
Methods ↓	I-AUC	I-F1	P-AUC	P-F1	PRO	I-AUC	I-F1	P-AUC	P-F1	PRO
PaDiM [8]	76.6±3.1	88.2±1.1	89.3±0.9	40.2±2.1	73.3±2.0	62.8±5.4	75.3±1.2	89.9±0.8	17.4±1.7	64.3±2.4
PatchCore [28]	83.4±3.0	90.5±1.5	92.0±1.0	50.4±2.1	79.7±2.0	79.9±2.9	81.7±1.6	95.4±0.6	38.0±1.9	80.5±2.5
WinCLIP+ [18]	93.1±2.0	93.7±1.1	95.2±0.5	55.9±2.7	87.1±1.2	83.8±4.0	83.1±1.7	96.4±0.4	41.3±2.3	85.1±2.1
AnomalyGPT [13]	94.1±1.1	-	95.3±0.1	-	-	87.4±0.8	-	96.2±0.1	-	-
PromptAD [20]	94.6±1.7	-	**95.9±0.5**	-	-	86.9±2.3	-	96.7±0.4	-	-
AnomalyAny (Ours)	**94.9±0.4**	**94.7±0.4**	95.4±0.2	**57.3±0.0**	**91.9±0.0**	**89.7±0.8**	**85.8±0.5**	**97.7±0.4**	**43.2±0.4**	**92.5±0.1**

Table 2. **Comparison of 1-shot anomaly detection on MVTec AD and VisA.** Results are reported over 5 runs. The best results are in **bold**, and the second-best results are underlined.

Metrics →	MVTec AD					VisA				
Methods ↓	I-AUC	I-F1	P-AUC	P-F1	PRO	I-AUC	I-F1	P-AUC	P-F1	PRO
AnomalyDiffusion [17]	94.4±0.3	94.4±0.2	95.3±0.5	57.3±3.0	92.2±1.0	-	-	-	-	-
DRAEM [33]	93.6±0.3	94.2±0.4	95.1±0.1	56.0±0.9	91.8±0.1	86.0±0.7	83.0±0.9	97.5±0.1	42.6±0.7	92.6±0.6
NSA [29]	94.0±0.5	94.2±0.3	95.1±0.1	56.1±0.5	91.8±0.2	86.2±2.0	83.1±1.2	97.4±0.1	40.8±0.5	92.3±0.3
RealNet [35]	92.7±0.7	93.6±0.3	95.1±0.1	56.3±1.3	91.7±0.1	86.0±1.4	82.9±1.1	97.5±0.2	41.9±1.8	**92.8±0.3**
AnomalyAny (Ours)	**94.9±0.4**	**94.7±0.4**	**95.4±0.2**	**57.3±0.0**	**91.9±0.0**	**89.7±0.8**	**85.8±0.5**	**97.7±0.4**	**43.2±0.4**	92.5±0.1

Table 3. **Comparisons between 1-shot anomaly detection performance with different anomaly generation methods.** Since AnomalyDiffusion utilizes anomalous data for training and results in data leakage, we exclude it from ranking. Results are reported over 5 runs. The best results are in **bold**, and the second-best results are underlined.

our method's great potential in personalized anomaly generation with arbitrary normal patterns and anomalous descriptions, and may greatly contribute to the enhancement of anomaly detection models for a specific category even without real collected anomalous samples. Additional generation results are provided in the supplementary material.

Comparison with other anomaly generation methods. We compare the generation quality of AnomalyAny on MVTec AD with several existing anomaly generation methods, including DRAEM [33], NSA [29], RealNet [35], and AnomalyDiffusion [17]. Among these methods, AnomalyDiffusion trains on $1/3$ anomalous data from the test set. Fig. 7 presents the generated anomalies of these methods. We include more examples in the supplementary material. The examples clearly show that DRAEM, NSA, and RealNet introduce inauthentic patterns, while AnomalyDiffusion and AnomalyAny can generate more realistic anomalous samples. Notably, AnomalyDiffusion and RealNet rely on available anomalous training data, while AnomalyAny generates without training and generalizes to unseen object types and anomalies during inference.

We present quantitative evaluations of generation quality in Tab. 1. For NSA, RealNet, and AnomalyAny, we condition on MVTec AD training images to generate anomaly samples. As DRAEM crops random textures from external datasets, we do not include it for comparison [17]. The result demonstrates that our model generates anomaly data with both the highest quality and diversity. Additionally, a user study on anomaly generation quality is included in the supplementary material, providing further evidence of the superior generation quality achieved by AnomalyAny.

5.3. Anomaly Detection Results

The key advantage of our method lies in its capability to generate unseen anomalies without needing any normal or abnormal samples for training, making it highly effective under scenarios with data scarcity. To further demonstrate the effectiveness of our method, we evaluate it under the challenging 1-shot anomaly detection scenario, where only one single normal sample is accessible and no abnormal samples are available. Following the method proposed in [13], we condition on a single normal image to generate 100 random anomalous samples. We use the generated samples with the attention map of the anomaly token c_j as the corresponding anomaly mask to facilitate anomaly detection training. Further details on the generation process and anomaly detection training are provided in the supplementary material. Tab. 2 presents the comparison results between our AnomalyAny and existing few-shot AD methods, including two full-shot methods [8, 28] in the few-shot settings, and three CLIP-based few-shot methods [13, 18, 20]. As can be seen, our generated anomalous samples consistently outperform other alternatives in the majority of the metrics.

To further compare with other anomaly generation methods under the data scarcity scenarios, we follow the same 1-shot setup and substitute the synthetic training data with anomalous samples generated by alternative methods. For DRAEM, NSA, and RealNet, we condition on the same single normal sample as for our method and generate 100 random anomalous samples for training. We report the results in Tab. 3. The results indicate that anomalous samples generated by AnomalyAny yield the best detection results. Ad-

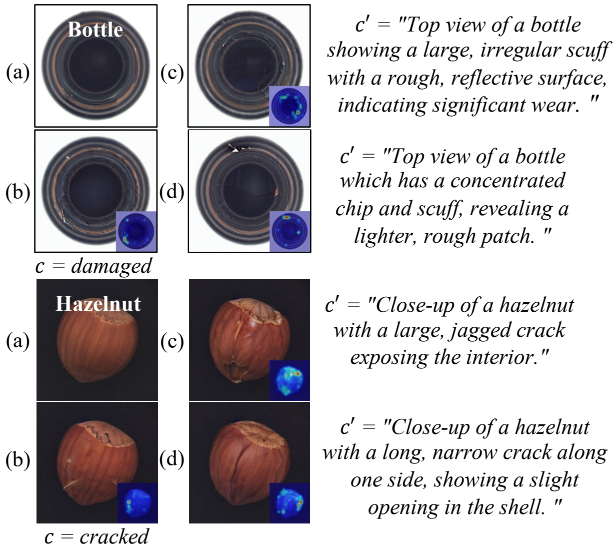


Figure 8. **Ablations on different types of prompt guidance.** (a) Normal image, (b) Generation results given anomaly type prompt c and (c)(d) Generation results given c and detailed prompt c'.

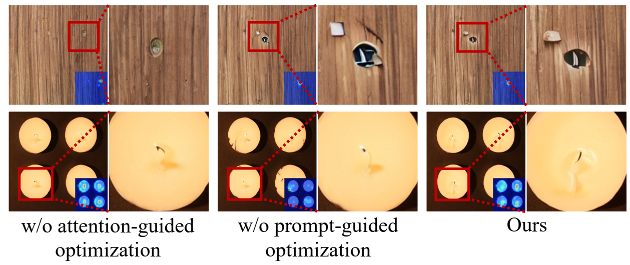


Figure 9. **Ablation on optimization strategies.** Anomaly types are *"hole"* and *"melted"* respectively.

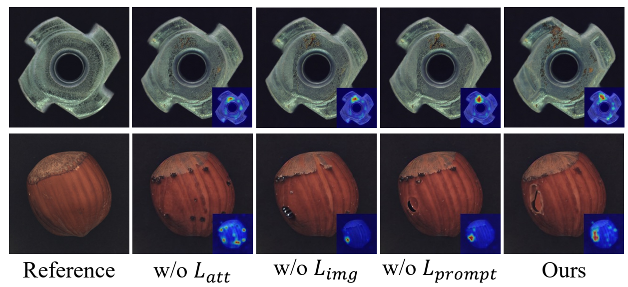


Figure 10. **Ablation on prompt-guided optimization objectives.** Anomaly types are *"rust"* and *"hole"* respectively.

ditionally, we compare with AnomalyDiffusion, which has seen all the training samples and some of the test anomalous samples and thus introduced data leakage. Despite not being trained with test set anomalous samples, our method produces comparable anomaly detection performance.

5.4. Ablations

Ablation on different types of prompt guidance. We compare the generation results under different types of guidance in our AnomalyAny. The results in Fig. 8 demonstrate that our method reflects the semantics of the anomaly prompts c generated by GPT-4. By providing the detailed prompt description c', we are able to get more fine-grained and realistic generation results with more diversity.

Ablation on attention-guided & prompt-guided optimization. We evaluate the effectiveness of attention-guided optimization and prompt-guided optimization separately, as illustrated in Fig. 3 (d-f), with additional results shown in Fig. 9. Our results show that attention guidance successfully enforces the generation of the specified anomaly description, ensuring it is not overlooked by SD. Furthermore, detailed prompt guidance enhances the realism of the generated samples. We also analyze the impact of different loss terms in Eq. (11) and Eq. (13) for optimizing z and $\tau(c)$ respectively, as shown in Fig. 10. The attention loss term improves focus on the anomaly region, while prompt-based terms help enhance generation quality. Notably, when the anomaly type is unambiguous (e.g., *"rust"*), prompt guidance has a reduced impact on the generation quality. Additional visual results from the ablation study are provided in the supplementary material.

6. Conclusion

In this paper, we propose AnomalyAny, a novel framework that utilizes SD for unseen anomaly generation, allowing users to generate realistic anomalous samples from arbitrary normal images of objects and anomaly text descriptions. Our framework incorporates attention-guided anomaly optimization to direct SD's attention toward anomaly concepts and prompt-guided anomaly refinement to enhance the authenticity of generated samples. Compared to existing methods, AnomalyAny achieves authentic unseen anomaly generation without additional training. Extensive experimental results demonstrate the effectiveness of AnomalyAny in generating high-quality anomalies that improve the performance of downstream anomaly detection tasks.

Currently, AnomalyAny relies on normal image guidance and textual descriptions for anomaly generation. While text prompts offer flexibility, they can sometimes lack sufficient guidance due to CLIP's limited understanding of lengthy and specialized descriptions as well as the precision required for complex anomalies. Future work could address this by incorporating additional inputs, such as one-shot anomalous image. This would increase AnomalyAny's versatility, enabling more precise anomaly generation across various applications. With its demonstrated generation capabilities and ability to generalize to unseen anomalies, AnomalyAny has significant potential to advance foundation models for anomaly detection.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[2] 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. 6

[3] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021. 1, 6

[4] Yunkang Cao, Xiaohao Xu, Zhaoge Liu, and Weiming Shen. Collaborative discrepancy optimization for reliable image anomaly localization. *IEEE Transactions on Industrial Informatics*, pages 1–10, 2023. 2

[5] Yunkang Cao, Xiaohao Xu, Jiangning Zhang, Yuqi Cheng, Xiaonan Huang, Guansong Pang, and Weiming Shen. A survey on visual anomaly detection: Challenge, approach, and prospect. *arXiv preprint arXiv:2401.16402*, 2024. 1

[6] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 4

[7] 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

[8] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 7

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

[10] Hao Dong, Gaëtan Frusque, Yue Zhao, Eleni Chatzi, and Olga Fink. NNG-Mix: Improving Semi-supervised Anomaly Detection with Pseudo-anomaly Generation. *arXiv preprint arXiv:2311.11961*, 2023. 2

[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. 2

[12] Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17481–17490, 2023. 1

[13] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1932–1940, 2024. 2, 7

[14] Guan Gui, Bin-Bin Gao, Jun Liu, Chengjie Wang, and Yunsheng Wu. Few-shot anomaly-driven generation for anomaly classification and segmentation. In *European Conference on Computer Vision (ECCV 2024)*, pages –, 2024. 2

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[16] Jie Hu, Yawen Huang, Yilin Lu, Guoyang Xie, Guannan Jiang, and Yefeng Zheng. Anomalyxfusion: Multimodal anomaly synthesis with diffusion. *arXiv preprint arXiv:2404.19444*, 2024. 6

[17] Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8526–8534, 2024. 1, 2, 6, 7

[18] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19606–19616, 2023. 1, 2, 6, 7

[19] 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. 2

[20] Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2024. 2, 7

[21] Dongyun Lin, Yanpeng Cao, Wenbin Zhu, and Yiqun Li. Few-shot defect segmentation leveraging abundant defect-free training samples through normal background regularization and crop-and-paste operation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 2

[22] Ruiying Lu, YuJie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. In *Advances in Neural Information Processing Systems*, pages 8487–8500, 2023. 1, 2

[23] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10743–10752, 2021. 6

[24] OpenAI, Josh Achiam, Steven Adler, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024. 1

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 2

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 5

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3

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

[29] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *European Conference on Computer Vision*, pages 474–489. Springer, 2022. 2, 6, 7

[30] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 2

[31] Shuai Yang, Zhifei Chen, Pengguang Chen, Xi Fang, Yixun Liang, Shu Liu, and Yingcong Chen. Defect spectrum: A granular look of large-scale defect datasets with rich semantics. In *European Conference on Computer Vision*, pages 187–203. Springer, 2025. 2

[32] 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*, pages 4571–4584, 2022. 1, 2

[33] 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, 7

[34] Gongjie Zhang, Kaiwen Cui, Tzu-Yi Hung, and Shijian Lu. Defect-gan: High-fidelity defect synthesis for automated defect inspection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2524–2534, 2021. 2, 6

[35] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 6, 7

[36] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022. 6