Training-Free Industrial Defect Generation with Diffusion Models

Ruyi Xu[1] Yen-Tzu Chiu[1] Tai-I Chen[1,2] Oscar Chew[2] Yung-Yu Chuang[1] Wen-Huang Cheng[1†]
[1]National Taiwan University, [2]ASUS
{xuruyi, irisqiu0106}@cmlab.csie.ntu.edu.tw, {cyy, wenhuang}@csie.ntu.edu.tw, {lawrence2_chen, oscar_chew}@asus.com

Abstract

Anomaly generation has become essential in addressing the scarcity of defective samples in industrial anomaly inspection. However, existing training-based methods fail to handle complex anomalies and multiple defects simultaneously, especially when only a single anomaly sample is available per defect type. To address this issue, we propose TF-IDG, a novel training-free defect generation framework capable of generating diverse anomaly samples in a one-shot setting. We propose a Feature Alignment strategy that provides fine-grained appearance guidance by minimizing the distributional gap between generated and real defects with high complexity. Additionally, we introduce an Adaptive Anomaly Mask mechanism to mitigate the issue of defects with small regions being ignored during the generation process, enhancing consistency between synthetic defects and their corresponding masks. Finally, we incorporate a Texture Preservation module that extracts background information from anomaly-free images, ensuring that the visual properties of synthetic defects are seamlessly integrated into the image. Extensive experiments demonstrate the effectiveness of our method in generating accurate and diverse anomalies, further leading to superior performance in downstream anomaly inspection tasks. Our code is available at https://github.com/rubymiaomiao/TF-IDG.

1. Introduction

Industrial anomaly inspection algorithms play crucial roles in manufacturing. Anomaly detection involves assessing an entire image to determine whether any defects are present, while anomaly localization goes further by specifying the exact defective areas on the image. Anomaly classification, a more complex task, identifies specific defect types with corresponding characteristics.

Despite the importance of this field, the effectiveness and robustness of machine learning-based industrial anomaly inspection are often limited, primarily due to the scarcity of defective samples from a real-world production line. In this case, numerous studies have explored unsupervised methods [27, 36, 51, 55] and few-shot supervised approaches [7, 49, 53], which mainly leverage abundant normal samples or a small number of anomaly samples to construct the decision boundaries. Although these approaches perform reasonably well in anomaly detection, their ability to locate and classify versatile anomalies is challenged due to the inadequate integration of real-world defects and the sparse sampling of anomaly distributions, which limits the capture of full variability and increases model complexity to satisfy diversity requirements. A promising solution is synthesizing realistic anomaly samples, enabling more effective algorithms available in a supervised setting.

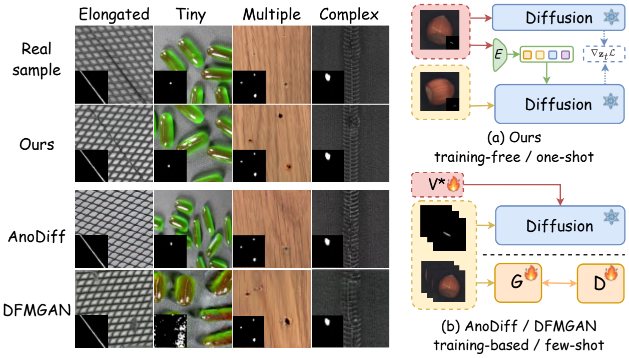

Figure 1. **Left:** Comparison of defect generation quality across four characteristics: elongated, tiny, multiple, and complex defects. Our method offers a clear advantage over existing approaches in handling these types. **Right:** Different frameworks for defect generation. Our model can adapt to limited training data and generate defects for multiple classes with a unified framework.

Defect generation methods can be categorized into two main types: (1) Manual editing: Approaches such as Cut-Paste and Crop-Paste [23, 25, 39] paste abnormal or unrelated textures onto target images but fail to capture the structural coherence and contextual realism of real-world defects, limiting their potential to improve model accuracy. (2) Generative models: GANs and diffusion models are widely applied, but their direct use with general image priors fails to generate realistic defects due to distributional gaps. GAN-based methods [31, 52] generally demand large defect datasets. DFMGAN [8] enables few-shot fine-tuning on a small number of samples but suffers from misalignment between defects and masks. AnoDiff [20] fine-tunes textual embeddings to learn anomaly appearance and positioning. However, due to parameter size constraints and training sample requirements, it struggles with subtle defects and object reconstruction. In the extreme case where only one anomaly sample is available, training-based methods struggle to capture diverse defect characteristics, further restricting the generation ability. This indicates the need for a training-free approach to generate realistic anomalies.

Here, we propose TF-IDG, a training-free method using an image-to-image framework to generate defects. Our approach builds upon Anydoor [3], providing an intuitive, zero-shot image-editing framework based on reference images, eliminating the need for fine-tuning textual space or supplementary models to learn each defect type. Leveraging the self-supervised model DINOv2 [33], we extract expressive features from reference defects while leveraging its robust instance retrieval capability to enhance defect diversity. ControlNet [54] additionally contributes by preserving the shape and fine-grained details of defects. This architecture facilitates rapid extraction of diverse defect visual features and ensures precise alignment between synthetic anomalies and their corresponding masks, making it highly well-suited to industrial defect generation.

The original Anydoor has notable limitations: as shown in Fig. 3, it struggles with geometric harmonization in complex object reconstructions [44], limits multi-object synthesis, and often fails to integrate defects seamlessly into backgrounds. To mitigate these limitations, our pipeline TF-IDG enhances realism by focusing on texture preservation and precise alignment to defect features. We introduce a feature alignment strategy calculating the minimum distribution distance between synthetic and real defects, pulling in the real anomalous structural features, and then generating anomalies in the mask area in each denoise step. This novel strategy allows our model to capture the abnormal visual structure of complex objects.

In cases where small or subtle defects are occasionally ignored during multi-defect generation, we incorporate an additional guidance function to identify these missed regions and align them with real defect characteristics. Additionally, we observe that image embedding guiding generation may sometimes neglect background factors, resulting in synthetic anomalies inconsistent with the object's textures or colors; we introduce a Texture Preservation module, which employs Adaptive Instance Normalization (AdaIN) [21] and Dual Source Attention. It preserves the appearance, color, and texture of the source image, enabling the generation of more cohesive and realistic anomalies.

As illustrated in Fig. 1, our method effectively tackles the challenges of generating complex and fine-grained defects, pioneering the production of realistic industrial anomalies within a training-free framework. By addressing the scarcity of industrial anomaly samples and eliminating the computational overhead of training, our approach reduces model redundancy and provides a significant advantage for practical applications. In summary, our contributions are as follows:

- We propose TF-IDG, a novel training-free framework for industrial defect generation that synthesizes diverse, fine-grained, and structurally aligned anomaly samples without requiring any model retraining.
- We introduce a gradient-guided optimization loss that integrates Feature Alignment strategy, Adaptive Anomaly Mask module, and Texture Preservation to ensure appearance fidelity, enhanced coverage of subtle defects, and realistic integration of anomalies.
- Extensive experiments demonstrate that TF-IDG surpasses existing anomaly generation models, achieving superior generation quality and enhancing performance in downstream anomaly inspection tasks.

2. Related Work

2.1. Image Editing with T2I Diffusion Models

The advancement of large-scale text-to-image (T2I) models, primarily based on the latent diffusion model (LDM) [34], has established a robust foundation for data augmentation across various fields. Numerous studies have further adapted diffusion models to specific domains through fine-tuning or editing techniques, advancing controllable generation beyond general text prompts toward fine-grained structural, semantic, and user-intent guidance.

Early works [6, 18] achieved category-conditioned generation using classifier gradients, while classifier-free guidance [16] enabled more flexible control without external models by leveraging the difference between conditional and unconditional predictions. More recent efforts integrate richer conditional inputs such as masks, depth, pose, or sketches via lightweight adapters or architectural modifications [24, 46, 54] to enhance spatial and semantic control. Parallel to structural guidance, personalization methods adapt pretrained diffusion models to new concepts. Fine-tuning approaches like DreamBooth [37] and Textual Inversion [10] bind custom tokens or embeddings using a few reference images. In contrast, self-supervised methods such as Paint-by-Example [47], ObjectStitch [43], and Anydoor [3]inject style and identity directly from examples without updating model weights. Our framework builds on Anydoor [3], which transfers custom objects into backgrounds using a mask, combining DINOv2 [33] for feature retrieval and ControlNet [54] for structure preservation. However, it struggles with geometric alignment in complex scenes, limits multi-defect synthesis, and fails to harmonize defects with backgrounds. Our method overcomes these issues, improving realism and structural coherence.

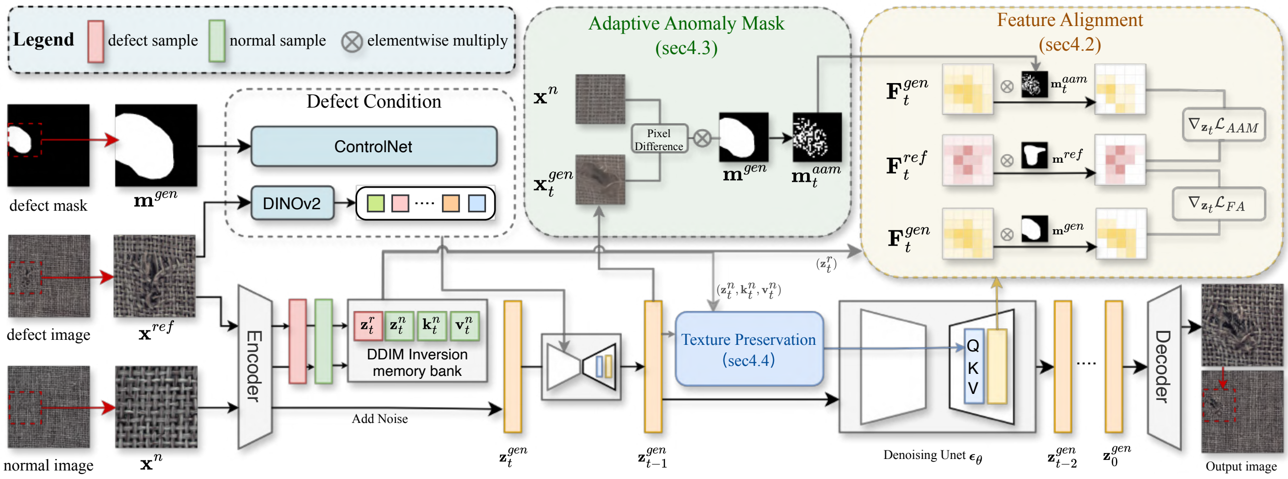

Figure 2. Illustration of TF-IDG model. Given a normal sample, TF-IDG aims to use three modules to improve the quality of generated defects in the target mask $\mathbf{m}^{gen}$ region during the inference process. First, $\mathcal{L}_{FA}$ in Feature alignment is used to guide the generated feature $\mathbf{F}_t^{gen}$ to be closer to the real feature $\mathbf{F}_t^{ref}$ (Sec. 4.2). Then, $\mathcal{L}_{AAM}$ locally optimizes $\mathbf{F}_t^{gen}$ within the defect-misaligned region, denoted as $\mathbf{m}_t^{aam}$ and defined by the Adaptive Anomaly Mask (Sec. 4.3). Finally, the Texture Preservation module was introduced to enhance the coherence of generated images by incorporating background information into the synthesis process in later iterations (Sec. 4.4).

2.2. Anomaly Generation

Anomaly generation is an effective strategy to augment defective industrial datasets.Various anomaly synthesis methods generate abnormal samples and corresponding annotations to address data scarcity and enhance the performance of downstream industrial inspection tasks.

Early efforts in anomaly synthesis primarily relied on hand-crafted strategies, such as Crop-Paste [25], DRAEM [51], and PRN [53], which simply crop and paste anomalies from anomaly images to normal ones, lacking fidelity and diversity. To address these limitations, generative models have been widely used to learn complex defect distributions and synthesize realistic anomalies. Among GAN-based methods, SDGAN [31], MultistageGAN[26], DefectGAN [52], and DFMGAN [8], DFMGAN supports few-shot anomaly generation with masks via a two-stage strategy: pretraining on normal images and fine-tuning on limited anomalies. Recent diffusion models have outperformed GAN-based approaches in generation quality [6]. Controllable [45] applies Dreambooth [37] to learn object concepts. DualAnoDiff [22] employs LoRAs [19] for learning both overall defective images and defect concepts. Defect Spectrum [48] enhances realism via multi-scale refinement. AnoDiff [20] and AnoGen [12] learn anomaly embeddings through Textual-Inversion [10], but their limited embedding space hinders learning complex objects. While RealNet [55] perturbs denoising variance to generate global anomaly maps without training, it tends to produce less realistic defects. In contrast, our method synthesizes semantically meaningful anomalies in a training-free manner, making it more practical for real-world use.

3. Preliminary

3.1. Diffusion-based Image Generation

Diffusion-based image generation iteratively removes noise from an initially noisy image $\mathbf{z}_t$ through a denoising network $\boldsymbol{\epsilon}_\theta$. At each time step t, the model refines $\mathbf{z}_t$ by subtracting the conditional estimated noise $\boldsymbol{\epsilon}_\theta\left(\mathbf{z}_t, t, \mathbf{c}\right)$, which can be equivalently defined under the score-based generative modeling framework grounded in Stochastic Differential Equations [42] [41] and is formalized as:

$$\boldsymbol{\epsilon}_\theta\left(\mathbf{z}_t, t, \mathbf{c}\right) = -\sqrt{1-\bar{\alpha}_t}\nabla_{\mathbf{z}_t} \log p_t\left(\mathbf{z}_t \mid \mathbf{c}\right). \tag{1}$$

3.2. Score-based guidance

Conditional diffusion models extend the unconditional formulation [17] by modeling the conditional distribution $p_t(\mathbf{z}_t \mid \mathbf{c})$, enabling control over generated content. The corresponding conditional score function can be decomposed using Bayes' Theorem as:

$$\begin{aligned} \nabla_{\mathbf{z_t}} \log p_t\left(\mathbf{z}_t \mid \mathbf{c}\right) \propto & \nabla_{\mathbf{z}_t} \log p_t\left(\mathbf{z}_t\right) \\ & + \nabla_{\mathbf{z}_t} \log p_t\left(\mathbf{c} \mid \mathbf{z}_t\right), \end{aligned} \tag{2}$$

where $\nabla_{\mathbf{z}_t} \log p_t(\mathbf{c} \mid \mathbf{z}_t)$ is the gradient of the log-posterior with respect to the condition $\mathbf{c}$, which can be practically approximated by the gradient of a differentiable energy function $\mathcal{L}_{\mathbf{c}}(\mathbf{z}_t, t, \mathbf{c})$ that quantifies the alignment between the sample and the target condition. Consequently, Eq. (1) can be reformulated as:

$$\hat{\boldsymbol{\epsilon}}_\theta\left(\mathbf{z}_t, t, \mathbf{c}\right) = \boldsymbol{\epsilon}_\theta\left(\mathbf{z}_t, t, \mathbf{c}\right) + \eta\sigma_t\nabla_{\mathbf{z}_t}\mathcal{L}_{\mathbf{c}}\left(\mathbf{z}_t, t, \mathbf{c}\right), \tag{3}$$

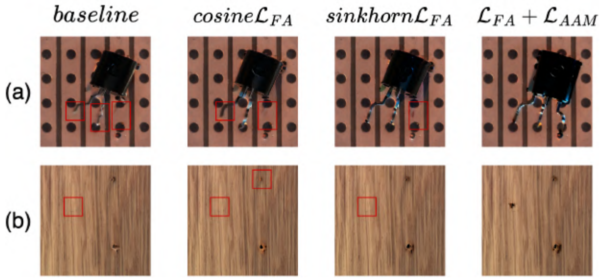
baseline
cosine$\mathcal{L}_{FA}$
sinkhorn$\mathcal{L}_{FA}$
$\mathcal{L}_{FA}+\mathcal{L}_{AAM}$
(a)
(b)

Figure 3. The effect of different gradient guidance on the appearance of defects.

where η is a scaling factor controlling guidance strength, σ_t is the noise schedule parameter at time step t. Based on this equation, we design two guidance functions to enhance the semantic accuracy of defect representation in Sec. 4.3.

4. Method

As shown in Fig. 2, given a normal (background) sample $\mathbf{x}^n$, a defect reference image $\mathbf{x}^{ref}$ with its corresponding segmentation mask $\mathbf{m}^{ref}$, and a target mask $\mathbf{m}^{gen}$ specifying the region of the generated defect within $\mathbf{x}^n$, our goal is to generate defects that are semantically consistent and precisely aligned with $\mathbf{m}^{gen}$ while visually harmonious with the background.

4.1. Overall Architecture

We first crop and enlarge defect regions from all inputs to focus the generation process on fine details. The generated defects are then seamlessly integrated back into the input normal image, addressing the challenges of tiny defect synthesis. Next, following the approach in [3], we treat defects in $\mathbf{x}^{ref}$ as the target for customization and feed $\mathbf{x}^{ref}$ into DINOv2 to extract image features, which are then projected into the embedding space via a linear layer to serve as the image condition for guiding diffusion toward the desired defect appearance. In parallel, $\mathbf{m}^{gen}$ is input into ControlNet to provide shape guidance, ensuring that generated defects $\mathbf{x}^{gen}$ align precisely with $\mathbf{m}^{gen}$ whose shape and position convey key information about defect semantics. Furthermore, an accurate $\mathbf{m}^{gen}$ is essential for generating defects, as its shape and position convey key information about defect semantics. For each anomaly type, we predefine the affected object area and randomly position augmented real masks within it, enhancing the alignment between defect location, shape, and anomaly characteristics.

In the subsequent generation sampling, TF-IDG enhances the quality of defects generated in the target mask region $\mathbf{m}^{gen}$ by employing three key modules: *Feature Alignment*, *Adaptive Anomaly Mask* and *Texture Preservation*. The Feature Alignment module leverages the rich visual features from pre-trained diffusion U-Net [35] to establish feature correspondences. This guides the generated defect features $\mathbf{F}_t^{gen}$ to align more closely with the real defect features $\mathbf{F}_t^{ref}$ during the early denoising steps, improving semantic consistency. The Adaptive Anomaly Mask module calculates the difference between generated defects and normal samples within $\mathbf{m}^{gen}$ and converts it into $\mathbf{m}_t^{aam}$, providing appearance guidance for areas that might otherwise be neglected during generation. Finally, the Texture Preservation module enhances the seamless integration of generated defects with the background by incorporating contextual information into the synthesis process of subsequent iterations. Further details are discussed below.

4.2. Feature Alignment

When applying the pretrained zero-shot image-editing framework [3] to defect generation, we observe that the self-supervised model DINOv2 excels at retaining discriminative features. Also, extensive pretraining on large-scale datasets endows DINOv2 with powerful instance retrieval abilities, enhancing the diversity of generated defects. However, in industrial contexts, DINOv2 exhibits biased feature retrieval, resulting in deviations between the generated defects and real defects. To address this limitation, we propose an approach to mitigate such inconsistencies.

Building on insights from [9, 28, 29], we propose a feature alignment strategy to optimize the features of generated defects using image embeddings as guidance. Specifically, we construct a memory bank to store the intermediate latent $\mathbf{z}_t^n$ and $\mathbf{z}_t^{ref}$ obtained from $\mathbf{x}^n$ and $\mathbf{x}^{ref}$ through DDIM Inversion process [40] at each inversion step t. Additionally, we store the self-attention layer's key $\mathbf{k}_t^n$ and value $\mathbf{v}_t^n$ corresponding to $\mathbf{z}_t^n$, which are used to guide the subsequent generation process. In the first n steps of diffusion generation, we extract intermediate features $\mathbf{F}_t^{gen}$ and $\mathbf{F}_t^{ref}$ from $\mathbf{z}_t^{gen}$ and $\mathbf{z}_t^{ref}$ using the UNet denoiser $\boldsymbol{\epsilon}_\theta$.

Following Eq. (3), the energy function $\mathcal{L}_{FA}$ is then constructed by calculating the correspondence between $\mathbf{F}_t^{gen}$ and $\mathbf{F}_t^{ref}$ within the regions defined by $\mathbf{m}^{gen}$ and $\mathbf{m}^{ref}$. Initially, as in previous studies [29], we use cosine similarity computed over average features within the masked regions to measure the similarity between features, defined as follows:

$$\mathrm{CosSim} = \cos\left(\frac{\sum \mathbf{F}_t^{gen}[\mathbf{m}^{gen}]}{\sum \mathbf{m}^{gen}}, \mathrm{sg}\left(\frac{\sum \mathbf{F}_t^{ref}[\mathbf{m}^{ref}]}{\sum \mathbf{m}^{ref}}\right)\right), \tag{4}$$

where sg is the gradient clipping operation. The loss is defined as a stabilized inverse transformation of the cosine similarity $\mathcal{L}_{FA} = 1/(\alpha + \beta \cdot \mathrm{CosSim}(\cdot))$.

Due to the diversity of defects, $\mathbf{m}^{ref}$ is not necessarily equal to $\mathbf{m}^{gen}$, allowing for the generation of defects with varying shapes while maintaining the same defect type. As a result, we average the features within the specified regions as a representation of defect appearance to align dimensions. However, we found that this approach has limited effectiveness. For example, in the second image of Fig. 3(a), the transistor pins show only slight improvement and generally remain blurry. Averaging compresses and aggregates feature details, leading to the loss of critical information about local feature variations.

To address structural misalignment between generated and reference features, we adopt the Sinkhorn Distance [4], which formulates alignment as an optimal transport problem rather than relying on pointwise or global similarity. By minimizing the transport cost between feature distributions, it captures fine-grained spatial relationships and preserves structural consistency in short-step diffusion optimization. This property is particularly beneficial in defect generation, where diverse shapes must still correspond to the correct defect type. Specifically, we first extract the feature vectors from $\mathbf{m}^{ref}$ and $\mathbf{m}^{gen}$ at each step:

$$\begin{aligned}\mathbf{V}^{ref} &= \left\{\mathbf{F}_t^{ref}(i,j) \mid \mathbf{m}^{ref}(i,j) = 1\right\}, \\ \mathbf{V}^{gen} &= \{\mathbf{F}_t^{gen}(i,j) \mid \mathbf{m}^{gen}(i,j) = 1\}.\end{aligned} \tag{5}$$

Next, we calculate the cosine similarity between each vector in the two feature sets as follows:

$$\mathbf{C}_{ij} = 1 - \frac{\mathbf{V}_i^{ref} \cdot \mathbf{V}_j^{gen}}{\|\mathbf{V}_i^{ref}\|\|\mathbf{V}_j^{gen}\|}. \tag{6}$$

Eventually, the energy function $\mathcal{L}_{FA}$ is updated to:

$$\mathcal{L}_{FA} = \min_{P \in U(\mathbf{V}^{ref},\mathbf{V}^{gen})} \sum_{i,j} P_{ij}\mathbf{C}_{ij} - \frac{1}{\lambda}h(P), \tag{7}$$

where P represents the optimal transport plan, and $h(P)$ is the entropy regularization term, which controls the sparsity of the transport plan. This approach quantifies the relative transport cost across all feature pairs, capturing the global structure of the distributions. As a result, it enhances robustness to shape variation and prevents overfitting or structural collapse, enabling precise alignment of complex defects.

4.3. Adaptive Anomaly Mask guidance

Although the quality of defect features is improved by optimizing the overall defect features using the appearance alignment loss $\mathcal{L}_{FA}$, smaller or narrower anomaly regions are often overlooked, particularly when multiple abnormal regions exist in an image or the shape of $\mathbf{m}^{gen}$ is irregular, resulting in anomalies that do not fully conform to the mask, such as the small transistor pin in Fig. 3(a) and the small hole in the wood in Fig. 3(b).

We aim to identify which regions are overlooked during generation and analyze the underlying causes inspired by [20]. Upon closer examination of the edit gradient during defect generation (as shown in Fig. 4), we observe that as diffusion sampling progresses, activations for small defects diminish too rapidly and prematurely. This reduction weakens the impact of appearance guidance in these areas. To address it, we designed an Adaptive Anomaly Mask module (AAM), which calculates the pixel-level difference between generated defects and normal samples, producing $\mathbf{m}_t^{aam}$ to provide appearance guidance for previously overlooked regions. Specifically, given latent variable $\mathbf{z}_t^{gen}$ at timestamp $t \in [T, \cdots, 0]$, the generated image $\mathbf{x}_t^{gen}$ is obtained via:

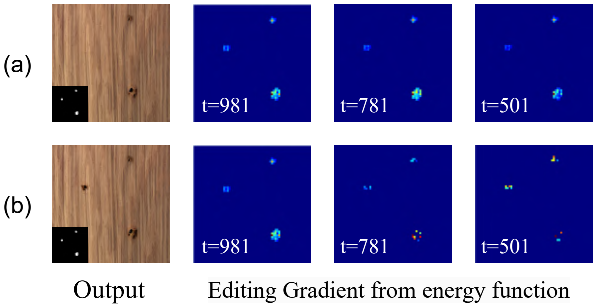

Figure 4. Visualization of gradient guidance at different timesteps. (a) shows the guidance effect of $\mathcal{L}_{FA}$, while (b) illustrates the guidance effect of $\mathcal{L}_{AAM}$.

$$\mathbf{x}_t^{gen} = \mathcal{D}\left(\left(\mathbf{z}_t^{gen} - \sigma_t\hat{\boldsymbol{\epsilon}}_t\right)/\alpha_t\right), \tag{8}$$

where $\hat{\boldsymbol{\epsilon}}_t = \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t^{gen}, t)$, (σ_t, α_t) are predefined diffusion scalars, and $\mathcal{D}$ is the decoder that maps the latent variable back to the image space. Then we can construct a difference map $\mathbf{m}_t^{aam}$ to highlight discrepancies between the generated image and a normal sample within the mask $\mathbf{m}^{gen}$:

$$\mathbf{D}_t^{map} = \left(\mathbf{m}^{gen} \odot (\mathbf{x}^n - \mathbf{x}_t^{gen})\right)^2. \tag{9}$$

Using this difference map, the module then computes an adaptive mask $\mathbf{m}_t^{aam}$ at each pixel location (i,j) as follows:

$$(\mathbf{m}_t^{aam})_{i,j} = \begin{cases} 1, & \text{if } (\mathbf{D}_t^{map})_{i,j} < \tau \\ 0, & \text{otherwise} \end{cases}, \tag{10}$$

where τ is the threshold for evaluating the boundaries of similar regions. Therefore, $\mathbf{m}_t^{aam}$ indicates the areas where the generated anomalies have not changed significantly, highlighting the region where enhanced appearance guidance is needed. We then apply Eq. (7) to reinforce feature consistency between the reference appearance $\mathbf{V}^{ref}$ and the adaptively masked anomaly representation $\mathbf{V}^{aam}$:

$$\mathcal{L}_{AAM} = d_C^\lambda\left(\mathbf{V}^{ref}, \mathbf{V}^{aam}\right), \tag{11}$$

where d_C^λ denotes the Sinkhorn Distance between $\mathbf{V}^{ref}$ and $\mathbf{V}^{aam}$, as defined in Eq. (7). $\mathbf{V}^{aam}$ is the feature sets extracted from the $\mathbf{F}_t^{gen}$ region in $\mathbf{m}^{gen}$ following Eq. (5). Lastly, the total loss containing global and local objectives is expressed as $\mathcal{L}_{total} = \mathcal{L}_{FA} + \mathcal{L}_{AAM}$ and the corresponding modified noise prediction is defined as:

$$\begin{aligned}\hat{\boldsymbol{\epsilon}}_\theta\left(\mathbf{z}_t, \mathbf{z}_t^{ref}, t, \mathbf{c}\right) = \boldsymbol{\epsilon}_\theta\left(\mathbf{z}_t, t, \mathbf{c}\right) + \\ \eta\sigma_t\nabla_{\mathbf{z}_t}\mathcal{L}_{total}\left(\mathbf{z}_t, \mathbf{z}_t^{ref}, t\right),\end{aligned} \tag{12}$$

where $\mathbf{z}_t$ is a simplification of $\mathbf{z}_t^{gen}$, η is the scaling constant that controls the relative strength of loss guidance.

In practice, we observe that applying gradient guidance for every timestep t affects the refinement of details and introduces noticeable artifacts. Hence, we constrain the application of Eq. (12) to the range $t \in [T, \cdots, T_n]$, allowing the model to adjust texture generation during the later iterations. Here, n is empirically set to three-fifths of the total number of sampling steps. By applying gradient correction, the model reinforces feature consistency in small and narrow regions. This is demonstrated by examples such as a misplaced transistor and multiple subtle holes in Fig. 3.

4.4. Texture Preservation

Beyond appearance and shape, inconsistencies in texture or color between defects and their background can cause the defects to appear "artificially placed" on the normal image, rather than seamlessly integrated. Unlike typical inpainting models that emphasize edge blending, industrial defects involve damage or displacement, requiring the generation to reflect contextual variations in the input normal samples. This creates a task gap between generic object generation and defect synthesis, where nuanced interactions between the defect and its surrounding context are essential.

To bridge this gap, we introduce a Texture Preservation mechanism to enhance the seamless integration of generated defects with their background. Specifically, inspired by the prior work [5, 15, 29], we inject reference information into the self-attention module. We extract $\mathbf{k}_t^n$ and $\mathbf{v}_t^n$, which encode background style information from the DDIM inversion memory bank mentioned in Sec. 4.2, and incorporate them into the self-attention calculations throughout the iterative process. The attention module is defined as follows:

$$Att(\mathbf{z}_t^{gen}) = \text{Softmax}\left(\frac{\boldsymbol{Q}^{gen}}{\sqrt{d}}\begin{bmatrix}\mathbf{k}_t^{gen}\\ \mathbf{k}_t^n\end{bmatrix}^\top\right)\begin{bmatrix}\mathbf{v}_t^{gen}\\ \mathbf{v}_t^n\end{bmatrix}, \tag{13}$$

where $t \in [T_i, \cdots, 0]$, i is a predefined parameter, typically set during the latter stages of the sampling process, ensuring that texture style information is effectively provided to the diffusion model during the detail refinement phase.

We observed that modifications applied only at the attention level were insufficient due to the low resolution of its semantic layer, which hindered fine-grained editing [13, 30], particularly in achieving adequate color modification intensity. Therefore, we employ an AdaIN operation [21] at the latent space level for color correction, formulated as:

$$\mathbf{z}_t^{gen} \leftarrow \text{AdaIN}\left(\mathbf{m} \odot \mathbf{z}_t^{gen}, \mathbf{m} \odot \mathbf{z}_t^n\right) + (1-\mathbf{m}) \odot \mathbf{z}_t^n, \tag{14}$$

where we omit the subscript gen of $\mathbf{m}^{gen}$ for simplicity. Here, $\mathbf{z}_t^{gen}$ is adjusted to align with $\mathbf{z}_t^n$, ensuring the color distributions are matched for seamless integration.

5. Experiments

5.1. Experiment Settings

Dataset. We evaluate our approach on MVTec AD [1] and VisA [56] datasets, two widely recognized benchmarks in industrial anomaly detection. To simulate realistic industrial inspection scenarios, we adopt the following two protocols: (1) **Few-shot setting** means that one-third of the images in each category of defects are used as reference sets to generate images, the remaining two-thirds serving as test sets. (2) **One-shot setting** simulates the challenge of generating new defect types. In this setting, only one anomaly sample per defect category is used as the reference. Our method is applicable to both settings. We add the symbol † to denote our model in the one-shot setting.

Metric. For defect generation, we introduce "Local IS" which measures the Inception Score (IS) [38] within defect regions cropped from the mask, addressing the insensitivity of standard IS in capturing fine-grained and small defects. We further adopt IC-LPIPS [32] to quantify diversity across generated anomaly clusters. To evaluate detection and localization performance for anomaly inspection, we employ the Area Under the Receiver Operating Characteristic curve (AUROC), Average Precision (AP), F1-score at optimal threshold (F1-max), and the Area Under the Precision-Recall curve (Pro).

Implementation Details. We choose Stable Diffusion V2.1 [34] and ControlNet [54] trained by Anydoor [3] as the backbone. We configure the parameters as follows: the ControlNet scale is set to 1, the DINOv2 guidance scale to 4.5, and total sampling steps $T = 50$ across all experiments, with denoising steps $T_n = 30$ and $T_i = 25$. Additionally, the scaling constant η is set to 4×10^{-2}, and we use the DDIM scheduler [40] during the denoising phase. We generate 1,000 anomalous image-mask pairs for each type of anomaly to conduct the subsequent experiments. A more detailed analysis of parameter selection is provided in the supplementary material.

5.2. Comparison in Anomaly Generation

We mainly focus on generating high-quality anomaly images and compare our method with five representative approaches for anomaly generation: Crop-Paste [25] (handcrafted synthesis), DFMGAN [8] (fine-tuning with StyleGAN2), Anodiff [20], AnoGen [12] (fine-tuning with diffusion models) and DIAG [11] (training-free method).

Anomaly generation quality. As shown in Tab. 1, our method significantly outperforms existing approaches in Local IS, indicating that our generated anomalies are both diverse and semantically meaningful. While the improvement in the IC-LPIPS score isn't as notable, the qualitative results in Fig. 5 show significant enhancements in generating small, narrow, and structurally complex defects, thread defects in grids, as well as intricate anomalies in transistors

Category	Crop-Paste [25]		DFMGAN [8]		AnoDiff [20]		AnoGen [12]		DIAG [11]		Ours	
	Local IS ↑	IC-L ↑	Local IS ↑	IC-L ↑	Local IS ↑	IC-L	Local IS ↑	IC-L ↑	Local IS ↑	IC-L ↑	Local IS ↑	IC-L ↑
MVTec AD	2.73	0.14	2.80	0.20	2.77	**0.32**	2.84	**0.32**	2.60	0.28	**3.32**	**0.32**
VisA	3.65	0.29	1.53	0.21	2.52	**0.30**	3.28	**0.30**	2.86	0.29	**3.90**	**0.30**

Table 1. **Comparison of generation quality with Local IS and IC-LPIPS on MVTec AD and VisA.** The best results for each metric are marked in **bold**, and the second-best results are marked with underline.

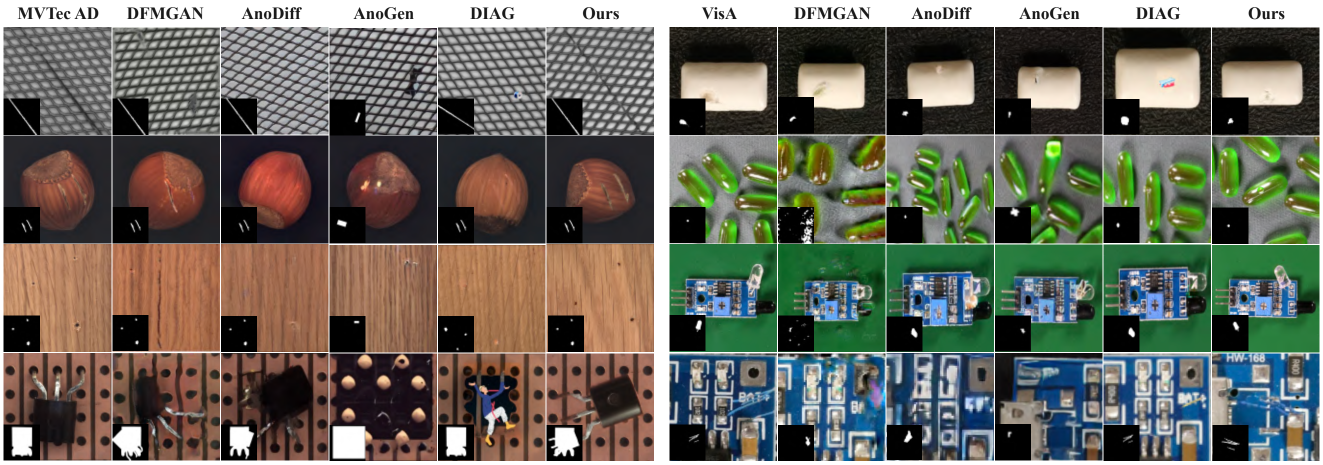

Figure 5. Selected qualitative results of the DFMGAN, AnoDiff, AnoGen, DIAG and our TF-IDG.

Category	Crop-Paste	DFMGAN	AnoDiff	DIAG	Ours	Ours†
bottle	52.71	56.59	90.70	84.13	95.53	**98.41**
cable	32.81	45.31	67.19	66.67	**85.94**	77.17
capsule	32.89	37.23	66.67	49.54	**73.33**	58.72
carpet	27.96	47.31	58.06	57.30	**80.65**	69.66
grid	28.33	40.83	**42.50**	46.74	40.00	31.58
hazelnut	59.03	81.94	85.42	90.00	**93.75**	91.43
leather	34.39	49.73	61.90	53.26	**85.71**	80.43
metal nut	59.89	64.58	59.38	68.57	60.94	**70.97**
pill	26.74	29.52	**59.38**	29.08	57.69	46.88
screw	28.81	37.45	48.15	27.73	**97.53**	70.59
tile	68.42	74.85	84.21	44.05	**94.74**	85.96
transistor	41.67	52.38	60.71	52.50	**89.29**	85.00
wood	47.62	49.21	71.43	53.33	**80.95**	71.67
zipper	26.42	27.64	69.51	15.97	**81.71**	68.42
Average	40.55	49.61	66.09	52.78	**79.84**	71.92

Table 2. Comparison of anomaly classification on MVTec AD by training a ResNet-34 model on the generated data.

and PCB. In contrast, many training-based methods tend to produce unrealistic artifacts or structurally implausible defects. Furthermore, the anomaly classification accuracy in Tab. 2 confirms that our generated defects closely match the semantic characteristics of each anomaly category, helping avoid significant distribution shifts while preserving essential structural features. As a result, it leads to more effective dataset augmentation for downstream tasks.

Anomaly generation for anomaly inspection. We use 1,000 images generated by the aforementioned methods to train a model with the U-Net [35] to derive confidence scores for anomaly localization and detection (Tab. 3), and train a ResNet-34 [14] to evaluate classification accuracy (Tab. 2). Experimental results demonstrate that our method outperforms other generative models in the few-shot setting, achieving a 6.0% improvement in image-level AUROC and a 3.1% increase in Pro score on the VisA dataset, which features more complex backgrounds and smaller defects. In

Dataset	Method	Image-level			Pixel-level			
		AUC	AP	F_1-max	AUC	AP	F_1-max	Pro
MVTec	Crop-Paste	94.7	97.9	95.2	91.6	66.0	83.4	64.8
	DFMGAN	87.2	94.8	94.7	90.0	62.7	62.1	76.3
	AnoDiff	99.2	99.7	98.7	**99.1**	81.4	76.3	94.0
	AnoGen	98.1	98.9	96.9	97.5	68.8	66.6	92.1
	DIAG	93.8	96.4	93.3	94.4	59.7	58.6	86.7
	Anydoor	95.5	98.1	95.7	94.7	70.0	66.9	89.7
	Ours†	98.2	99.0	97.1	97.4	75.1	69.8	91.4
	Ours	**99.6**	**99.8**	**98.8**	**99.1**	**84.1**	**78.3**	**95.8**
VisA	Crop-Paste	83.5	81.1	78.0	90.8	24.3	30.2	68.4
	DFMGAN	86.5	86.6	80.2	90.6	27.6	33.2	76.5
	AnoDiff	91.4	92.5	86.6	97.4	50.1	51.4	85.1
	AnoGen	90.7	93.1	88.0	88.0	26.7	35.1	69.3
	DIAG	91.3	91.5	84.6	94.9	33.2	38.9	87.6
	Anydoor	91.5	91.6	85.2	95.4	36.5	40.5	88.5
	Ours†	93.7	93.6	88.3	97.1	44.3	47.4	90.6
	Ours	**97.4**	**97.5**	**90.7**	**97.6**	**59.4**	**58.9**	**91.6**

Table 3. Quantitative Results on MVTec AD and VisA by training a U-Net on the generated data in the few-shot setting.

k-shot	Crop-Paste [25]					AnoDiff					Anydoor					Ours				
	i-AUC	i-AP	p-AUC	p-AP	Acc	i-AUC	i-AP	p-AUC	p-AP	Acc	i-AUC	i-AP	p-AUC	p-AP	Acc	i-AUC	i-AP	p-AUC	p-AP	Acc
1	93.2	96.6	88.8	55.7	46.3	93.6	97.9	96.3	68.5	54.2	90.3	95.7	91.1	54.8	61.6	**98.2**	**99.0**	**97.4**	**75.1**	**71.9**
3	93.4	97.4	91.4	62.3	52.3	98.2	**99.4**	98.1	76.4	64.2	93.7	97.8	94.7	63.1	65.4	**98.4**	**99.4**	**98.6**	**81.5**	**76.2**
5	94.7	97.9	91.6	66.0	59.7	98.6	**99.6**	98.6	78.1	66.3	94.2	97.9	94.8	61.8	66.0	**99.3**	**99.6**	**98.8**	**82.6**	**78.7**

Table 4. Impact of the number of seen anomalies on MVTec AD.

Method	MVTec AD			VisA		
	i-AUC	p-AUC	Pro	i-AUC	p-AUC	Pro
GLASS [2]	**99.9**	99.3	96.8	98.8	98.8	92.2
Ours + GLASS	**99.9**	**99.4**	**97.2**	**99.5**	**98.9**	**95.3**
UniAD [50]	96.5	96.8	90.7	88.8	98.3	85.5
Ours + UniAD	**98.9**	**97.1**	**91.5**	**94.3**	**98.6**	**86.9**

Table 5. Quantitative Results on MVTec AD and VisA for one-to-one model and multi-class model.

Method			Metric		
$sinkhorn\mathcal{L}_{FA}$	$\mathcal{L}_{AAM}$	TP	L-IS	Acc	p-AUC
			3.10	65.41	94.7
✓			3.27	78.33	98.2
✓	✓		3.30	79.15	99.0
✓	✓	✓	**3.32**	**79.84**	**99.1**

Table 6. **Ablation study**. Quantitative evaluation of each component's contribution using metrics from three different perspectives.

classification tasks, our models—both few-shot and one-shot—consistently outperform other competing methods. Our few-shot model improves average accuracy by 13.75% compared to the second-best method, while our one-shot model surpasses all competing methods even though most of them are trained in the few-shot setting.

5.3. Comparison with Anomaly Detection Models

To assess whether our generated anomaly samples enhance detection performance, we perform comparative experiments by integrating our approach to state-of-the-art and lightweight anomaly detection models, including the one-to-one AD model GLASS [2] and the multi-class AD model UniAD [50]. For GLASS, we replace the hand-crafted anomaly synthesis with images generated by our method while keeping all other modules unchanged. For UniAD, we incorporate the anomaly features extracted only from the generated images into the training process to enhance model reconstruction. As shown in Tab. 5, Ours+GLASS achieves a 3.4% improvement in Pro on VisA, while Ours+UniAD sees a 5.4% increase in i-AUC on VisA and a 2.5% gain on MVTec AD, further verifying the effectiveness and practical significance of our method.

5.4. Ablation Study

As shown in Tab. 6, we evaluate the impact of each module based on generation quality and performance of anomaly classification and localization under the few-shot setting. The results indicate that introducing the gradient-guided Sinkhorn loss significantly improves classification accuracy by refining the optimization process. The AAM module further enhances localization accuracy, and Fig. 3 illustrates each module's effect on the generation process. Finally, the Texture Preservation (TP) module enhances image authenticity, leading to an overall improvement in model performance. These ablation results clearly highlight the effectiveness of each individual component.

Impact of the number of seen anomalies. Since real anomaly data are often scarce and critical to industrial inspections, we examine how the number of visible anomalies affects model performance. We use Crop-Paste, AnoDiff [20], and our backbone method Anydoor [3] as baselines. A detailed review of Tab. 4 reveals that increasing the number of seen anomalies enhances the model's performance substantially. This improvement also suggests the limitation of the one-shot setting, which tends to bias the learned distribution towards the reference anomaly features, harming the model's generalization capability. Nonetheless, our model consistently surpasses the baseline methods across all experiments of seen anomalies, showing that our pipeline excels on both one-shot and few-shot settings. Furthermore, the overall variance in classification accuracies is notably smaller when applying our pipeline, which indicates that our method is resilient and robust in adapting to varying numbers of seen anomalies.

6. Conclusion

In this work, we propose TF-IDG, a training-free framework for industrial anomaly generation based on feature alignment optimization. Our method integrates three components: the Feature Alignment Module leverages gradient guidance to bridge the distribution gap between synthetic and real defects. The Adaptive Anomaly Mask Module ensures that small or subtle defects are preserved during multi-defect generation, while the Texture Preservation Module maintains the original image's color and texture for enhanced realism. Experiments on MVTec AD and VisA show that TF-IDG outperforms prior methods and significantly enhances downstream inspection performance.

Acknowledgements. This work is partially supported by the National Science and Technology Council, Taiwan, under Grant: NSTC-112-2628-E-002-033-MY4, and was financially supported in part by the Center of Data Intelligence: Technologies, Applications, and Systems, National Taiwan University (Grants: 114L900901/114L900902/114L900903), from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education, Taiwan.

References

[1] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *IJCV*, 129(4):1038–1059, 2021. 6

[2] Qiyu Chen, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang. A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. In *ECCV*, pages 37–54. Springer, 2024. 8

[3] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, pages 6593–6602, 2024. 2, 4, 6, 8

[4] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013. 5

[5] Yingying Deng, Xiangyu He, Fan Tang, and Weiming Dong. Z*: Zero-shot style transfer via attention reweighting. In *CVPR*, pages 6934–6944, 2024. 6

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 2, 3

[7] Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *CVPR*, pages 7388–7398, 2022. 1

[8] Yuxuan Duan, Yan Hong, Li Niu, and Liqing Zhang. Few-shot defect image generation via defect-aware feature manipulation. In *AAAI*, pages 571–578, 2023. 2, 3, 6, 7

[9] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *NeurIPS*, pages 16222–16239, 2023. 4

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*. 2, 3

[11] Federico Girella, Ziyue Liu, Franco Fummi, Francesco Setti, Marco Cristani, and Luigi Capogrosso. Leveraging latent diffusion models for training-free in-distribution data augmentation for surface defect detection. In *2024 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–7. IEEE, 2024. 6, 7

[12] Guan Gui, Bin-Bin Gao, Jun Liu, Chengjie Wang, and Yunsheng Wu. Few-shot anomaly-driven generation for anomaly classification and segmentation. In *ECCV*, pages 210–226. Springer, 2024. 3, 6, 7

[13] Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. In *CVPR*, 2024. 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7

[15] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *CVPR*, pages 4775–4785, 2024. 6

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 3

[18] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 2

[19] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021. 3

[20] 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 *AAAI*, pages 8526–8534, 2024. 2, 3, 5, 6, 7, 8

[21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 2, 6

[22] Ying Jin, Jinlong Peng, Qingdong He, Teng Hu, Hao Chen, Jiafu Wu, Wenbing Zhu, Mingmin Chi, Jun Liu, Yabiao Wang, and Chengjie Wang. Dualanodiff: Dual-interrelated diffusion model for few-shot anomaly image generation. *arXiv preprint arXiv:2408.13509*, 2024. 3

[23] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, pages 9664–9674, 2021. 1

[24] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, pages 22511–22521, 2023. 2

[25] 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 *ICME*, pages 1–6, 2021. 1, 3, 6, 7, 8

[26] Juhua Liu, Chaoyue Wang, Hai Su, Bo Du, and Dacheng Tao. Multistage gan for fabric defect detection. *IEEE Transactions on Image Processing*, 29:3388–3400, 2020. 3

[27] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *CVPR*, pages 20402–20411, 2023. 1

[28] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *CVPR*, pages 7465–7475, 2024. 4

[29] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. In *ICLR*, 2024. 4, 6

[30] Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. In *CVPR*, pages 8100–8110, 2024. 6

[31] 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. 1, 3

[32] 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 *CVPR*, pages 10743–10752, 2021. 6

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 6

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4, 7

[36] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, pages 14318–14328, 2022. 1

[37] 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 *CVPR*, pages 22500–22510, 2023. 2, 3

[38] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 6

[39] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *ECCV*, pages 474–489. Springer, 2022. 1

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 4, 6

[41] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 32, 2019. 3

[42] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*. 3

[43] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *CVPR*, pages 18310–18319, 2023. 2

[44] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel Aliaga. Imprint: Generative object compositing by learning identity-preserving representation. In *CVPR*, pages 8048–8058, 2024. 2

[45] Gabriele Valvano, Antonino Agostino, Giovanni De Magistris, Antonino Graziano, and Giacomo Veneri. Controllable image synthesis of industrial data using stable diffusion. In *WACV*, pages 5354–5363, 2024. 3

[46] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *CVPR*, pages 22428–22437, 2023. 2

[47] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18381–18391, 2023. 2

[48] Shuai Yang, Zhifei Chen, Pengguang Chen, Xi Fang, Yixun Liang, Shu Liu, and Yingcong Chen. Defect spectrum: a granular look of large-scale defect datasets with rich semantics. In *ECCV*, pages 187–203. Springer, 2024. 3

[49] Xincheng Yao, Ruoqi Li, Jing Zhang, Jun Sun, and Chongyang Zhang. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *CVPR*, pages 24490–24499, 2023. 1

[50] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *NeurIPS*, 35:4571–4584, 2022. 8

[51] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, pages 8330–8339, 2021. 1, 3

[52] Gongjie Zhang, Kaiwen Cui, Tzu-Yi Hung, and Shijian Lu. Defect-gan: High-fidelity defect synthesis for automated defect inspection. In *WACV*, pages 2524–2534, 2021. 1, 3

[53] Hui Zhang, Zuxuan Wu, Zheng Wang, Zhineng Chen, and Yu-Gang Jiang. Prototypical residual networks for anomaly detection and localization. In *CVPR*, pages 16281–16291, 2023. 1, 3

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2, 6

[55] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *CVPR*, pages 16699–16708, 2024. 1, 3

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