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.

DefectFill: Realistic Defect Generation with Inpainting Diffusion Model for Visual Inspection

Jaewoo Song[1,2] Daemin Park[1] Kanghyun Baek[3] Sangyub Lee[3]
Jooyoung Choi[1] Eunji Kim[1] Sungroh Yoon[1,3,4,*]
[1]Department of Electrical and Computer Engineering, Seoul National University
[2]Global Technology Research, Samsung Electronics
[3]IPAI, [4]AIIS, ASRI, INMC, ISRC, Seoul National University
{woo.song, eoalsqkr12, qor6271, nickyub, jy_choi, kce407, sryoon}@snu.ac.kr

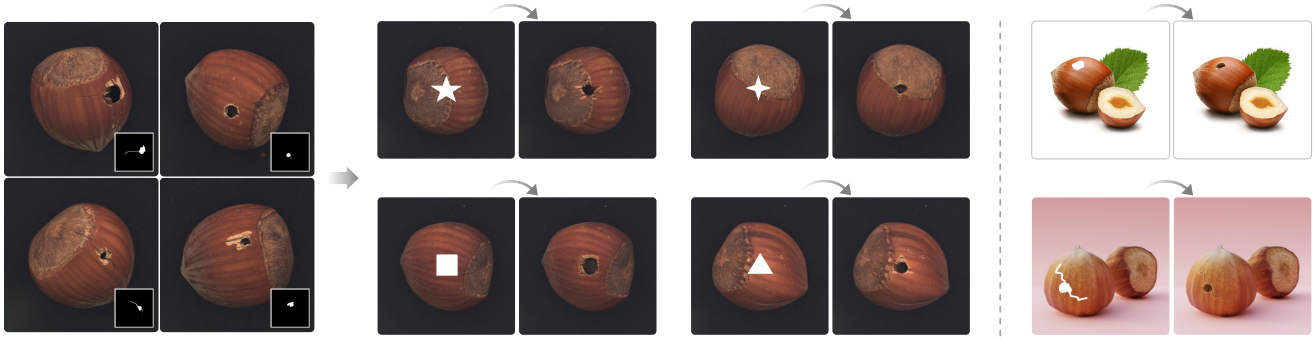


Figure 1. Given a few reference image-mask pairs of a defect (*e.g.* "hole" of a hazelnut), DefectFill learns the defect and realistically fill it onto defect-free objects in desired shapes (*e.g.* star, square, *etc.*), generating new defect images that integrate naturally with the objects. These generated images are then used for visual inspection tasks.

Abstract

Developing effective visual inspection models remains challenging due to the scarcity of defect data. While image generation models have been used to synthesize defect images, producing highly realistic defects remains difficult. We propose DefectFill, a novel method for realistic defect generation that requires only a few reference defect images. It leverages a fine-tuned inpainting diffusion model, optimized with our custom loss functions incorporating defect, object, and attention terms. It enables precise capture of detailed, localized defect features and their seamless integration into defect-free objects. Additionally, our Low-Fidelity Selection method further enhances the defect sample quality. Experiments show that DefectFill generates high-quality defect images, enabling visual inspection models to achieve state-of-the-art performance on the MVTec AD dataset.

*Correspondence to: Sungroh Yoon (sryoon@snu.ac.kr)

1. Introduction

Automating inspection on manufacturing lines is a crucial step in advancing smart factories. In this context, visual inspection focused on defect detection is a critical application for AI models. With substantial amounts of defective data, high-performance models can be developed through supervised learning [6]. However, collecting large quantities of defective data is challenging in real-world settings. For example, in newly established production lines or semiconductor processes with exceptionally low defect rates, it may be difficult or even impossible to acquire enough data.

To overcome the limited availability of defect data, various approaches have been developed, including out-of-distribution (OOD) techniques [21] and anomaly detection (AD) [26], which only use non-defective data, as well as active learning [20] and semi-supervised learning [27] with limited defective data. However, these methods have limitations: defect criteria vary across different problems and often require domain expertise, and they struggle to clas-

sify defect types accurately. To address these issues, some methods propose generating defect images to train visual inspection models [7, 14, 22, 36]. Yet, a key problem remains: defect images generated by existing methods appear unrealistic, lacking the clarity and natural details of real-world defects, which limits their practical effectiveness.

In this paper, we focus on generating realistic defect images to improve the accuracy of visual inspection tasks. To achieve this, we address two key considerations: (1) precisely capturing defect details and (2) seamlessly incorporating these defect features into defect-free images.

We introduce DefectFill, a novel approach for generating realistic and detailed defect images using abundant normal images along with a few reference defect samples. We leverage a pre-trained inpainting diffusion model [24] to remove certain areas of a defect-free image and naturally fill those areas with defects. However, accurately filling defects is challenging, as these features often have entirely different textures or appearances compared to the original object. Therefore, we introduce three loss functions: *defect loss* to capture detailed features of the defect itself, *object loss* to establish the semantic relationship between the defect and the object, and *attention loss* to ensure the word representing the defect focuses precisely on the defect area. These carefully designed loss functions are essential for generating realistic defect images, enabling defects to be naturally and authentically "filled" within objects. To further refine samples, we implement the Low-Fidelity Selection method, which filters out generated images that fail to represent defects clearly, ensuring only high-quality samples are used.

Through extensive experiments, we demonstrate our model's ability to generate realistic defect images that outperform state-of-the-art methods in both qualitative and quantitative evaluations. Finally, by leveraging our high-quality generated defect images, we improve performance in visual inspection downstream tasks such as anomaly classification and localization, showing that DefectFill effectively addresses the shortage of defect data.

Our main contributions include: (1) pioneering the use of an inpainting diffusion model for generating defect images, (2) designing novel loss functions that enable the model to learn embedded defect characteristics within the context of the object, thereby generating realistic defects, (3) introducing the Low-Fidelity Selection method which is used to further enhance the quality of generated samples, and (4) demonstrating that our realistic defect images significantly improve performance in downstream tasks.

2. Related Work

2.1. Anomaly generation

Various approaches have been proposed to mitigate the scarcity of defective data by generating synthetic defects [17, 18, 35, 37]. Crop-Paste [18] and CutPaste [17] synthesize data by extracting in-distribution image patches and repositioning them, while PRN [37] and DRAEM [35] incorporate out-of-distribution elements into normal images to generate additional synthetic anomalies. Since these methods solely rely on data augmentation, their ability to generate truly novel defects remains limited, thus constraining diversity. Additionally, the defects synthesized using cross-distribution images often lack realism.

Recent research has shifted toward direct defect image generation using Generative Adversarial Networks (GANs) [9], including methods like SDGAN [22] and Defect-GAN [36]. However, these approaches require large and diverse defect datasets, which limits their applicability in data-scarce scenarios. DFMGAN [7] addresses this limitation by enabling defect image generation from a small number of reference images, by exploiting a pre-trained StyleGAN2 [15]. Nonetheless, it demands lengthy training times and struggles with generating realistic defects. In contrast to GAN-based models, studies using powerful text-to-image diffusion models [24] have shown promising results. AnomalyDiffusion [14] optimizes word vectors to disentangle the intrinsic characteristics of defects from their positional information, allowing defects to be generated at any specified location. However, these word vectors still fall short in capturing the fine structural details of defects [8], resulting in defects that lack realism.

2.2. Personalization

Leveraging the text-to-image capabilities of diffusion models, personalization research has emerged to learn new objects unknown to these models. This learning process uses a few reference images to enable a unique word token [V^*] to represent the new target concept. Once the concept is learned, prompts containing the [V^*] token can be used to generate new images of this concept. Most studies primarily focus on learning a main object that occupies most of the image, either by optimizing the unique word token [8] or fine-tuning the diffusion model [16, 28].

In contrast, CLiC [29] focuses on learning local concepts rather than the main object and employs cross-attention guidance [5] to transfer these local concepts. We draw inspiration from this approach, though it is primarily designed for realistic scenarios where the target object can naturally exhibit these concepts, unlike defect images. In addition to the previously mentioned studies, there has been an effort to use inpainting diffusion models to learn concepts [33]. This approach focuses on learning a single target image alongside reference images, solely for inpainting that target.

Related to these studies, we aim to learn a defect concept anomalous to objects and generate diverse, realistic defect images to enhance the performance of downstream tasks.

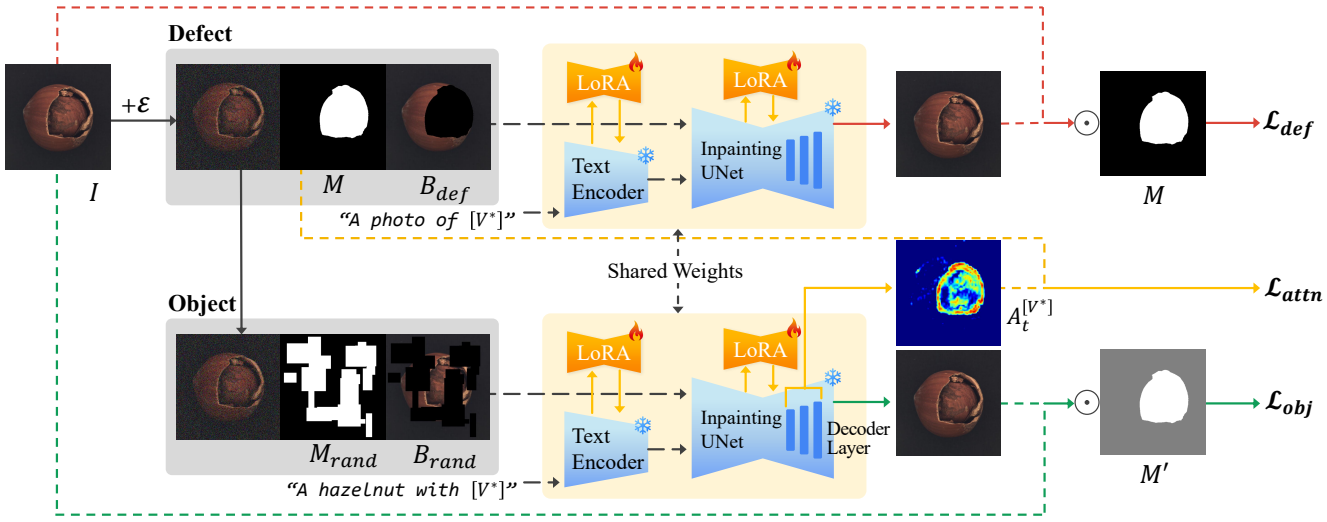


Figure 2. **Defect learning overview.** To fine-tune the inpainting diffusion model, we compute three types of loss ($\mathcal{L}_{def}$, $\mathcal{L}_{attn}$, and $\mathcal{L}_{obj}$) using an image I and a defect mask M. The image I is duplicated, with each copy combined with different masks (M and M_{rand}) and prompts ($\mathcal{P}_{def}$: *"A photo of [V*]"*, and $\mathcal{P}_{obj}$: *"A hazelnut with [V*]"*) as inputs to the model. The model prediction using the defect prompt $\mathcal{P}_{def}$ (upper pipeline) is used to compute $\mathcal{L}_{def}$ and, while the prediction using the object prompt $\mathcal{P}_{obj}$ (lower pipeline) is used to compute $\mathcal{L}_{attn}$ and $\mathcal{L}_{obj}$.

3. Methods

We introduce DefectFill, a novel method for generating diverse and realistic defect images. By fine-tuning a pre-trained inpainting diffusion model, DefectFill efficiently learns defect concepts using only a limited set of reference defect image-mask pairs. During inference, it fills the defect feature into specific areas of defect-free images, thereby enabling the generation of high-quality defect images that enhance performance in visual inspection tasks.

The following sections cover the background on inpainting diffusion models (Sec.3.1), followed by a formal description of our method for learning defects (Sec.3.2) and generating defect images with the learned defects (Sec.3.3). Subsequently, we describe how the generated images can be applied to downstream tasks (Sec.3.4).

3.1. Preliminaries

Latent Diffusion Models. Latent Diffusion Models (LDMs) [24] are a class of diffusion models [12, 30, 32] specifically designed to enhance efficiency by reducing computational complexity. An LDM consists of an encoder $\mathcal{E}$ that maps image I to a latent space $x_0 = \mathcal{E}(I)$, a decoder $\mathcal{D}$ that reconstructs images as $I = \mathcal{D}(x_0)$, and a diffusion model operating in the latent space. The encoder and decoder are pre-trained to accurately reconstruct images from their latent representations such that $\mathcal{D}(\mathcal{E}(I)) = I$, while the diffusion model is trained to predict the noise that needs to be removed from a noisy latent representation.

The forward process of the diffusion model gradually adds Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ to the latent image:

$$x_t = \sqrt{\alpha_t} x_0 + \left(\sqrt{1 - \alpha_t}\right) \epsilon, \tag{1}$$

where $\{\alpha_t\}_{t=1}^{T}$ is a noise scheduler that determines the proportion of noise added at each timestep t. The reverse process reconstructs the latent image from the noisy input x_t. The diffusion model can incorporate a text prompt $\mathcal{P}$ as conditioning, which is encoded by a text encoder τ_θ, and is trained using the following objective:

$$\mathcal{L} = \mathbb{E}_{x_t, t, \epsilon} \left\| \epsilon_\theta(x_t, t, \tau_\theta(\mathcal{P})) - \epsilon \right\|_2^2. \tag{2}$$

Inpainting Diffusion Models. Inpainting Diffusion Models are fine-tuned versions of LDM, specifically designed to fill content within masked areas. These models learn the inpainting task using both a mask M and a background image B where the masked areas are removed. Specifically, the image I and the background B are mapped into the latent space through the encoder, resulting in $x_0 = \mathcal{E}(I)$ and $b = \mathcal{E}(B)$. Gaussian noise is then added to x_0 at a specific timestep t, producing x_t. Subsequently, $x_t^{inpaint}$ is constructed by concatenating x_t, b, and M as follows:

$$x_t^{inpaint} = \text{concat}\left(x_t, b, M\right), \tag{3}$$

and is used for training diffusion models via Eq. (2). This process allows the LDM to learn how to accurately fill the masked areas with appropriate content.

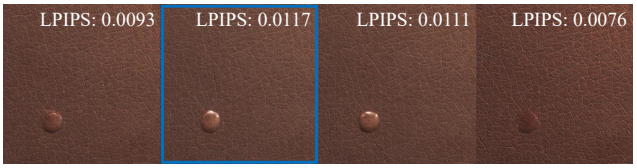


Figure 3. **Low-Fidelity Selection (LFS) for defect of leather's glue.** LFS automatically selects the defect image with the most pronounced expression (blue box) by identifying the sample with the lowest fidelity (highest LPIPS score) in the masked area.

3.2. Learning Defect

We use a stable diffusion inpainting model [24] to leverage its prior knowledge for seamlessly "filling" masked areas with desired defects. To train the model to understand the concept of defects, we fine-tune it using a small set of reference defect images I paired with defect masks M. This fine-tuning enables the model to associate the word token $[V^*]$ with defects. Specifically, to efficiently learn various defects while avoiding overfitting [10], we fine-tune only the text encoder τ_θ and the attention layers by using LoRA [13].

More precisely, we aim to achieve three goals to effectively learn local defects: (1) recognizing defects that are not the main object of the image but rather local features dependent on it, (2) understanding the semantic relationship between the defect and the main object to ensure natural blending, and (3) ensuring the word token $[V^*]$ corresponds to the defect region of the object. To achieve these goals, we propose three loss terms: defect, object, and attention loss, as illustrated in the overall training scheme shown in Fig. 2

Defect Loss. The key loss term, defect loss $\mathcal{L}_{def}$, directly learns the detailed features of the defect concept. By guiding the model to focus exclusively on the intrinsic features of the defect, it enables inpainting of even unusual features that would not typically appear in the object.

First, we sample a timestep $t \sim p(t)$ from the model's timestep distribution and obtain the noise latent x_t using Eq. (1). Next, we prepare the defect mask M for I and generate a background image $B_{def} = (1 - M) \odot I$ where the defect area is masked out. The latent $b_{def} = \mathcal{E}(B_{def})$ is then concatenated with x_t and M to form x_t^{def}:

$$x_t^{def} = \text{concat}\left(x_t, b_{def}, M\right), \tag{4}$$

which serves as input to the model.

To ensure the prompt focuses exclusively on the defect, we define it as $\mathcal{P}_{def}$ = *"A photo of $[V^*]$"*. The text encoder τ_θ encodes this prompt to generate the text condition embedding $c^{def} = \tau_\theta(\mathcal{P}_{def})$. Using these inputs, we optimize the $\mathcal{L}_2$ loss with respect to noise ϵ to reconstruct x_0, but we compute the loss only over the masked region M to avoid reconstructing the background:

$$\mathcal{L}_{def} = \mathbb{E}_{x_t^{def},t,\epsilon}\left[\left\| M \odot (\epsilon - \epsilon_\theta(x_t^{def}, t, c^{def}))\right\|_2^2\right]. \tag{5}$$

Object Loss. The object loss $\mathcal{L}_{obj}$ learns both the defect and its relationship to the object in which it appears. This ensures the defect blends naturally within the object.

The $\mathcal{L}_{obj}$ term shares the same sampled values for ϵ, t, and x_t as the defect loss. To capture the full semantic context of the object, we create a mask with 30 random boxes, M_{rand}, and train the model to fill in the occluded information across the entire image. Similar to the defect loss, we obtain the conditioning background $B_{rand} = (1 - M_{rand}) \odot I$ and its latent $b_{rand} = \mathcal{E}(B_{rand})$. This b_{rand} is then concatenated with x_t and M_{rand} to form x_t^{obj}:

$$x_t^{obj} = \text{concat}\left(x_t, b_{rand}, M_{rand}\right). \tag{6}$$

To express the object's possession of the defect, we set the prompt as $\mathcal{P}_{obj}$ = *"A [Object] with $[V^*]$"* and obtain the text embedding $c^{obj} = \tau_\theta(\mathcal{P}_{obj})$. Although it is essential to learn the semantic context of the defect within the object, capturing the fine details of the defect itself is also crucial for authentic inpainting. To address this, we apply a weight of 1 to the defect mask areas and a weight of α, less than 1, to the background areas, producing an adjusted mask M':

$$\mathcal{L}_{obj} = \mathbb{E}_{x_t^{obj},t,\epsilon}\left[\left\| M' \odot (\epsilon - \epsilon_\theta(x_t^{obj}, t, c^{obj}))\right\|_2^2\right], \quad M' = M + \alpha \cdot (1 - M). \tag{7}$$

Attention Loss. We also utilize cross-attention maps from the forward pass for $\mathcal{L}_{obj}$. The maps for a specific token represent the layout of the corresponding object, allowing the model to focus more precisely on that region. This helps the model better attend to the defect's features, resulting in higher-fidelity defect generation. Since the encoder in the UNet [25] does not effectively represent the layout of the corresponding token object [4], we use only decoder-layer maps. To handle varying spatial sizes across decoder layers, we resize them to match the latent size, then average those of the $[V^*]$ token to obtain $A_t^{[V^*]}$. Finally, we compute the $\mathcal{L}_2$ loss with the defect mask M, increasing values in the defect region while reducing them in the background:

$$\mathcal{L}_{attn} = \mathbb{E}\left[\left\| A_t^{[V^*]} - M\right\|_2^2\right]. \tag{8}$$

DefectFill Loss. Finally, we fine-tune the model using a linear combination of these three loss terms:

$$\mathcal{L}_{ours} = \lambda_{def} \cdot \mathcal{L}_{def} + \lambda_{obj} \cdot \mathcal{L}_{obj} + \lambda_{attn} \cdot \mathcal{L}_{attn}. \tag{9}$$

The weights for each term are set to 0.5, 0.2, and 0.05, based on experiments that account for the scale of each loss.

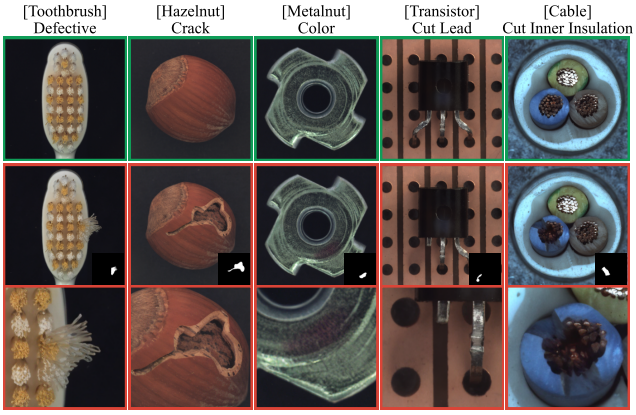


Figure 4. **Generated Defects by DefectFill.** The first row displays the normal images (green boxes), while the second row shows the generated defect images along with their masks, and the third row provides zoomed-in views of the defects (red boxes). The zoomed images highlight the realistic and detailed rendering of the defects.

3.3. Generating Defect

Sampling. After fine-tuning the inpainting diffusion model with our DefectFill loss to learn the defect concept, we utilize a widely adopted diffusion-based inpainting pipeline [1, 32] to generate diverse defect samples. Specifically, as input, we provide a defect-free image I along with a mask M indicating the exact area intended for defect placement. At each inference step t, we replace the latent representation's background area outside the mask with the latent of the defect-free image diffused with Eq. (1). This approach ensures that the model modifies only the masked region while preserving the background that should remain unchanged. This approach maintains the structure of the original image, allowing for seamless integration of defects without affecting the overall image quality.

Low-Fidelity Selection. Finally, we propose an additional method for selecting samples where the defect is more accurately filled. Since the diffusion model generates diverse samples depending on the initial latent inputs, and due to the nature of the inpainting diffusion model, the masked area is occasionally overly reconstructed, resulting in lower-quality defect. To mitigate this issue, we select the least reconstructed image from the eight samples generated using the same normal image I and defect mask M. This selection is based on a reconstruction metric (*e.g.* PSNR, SSIM [34], LPIPS [38]) measured only within the masked region (as shown in Fig. 3). This simple yet effective process filters out unclear cases and improves defect generation quality. In particular, for downstream tasks using generated defect images, this approach allows us to automatically select high-quality defects samples without manual effort. In our case, we employ LPIPS as the reconstruction metric.

3.4. Applying to Visual Inspection

The generated high-quality defect images are used to train a visual inspection model. First, we learn the concept for each defect category (Sec.3.2) and generate defective images (Sec.3.3) for each category. After that, for classification, we train standard classification models (*e.g.* ResNet [11]) using the generated images labeled by defect category. For localization, we train segmentation models (*e.g.* UNet [25]) with normal and synthesized defect images along with their corresponding masks, optimizing with focal loss [19].

4. Experiments

Dataset. We evaluate DefectFill on the MVTec AD Dataset [2], which consists of 15 industrial objects with multiple defect categories. Each category contains hundreds of normal images and approximately 20 defect images with masks. Instead of traditional anomaly detection, we generate defect images by training on one-third of the defect image-mask pairs and applying the model to the remaining two-thirds of masks with normal images. For reliable quantitative results, we evaluate on 10 objects, while all objects are used for qualitative analysis.

Implementation Details. Our approach leverages the Stable-Diffusion-2-inpainting model [24], fine-tuning the text encoder and UNet's attention layers with LoRA (rank 8) [13]. We use a learning rate of 2e-4 for the UNet and 4e-5 for the text encoder. Inference is conducted with a DDIM [31] scheduler with 50 denoising steps. Additional details are provided in the supplementary materials.

Metric. We evaluate defect generation quality using Kernel Inception Distance (KID) [3] for quality and IC-LPIPS [23] for diversity, excluding FID and IS due to their limitations on smaller or unreferenced datasets. For defect inspection, we measure classification accuracy, Area Under the ROC Curve (AUROC), Average Precision (AP), F_1-max, and Per Region Overlap (PRO).

Baselines. We compare our method against two state-of-the-art defect generation methods: DFMGAN [7], a two-stage GAN-based approach, and AnomalyDiffusion [14], a text-to-image diffusion model that disentangles the appearance and spatial attributes of defects.

4.1. Defect Generation Evaluation

Qualitative Results. The generated results are shown in Fig. 4. The first row displays the normal images, the second row shows the generated defect images using the mask in the lower right, and the third row provides a zoomed-in view of the generated defects. Despite using custom-drawn

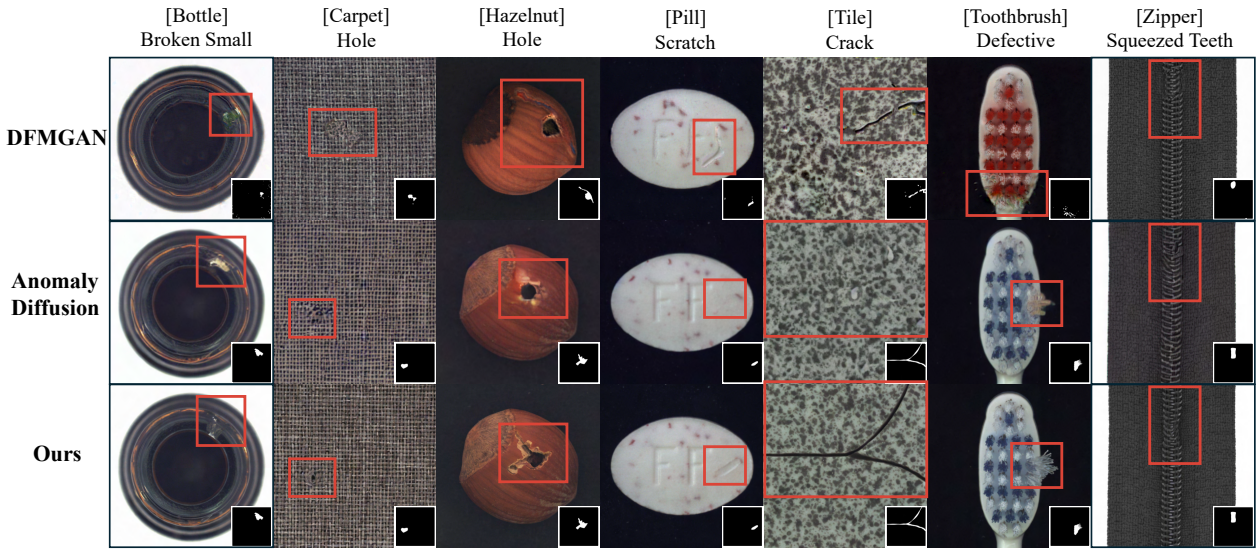


Figure 5. **Defect Generation Comparisons.** This figure compares the quality of defect images generated by our method (bottom row) with baseline approaches. Our method produces the most realistic results, with defects that blend seamlessly into the objects.

Objects	DFMGAN*		DFMGAN†		AnoDiff‡		Ours	
	KID↓	IC-LPIPS↑	KID↓	IC-LPIPS↑	KID↓	IC-LPIPS↑	KID↓	IC-LPIPS↑
bottle	70.90	0.12	76.75	0.15	131.52	**0.17**	**30.99**	0.12
capsule	40.63	0.10	182.83	0.17	54.97	**0.18**	**5.60**	**0.18**
carpet	**25.14**	0.13	46.02	**0.25**	149.83	0.22	50.37	0.22
hazelnut	21.16	0.24	30.94	**0.33**	50.61	0.31	**1.13**	0.31
leather	75.85	0.17	501.61	**0.51**	244.75	0.40	**74.66**	0.30
pill	123.52	0.16	70.15	**0.23**	77.69	**0.23**	**8.76**	**0.23**
tile	85.28	0.22	254.82	0.25	314.30	**0.48**	**45.14**	0.44
toothbrush	46.49	0.18	61.43	**0.20**	73.31	0.18	**3.19**	0.15
wood	68.13	0.34	406.61	**0.35**	83.94	**0.35**	**4.72**	**0.35**
zipper	78.08	**0.27**	35.74	0.26	126.65	0.24	**34.91**	0.20

Table 1. **Generation Comparison.** This table presents the average KID and IC-LPIPS scores, computed from 1,000 generated images per defect category and averaged across all categories for each object. Our method achieves the best KID scores for all objects except carpet and the highest IC-LPIPS scores for capsule, pill, and wood. DFMGAN*: scores taken directly from the paper. DFMGAN†: scores reproduced by us. AnoDiff‡: scores measured from the generated dataset (poor samples filtered) on their official page.

masks which are unseen during training, the model generates authentic and well-aligned defects. Notably, for the hazelnut, the model produces a realistic defect that aligns with the object's semantics, even with an unrealistic mask shape for the crack, demonstrating its strong generalization ability. Additionally, the detailed texture within the hazelnut is observable and highlights the realism of the defects.

Fig. 5 presents a qualitative comparison with the baselines. For AnomalyDiffusion, we use the same normal image and mask, while DFMGAN cannot use the same base image as it generates both normal and defect images directly. In the hazelnut case, both baselines struggle with the texture around the hole, whereas our method produces realistic defects that blend seamlessly with the object texture, handling irregular mask shapes and demonstrating DefectFill's robustness. In cases such as carpet and tile, where defects are small or thin, the baselines either fail to capture them accurately or omit them entirely, while our model generates well-defined defects. For the toothbrush, DFMGAN blurs the masked area, and AnomalyDiffusion generates defects with colors misaligned with the object context. In contrast, our model produces a realistic blueish defect that reflects the object's context (similar to how the toothbrush in Fig. 4 appears yellowish). This demonstrates our model's ability to integrate object semantics into defect generation.

Objects	DFMGAN†				AnoDiff*				AnoDiff‡				Ours			
	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO
bottle	0.96	0.80	0.74	0.84	0.99	0.94	0.87	0.94	0.99	0.91	0.83	0.94	**1.00**	**0.96**	**0.90**	**0.97**
capsule	0.74	0.04	0.09	0.34	0.99	0.57	0.60	0.95	0.98	0.41	0.44	0.84	**1.00**	**0.75**	**0.69**	**0.96**
carpet	0.95	0.62	0.60	0.84	**0.99**	0.81	0.75	0.92	0.97	0.74	0.68	0.82	**0.99**	**0.92**	**0.86**	**0.96**
hazelnut	**1.00**	0.94	0.87	0.96	**1.00**	0.97	0.91	0.97	**1.00**	0.96	0.90	0.97	**1.00**	**0.99**	**0.94**	**0.99**
leather	0.96	0.56	0.56	0.84	**1.00**	0.80	0.71	**0.98**	**1.00**	0.80	0.72	**0.98**	**1.00**	**0.91**	**0.83**	**0.98**
pill	0.99	0.89	0.86	0.91	**1.00**	0.97	0.91	0.97	**1.00**	0.97	0.91	0.95	**1.00**	**0.98**	**0.93**	**0.98**
tile	0.99	0.94	0.87	0.96	0.99	0.94	0.86	0.96	0.99	0.95	0.87	0.97	**1.00**	**0.97**	**0.90**	**0.98**
toothbrush	0.98	0.60	0.61	0.89	**0.99**	0.77	0.73	0.91	**0.99**	0.71	0.68	0.91	**0.99**	**0.89**	**0.82**	**0.94**
wood	0.72	0.37	0.41	0.73	0.99	0.85	0.75	0.94	0.97	0.79	0.72	0.89	**1.00**	**0.93**	**0.86**	**0.98**
zipper	0.99	0.82	0.76	0.95	0.99	0.86	0.79	0.96	**1.00**	0.87	0.80	**0.97**	**1.00**	**0.90**	**0.84**	**0.97**

Table 2. **Localization Comparison.** The table presents AUROC, AP, F_1-max, and PRO scores for localization evaluation using a UNet trained on generated defect images. Our method achieves the highest performance across all metrics and objects. AnoDiff*: scores reported in the paper. The others: described in Tab. 1.

Objects	DFMGAN*	DFMGAN†	AnoDiff*	AnoDiff‡	Ours
bottle	56.59	63.41	90.70	95.35	**97.56**
capsule	37.27	25.00	66.67	45.33	**87.50**
carpet	47.31	42.11	58.06	64.52	**87.72**
hazelnut	81.94	86.96	85.42	89.58	**100.00**
leather	49.73	32.20	61.90	65.08	**93.22**
pill	29.52	44.44	59.38	64.58	**97.53**
tile	74.85	81.82	84.21	96.49	**100.00**
wood	49.02	45.16	71.43	78.57	**100.00**
zipper	27.64	45.45	69.51	85.37	**90.91**

Table 3. **Classification Comparison.** The table shows classification accuracy (%) when a ResNet-34 is trained on generated defect images for defect category prediction. Our method achieves the highest performance across all objects. AnoDiff*: scores reported in the paper. The others: described in Tab. 1.

Quantitative Results. Tab. 1 compares the KID and IC-LPIPS scores of our method with baseline approaches across various objects. For evaluation, we generate 1,000 images for each defect category within each object, ensuring that all metrics, including KID, are calculated using only defect images excluded from the training set. This approach is necessary as KID often produces overly optimistic values when models overfit and are evaluated on training data.

Our method outperforms the baselines in KID scores across most objects. For IC-LPIPS, it also achieves the best scores on three objects (capsule, pill, wood). In the case of leather, DFMGAN† and AnoDiff‡ score significantly higher, but this is primarily due to their generation of diverse yet low-quality samples across various masks. The high KID values for these methods further confirm that the quality of their generated defect images is low.

4.2. Visual Inspection Evaluation

To demonstrate that the realistic images generated by DefectFill can enhance performance in downstream visual inspection tasks, we apply it to two tasks: classification and localization. Following the experimental setup of AnomalyDiffusion [14], we use ResNet-34 [11] for classification and UNet [25] for localization. As outlined in the quantitative results (Sec. 4.1), we generate 1,000 defects per category and train the models on this data. Testing is conducted on the remaining two-thirds of the dataset.

Classification. As shown in Tab.3, our method achieves higher classification accuracy across all objects compared to the baselines. Notably, there is a significant improvement for objects with small defect areas, which are typically challenging to generate meaningful defects for, such as capsule ($66.67\% \rightarrow 87.50\%$) and pill ($64.58\% \rightarrow 97.53\%$).

Localization. The UNet is trained to predict defect locations, and the predictions are evaluated using various metrics. As shown in Tab. 2, our model achieves the best performance across all metrics and objects. The capsule is a particularly challenging object for localization, yet our model significantly outperforms the baseline with a notable improvement in AP score ($0.57 \rightarrow 0.75$).

4.3. Ablation Studies

Inpainting Ablation. We conduct an ablation study to evaluate the impact of leveraging the inpainting diffusion model and our defect-specific loss tailored for this model (Fig. 6). As mentioned in Sec. 2, CLiC [29] is a method that learns local concepts without using the inpainting diffusion model. However, the generated results tend to focus on reconstruction rather than creating actual holes (left image). This is because, unlike general local concepts, the defect concept we aim to learn is an unusual concept unknown to the model's prior. When applying our defect-specific loss (middle image), the model better learns the defect features, resulting in more accurately formed holes. However, the

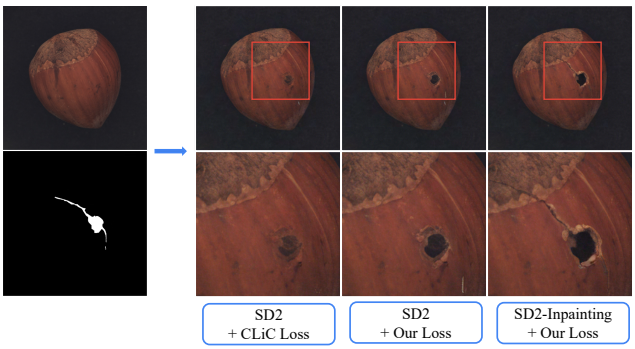


Figure 6. **Inpainting Ablation.** Ablation study comparing three setups: applying CLiC [29] loss to vanilla Stable Diffusion (SD2+CLiC Loss), replacing CLiC with our loss (SD2+Our Loss), and our full approach (SD2-Inpainting+Our Loss). Using the inpainting model with our loss is necessary to produce realistic defects that align well with both the mask and the object.

thin regions of the mask are still neglected, and the texture around the hole doesn't blend well with the surrounding hazelnut texture. Finally, by leveraging the inpainting diffusion model's strong prior for filling, we generate realistic defects that blend naturally with their surroundings (as shown by the light brown texture around the hole in the right image), and aligning with the thin mask regions.

Loss Ablation. As described in Sec. 3.2, we structure our loss function with three terms to achieve three specific goals. To illustrate the contribution of each term, we perform an ablation study. Fig. 7 shows the defect generation results when each loss term is omitted during training. When the defect loss is excluded, the model tends to reconstruct rather than generate defects. This occurs because the inpainting diffusion model fails to learn the distinctive characteristics of defects and instead fills the masked area with just plausible context. Without the object loss, the model lacks semantic alignment with the object, leading to unnatural defect generation. For example, the middle section of a zipper may appear fused, or a hole may look like it's placed on a carpet rather than genuinely puncturing it. Lastly, when the attention loss is omitted, the model struggles to focus accurately on the defect mask area, resulting in lower defect fidelity (*e.g.* an awkwardly split zipper or an incomplete hole). Finally, by combining all loss terms, we achieve realistic defects seamlessly filled onto objects.

Low Fidelity Selection. Our simple yet effective Low-Fidelity Selection method enables high-quality defect sampling without human effort. As shown in Fig. 3, it intuitively selects qualitatively good samples. Additionally, as reported in Tab. 4, it improves both the quality (KID) and the diversity (IC-LPIPS) of generated defects.

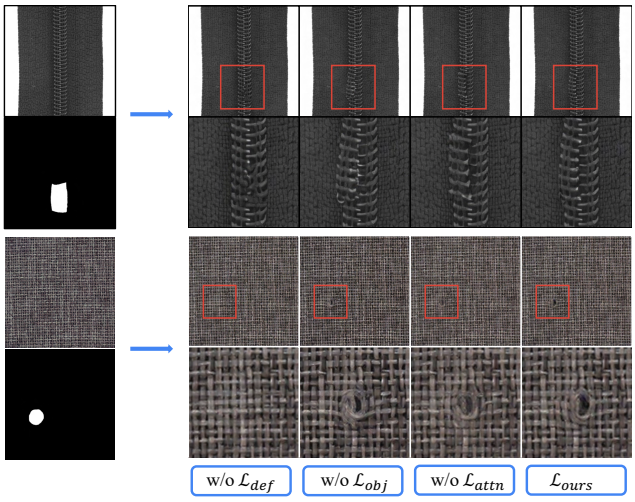


Figure 7. **Loss Ablation.** This figure illustrates the impact of each loss term on generated defect quality. We show the results when each loss term is individually removed during fine-tuning, as well as the result when all terms are used together. Utilizing all loss terms results in realistic defects that align well with the context.

Ours w/o LFS		Ours	
KID↓	IC-LPIPS↑	KID↓	IC-LPIPS↑
28.485	0.247	**25.947**	**0.250**

Table 4. **Generation Comparison with Low-Fidelity Selection.** The application of LFS shows improvements in quality (KID) and diversity (IC-LPIPS). The values represent averages calculated for each defect category, and then averaged across objects.

5. Conclusions

In this work, we present DefectFill, a novel approach that fine-tunes an inpainting diffusion model to generate realistic and high-fidelity defect images. Our method achieves state-of-the-art performance in both generation quality and visual inspection tasks on the MVTec AD dataset, demonstrating its effectiveness even when limited reference samples are available. These strengths make DefectFill particularly well-suited for widespread industrial applications, especially in scenarios where defect images are scarce.

Limitations. While our method excels at generating localized defects—a common real-world scenario—it is less effective for global structural defects that affect the entire object, such as misalignment. This limitation arises because our inpainting-based approach, which focuses on local masked regions. Addressing such global defects remains an area for future research, though our method already robustly handles the majority of practical defect cases, where localized defects are the primary concern.

Acknowledgement This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [No. 2022R1A3B1077720], Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2024.

References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42 (4):1–11, 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. 5

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 5

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023. 4

[5] 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. 2

[6] Wangzhe Du, Hongyao Shen, Jianzhong Fu, Ge Zhang, Xuanke Shi, and Quan He. Automated detection of defects with low semantic information in x-ray images based on deep learning. *Journal of Intelligent Manufacturing*, 32:141–156, 2021. 1

[7] 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, 5

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[10] Cusuh Ham, Matthew Fisher, James Hays, Nicholas Kolkin, Yuchen Liu, Richard Zhang, and Tobias Hinz. Personalized residuals for concept-driven text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8186–8195, 2024. 4

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 7

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

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4, 5

[14] 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. 2, 5, 7

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020. 2

[16] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 2

[17] 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

[18] 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

[19] T Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017. 5

[20] Xiaoming Lv, Fajie Duan, Jia-Jia Jiang, Xiao Fu, and Lin Gan. Deep active learning for surface defect detection. *Sensors*, 20(6):1650, 2020. 1

[21] Ibrahima J Ndiour, Nilesh A Ahuja, and Omesh Tickoo. Subspace modeling for fast out-of-distribution and anomaly detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3041–3045. IEEE, 2022. 1

[22] Shuanlong Niu, Bin Li, Xinggang Wang, and Hui Lin. Defect image sample generation with gan for improving defect recognition. *IEEE Transactions on Automation Science and Engineering*, 17(3):1611–1622, 2020. 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. 5

[24] 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. 2, 3, 4, 5

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 4, 5, 7

[26] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2022. 1

[27] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. *arXiv preprint arXiv:1906.02694*, 2019. 1

[28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, pages 22500–22510, 2023. 2

[29] Mehdi Safaee, Aryan Mikaeili, Or Patashnik, Daniel Cohen-Or, and Ali Mahdavi-Amiri. Clic: Concept learning in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6924–6933, 2024. 2, 7, 8

[30] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[32] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3, 5

[33] Luming Tang, Nataniel Ruiz, Qinghao Chu, Yuanzhen Li, Aleksander Holynski, David E Jacobs, Bharath Hariharan, Yael Pritch, Neal Wadhwa, Kfir Aberman, et al. Realfill: Reference-driven generation for authentic image completion. *ACM Transactions on Graphics (TOG)*, 43(4):1–12, 2024. 2

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[35] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draema discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8330–8339, 2021. 2

[36] 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

[37] Hui Zhang, Zuxuan Wu, Zheng Wang, Zhineng Chen, and Yu-Gang Jiang. Prototypical residual networks for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16281–16291, 2023. 2

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5