AnoStyler: Text-Driven Localized Anomaly Generation via Lightweight Style Transfer

Yulim So, Seokho Kang*

Sungkyunkwan University
{you0715, s.kang}@skku.edu

Abstract

Anomaly generation has been widely explored to address the scarcity of anomaly images in real-world data. However, existing methods typically suffer from at least one of the following limitations, hindering their practical deployment: (1) lack of visual realism in generated anomalies; (2) dependence on large amounts of real images; and (3) use of memory-intensive, heavyweight model architectures. To overcome these limitations, we propose *AnoStyler*, a lightweight yet effective method that frames zero-shot anomaly generation as text-guided style transfer. Given a single normal image along with its category label and expected defect type, an anomaly mask indicating the localized anomaly regions and two-class text prompts representing the normal and anomaly states are generated using generalizable category-agnostic procedures. A lightweight U-Net model trained with CLIP-based loss functions is used to stylize the normal image into a visually realistic anomaly image, where anomalies are localized by the anomaly mask and semantically aligned with the text prompts. Extensive experiments on the MVTec-AD and VisA datasets show that *AnoStyler* outperforms existing anomaly generation methods in generating high-quality and diverse anomaly images. Furthermore, using these generated anomalies helps enhance anomaly detection performance.

Code — https://github.com/yulimso/AnoStyler
Extended version — http://arxiv.org/abs/2511.06687

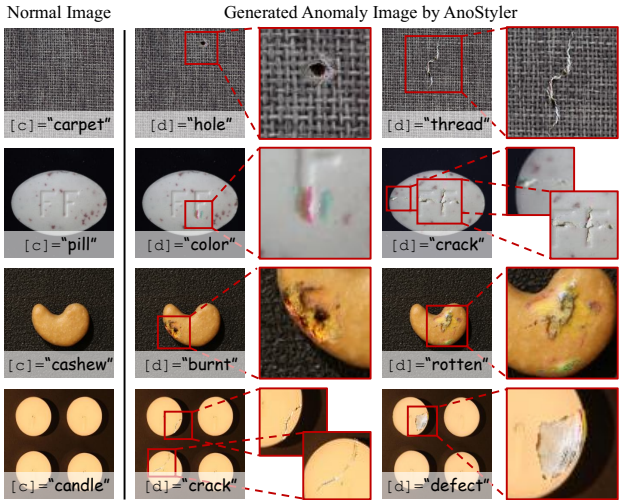

Figure 1: Examples of anomaly images generated by AnoStyler. Given a normal image and a category-defect pair (`[c]`, `[d]`), our method generates visually realistic and semantically aligned anomalies.

1 Introduction

Anomaly detection aims to identify patterns or regions in an image that deviate from the learned notion of normality (Li et al. 2025). Due to the rarity and diversity of anomalies, unsupervised learning (Defard et al. 2021; Roth et al. 2022; Batzner, Heckler, and König 2024; Hyun et al. 2024; Wu et al. 2025; Fang et al. 2025) on normal images has emerged as the dominant paradigm. Despite their success, these methods lack the capacity to model diverse real-world anomaly distributions, which limits their performance, particularly in complex or unseen domains (Cui, Liu, and Lian 2023; Cao, Zhu, and Pang 2023). This underscores the necessity of generating realistic and diverse anomaly images as alternative supervisory signals, leading to active research on *Anomaly Generation* to mitigate the scarcity of real anomaly images.

Existing anomaly generation methods can be broadly categorized into two paradigms: few-shot and zero-shot methods. Few-shot methods typically train generative models to synthesize anomaly images based on a few real anomaly images, assuming access to these real anomalies (Zhang et al. 2021; Duan et al. 2023; Hu et al. 2024; Gui et al. 2024). Zero-shot methods, on the other hand, operate under a more challenging yet practical scenario where no anomaly images are available, thus have emerged as promising research directions. Earlier zero-shot methods relied on direct manipulation of normal images using handcrafted operations (Lin et al. 2021; Li et al. 2021; Zavrtanik, Kristan, and Skočaj 2021; Schlüter et al. 2022) , whereas more recent methods either employ generative models trained solely on normal images to synthesize anomalies by perturbing the generation process (Zhang, Xu, and Zhou 2024) or leverage pre-trained generative models guided by text prompts (Sun et al. 2025).

While effective, existing methods still face one or more of

*Corresponding author.

the following limitations that restrict their practical applicability. First, the generated anomalies lack visual realism and semantic fidelity, especially for methods that directly manipulate normal images (Li et al. 2021; Zavrtanik, Kristan, and Skočaj 2021; Schlüter et al. 2022). Using such low-quality anomalies may result in poor generalization to real-world anomalies. Second, the generation process necessitates access to large amounts of real normal images (Zhang, Xu, and Zhou 2024), and further requires even a small number of real anomalies in the case of few-shot methods (Duan et al. 2023; Hu et al. 2024; Gui et al. 2024), making it challenging in scenarios where collecting real data is costly or difficult. Third, the generation process depends on memory-intensive heavyweight architectures like diffusion models (Duan et al. 2023; Hu et al. 2024; Gui et al. 2024; Zhang, Xu, and Zhou 2024; Sun et al. 2025), rendering it impractical for real-time or resource-constrained scenarios.

In this work, we propose *AnoStyler*, a lightweight yet effective zero-shot anomaly generation method that simultaneously addresses the aforementioned limitations. AnoStyler frames anomaly generation as a text-driven style transfer task, in which a normal image is transformed into an anomaly image by locally modifying its visual attributes while preserving its overall structural content. Compared to mainstream methods that leverage generative models (Goodfellow et al. 2014; Ho, Jain, and Abbeel 2020), style transfer (Gatys, Ecker, and Bethge 2016) can better preserve the overall content of the original image while injecting localized anomalies, making it inherently more suitable for anomaly generation. Nevertheless, it remains underexplored in this context; to the best of our knowledge, we present the first approach that effectively leverages style transfer for this purpose. AnoStyler integrates shape-guided masking and state-aware prompt generation with tailored losses to achieve semantically aligned anomaly stylization. As illustrated in Figure 1, AnoStyler generates realistic and diverse anomalies that are semantically consistent with text prompts while preserving the global structure of the input image. Leveraging these components, AnoStyler achieves state-of-the-art performance among zero-shot methods on MVTec-AD and VisA, and its generated anomalies significantly enhance downstream anomaly detection.

Our main contributions are summarized as follows:

- We propose a method that generates high-quality and realistic anomalies in various types, with precise semantic alignment to text prompts.
- We design a zero-shot anomaly generation framework that requires only a single normal image to synthesize each anomaly, removing the dependency on large collections of normal or anomaly images that hinder scalability.
- We introduce a lightweight model architecture that enables computationally and memory-efficient anomaly generation while preserving competitive performance.

2 Related Work

2.1 Anomaly Generation

Depending on the availability of anomalies in the training data, anomaly generation can be formulated as either a few-shot or zero-shot task. Few-shot methods leverage a small number of anomaly images to train generative models, such as DFMGAN (Duan et al. 2023) based on generative adversarial networks (GANs) (Goodfellow et al. 2014), and AnoDiff (Hu et al. 2024) and AnoGen (Gui et al. 2024) based on diffusion models (Ho, Jain, and Abbeel 2020), which are trained using a large number of normal images in addition to the few anomaly images.

In contrast, zero-shot methods operate under the assumption that no anomaly images are available. Heuristic methods generate synthetic anomalies by directly manipulating normal images using handcrafted operations. For example, CutPaste (Li et al. 2021) and NSA (Schlüter et al. 2022) use cut-and-paste operation, and DRAEM (Zavrtanik, Kristan, and Skočaj 2021) uses external texture injection. Recent studies have proposed methods that leverage generative models to produce more realistic anomalies with enhanced semantic alignment and visual fidelity. RealNet (Zhang, Xu, and Zhou 2024) utilizes a diffusion model trained on normal images and generates anomalies by perturbing the denoising process. AnomalyAny (Sun et al. 2025) guides Stable Diffusion using text prompts to generate anomaly images. While they are effective, they often incur high computational costs and memory usage due to their reliance on diffusion models.

2.2 Style Transfer

Style transfer aims to generate a new image that combines the structural content of an input with the texture or style of a reference. The seminal work on neural style transfer (Gatys, Ecker, and Bethge 2016) presented an optimization-based method that uses feature statistics from a pre-trained CNN to separately model content and style. Subsequent methods improved efficiency and generality by introducing feature-level transformations, such as AdaIN (Huang and Belongie 2017) and WCT (Li et al. 2017, 2018).

With the advent of CLIP (Radford et al. 2021), a new line of research has emerged that replaces reference style images with natural language prompts. StyleCLIP (Patashnik et al. 2021) maps text embeddings to latent directions in StyleGAN (Karras, Laine, and Aila 2019), enabling text-driven manipulation within the learned image manifold. CLIPstyler (Kwon et al. 2022) performs text-conditioned style transfer by optimizing patch-wise CLIP loss, achieving localized and semantically meaningful stylization without style images. Building on CLIPstyler, subsequent research has advanced text-driven and object-centric style editing, exploring approaches for fine-grained and localized stylization. Recent studies (Kamra, Mastan, and Gupta 2023; Ganugula et al. 2023; Singh et al. 2024; Chen et al. 2024) incorporate mechanisms such as foreground–background separation, segmentation masks, and semantic guidance to selectively apply distinct styles to specific regions or objects.

Following this line of research, we apply text-driven localized style transfer to zero-shot anomaly generation. With CLIP-based losses, AnoStyler efficiently adds text-guided localized anomalies while preserving overall content.

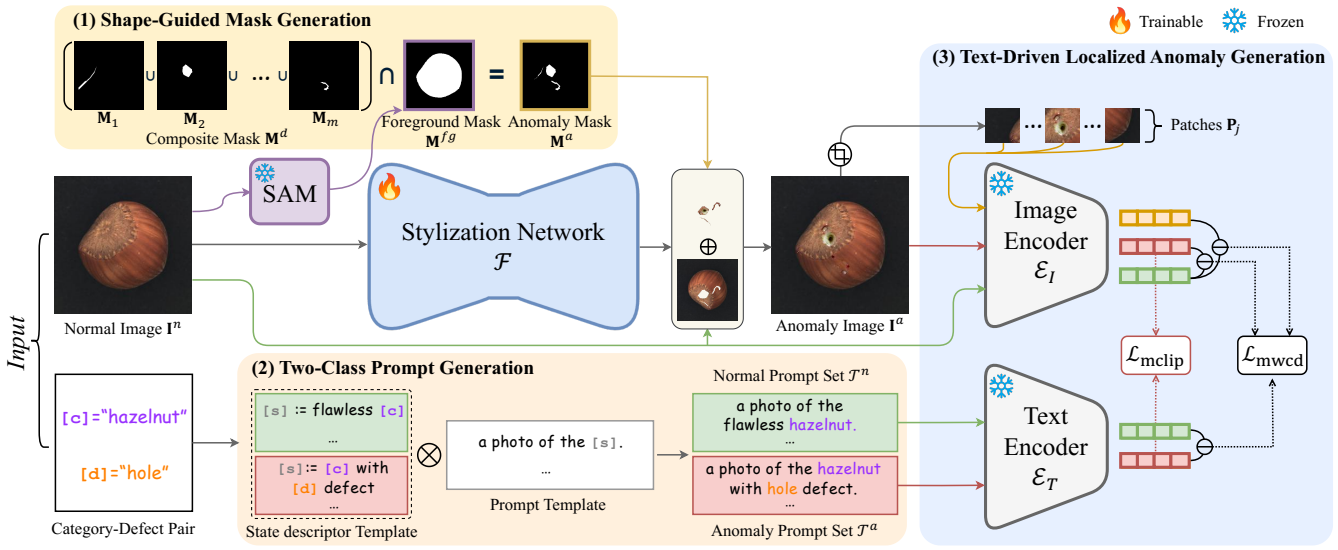

Figure 2: Overall framework of AnoStyler. (1) **Shape-Guided Mask Generation**: A union of primitive masks $\mathbf{M}_1, \ldots, \mathbf{M}_m$ from Meta-Shape Priors is intersected with the foreground mask $\mathbf{M}^{fg}$ to obtain the anomaly mask $\mathbf{M}^a$. (2) **Two-Class Prompt Generation**: Structured text prompt templates are filled with the category-defect pair (`[c]`, `[d]`) to form normal and anomaly prompt sets $\mathcal{T}^n$ and $\mathcal{T}^a$. (3) **Text-Driven Localized Anomaly Generation**: Guided by $\mathbf{M}^a$, $\mathcal{T}^n$, and $\mathcal{T}^a$, the stylization network $\mathcal{F}$ is trained to stylize the masked regions of the input image $\mathbf{I}^n$ as anomalies, resulting in the synthetic anomaly image $\mathbf{I}^a$.

3 Method

3.1 Overview of AnoStyler

AnoStyler frames anomaly generation as a text-guided style transfer process that transforms a normal image by injecting localized anomalies. The inputs to the generation process are a real normal image $\mathbf{I}^n \in \mathbb{R}^{C\times H\times W}$, where C, H, and W denote the number of channels, image height, and width, respectively, as well as its category label `[c]` and expected defect type `[d]` in text format. Using the image $\mathbf{I}^n$ as the content and the category-defect pair (`[c]`, `[d]`) as a reference style, the generation process outputs a generated anomaly image $\mathbf{I}^a \in \mathbb{R}^{C\times H\times W}$.The generation process, which generates one synthetic anomaly image at a time from one real normal image, consists of three sequential steps, as illustrated in Figure 2.

3.2 Model Architecture

The stylization network $\mathcal{F}$ is modeled as a lightweight U-Net (Kwon et al. 2022) that takes a normal image $\mathbf{I}^n$ as input and generates a stylized output $\mathbf{I}^a$ serving as a synthetic anomaly image. A pre-trained CLIP model (Radford et al. 2021) is incorporated to guide anomaly generation by $\mathcal{F}$. The text encoder $\mathcal{E}_T$ is a Transformer encoder that embeds text prompts in $\mathcal{T}^n$ or $\mathcal{T}^a$. The image encoder $\mathcal{E}_I$ is a Vision Transformer that embeds images $\mathbf{I}^n$, $\mathbf{I}^a$, or patches $\mathbf{P}_j$ extracted from $\mathbf{I}^a$. These embeddings are used to compute CLIP-based loss functions that enforce semantic consistency between the generated image and the given text prompt. During training, only the stylization network $\mathcal{F}$ is trainable, while all parameters of the text encoder $\mathcal{E}_T$ and image encoder $\mathcal{E}_I$ are frozen.

3.3 Shape-Guided Mask Generation

We argue that a practical anomaly mask generation method should meet two essential criteria. First, it should produce visually plausible masks that generalize across diverse categories and domains without relying on category-specific assumptions. Previous methods often rely on simple geometric shapes like rectangles (Li et al. 2021; Schlüter et al. 2022) or Perlin noise (Zavrtanik, Kristan, and Skočaj 2021; Zhang, Xu, and Zhou 2024), which lack the fidelity required to mimic plausible anomalies. Second, it should be lightweight for scalable deployment. Previous methods that rely on dedicated heavyweight model like diffusion models (Duan et al. 2023; Hu et al. 2024) are not scalable and introduce significant computational overhead. To this end, we propose a generalizable category-agnostic and lightweight non-parametric procedure that leverages primitive geometric components.

Meta-Shape Priors. For the procedural generation of anomaly masks, we introduce meta-shape priors, which comprise three primitive shapes designed to capture a distinct class of geometric patterns for localized anomaly regions: *Line*, *Dot*, and *Freeform*. These three shape types are carefully chosen to cover a broad spectrum of plausible anomaly geometries observed in real-world data, while agnostic to specific categories or domains. Line-shaped masks are created by connecting multiple points along a sampled direction with variable thickness. Sinusoidal perturbations add curvature, effectively capturing straight or wavy line-shaped anomalies. Dot-shaped masks are generated by sampling points radially around a center and perturbing their radii with noise. They preserve a circular form while producing smooth, spiky, or irregular closed shapes. Freeform

masks are produced by simulating unconstrained random trajectories. Unlike the structured forms above, they generate irregular, topology-free regions resembling diffused or amorphous anomalies. Detailed procedures for generating masks of these three types are provided in Appendix B.

Anomaly Mask Generation. To generate an anomaly mask for the input image $\mathbf{I}^n$, we first combine multiple anomaly regions derived from meta-shape priors, then retain the parts that overlap with the foreground of the image. The assumption is that multiple anomaly regions may co-occur within a single image and can appear only in the foreground.

The number of anomaly regions m is sampled from a categorical distribution with exponentially decaying probabilities. The probability mass function is defined as:

$$P(m=i) = \frac{\exp(-\alpha i)}{\sum_{j=1}^{m_{\max}} \exp(-\alpha j)}, \quad i \in \{1, 2, \ldots, m_{\max}\}, \tag{1}$$

where $m_{\max}$ denotes the maximum number of anomaly regions and α is a decay coefficient that favors fewer anomalies while allowing for more complex compositions to occasionally occur within an image.

To generate a primitive mask $\mathbf{M}_i$ indicating an anomaly region, we randomly select one of the three meta-shape priors (*i.e.*, Line, Dot, or Freeform) and follow its mask generation procedure. After obtaining m masks, we take their union to allow multiple distinct regions to be represented within a single mask. The composite mask $\mathbf{M}^d$ is given by:

$$\mathbf{M}^d = \bigcup_{i=1}^{m} \mathbf{M}_i. \tag{2}$$

The final anomaly mask $\mathbf{M}^a$ is obtained by intersecting $\mathbf{M}^d$ with the foreground mask $\mathbf{M}^{fg}$ indicating the foreground region of an image. If the category label `[c]` for an image given is object-centric, where anomalies typically occur on object surfaces, we use the Segment Anything Model (SAM) (Kirillov et al. 2023) with a ViT-B backbone to generate the foreground mask. Specifically, we designate four corner points as positive prompts to guide SAM in extracting the background region, and then apply a negation operation to obtain the foreground mask $\mathbf{M}^{fg}$. If the category label `[c]` is texture-centric, such as fabrics or surfaces without distinct object boundaries, we treat the entire image as foreground by setting $\mathbf{M}^{fg} = \mathbf{1}^{H \times W}$. The final anomaly mask $\mathbf{M}^a$ is then given by:

$$\mathbf{M}^a = \mathbf{M}^d \cap \mathbf{M}^{fg}. \tag{3}$$

3.4 Two-Class Prompt Generation

We propose a generalizable template-based text prompt generation procedure by adapting three key techniques: two-class design, prompt templates, and prompt averaging. Two-class design (Jeong et al. 2023) constructs separate text prompts for the normal and anomaly states. Explicitly modeling each state provides clearer guidance on how an anomaly image should differ from a normal one. The use of structured prompt templates facilitates intuitive and scalable prompt generation by requiring only the specification of the category label and defect type, without requiring domain-specific knowledge or real anomaly images. Prompt averaging (Radford et al. 2021) generates multiple semantically equivalent prompts for the given context and averages their embeddings to guide anomaly generation. This reduces variability and potential bias introduced by any single prompt, resulting in more stable and consistent text conditioning.

Given the category-defect pair (`[c]`, `[d]`) for the input image $\mathbf{I}^n$, we first generate two-class state descriptors `[s]` for the normal and anomaly states using predefined state descriptor templates (Jeong et al. 2023). For example, the normal state is described by phrases such as `"flawless [c]"`, while the anomaly state is described by phrases such as `"[c] with [d] defect"`. Next, each state descriptor `[s]` is inserted into predefined prompt templates to form full text prompts such as `"a photo of the [s]"` (Kwon et al. 2022). If the category label `[c]` or the defect type `[d]` is not provided, the default token `"sample"` or `"defect"` can be used, respectively. This yields a normal prompt set $\mathcal{T}^n$ and an anomaly prompt set $\mathcal{T}^a$. The complete list of templates used is provided in Appendix C. By prompt averaging, the embeddings of $\mathcal{T}^n$ and $\mathcal{T}^a$ are obtained using the clip encoder $\mathcal{E}_T$ as:

$$\mathbf{h}_T^n = \frac{1}{|\mathcal{T}^n|} \sum_{T_i \in \mathcal{T}^n} \mathcal{E}_T(T_i); \quad \mathbf{h}_T^a = \frac{1}{|\mathcal{T}^a|} \sum_{T_i \in \mathcal{T}^a} \mathcal{E}_T(T_i). \tag{4}$$

3.5 Text-Driven Localized Anomaly Generation

The anomaly mask $\mathbf{M}^a$ and the prompt sets $\mathcal{T}^n$ and $\mathcal{T}^a$ are used to guide the stylization network $\mathcal{F}$ to transform the input image $\mathbf{I}^n$ by incorporating localized anomalies. We introduce two loss terms specifically designed to promote the spatial localization and semantic alignment of synthetic anomalies: the *Mask-Weighted Co-Directional Loss* $\mathcal{L}_{\text{mwcd}}$ and the *Masked CLIP Loss* $\mathcal{L}_{\text{mclip}}$. Following CLIPstyler, we additionally adopt the *Content Loss* $\mathcal{L}_c$ (Gatys, Ecker, and Bethge 2016) and the *Total Variation Loss* $\mathcal{L}_{\text{tv}}$ (Rudin, Osher, and Fatemi 1992) to ensure the semantic fidelity and spatial smoothness of the generated image. The training objective for $\mathcal{F}$ is given by:

$$\mathcal{L} = \mathcal{L}_{\text{mwcd}} + \lambda_{\text{mclip}} \cdot \mathcal{L}_{\text{mclip}} + \lambda_{\text{c}} \cdot \mathcal{L}_{\text{c}} + \lambda_{\text{tv}} \cdot \mathcal{L}_{\text{tv}}, \tag{5}$$

where λ_{mclip}, λ_{c}, and λ_{tv} are weights that control the strengths of each loss term.

After the stylization network $\mathcal{F}$ is trained, the stylized image $\mathbf{I}^a$, which serves as a generated anomaly image, is generated by compositing the network output $\mathcal{F}(\mathbf{I}^n)$ with the original input image $\mathbf{I}^n$ using the anomaly mask $\mathbf{M}^a$ as a binary weight matrix:

$$\mathbf{I}^a = \mathcal{F}(\mathbf{I}^n) \odot \mathbf{M}^a + \mathbf{I}^n \odot (1 - \mathbf{M}^a), \tag{6}$$

where $\odot$ is the element-wise product operator. This ensures that style transfer is applied exclusively within the masked regions while preserving the surrounding background.

Mask-Weighted Co-Directional Loss. The concept of directional supervision in CLIP embedding space, originally

introduced in StyleGAN-NADA (Gal et al. 2022) and further explored in CLIPstyler (Kwon et al. 2022), is the basis of the loss $\mathcal{L}_{\text{mwcd}}$. The loss is computed as a weighted sum of a global directional alignment term $\mathcal{L}_{\text{gdir}}$ and a patch-wise directional alignment term $\mathcal{L}_{\text{pdir}}$, defined as:

$$\mathcal{L}_{\text{mwcd}} = \lambda_{\text{gdir}} \cdot \mathcal{L}_{\text{gdir}} + \lambda_{\text{pdir}} \cdot \mathcal{L}_{\text{pdir}}, \tag{7}$$

where λ_{gdir} and $\mathcal{L}_{\text{pdir}}$ are weights assigned to each term. We modify the original loss to place greater emphasis on anomaly regions indicated by the anomaly mask $\mathbf{M}^a$. The modified loss terms incorporating the mask $\mathbf{M}^a$ and the mask-guided anomaly image $\mathbf{I}^a$ are detailed as follows.

The first term $\mathcal{L}_{\text{gdir}}$ measures global directional alignment between images and text prompts by comparing their respective shifts from the normal to the anomaly state in the CLIP embedding space. The directional shifts of the image and prompt embeddings are computed as:

$$\Delta\mathbf{h}_I = \mathbf{h}_I^a - \mathbf{h}_I^n = \mathcal{E}_I(\mathbf{I}^a) - \mathcal{E}_I(\mathbf{I}^n); \tag{8}$$

$$\Delta\mathbf{h}_T = \mathbf{h}_T^a - \mathbf{h}_T^n. \tag{9}$$

The term $\mathcal{L}_{\text{gdir}}$ is then defined as the cosine distance between $\Delta\mathbf{h}_I$ and $\Delta\mathbf{h}_T$:

$$\mathcal{L}_{\text{gdir}} = 1 - \frac{\Delta\mathbf{h}_I \cdot \Delta\mathbf{h}_T}{\|\Delta\mathbf{h}_I\| \cdot \|\Delta\mathbf{h}_T\|}. \tag{10}$$

Minimizing this term encourages the global semantic shift from the normal image $\mathbf{I}^n$ to the generated image $\mathbf{I}^a$ to align with the intended transformation described by the two-class prompt sets $\mathcal{T}^n$ and $\mathcal{T}^a$.

The second term $\mathcal{L}_{\text{pdir}}$ is a patch-wise extension of $\mathcal{L}_{\text{gdir}}$ designed to further refine alignment at a finer scale. It measures patch-wise directional alignment by operating on image patches extracted from $\mathbf{I}^a$, thereby providing localized semantic guidance in contrast to global alignment over the whole image. Here, we extract l patches $\{\mathbf{P}_j\}_{j=1}^l$ from $\mathbf{I}^a$ via random cropping. For each patch $\mathbf{P}_j$, we measure the local directional shift from the input image $\mathbf{I}^n$ as:

$$\Delta\mathbf{h}_{P_j} = \mathbf{h}_{P_j}^a - \mathbf{h}_I^n = \mathcal{E}_I(\tau(\mathbf{P}_j)) - \mathcal{E}_I(\mathbf{I}^n), \tag{11}$$

where τ denotes a random perspective transformation applied to induce patch-wise geometric variation. The term $\mathcal{L}_{\text{pdir}}$ is then defined as:

$$\mathcal{L}_{\text{pdir}} = \frac{1}{\sum_{j=1}^l r_j} \sum_{j=1}^{l} r_j \left(1 - \frac{\Delta\mathbf{h}_{P_j} \cdot \Delta\mathbf{h}_T}{\|\Delta\mathbf{h}_{P_j}\| \cdot \|\Delta\mathbf{h}_T\|}\right), \tag{12}$$

where $r_j \in [0, 1]$ denotes the ratio of pixels in $\mathbf{P}_j$ that are covered by the anomaly mask $\mathbf{M}^a$. With this soft weighting, each patch $\mathbf{P}_j$ contributes proportionally to its overlap with $\mathbf{M}^a$, thereby placing greater emphasis on anomaly regions.

Masked CLIP Loss. The loss $\mathcal{L}_{\text{mclip}}$ is designed to further enforce semantic alignment between the localized anomalies in the generated anomaly image $\mathbf{I}^a$ and the target semantics described by the anomaly prompt set $\mathcal{T}^a$. Specifically, the loss is defined as the cosine distance between the masked region of the generated image $\mathbf{I}^a$ and the anomaly prompt set $\mathcal{T}^a$ in the CLIP embedding space:

$$\mathcal{L}_{\text{mclip}} = 1 - \frac{\mathcal{E}_I(\mathbf{I}^a \odot \mathbf{M}^a) \cdot \mathbf{h}_T^a}{\|\mathcal{E}_I(\mathbf{I}^a \odot \mathbf{M}^a)\| \cdot \|\mathbf{h}_T^a\|}. \tag{13}$$

Method	MVTec-AD		VisA	
	IS	IC-L	IS	IC-L
Few-Shot Anomaly Generation				
DFMGAN* (Duan et al. 2023)	1.72	0.20	1.48	0.28
AnoDiff* (Hu et al. 2024)	1.80	<u>0.32</u>	1.50	<u>0.29</u>
AnoGen† (Gui et al. 2024)	1.77	0.27	1.40	0.22
Zero-Shot Anomaly Generation				
CutPaste† (Li et al. 2021)	1.76	0.22	1.52	0.26
DRAEM† (Zavrtanik et al. 2021)	1.76	0.25	1.50	0.25
NSA$^\circ$ (Schlüter et al. 2022)	1.44	0.26	1.42	0.19
RealNet$^\circ$ (Zhang et al. 2024)	1.64	0.22	<u>1.53</u>	<u>0.29</u>
AnomalyAny$^\circ$ (Sun et al. 2025)	<u>2.02</u>	**0.33**	1.41	0.19
AnoStyler (ours)	**2.04**	<u>0.32</u>	**1.55**	**0.32**

Table 1: Comparison of anomaly generation on MVTec-AD and VisA. For each metric, the best and second-best scores are shown in **bold** and <u>underlined</u>. †: Re-implemented on both datasets; $\star$ and $\circ$: Results on MVTec-AD taken from (Hu et al. 2024) and (Sun et al. 2025), respectively, with results on VisA re-implemented for a fair comparison.

Restricting this loss to the regions specified by the anomaly mask $\mathbf{M}^a$ ensures that semantic alignment is focused only on these anomaly regions, without affecting the background of the input image.

4 Experiments

4.1 Experiment Settings

Datasets. We conducted experiments on two representative benchmark datasets for industrial visual anomaly detection: MVTec-AD (Bergmann et al. 2019) and VisA (Zou et al. 2022). MVTec-AD consists of 5,354 high-resolution images across 10 object and 5 texture categories, each associated with 1 to 7 defect types, which is curated to support the detection of subtle local anomalies in controlled settings. VisA consists of 10,821 images spanning 12 object categories, and it captures more complex scenes with multi-object arrangements and diverse structural variations. For both datasets, each image is paired with a ground-truth pixel-wise anomaly mask. In the experiments, all images and masks were resized to 512×512. Detailed information about the benchmark datasets is provided in Appendix A.

Implementation Details. The stylization network $\mathcal{F}$ had a U-Net architecture, consisting of three downsampling blocks and three upsampling blocks, as used in Kwon et al. (2022). For the image encoder $\mathcal{E}_I$ and text encoder $\mathcal{E}_T$, we used the pre-trained CLIP ViT-B/32 model. Detailed hyperparameter configurations of AnoStyler are provided in Appendix D. All experiments were conducted on a single NVIDIA RTX 2080Ti GPU with 11 GB of memory. All experiments were repeated five times with different random seeds, and the average performance is reported.

Baselines. Our method AnoStyler was compared with a diverse set of zero-shot anomaly generation methods, including CutPaste (Li et al. 2021), DRAEM (Zavrtanik, Kris-

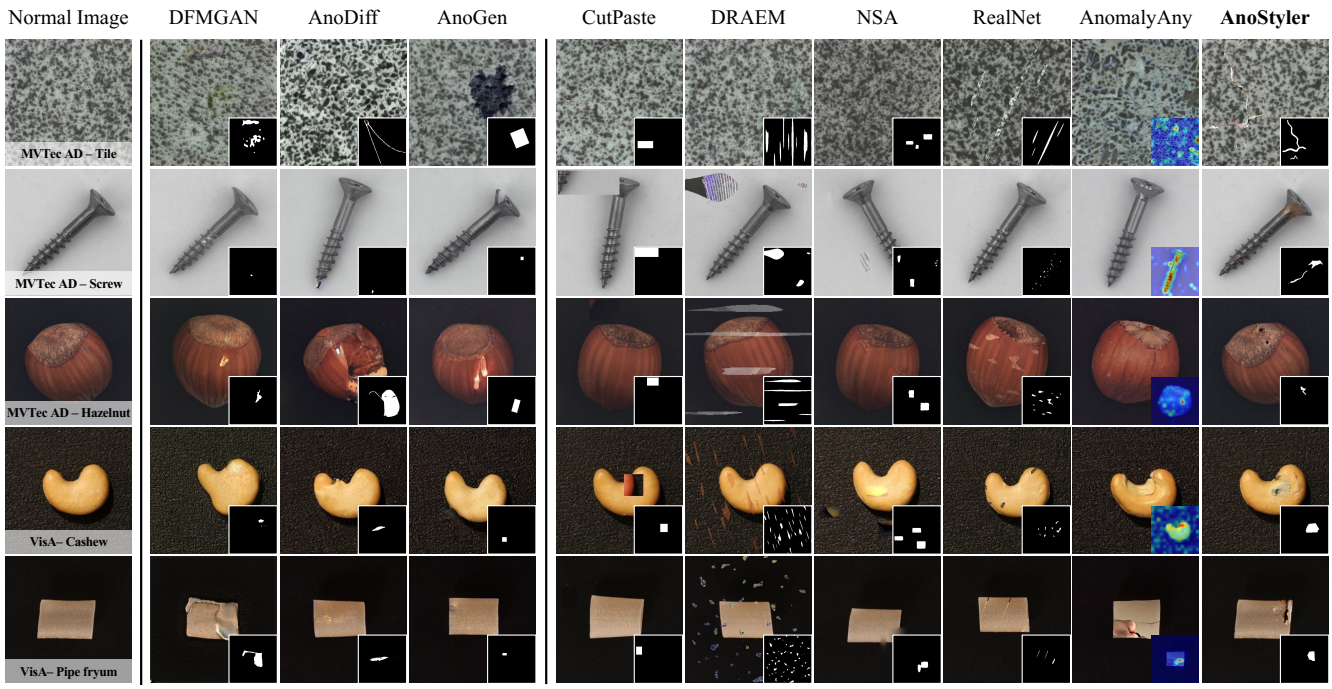

Figure 3: Comparison of generated anomaly images and their corresponding anomaly masks on MVTec-AD and VisA. AnoStyler generates visually realistic anomalies that align well with the masks.

tan, and Skočaj 2021), NSA (Schlüter et al. 2022), RealNet (Zhang, Xu, and Zhou 2024), and AnomalyAny (Sun et al. 2025). For broader comparison, we also considered few-shot methods, including DFMGAN (Duan et al. 2023), AnoDiff (Hu et al. 2024), AnoGen (Gui et al. 2024). It should be noted that few-shot methods have an inherent advantage in terms of performance, as they benefit from access to a small number of real anomaly images.

4.2 Anomaly Generation Results

We first evaluated the effectiveness of AnoStyler and the baseline methods in anomaly generation. Following the evaluation protocol of (Hu et al. 2024), we generated 1,000 anomaly images per category for each method. The Inception Score (IS) (Salimans et al. 2016) and the Intra-Cluster pairwise LPIPS distance (IC-L) (Ojha et al. 2021) were used to quantitatively evaluate generation quality and diversity.

Table 1 presents the quantitative performance comparison for anomaly generation, with results averaged across all categories within each benchmark dataset. The full category-wise results of AnoStyler are provided in Appendix E. AnoStyler achieved the highest IS and the second-highest IC-L on MVTec-AD, and obtained the best scores for both metrics on VisA, indicating superior visual quality and diversity in its images. The qualitative results are presented in Figure 3. The images generated by AnoStyler demonstrate competitive visual realism compared to few-shot methods that have access to real anomaly images, including DFMGAN, AnoDiff, and AnoGen. In contrast, heuristic zero-shot methods, including CutPaste, DRAEM, and NSA, tend to produce unrealistic anomalies. Among generative zero-shot methods, RealNet also suffers from a lack of realism, while AnomalyAny generates more realistic anomalies but often introduces artifacts and overly smoothed details in certain categories. Additional qualitative examples for all categories generated by AnoStyler are provided in Appendix F.

4.3 Anomaly Detection Results

We evaluated the methods based on their impact on the performance of the downstream anomaly detection task The evaluation protocol followed those used in Zavrtanik, Kristan, and Skočaj (2021) and Hu et al. (2024). For each category, we generated 500 anomaly images and used them with real normal images to train a U-Net model, which takes an image as input and predicts its anomaly mask. Anomaly detection performance was assessed on the test sets from each benchmark dataset. We used Area Under ROC Curve (AUROC), Average Precision (AP), and F1-score at both the image level (I-AUC, I-AP, I-F1) and pixel level (P-AUC, P-AP, P-F1) . Additionally, we included Per-Region Overlap (PRO) to evaluate region-level localization performance.

Table 2 presents the results averaged across categories within each benchmark dataset. The full category-wise results of AnoStyler are provided in Appendix E. AnoStyler achieved state-of-the-art anomaly detection performance on both MVTec-AD and VisA under zero-shot settings. Remarkably, despite having no access to real anomaly images during the generation of synthetic anomalies, AnoStyler yielded performance comparable with or even superior to few-shot baselines. This highlights the effectiveness

Method	MVTec-AD							VisA						
	I-AUC	I-AP	I-F1	P-AUC	P-AP	P-F1	PRO	I-AUC	I-AP	I-F1	P-AUC	P-AP	P-F1	PRO
Few-Shot Anomaly Generation														
DFMGAN⋆ (Duan et al. 2023)	87.2	94.8	94.7	90.0	62.7	62.1	76.3	83.7	85.7	80.3	90.6	31.0	34.6	74.9
AnoDiff⋆ (Hu et al. 2024)	99.2	99.7	98.7	99.1	81.4	76.3	94.0	86.9	89.1	85.4	93.2	33.0	36.8	79.0
AnoGen† (Gui et al. 2024)	98.7	99.6	97.7	96.9	73.6	66.7	90.7	90.4	92.4	87.2	89.1	31.6	33.4	73.8
Zero-Shot Anomaly Generation														
CutPaste† (Li et al. 2021)	89.8	92.1	89.8	88.2	51.9	50.7	76.4	86.3	86.9	87.1	88.4	32.2	39.6	77.7
DRAEM⋆ (Zavrtanik et al. 2021)	94.6	97.0	94.4	92.2	54.1	53.1	83.1	91.8	92.9	88.6	91.4	29.5	37.2	81.9
NSA† (Schlüter et al. 2022)	93.0	95.6	91.6	92.0	52.6	52.5	82.2	87.3	89.8	84.2	92.6	26.1	34.2	74.2
RealNet† (Zhang et al. 2024)	95.2	97.0	95.3	94.0	57.7	56.6	85.2	92.6	93.8	89.2	92.2	**33.3**	**41.0**	83.0
AnomalyAny† (Sun et al. 2025)	95.2	96.9	96.3	89.0	62.7	59.9	84.7	88.9	86.2	85.9	90.4	31.2	33.0	**84.6**
AnoStyler (ours)	**98.0**	**99.0**	**97.0**	**94.4**	**62.9**	**60.7**	**88.3**	**93.9**	**95.3**	**90.1**	**93.8**	31.4	36.4	84.3

Table 2: Comparison of anomaly detection on MVTec-AD and VisA. All scores are shown in percentages. For each metric, the best and second-best scores among zero-shot methods are indicated in **bold** and underlined. †: Re-implemented on both datasets; ⋆: Results on MVTec-AD taken from (Hu et al. 2024), with results on VisA re-implemented for a fair comparison.

#	$\mathcal{L}_{\text{gdir}}$	$\mathcal{L}_{\text{pdir}}$	$\mathcal{L}_{\text{mclip}}$	IS	IC-L	I-AUC	P-AUC
(a)				1.70	0.25	88.2	85.7
(b)	✓			1.86	0.29	95.2	92.5
(c)	✓	✓		1.96	0.30	96.7	93.2
(d)	✓	✓	✓	**2.04**	**0.32**	**98.0**	**94.4**

Table 3: Ablation study on the effect of the proposed loss components in AnoStyler on the MVTec-AD dataset. The results on the VisA dataset are provided in Appendix E.

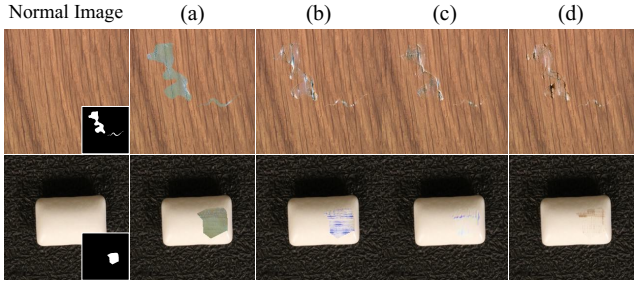

Figure 4: Qualitative results corresponding to the loss configurations presented in Table 3. The first and second rows use category labels `[c] = "wood"` and `"chewinggum"`, respectively, with defect type `[d] = "scratch"`.

of AnoStyler for downstream anomaly detection. Additional statistical significance analysis is provided in Appendix G.

4.4 Ablation Study

We conducted an ablation study to analyze the contributions of the proposed loss components: modifying $\mathcal{L}_{\text{gdir}}$, modifying $\mathcal{L}_{\text{pdir}}$, and adding $\mathcal{L}_{\text{mclip}}$, relative to the baseline training objective (Kwon et al. 2022). Table 3 presents the quantitative results. Directly applying the baseline objective, denoted as (a), resulted in degraded performance. The proposed modifications to $\mathcal{L}_{\text{gdir}}$ and $\mathcal{L}_{\text{pdir}}$, as in (b) and (c), led to performance improvements, highlighting their respective contributions. The best performance was achieved with (d), which corresponds to the full AnoStyler training objective incorporating all three proposed components, demonstrating their combined effectiveness. As shown in the qualitative examples in Figure 4, while (a) produces unrealistic anomalies, (b), (c), and (d) show gradual improvements in both visual realism and semantic alignment with the text prompt.

4.5 Computational Efficiency

AnoStyler comprises 263M parameters in total, including 91M for SAM, 0.61M for the stylization network $\mathcal{F}$, 151M for the CLIP encoders $\mathcal{E}_T$ and $\mathcal{E}_I$, and 20M for the feature extractor computing $\mathcal{L}_{\text{c}}$, whereas diffusion-based baselines such as AnoDiff, AnoGen, RealNet, and AnomalyAny each contain over 1B parameters, making AnoStyler much more compact. Among these methods, AnomalyAny is the only one that generates anomalies without large amounts of real images, serving as the most comparable baseline. Generating one anomaly image requires 9.5 TFLOPs for AnoStyler versus 22.8 TFLOPs for AnomalyAny, demonstrating its substantially lower computational cost.

5 Conclusion

In this paper, we proposed *AnoStyler*, a novel zero-shot anomaly generation method that generates synthetic anomaly images through text-guided style transfer from a single normal image. It produces high-quality, diverse anomalies with strong visual realism and semantic alignment to text prompts, without requiring access to large amounts of real images. In addition, its lightweight architecture enables computationally and memory-efficient anomaly generation. In experimental results on the MVTec-AD and VisA benchmarks, AnoStyler not only outperformed existing zero-shot methods but also achieved performance comparable to few-shot methods in both anomaly generation and downstream anomaly detection. These findings suggest that AnoStyler offers a text-driven alternative to recent mainstream methods that demand substantial computational resources, making it well suited for real-world deployment.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant (No. RS-2023-00207903) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No. RS-2025-02214591, Development of an Innovative AI Agent for Worker-Friendly Autonomous Manufacturing), funded by the Korea government (MSIT; Ministry of Science and ICT).

References

Batzner, K.; Heckler, L.; and König, R. 2024. EfficientAD: Accurate Visual Anomaly Detection at Millisecond-Level Latencies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 128–138.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9592–9600.

Cao, T. T.; Zhu, J.; and Pang, G. 2023. Anomaly Detection Under Distribution Shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6511–6523.

Chen, J.; Rong, P.; Sun, J.; Li, C.; Li, X.; and Lv, H. 2024. Soulstyler: Using Large Language Model to Guide Image Style Transfer for Target Object. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 4015–4020.

Cui, Y.; Liu, Z.; and Lian, S. 2023. A Survey on Unsupervised Anomaly Detection Algorithms for Industrial Images. *IEEE Access*, 11: 55297–55315.

Defard, T.; Setkov, A.; Loesch, A.; and Audigier, R. 2021. PaDiM: a Patch Distribution Modeling Framework for Anomaly Detection and Localization. In *Proceedings of the International Conference on Pattern Recognition*, 475–489.

Duan, Y.; Hong, Y.; Niu, L.; and Zhang, L. 2023. Few-Shot Defect Image Generation via Defect-Aware Feature Manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 571–578.

Fang, Q.; Su, Q.; Lv, W.; Xu, W.; and Yu, J. 2025. Boosting Fine-Grained Visual Anomaly Detection with Coarse-Knowledge-Aware Adversarial Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7943–7951.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. *ACM Transactions on Graphics*, 41(4): 141.

Ganugula, P.; Kumar, Y. S. S. S. S.; Reddy, N. K. S.; Chellingi, P.; Thakur, A.; Kasera, N.; and Anand, C. S. 2023. MOSAIC: Multi-Object Segmented Arbitrary Stylization Using CLIP. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 892–903.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image Style Transfer Using Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, 2672–2680.

Gui, G.; Gao, B.-B.; Liu, J.; Wang, C.; and Wu, Y. 2024. Few-Shot Anomaly-Driven Generation for Anomaly Classification and Segmentation. In *Proceedings of the European Conference on Computer Vision*, 210–226.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, 6840–6851.

Hu, T.; Zhang, J.; Yi, R.; Du, Y.; Chen, X.; Liu, L.; Wang, Y.; and Wang, C. 2024. AnomalyDiffusion: Few-Shot Anomaly Image Generation with Diffusion Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8526–8534.

Huang, X.; and Belongie, S. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.

Hyun, J.; Kim, S.; Jeon, G.; Kim, S. H.; Bae, K.; and Kang, B. J. 2024. ReConPatch: Contrastive Patch Representation Learning for Industrial Anomaly Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2052–2061.

Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; and Dabeer, O. 2023. WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19606–19616.

Kamra, C. G.; Mastan, I. D.; and Gupta, D. 2023. SEM-CS: Semantic CLIPStyler for Text-Based Image Style Transfer. In *Proceedings of the IEEE International Conference on Image Processing*, 395–399.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4015–4026.

Kwon, H.; Kim, D.; Choi, Y.; Kim, J.; and Ha, J.-W. 2022. CLIPstyler: Image Style Transfer with a Single Text Condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18062–18071.

Li, C.-L.; Sohn, K.; Yoon, J.; and Pfister, T. 2021. CutPaste: Self-Supervised Learning for Anomaly Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9664–9674.

Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; and Yang, M.-H. 2017. Universal Style Transfer via Feature Transforms. In *Advances in Neural Information Processing Systems*, 386–396.

Li, Y.; Liu, M.-Y.; Li, X.; Yang, M.-H.; and Kautz, J. 2018. A Closed-Form Solution to Photorealistic Image Stylization.

In *Proceedings of the European Conference on Computer Vision*, 453–468.

Li, Z.; Yan, Y.; Wang, X.; Ge, Y.; and Meng, L. 2025. A Survey of Deep Learning for Industrial Visual Anomaly Detection. *Artificial Intelligence Review*, 58(279).

Lin, D.; Cao, Y.; Zhu, W.; and Li, Y. 2021. Few-Shot Defect Segmentation Leveraging Abundant Normal Training Samples Through Normal Background Regularization and Crop-and-Paste Operation. In *Proceedings of the IEEE International Conference on Multimedia and Expo*.

Ojha, U.; Li, Y.; Lu, J.; Efros, A. A.; Lee, Y. J.; Shechtman, E.; and Zhang, R. 2021. Few-Shot Image Generation via Cross-Domain Correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10743–10752.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2085–2094.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards Total Recall in Industrial Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328.

Rudin, L. I.; Osher, S.; and Fatemi, E. 1992. Nonlinear Total Variation Based Noise Removal Algorithms. *Physica D: Nonlinear Phenomena*, 60(1–4): 259–268.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, 2234–2242.

Schlüter, H. M.; Tan, J.; Hou, B.; and Kainz, B. 2022. Natural Synthetic Anomalies for Self-Supervised Anomaly Detection and Localization. In *Proceedings of the European Conference on Computer Vision*, 474–489.

Singh, S.; Jandial, S.; Shahid, S.; and Java, A. 2024. LEAST: "Local" text-conditioned image style transfer. In *Proceedings of the CVPR Workshop on AI for Content Creation*.

Sun, H.; Cao, Y.; Dong, H.; and Fink, O. 2025. Unseen Visual Anomaly Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25508–25517.

Wu, H.; Xu, Q.; Li, Z.; Liang, S.; Chen, L.; Zhao, H.; Lin, D.; Xu, W.; and Wei, Y. 2025. DFM: Differentiable Feature Matching for Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15224–15233.

Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2021. DRAEM – A Discriminatively Trained Reconstruction Embedding for Surface Anomaly Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8330–8339.

Zhang, G.; Cui, K.; Hung, T.; and Lu, S. 2021. Defect-GAN: High-Fidelity Defect Synthesis for Automated Defect Inspection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2524–2534.

Zhang, X.; Xu, M.; and Zhou, X. 2024. RealNet: A Feature Selection Network with Realistic Synthetic Anomaly for Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16699–16709.

Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; and Dabeer, O. 2022. SPot-the-Difference Self-Supervised Pre-training for Anomaly Detection and Segmentation. In *Proceedings of the European Conference on Computer Vision*, 392–408.