UniScale: Arbitrary-Scale Industrial Anomaly Generation

Shilei Zeng*, Linxin Guan*, Xurui Li, Yaohan Tang, and Yu Zhou✉

School of Electronic Information and Communications,
Huazhong University of Science and Technology
{shlzeng,lxguan,xrli_plus,yhtang_,yuzhou}@hust.edu.cn

Abstract. Industrial anomaly inspection faces a major challenge due to the lack of real-world anomaly samples. While generative models are used to create anomaly data, existing methods still struggle when handling small-scale anomalies. This failure occurs because extreme downsampling in diffusion models causes the information of small anomalies to be lost in the latent space. To address this, we introduce UniScale, a unified training and inference framework for high-fidelity industrial anomaly generation across arbitrary scales. During training, we introduce an Error-Suppressed Multi-Scale Training (EMT) strategy, which enables the model to learn the rich location-aware textures of anomalies, while suppressing upsampling-induced interpolation errors in texture acquisition, ensuring the model is capable of learning small-scale anomalies, while remaining effective for regular scale anomalies. For inference, we propose Generation-then-Fusion Denoising. It decouples anomaly generation from background integration, preventing small anomalies from being overwhelmed. Extensive experiments demonstrate that our method outperforms state-of-the-art competitors in both anomaly generation quality and downstream detection performance. It achieves a relative IS(a) improvement of 45.86% (from 1.81 to 2.64) on VisA and 37.70% (from 1.22 to 1.68) on MVTec AD 2, while also improving the downstream pixel-level IoU by 4.22% on VisA and AUROC by 6.55% on MVTec AD 2. Code is available at https://github.com/HUST-SLOW/UniScale.

Keywords: Industrial anomaly generation · Multi-scale learning

1 Introduction

Industrial anomaly inspection plays a vital role in modern manufacturing, yet anomalous samples are inherently scarce in real-world scenarios. This scarcity severely limits the development of reliable inspection systems, as collecting sufficient anomaly data is both challenging and costly. To alleviate data insufficiency, recent generative approaches such as AnomalyDiffusion [14], SeaS [4], and DualAnoDiff [17] have shown promise in generating anomaly data.

* Contributed Equally

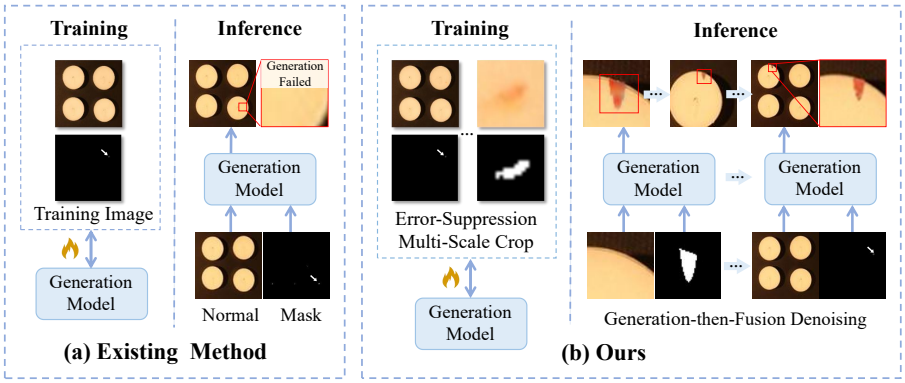

Fig. 1: (a) Existing anomaly generation methods often fail to generate small-scale anomalies. In contrast, our method (b) proposes an Error-Suppressed Multi-Scale Training Strategy and Generation-then-Fusion Denoising, successfully generating small-scale anomalies.

However, despite recent progress, these methods struggle to handle **small-scale anomalies**. Given the typical operating resolution in anomaly generation tasks (e.g., 512×512), we define anomalies smaller than 32×32 pixels as small-scale. Our statistical analysis shows that such anomalies are very common, accounting for 56.33% of anomaly images in the VisA dataset. The main challenge lies in the extreme downsampling required by the overall model architecture. Specifically, this information loss is a compounded effect caused by both the VAE, which initiates an $8\times$ downsampling, and the U-Net denoising network, which introduces further spatial compression. As noted in Prompt-to-Prompt [12], the semantics of text-to-image models are mainly controlled by attention maps, which are most effective at low resolutions (e.g., 16×16). This results in a significant $32\times$ downsampling factor compared to the original input. Consequently, the information carried by small-scale anomalies is severely compressed or completely lost in the latent space, making it more challenging for the model to learn these fine details. Because of this severe information loss, existing methods trained directly on original full-resolution images consistently fail to generate these small anomalies, as visually demonstrated in Fig. 1(a).

To address these challenges, we propose UniScale, a unified training and inference framework for industrial anomaly generation across arbitrary scales, as shown in Fig.1(b). In the training stage, we propose an Error-Suppressed Multi-Scale Training (EMT) strategy, which contains a Multi-Scale Cropping (MSC) operation and a Joint Scale (JS) loss. EMT extracts anomaly features across varying-scale crops, optimizes them together, enabling the text embeddings to learn the rich textures of anomalies that satisfy location constraints. In addition, we propose a Cross-Scale Perception (CSP) loss, which is designed to alleviate the interpolation errors caused by upsampling during texture acquisition. By enforcing consistency across scales, CSP loss ensures the text embeddings learn authentic anomaly features, resulting in high-quality anomaly learning.

In the inference stage, we propose Generation-then-Fusion Denoising. By separating anomaly generation from its fusion with the product, it preserves fine details while achieving seamless integration with the product, resulting in high-quality anomaly generation across scales. Experiments show that our method significantly outperforms existing anomaly generation methods on the VisA and MVTec AD 2 datasets. Regarding anomaly generation quality, our approach achieves a remarkable relative IS(a) improvement of 45.86% (from 1.81 to 2.64) on the VisA dataset and 37.70% (from 1.22 to 1.68) on the MVTec AD 2 dataset. Furthermore, downstream segmentation models trained on our generated data exhibit substantial gains over those trained with existing generation methods. Specifically, our method improves the pixel-level IoU by 4.22% on VisA, and achieves a significant performance leap on the more challenging MVTec AD 2 dataset, improving the pixel-level AUROC by 6.55% compared to the second-best method.

In summary, our contributions are threefold:

- We introduce a unified framework for high-fidelity industrial anomaly generation across scales. To the best of our knowledge, this is the first approach capable of handling the particularly challenging task of small-scale industrial anomaly generation.
- We propose an Error-Suppressed Multi-Scale Training strategy. It jointly learns fine-grained textures and location-aware anomaly appearance, and suppresses the interpolation errors introduced by the upsampling in texture acquisition. In addition, we introduce a Generation-then-Fusion Denoising strategy, which decouples the anomaly generation from background integration, and prevents small anomalies from being ignored during inference, ensuring high-fidelity detail preservation.
- Our approach achieves superior performance on the VisA and MVTec AD 2 datasets. Notably, it sets new state-of-the-art records in both generation quality (with a relative IS(a) increase of 45.86% and 37.70%, respectively) and downstream anomaly detection (improving pixel-level IoU by 4.22% on VisA, and pixel-level AUROC by 6.55% on MVTec AD 2), confirming its effectiveness for industrial inspection.

2 Related Work

2.1 Industrial Anomaly Generation

Early methods [5,19,29,35] primarily relied on data augmentation techniques to synthesize pseudo-anomaly images. However, the significant differences between the generated anomalies and the real ones cause low authenticity. In current studies, some approaches [6,23,36] employ generation models to enrich the anomalies. Some approaches [20, 38] leverage diffusion models to empower reconstruction-based anomaly detection. The DFMGAN [6] is based on StyleGAN2 [18], maintaining better structural integrity of the product while generating high-quality anomaly images. Nevertheless, due to insufficient training samples, this method

is prone to overfitting. Further, some methods [4, 10, 14, 17] achieve anomaly generation based on diffusion models. SeaS [4] and DualAnoDiff [17] train diffusion models to learn complete anomaly images, which are capable of generating anomaly images and their corresponding annotations. AnomalyDiffusion [14] decouples anomaly region information into appearance features and positional priors to generate anomaly regions. However, existing generative approaches remain limited in addressing small-scale anomalies, which are common in industrial inspection. To overcome such a limitation, we propose an Error-Suppressed Multi-Scale Training strategy. This strategy ensures the model learns both the fine details of small anomalies and their location-dependent appearances, while remaining effective for regular-scale anomalies. This achieves faithful generation of small-scale anomalies, without compromising the quality of regular-scale anomaly generation.

2.2 Subject Tuning in Diffusion Models

Subject tuning has recently emerged as a powerful paradigm in diffusion models [**?**, 13, 31], aiming to achieve precise text–image binding for specific objects or concepts. It focuses on preserving the identity of a given subject, even under novel prompts or diverse contexts. Representative approaches [2, 3, 7, 15, 30, 39, 40] extend earlier personalization techniques, such as Textual Inversion [9] and DreamBooth [27]. Tuning and generating small-scale subjects remains a persistent challenge for standard diffusion models, as the fine-grained texture details of diminutive subjects are difficult to learn. To mitigate this, early strategies [8, 16, 32] relied on crop-and-magnify paradigms, such as extracting sparse patches [16] or utilizing random crops [8]. However, purely crop-based fine-tuning isolates small targets, disrupting the seamless integration with the background. To address this, recent approaches, such as Text2Traffic [22] and SOEDiff [33], attempt to jointly train diffusion models on both magnified local patches and original images. Nevertheless, simply combining these scales is insufficient, as magnifying small crops inevitably introduces interpolation errors. Our method employs Cross-Scale Perception (CSP) loss to suppress the error. By enforcing consistency across scales, the CSP loss ensures that the learnable text embeddings are encouraged to learn consistent anomaly appearances across scales. To the best of our knowledge, this critical challenge of small-subject tuning has not yet been addressed in the field of industrial anomaly generation.

3 Preliminaries

Stable Diffusion [26] is a text-conditioned latent diffusion framework. An input image x_0 is first encoded into a latent representation z_0 using a VAE encoder $\mathcal{E}(\cdot)$, which downsamples the image by a factor of 8 into a compact latent space to reduce computational cost. Then random noise $\epsilon \sim N(0, I)$ is added at discrete timesteps t, producing noisy latent states: $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$. Meanwhile, a text prompt P is encoded by a CLIP text encoder into a conditional vector

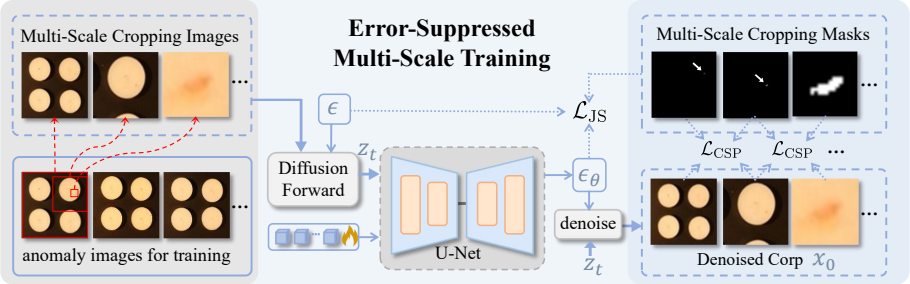

Fig. 2: Pipeline of Error-Suppressed Multi-Scale Training Strategy. We apply a Multi-Scale Cropping strategy to construct a supervision set. Based on this set, we propose a Joint-Scale (JS) loss to bind the anomaly to the text embeddings with a Cross-Scale Perception (CSP) loss to mitigate the interpolation error.

e, which guides the U-Net denoising network. The training objective minimizes the discrepancy between true and predicted noise:

$$L_{\text{SD}} = \mathbb{E}_{z_0=\mathcal{E}(x_0),P,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, e)\|_2^2\right]. \tag{1}$$

Within the U-Net, cross-attention layers enable direct interaction between textual embeddings and image latents, establishing a text-image binding, while self-attention layers capture dependencies among image patches to preserve structural coherence. This attention-based mechanism ensures semantic alignment between prompts and visual structures during denoising.

Blended Diffusion [1] performs local image editing guided by text. It iteratively combines generated content with the original image during the denoising process. At each sampling step, the model blends the text-guided latent prediction $\hat{z}_t$ with the latent representation of the original image z_t. This blending is performed within a binary mask m after the forward diffusion process. This spatial fusion is formulated as,

$$z_t = \hat{z}_t \odot m + z_t \odot (1 - m) \tag{2}$$

This process ensures that the edited region conforms to the textual prompt while preserving seamless coherence with unmodified image areas.

4 Method

We introduce **UniScale**, a unified training and inference framework for high-fidelity industrial anomaly generation across arbitrary scales. On the training side, as shown in Fig. 2, we introduce the Error-Suppressed Multi-Scale Training strategy (Sec. 4.1), which learns the rich location-aware textures of the anomalies, while suppressing upsampling-induced interpolation errors in texture acquisition. On the inference side, as shown in Fig. 5, we propose Generation-then-Fusion Denoising (GFD) (Sec. 4.2), which achieves faithful anomaly construction and seamless background integration.

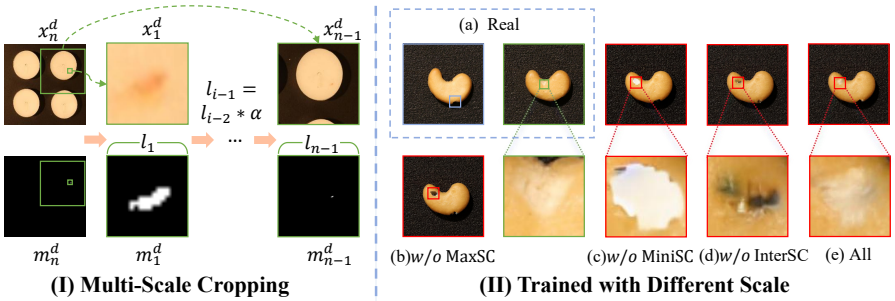

Fig. 3: Multi-Scale Cropping and Training Visualization. (I): Our proposed Multi-Scale Cropping. (II): Real VisA dataset images of one anomaly type (blue, green) and anomaly generation images trained with different scale crops (red).

4.1 Error-Suppressed Multi-Scale Training Strategy

In this section, we first introduce a Multi-Scale Cropping strategy, shown in Fig. 3 (I), to enable the text embeddings to learn rich textures of the anomalies that satisfy location constraints. And then we incorporate a Cross-Scale Perception (CSP) loss to alleviate the interpolation error introduced during texture acquisition.

Maximum Scale Crop. Given an input RGB anomaly image, we resize it to the generation resolution $\mathcal{S}$ and get $x_n^d \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}\times 3}$ and its corresponding mask $m_n^d \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}\times 3}$, thereby obtaining the Maximum Scale Crop (MaxSC). It corresponds to the standard full-image input [4, 14], and enables the model to learn the location-aware anomaly appearance through the attention mechanism in U-Net, which is critical for correct generation. As shown in Fig. 3(II)(a), the appearance of breakage depends on its location. At the edges (blue box), it appears black, which corresponds to the background texture exposed by product damage. While inside the product (green box), it shows a whitish texture related to the product itself. Without MaxSC, the model tends to learn isolated texture patterns, ignoring the location-aware characteristics of the anomalies, which may lead to incorrect generation. As illustrated in Fig. 3(II)(b), the model incorrectly generates a black background texture in the center of the product, instead of producing the whitish texture associated with the product. Therefore, MaxSC is indispensable for preserving the location awareness of anomalies. The limitation of MaxSC is that, due to the small size of the anomaly region, the learnable text embeddings are unable to capture sufficient abnormal texture information during text-image alignment.

Minimum Scale Crop. To complement the missing fine-grained texture details of anomalies for text embedding learning, we further introduce Minimum Scale Crop (MiniSC). We first locate the anomaly by extracting the minimum bounding box from the connected components of the mask m_n^d. Then we extend this bounding box into a square based on its longer side. This crop is then resized to the training resolution $\mathcal{S}$, yielding $x_1^d \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}\times 3}$, its mask $m_1^d \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}\times 3}$, and the corresponding side length l_1. The MiniSC plays a crucial role in preserving

anomaly visibility and fine-grained texture details in text embedding learning. As illustrated in Fig. 3(II)(c), when MiniSC is absent, the learned text embeddings tend to guide the generative model to produce blurry regions lacking texture, which differ significantly from the anomaly textures observed in real data in (II)(a). However, the problem with MiniSC is that it cannot observe the overall product, making it impossible for U-Net to learn the relationship between the anomalies and the product. This hinders the model from determining whether to generate an anomaly with the correct texture at the correct location. At the same time, since upsampling interpolation alters the original anomaly appearance, it consequently negatively impacts text embedding learning.

Intermediate Scale Crop. Although MaxSC and MiniSC are theoretically complementary, learning from them alone fails to integrate location constraints with texture information, which may lead to confused generations. For example, in Fig. 3(II)(d), the edge-specific texture and internal anomaly patterns are mixed, which may reduce the realism of the generated anomalies. Therefore, we further introduce Intermediate Scale Crop (InterSC), by expanding the cropping scope starting from the minimum scale to the maximum scale. Specifically, let l_1 be the side length of the initial minimum bounding box x_1^d. For each subsequent step i ($i > 1$), we enlarge the field of view by increasing the side length: $l_i = l_{i-1} \times \alpha$. The cropping operation is performed centered on the geometric center of the initial minimum bounding box, it is formulated as:

$$(x_i^d, m_i^d) = \text{Crop}((x_n^d, m_n^d), l_i), \quad l_i < \mathcal{S}, \tag{3}$$

The resulting region is then resized to the training resolution $\mathcal{S} \times \mathcal{S}$, producing a sequence of multi-scale inputs (x_i^d, m_i^d). InterSC progressively reduces the anomaly-to-background ratio in a controlled geometric progression, preserves fine-grained anomaly textures with less interpolation error, and serves as a crucial bridge between MaxSC and MiniSC. As shown in Fig. 3(II)(e), by using InterSC together with MaxSC and MiniSC, the generation model is enabled to learn how anomaly appearance depends jointly on texture and location. The choice of α is critical. When α is too large, the transition between scales becomes abrupt, and the intermediate crops fail to bridge texture and positional semantics, resulting in generations similar to those without intermediate crops. Conversely, if α is too small, the number of crops increases excessively, consuming more GPU memory. Therefore, α must be carefully chosen to balance smooth scale transition and GPU memory consumption.

Joint Scale Training. Through the Multi-Scale Cropping process, we construct a multi-scale image set $\{(x_1^d, m_1^d), \ldots, (x_n^d, m_n^d)\}$, ranging from a tightly focused anomaly patch to the full resolution image. The number of crops n is adaptively determined by the size of the anomaly. Since these three scales complement each other, learning from them together enables text embedding to learn rich textures that satisfy location constraints. Therefore, we aggregate the set $\{x_i^d\}_{i=1}^n$ into a single training batch, and optimize the following Joint-Scale (JS) loss according

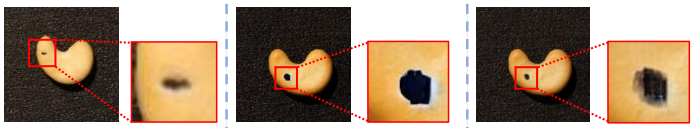

Fig. 4: Comparison of generated images w/o CSP loss and w/ (Ours). The model w/o CSP loss cannot generate faithful anomalies.

to the standard diffusion objective L_{SD} (Eq. 1):

$$\mathcal{L}_{\text{JS}} = \sum_{i=1}^{n} \|m_i^d \odot (\epsilon_i - \epsilon_\theta(z_{i,t}, t, e))\|_2^2. \tag{4}$$

Here, $z_{i,t}$ represents the noisy latent corresponding to the i-th crop x_i^d at timestep t, ϵ_i is the target noise, and e is the learnable text embedding shared across all scales.

Cross-Scale Perception Loss. As we mentioned earlier, interpolation distortions make anomalies appear differently at different scales, causing the text embeddings to learn inconsistent features and produce unrealistic generations. As shown in Fig. 4(b), training with JS loss alone produces artifacts that appear unrealistic and lack meaningful texture. To address this cross-scale inconsistency, we introduce a Cross-Scale Perception (CSP) loss. The key idea is to enforce perceptual similarity [37] between anomaly regions at adjacent scales, which are generated under the guidance of the text embeddings. In this way, the embeddings are encouraged to learn consistent anomaly features.

During training, for scale i, we first predict the noise $\epsilon_\theta(z_{i,t}, t, e)$ under the guidance of e. Then we use the predicted noise to denoise $z_{i,t}$, yielding the clean latent $z_{i,0}$. Then we decode $z_{i,0}$ using the VAE decoder, and get the predicted image $\hat{x}_i^d$, which is used as input for CSP loss. Then the CSP loss is computed by comparing the perceptual features of $\hat{x}_i^d$ and $\hat{x}_{i+1}^d$ within the anomaly regions. Specifically, we crop the anomaly region of higher-resolution scale predicted crops $\hat{x}_{i+1}^d$, use crop side length $l'_{i+1} = \mathcal{S} \times l_i / l_{i+1}$. Then we resize the crop to resolution $\mathcal{S}$, yielding $\hat{x}_{i+1}^{cd}$. Then we use CSP loss to align it to the lower-resolution scale predicted crops $\hat{x}_i^d$. The alignment is performed as:

$$\hat{x}_{i+1}^{cd} = \text{Crop}((\hat{x}_{i+1}^d, m_{i+1}^d), l'_{i+1}) \tag{5}$$

where $\hat{x}_{i+1}^d$ is the decoder output at scale $(i+1)$ generated from the latent $\hat{z}_{i+1,0}$. Then the CSP loss is defined as:

$$\mathcal{L}_{\text{CSP}} = \sum_{i=1}^{n-1} \mathcal{P}(m_i^d \odot \hat{x}_i^d, m_i^d \odot \hat{x}_{i+1}^{cd}) \tag{6}$$

where $\mathcal{P}$ denotes an off-the-shelf feature extractor used to compute perceptual similarity between anomaly regions at adjacent scales. By minimizing $\mathcal{L}_{\text{CSP}}$, the

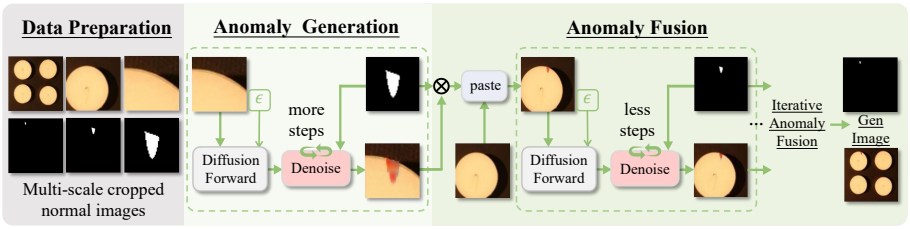

Fig. 5: Pipeline of the Inference. We first crop the normal image, and get the multi-scale normal crops. Then we perform Anomaly Generation and Anomaly Fusion on the crops, and get the final generated anomaly image.

learnable text embeddings are encouraged to learn consistent anomaly appearance across scales. This effectively alleviates the severe interpolation error and the inconsistency introduced by resizing artifacts during multi-scale cropping.

The overall training objective combines the Joint-Scale (JS) loss and the Cross-Scale Perception (CSP) loss:

$$\mathcal{L} = \mathcal{L}_{\text{JS}} + \lambda \cdot \mathcal{L}_{\text{CSP}}. \tag{7}$$

The hyperparameter λ controls the relative weight of the CSP loss in the final objective. During training, as in Textual Inversion [9], we use learnable tokens whose embeddings are optimized with $\mathcal{L}$.

4.2 Generation-then-Fusion Denoising

Existing text-to-image inpainting methods typically perform anomaly generation and background fusion within a single denoising stage. Due to the low resolution of cross-attention maps, the denoising process is dominated by large-scale background textures and structures. When generation and fusion are conducted simultaneously, small anomalies are easily ignored. This makes one-stage editing unsuitable for small-scale anomaly generation. To overcome this limitation, we introduce Generation-then-Fusion Denoising (GFD), illustrated in Fig. 5.

Stage 1: Data Preparation. Given a normal image x_n^g, we first generate a stochastic anomaly mask m_n^g using Perlin noise [25]. The mask m_n^g specifies the spatial regions where anomalies will be injected during subsequent generation. Similar to the procedure in Sec. 4.1, we extract the minimum bounding box from the connected components of m_n^g and extend it into a square. Then we apply the multi-scale cropping strategy (Eq. (3)) to x_n^g. Starting from this minimum bounding box, we obtain a set of multi-scale normal crops and masks $\{(x_1^g, m_1^g), \ldots, (x_n^g, m_n^g)\}$, which are used as inputs in the following stages.

Stage 2: Anomaly Generation. With the prepared crops and masks, we begin anomaly generation at the minimum crop scale ($i = 1$). The inputs are the normal crop x_1^g, the anomaly mask m_1^g, and the trained text embeddings e. We generate anomaly content within the masked region by applying $s = 50$ steps of Blended Diffusion (Eq. (2)). The output of this stage is the minimum-scale anomaly image $\hat{x}_1$, which is used as the input for the fusion stage.

Stage 3: Progressive Anomaly Fusion. For each subsequent scale ($i > 1$), the inputs are the anomaly image $\hat{x}_{i-1}$ and its mask m^g_{i-1} from the previous scale, together with the current normal crop x^g_i and mask m^g_i. First, $\hat{x}_{i-1}$ and m^g_{i-1} are downsampled by a factor of $1/\alpha$, producing $\hat{x}^{\downarrow}_{i-1} \in \mathbb{R}^{(\mathcal{S}/\alpha)\times(\mathcal{S}/\alpha)\times 3}$ and $m^{\downarrow}_{i-1} \in \mathbb{R}^{(\mathcal{S}/\alpha)\times(\mathcal{S}/\alpha)}$. The valid anomaly regions indicated by $m^{\downarrow}_{i-1}$ and the target mask m^g_i are identical in size. Next, we use $m^{\downarrow}_{i-1}$ to extract the anomaly pixels from $\hat{x}^{\downarrow}_{i-1}$. We then directly paste these pixels into x^g_i at the exact location specified by m^g_i. This pixel-level replacement is formulated as: $x^g_i(m^g_i = 1) \leftarrow \hat{x}^{\downarrow}_{i-1}(m^{\downarrow}_{i-1} = 1)$. This operation explicitly injects the anomaly into the higher-resolution crop. Since direct pasting may cause misalignment with the surrounding background, we perform diffusion forward on the pasted image and perform $s = 10$ denoising steps using m^g_i via Eq. (2), guided by trained text embeddings e. The output of this fusion step is the anomaly image $\hat{x}_i$. This procedure is repeated iteratively: at each level, the anomaly is injected, aligned, and fused. Finally, at the maximum scale ($i = n$), GFD produces the fully anomaly image $\hat{x}_n$.

GFD first constructs a faithful anomaly and then adopts a progressive fusion strategy. At each scale, the anomaly is reinforced and aligned to the token, ensuring that it remains distinguishable while being naturally integrated into the full-resolution image.

5 Experiments

5.1 Experimental setting

Datasets. We conduct our experiments on two established benchmarks, the VisA dataset [41] and MVTec AD 2 dataset [11]. VisA dataset consists of 12 objects spanning 3 distinct domains, featuring complex structures, multiple instances, and numerous small-scale anomalies. MVTec AD 2 dataset consists of 8 product categories, including 2 texture types and 6 object types. Both datasets contain a substantial proportion of small-scale anomalies (around 50%), making them well suited for assessing the effectiveness of our arbitrary-scale anomaly generation method.

Implementation Details. We fine-tune the pre-trained Stable Diffusion v1-4 [26]. We update only the text embedding parameters while keeping all other components frozen. For training, we use one-third of the real anomaly images for each anomaly type. A single unified generative model is trained for each product category to cover all corresponding anomaly types. The weighting coefficient λ is set to 0.001 to ensure that the two loss terms remain on a comparable scale. The crop expansion factor α is set to 4. For each anomaly type, we generate 1,000 synthetic anomaly image-mask pairs to facilitate downstream tasks. These settings remain consistent across all experiments.

Evaluation Metrics. Following SeaS [4], our evaluation contains two levels: whole images and anomaly regions, using 2 metrics: (1) Inception Score (IS) [28] and Intra-cluster pairwise LPIPS distance (IC-LPIPS) [24] for authenticity and

Table 1: Comparison on IS, IC-LPIPS (short for IC-L), IS(a) and IC-LPIPS(a) (short for IC-L(a)) on VisA and MVTec AD 2 datasets. Bold denotes the best performance.

Methods	VisA				MVTec AD 2			
	IS↑	IC-L↑	IS(a)↑	IC-L(a)↑	IS↑	IC-L↑	IS(a)↑	IC-L(a)↑
Crop&Paste [21]	1.24	0.22	–	–	1.37	0.26	–	–
DFMGAN [6]	1.25	0.25	1.38	0.05	1.39	0.29	1.04	0.05
AnomalyDiffusion [14]	1.26	0.25	1.33	0.04	1.40	0.31	1.02	0.04
DualAnoDiff [17]	1.26	0.25	1.80	0.06	1.46	0.31	1.19	0.05
SeaS [4]	1.27	**0.26**	1.81	0.06	1.42	0.32	1.22	0.05
Ours	**1.34**	**0.26**	**2.64**	**0.09**	**1.60**	**0.33**	**1.68**	**0.06**

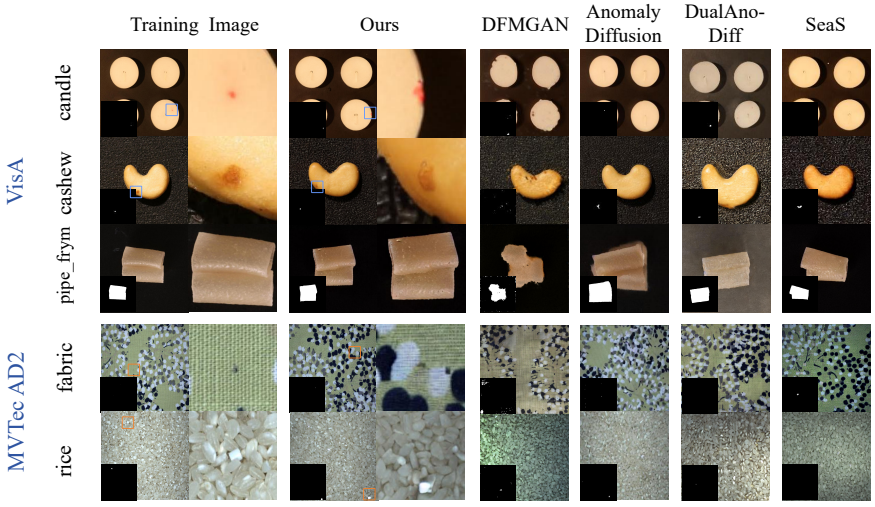

Fig. 6: Qualitative comparison on VisA and MVTec AD 2. Our method generates high-fidelity small-scale anomalies where existing methods fail, while also maintaining structural coherence for larger structural anomalies (e.g., *pipe_frym*).

diversity of whole generated images. (2) IS and IC-LPIPS calculated only in anomaly regions (denoted as IS(a) and IC-LPIPS(a)) for the authenticity and diversity of anomalies. For image-level and pixel-level anomaly detection, we use 3 metrics: Area Under Receiver Operating Characteristic curve (AUROC), Average Precision (AP) and F1-score at optimal threshold (F1-max). We also report Intersection over Union (IoU) for segmentation.

Comparison Methods. For the evaluation of anomaly generation quality, we compare our method against current state-of-the-art anomaly generation approaches, including Crop&Paste [21], DFMGAN [6], AnomalyDiffusion [14], DualAnoDiff [17] and SeaS [4]. To evaluate the generated anomaly image-mask pairs, we train the BiSeNetV2 [34] segmentation model using the data generated by DFMGAN, AnomalyDiffusion, DualAnoDiff, and SeaS. In line with the segmentation experimental setup of SeaS, we train a single unified supervised

Table 2: Comparison of generative models for anomaly detection (Image-level) and anomaly segmentation (Pixel-level). The best-performing result is in bold, the second-best result is underlined.

Generative Models	Image-level			Pixel-level			
	AUROC	**AP**	F_1**-max**	**AUROC**	**AP**	F_1**-max**	**IoU**
VisA							
DFMGAN	63.07	62.63	66.48	75.91	9.17	15.00	9.66
AnomalyDiffusion	76.11	77.74	73.13	89.29	34.16	37.93	15.93
DualAnoDiff	78.06	77.92	74.76	92.06	27.12	31.94	21.48
SeaS	<u>85.61</u>	<u>86.64</u>	<u>80.49</u>	<u>96.03</u>	<u>42.80</u>	<u>45.41</u>	<u>25.93</u>
Ours	**88.15**	**88.54**	**82.01**	**97.28**	**42.91**	**47.03**	**30.15**
MVTec AD 2							
DFMGAN	58.01	61.22	68.91	56.31	10.12	11.53	8.22
AnomalyDiffusion	61.11	63.65	71.15	59.36	13.20	15.33	9.59
DualAnoDiff	63.51	69.71	74.79	64.51	14.62	16.11	11.38
SeaS	<u>64.22</u>	<u>70.10</u>	<u>75.65</u>	<u>71.12</u>	<u>16.54</u>	<u>18.14</u>	<u>12.37</u>
Ours	**66.26**	**72.33**	**76.81**	**77.67**	**20.34**	**24.48**	**16.06**

segmentation model across all product categories. In all comparative experiments, we adopt each method's original prompt and mask configurations, as their specific prompt designs (e.g., SeaS [4]) and mask acquisition strategies (e.g., DualAnoDiff [17]) constitute their core contributions.

5.2 Comparison in Anomaly Image Generation

Quantitative Analysis. We compare our method with state-of-the-art anomaly generation methods in terms of fidelity (IS and IS(a)) and diversity (IC-LPIPS and IC-LPIPS(a)). As presented in Tab. 1, our method demonstrates superior performance on both the VisA and MVTec AD 2 datasets. Specifically, it achieves significant improvements in fidelity metrics, with an average IS(a) increase of 0.83 on VisA (from 1.81 to 2.64) and 0.46 on MVTec AD 2 (from 1.22 to 1.68). Furthermore, the diversity scores (IC-LPIPS and IC-LPIPS(a)) confirm that our model generates highly diverse anomalies.

Qualitative Analysis. We further assess the visual quality of the generated anomaly images. As shown in Fig. 6, existing methods struggle with small-scale anomalies. For categories such as *cashew* and *fabric*, existing methods often produce distorted artifacts or fail to generate anomalies altogether. In contrast, our method generates high-fidelity and realistic anomalies with clear details. Moreover, our model demonstrates robust generalization capabilities on larger scale structural anomalies (e.g., *pipe_frym*), successfully generating realistic structural anomalies.

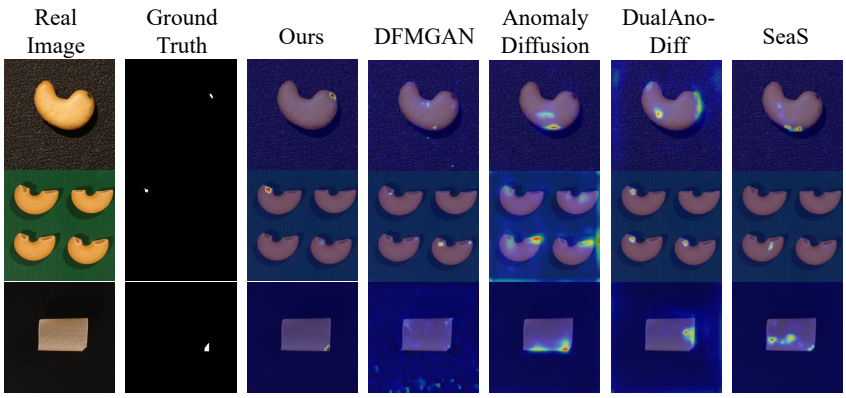

Fig. 7: Qualitative results of supervised anomaly segmentation.

5.3 Training supervised downstream models for anomaly segmentation and detection

We generate 1,000 image-mask pairs for each anomaly type and use them, along with all normal images in the original training sets, to train a unified supervised segmentation model. The models are tested on the remaining images not included in the training set.

Quantitative Analysis. As shown in Tab.2, our method outperforms state-of-the-art competitors on both the VisA and MVTec AD 2 datasets. Specifically, regarding the AUROC metric, our method achieves substantial improvements at both image and pixel levels. On the VisA dataset, our method sets new state-of-the-art records, reaching 88.15% image-level AUROC and 97.28% pixel-level AUROC. Furthermore, on the challenging MVTec AD 2 dataset, we observe remarkable gains, such as a 6.55% increase in pixel-level AUROC compared to the second-best method.

Qualitative Analysis. We visually compare the segmentation results of our method against other approaches in Fig. 7. While existing methods often fail to detect small-scale anomalies, segmentation models trained on our generated image–mask pairs produce precise masks that closely align with the ground truth. As illustrated, models trained with our generated image–mask pairs significantly reduce false negatives and ensure high-quality anomaly detection.

5.4 Ablation Study

Ablation on core components. Tab.3 presents the ablation study on the core modules of our framework. The baseline (a), trained without the Error-Suppressed Multi-Scale Training Strategy (EMT). In other words, it is trained on original images. It exhibits the lowest performance, with IS(a) dropping from 2.64 to 1.84 and Pixel-AP decreasing by over 12%. This indicates that original-resolution training fails to capture the anomaly details. However, training with

Table 3: Ablation on the core modules of our generation model.

Method	Metrics					
	IS(a)	IC-L(a)	AUROC	AP	F_1-max	IoU
(a) w/o EMT	1.84	0.05	93.72	30.76	35.00	19.04
(b) w/o CSP	2.01	0.05	97.22	31.98	37.18	27.07
(c) w/o GFD	2.08	0.05	96.03	33.62	39.51	27.97
(d) All (Ours)	**2.64**	**0.09**	**97.28**	**42.91**	**47.03**	**30.15**

Table 4: Ablation on Multi-Scale Cropping.

Method	Metrics					
	IS(a)	IC-L(a)	AUROC	AP	F_1-max	IoU
w/o MaxSC	2.58	0.07	96.03	33.62	39.51	27.97
w/o MiniSC	2.04	0.05	96.43	35.21	40.59	28.72
w/o InterSC	2.18	0.05	93.48	34.17	39.93	29.97
MaxSC only	1.84	0.05	93.72	30.76	35.00	19.04
MiniSC only	2.40	0.07	94.62	32.83	41.02	28.03
InterSC only	2.41	0.07	96.50	34.90	40.56	27.40
Ours	**2.64**	**0.09**	**97.28**	**42.91**	**47.03**	**30.15**

Table 5: Ablation on the cropping factor.

Method	Metrics						GPU Mem. (GB)
	IS(a)	IC-L(a)	AUROC	AP	F_1-max	IoU	
$\alpha = 2$	2.63	**0.09**	96.86	41.84	45.58	30.07	24.61
$\alpha = 4$ **(Ours)**	**2.64**	**0.09**	**97.28**	**42.91**	**47.03**	**30.15**	20.86
$\alpha = 8$	2.55	0.08	96.21	39.50	43.43	28.62	18.91
$\alpha = 16$	2.34	0.07	96.08	38.87	42.02	28.34	16.18

Table 6: Ablation on GFD.

Method	Metrics					
	IS(a)	IC-L(a)	AUROC	AP	F_1-max	IoU
(a) One-stage	2.08	0.05	96.03	33.62	39.51	27.97
(b) Gen + paste	2.52	0.08	97.02	40.40	44.60	27.65
(c) Gen + fusion@orig	2.16	0.05	96.44	34.28	40.12	28.41
(d) GFD (Ours)	**2.64**	**0.09**	**97.28**	**42.91**	**47.03**	**30.15**

Multi-Scale Cropping alone is insufficient. Without the CSP loss (b), generation fidelity remains suboptimal (IS(a) 2.01), confirming that suppressing interpolation error is critical for realistic texture generation. Moreover, replacing our GFD strategy with standard one-stage generation (c) leads to a significant decline in segmentation accuracy (AP 33.62% vs. 42.91%). This demonstrates that our GFD effectively prevents anomalies from being overwhelmed by the background, ensuring high-quality downstream segmentation.

Ablation on Multi-Scale Cropping. Tab.4 examines the contribution of each scale within our Multi-Scale Cropping strategy (MaxSC, MiniSC and InterSC). Removing the MaxSC results in the largest AP drop (42.91% to 33.62%), indicating that correctly aligning anomaly appearances with their positions is crucial. Excluding the MiniSC leads to the most significant decrease in IS(a) (2.64 to 2.04), confirming this scale as the primary source of fine-grained textures. Notably, removing the InterSC causes AUROC to fall from 97.28% to 93.48%, suggesting that InterSC is vital for bridging the gap between fine-grained texture details and location-aware anomaly appearance. Using only InterSC yields suboptimal performance, proving all three scales provide complementary information essential for capturing the dependency between textures and locations. These results demonstrate that all crops provide complementary information.

Ablation on cropping factor α. We further study the effect of the cropping factor α in our Multi-Scale Cropping. As shown in Tab. 5, a moderate factor ($\alpha = 4$) achieves the best overall performance. The choice of α balances smooth scale transitions with GPU memory consumption. If α is too large, the transition between scales becomes abrupt, failing to bridge texture and positional semantics and diluting the anomaly. Conversely, although a smaller α achieves comparable results to $\alpha = 4$, it over-crops the anomaly region and increases the number of crops, leading to excessive GPU memory consumption. Thus, $\alpha = 4$ strikes the optimal balance.

Ablation on Generation-then-Fusion Denoising. Tab.6 evaluates the design choices in our GFD module. One-stage generation (a) results in low AP as fine details are overwhelmed by the background. While direct pasting (b) improves fidelity, it yields the lowest IoU (27.65%) due to severe boundary misalignment. Performing fusion only at the original scale (c) fails to recover lost details, as seen in the drop in AP compared to our full model. In contrast, our progressive multi-scale fusion (d) achieves the best performance across all metrics. These results highlight that decoupling generation from fusion is essential for high-fidelity anomaly generation and accurate downstream localization.

6 Conclusion

In this paper, we propose UniScale, a unified training and inference framework for high-fidelity industrial anomaly generation across arbitrary scales. During training, we introduce an Error-Suppressed Multi-Scale Training (EMT) strategy to learn detailed location-aware textures and suppress interpolation errors. Additionally, our Generation-then-Fusion Denoising decouples anomaly generation from background fusion, resulting in high-quality anomaly generation across scales. Extensive experiments on the VisA and MVTec AD 2 datasets validate our method. We establish new state-of-the-art records, achieving relative IS(a) improvements of 45.86% and 37.70%.

7 Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No.62176098. The computation is completed on the HPC Platform of Huazhong University of Science and Technology.

References

1. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 18208–18218 (2022)
2. Chen, H., Zhang, Y., Wu, S., Wang, X., Duan, X., Zhou, Y., Zhu, W.: Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. arXiv preprint arXiv:2305.03374 (2023)
3. Chen, N., Huang, M., Chen, Z., Zheng, Y., Zhang, L., Mao, Z.: Customcontrast: A multilevel contrastive perspective for subject-driven text-to-image customization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2123–2131 (2025)
4. Dai, Z., Zeng, S., Liu, H., Li, X., Xue, F., Zhou, Y.: Seas: few-shot industrial anomaly image generation with separation and sharing fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23135–23144 (2025)
5. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)

6. Duan, Y., Hong, Y., Niu, L., Zhang, L.: Few-shot defect image generation via defect-aware feature manipulation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 571–578 (2023)
7. Fan, Z., Yin, Z., Li, G., Zhan, Y., Zheng, H.: Dreambooth++: Boosting subject-driven generation via region-level references packing. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 11013–11021 (2024)
8. Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Jia, X., Lin, L.: InstaGen: Enhancing object detection by training on synthetic dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
9. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
10. Gui, G., Gao, B.B., Liu, J., Wang, C., Wu, Y.: Few-shot anomaly-driven generation for anomaly classification and segmentation. In: European Conference on Computer Vision. pp. 210–226 (2024)
11. Heckler-Kram, L., Neudeck, J.H., Scheler, U., König, R., Steger, C.: The mvtec ad 2 dataset: Advanced scenarios for unsupervised anomaly detection. arXiv preprint arXiv:2503.21622 (2025)
12. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
14. Hu, T., Zhang, J., Yi, R., Du, Y., Chen, X., Liu, L., Wang, Y., Wang, C.: Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 8526–8534 (2024)
15. Huang, S., Gong, B., Feng, Y., Chen, X., Fu, Y., Liu, Y., Wang, D.: Learning disentangled identifiers for action-customized text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7797–7806 (2024)
16. Jian, Y., Yu, F., Singh, S., Stamoulis, D.: Stable diffusion for aerial object detection. In: NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI (2023)
17. Jin, Y., Peng, J., He, Q., Hu, T., Wu, J., Chen, H., Wang, H., Zhu, W., Chi, M., Liu, J., et al.: Dual-interrelated diffusion model for few-shot anomaly image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 30420–30429 (2025)
18. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
19. Li, C.L., Sohn, K., Yoon, J., Pfister, T.: Cutpaste: Self-supervised learning for anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9664–9674 (2021)
20. Li, Y., Feng, Y., Chen, B., Chen, W., Wang, Y., Hu, X., Sun, B., Qu, C., Zhou, M.: Vague prototype-oriented diffusion model for multi-class anomaly detection. In: Forty-first International Conference on Machine Learning (2024)
21. Lin, D., Cao, Y., Zhu, W., Li, Y.: 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. pp. 1–6 (2021)

22. Lv, F., Feng, H., Zhang, Z., Xia, C., Li, Y.: Text2traffic: A text-to-image generation and editing method for traffic scenes. arXiv preprint arXiv:2511.12932 (2025)
23. Niu, S., Li, B., Wang, X., Lin, H.: Defect image sample generation with gan for improving defect recognition. IEEE Transactions on Automation Science and Engineering **17**(3), 1611–1622 (2020)
24. Ojha, U., Li, Y., Lu, J., Efros, A.A., Lee, Y.J., Shechtman, E., Zhang, R.: Few-shot image generation via cross-domain correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10743–10752 (2021)
25. Perlin, K.: An image synthesizer. In: ACM Siggraph Computer Graphics. p. 287–296 (1985)
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
27. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
28. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in Neural Information Processing Systems **29** (2016)
29. Schlüter, H.M., Tan, J., Hou, B., Kainz, B.: Natural synthetic anomalies for self-supervised anomaly detection and localization. In: European Conference on Computer Vision. pp. 474–489 (2022)
30. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8543–8552 (2024)
31. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
32. Wang, T., Zhang, T., Zhang, B., Ouyang, H., Chen, D., Chen, Q., Wen, F.: Patch diffusion: Faster and more data-efficient training of diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
33. Wu, Y., Pan, Q., Zhao, Z., Wang, Z., Long, S., Liang, R.: Soediff: Efficient distillation for small object editing. ACM Transactions on Multimedia Computing, Communications and Applications **21**(11), 1–19 (2025)
34. Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. International Journal of Computer Vision **129**, 3051–3068 (2021)
35. Zavrtanik, V., Kristan, M., Skočaj, D.: Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8330–8339 (2021)
36. Zhang, G., Cui, K., Hung, T.Y., Lu, S.: Defect-gan: High-fidelity defect synthesis for automated defect inspection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2524–2534 (2021)
37. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
38. Zhang, X., Xu, M., Zhou, X.: 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. pp. 16699–16708 (2024)

39. Zhou, D., Li, Y., Ma, F., Zhang, X., Yang, Y.: Migc: Multi-instance generation controller for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6818–6828 (2024)
40. Zhu, C., Li, K., Ma, Y., He, C., Li, X.: Multibooth: Towards generating all your concepts in an image from text. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10923–10931 (2025)
41. Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: European Conference on Computer Vision. pp. 392–408 (2022)