Text-Guided Variational Image Generation for Industrial Anomaly Detection and Segmentation

Mingyu Lee[1,2] Jongwon Choi[1]*
[1]Dept. of Advanced Imaging, GSAIM, Chung-Ang University, Seoul, Korea
[2]Generative AI Team, LG CNS, Seoul, Korea
mglee@vilab.cau.ac.kr, choijw@cau.ac.kr

Abstract

We propose a text-guided variational image generation method to address the challenge of getting clean data for anomaly detection in industrial manufacturing. Our method utilizes text information about the target object, learned from extensive text library documents, to generate non-defective data images resembling the input image. The proposed framework ensures that the generated non-defective images align with anticipated distributions derived from textual and image-based knowledge, ensuring stability and generality. Experimental results demonstrate the effectiveness of our approach, surpassing previous methods even with limited non-defective data. Our approach is validated through generalization tests across four baseline models and three distinct datasets. We present an additional analysis to enhance the effectiveness of anomaly detection models by utilizing the generated images.

1. Introduction

Identifying anomalous components in industrial manufacturing, a task known as anomaly detection, has been a challenging but important problem to solve. Conventional methods for anomaly detection solve this problem by training the distribution of non-defective data in an attempt to identify defects [8, 21, 24–26, 28]. Recently, it has also been proposed to train invertible functions with the images to Gaussian distribution using the probabilistic method without adaptation to the target distribution [14, 22, 23, 29].

The efficacy of anomaly detection depends on the quantity and quality of available non-defective data, as these factors directly impact the model's ability to encompass the diverse spectra of target object appearances. Most of all, it is difficult to visually classify industrial defects since errors can vary from subtle changes as thin scratches to significant structural defects as missing components. Such difficulty naturally falls into the problem of out-of-distribution detection, in which a model must differentiate the samples obtained from the training data distribution from those outside of its range. As a result, it is essential to obtain various non-defective data that effectively represent the data distribution and are distinguished from defective cases.

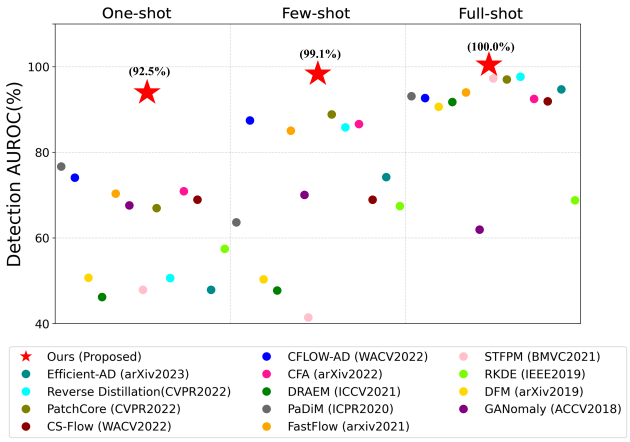


Figure 1. **Comparison with state-of-the-art baselines.** Our method generates non-defective images using a text-guided variational image generation method and utilizes the generated images as additional training data for anomaly detection. Ours outperforms the state-of-the-art methods across various settings, such as one-shot, few-shot (5 images), and full-shot training images. For comparison, we use the metal-nut class of MVTecAD dataset [3].

However, the industrial field suffers from various data issues. The first is the issue of imbalanced non-defective data, where the non-defective data itself can present a near-uniformity with only a few images exhibiting minor differences that qualify as acceptable defects. The second issue is related to the sensitivity of the data, especially in appearance and level of defectiveness, as they can vary significantly due to the type of machine used and the capturing conditions. The last one is the potential to wrongly label defective data as non-defective data by mistake during the excessive collection of non-defective data.

*Corresponding author.

To overcome the issues posed by the requirement of large non-defective data, we propose a text-guided variational image generation method for industrial anomaly detection and segmentation. To solve the problem of lack of diversity in the provided non-defective data, we extensively utilize text information about the target object learned through comprehensive text library documents to generate non-defective data images that are most similar to the input image. In addition, it is important to recognize that the normal distribution can also be influenced by the given set of non-defective image data, with its behavior being modulated by the variance predicted by our new framework of variational image generator. Integrating textual information and image data ensures that our proposed framework can generate several non-defective images while upholding the anticipated non-defective distribution derived from textual and image-based prior knowledge. By utilizing a bunch of non-defective images generated from our generator, we can employ our framework in various anomaly detection methods, thereby potentially enhancing their overall performance. We validated the proposed method through generalization tests across four baseline models and three distinct datasets. Interestingly, as shown in Fig. 1, even relying on a single or a few non-defective images, our method outperforms the previous methods requiring many non-defective data.

We can summarize our contribution as follows:

- To ensure generalization and robustness, we developed a variation-based image generator to predict and preserve the variance of the provided non-defective images.
- To solve the problem of lack of diversity in good product data, we developed a keyword-to-prompt generator that generates the best prompt by comparing text information about the target object, learned extensively through comprehensive text library documents, with the input image.
- To bridge the semantic gap arising from different modalities, we developed a text-guided knowledge integrator method in which latent image features are aligned with the text information of the target object.
- To validate the efficacy of our approach, we have merged our method into several state-of-the-art algorithms and tested extensively across various real-world industrial datasets, and the experimental results confirm that our framework shows impressive performance even with a single or a few non-defective images.

2. Related Works

Text-based Anomaly Detection. Most Anomaly detection methods include representation-based methods that extract discriminative features for patches and calculate distances between distributions [14, 29] and reconstruction-based methods [21, 30] using generative adversarial networks. With anomaly detection advances, benchmark dataset performance in full shot settings has become saturated due to [3] state-of-the-art research [1, 2, 27]. However, recent studies employ different settings and methods, such as text-based anomaly detection. For example, WinCLIP [15] generates pre-text prompts for the input image by mixing state words and text candidates. Then, it extracts multi-scale features of the input image through a binary mask and calculates a similarity score between the generated prompts to detect anomalies.

In contrast, our method focuses on generating non-defective images based on the similarity of the input image and the text-based prior knowledge. It improves performance by feeding the newly generated images of different attributes into the general anomaly detection methods.

Text-to-Image Generation. Recently, text-to-image generation models [16, 19] are widely studied, which predict or generate the next pixel values based on the given text and all previous predictions of pixel values. For example, DALL-E [20] takes text input from the GPT series model to predict the next pixels, which are then added to the initial image to be placed as the subsequent input along with the text. By repeating the process multiple times, DALL-E can generate an entire image. GAN [12] is known for high-quality and realistic image generation without a prompt, and Contrastive Language-Image Pre-training (CLIP) is another neural network that can determine how well a caption or prompt matches an image. Through VQGAN-CLIP [5], where VQGAN [10] creates an image, CLIP [19] checks whether the generated image from VQGAN matches the text. When the process is repeated, the created image gradually resembles the input text.

However, the previous text-guided image generation methods ignore the importance of the representation variance of the given image data, which is critical for generalizing the non-defective image data in anomaly detection tasks. In contrast to the previous text-based image generation methods, we aim to predict the variance of non-defective image distribution, which is utilized to avoid generating defective images while enlarging the variety of appearances in non-defective images.

3. Preliminary Analysis

Hypothesis. Data issues related to sensitivity, especially in appearance and level of defectiveness, can vary significantly due to the type of machine used and the capturing conditions. Thus, the similarity between the generated and original images is highly correlated with performance enhancement, and the visual variance of generated images also contributes to performance increases. In Fig. 2, the similarity between the generated and original images is highly correlated with the performance enhancement, and the variance of generated images also contributes to the performance increase. These results empirically validate our hypothesis

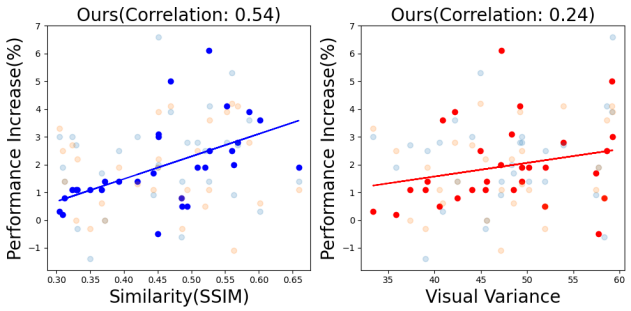


Figure 2. **Correlation of hypothesis.** We repeat the tests with different images to confirm our hypothesis. Performance enhancement is strongly connected with the similarity between the generated and original images, and the visual variance of created images also improves performance.

that the generated images should be similar to the appearance of the provided non-defective images while preserving their visual variance. Furthermore, we compared the correlation between the original image and the naive prompts (0.39), just words (0.35), and ours(0.54), which confirms the necessity of well-designed prompts in our hypothesis.

Preliminary Analysis. We repeat the preliminary experiments using various images with a simple scenario to verify our hypothesis in general. In the scenario, we compare the performance between the baseline anomaly detection model and the same model using one additional training image searched on the web or generated by an image generator. We consider a single *hazelnut* image from the MVTecAD dataset [3], utilizing Patchcore [21] for the baseline anomaly detection model.

We build various types of generated images : *Original Image*, *Web Image* [13], *Midjourney* [17], and *DALL-E* [20], *VQGAN-CLIP(only Word)*, and *VQGAN-CLIP(Naïve Prompt)*. First, *Original Image* utilizes the given single original image to train the anomaly detection model, while *Web Image* adds one additional image retrieved from the web by using a keyword of "Hazelnut." For the text-based image generator, we first generated the text captioning from the original hazelnut image by using GPT-4, and the result was: *"This image displays a single hazelnut with a textured, fibrous cap on top. The nut itself has a rich, one image was created using warm brown color with visible stripes and markings that suggest a natural origin. It rests against a dark background which serves to highlight the nut's detailed texture and organic shape."*. *Only Word* and *Naïve Prompt* utilize the prompts of "Hazelnut" and "A photo of a hazelnut", respectively. The three types of prompts were fed into *Midjourney*, *DALL-E*, and *VQGAN-CLIP* to generate the additional training image. Fig. 3 shows the additional training image from various generation schemes, presenting that the results highly depend on the types of image generation models and the prompts.

Table 1. **Preliminary experiment using Patchcore. [21]** The results the need for a well-designed generation model to boost anomaly detection performance by adding extra training images.

Method	Detection	Segmentation
Original Image	**88.3**	**94.3**
Web Image	84.8	93.1
Midjourney (by Generated Prompt)	82.7	93.3
DALL-E (by Generated Prompt)	85.7	91.9
VQGAN-CLIP (by Only word)	64.8	86.8
VQGAN-CLIP (by Naïve Prompt)	71.9	88.2
VQGAN-CLIP (by Generated Prompt)	81.4	91.7
Ours	**93.3**	**94.6**

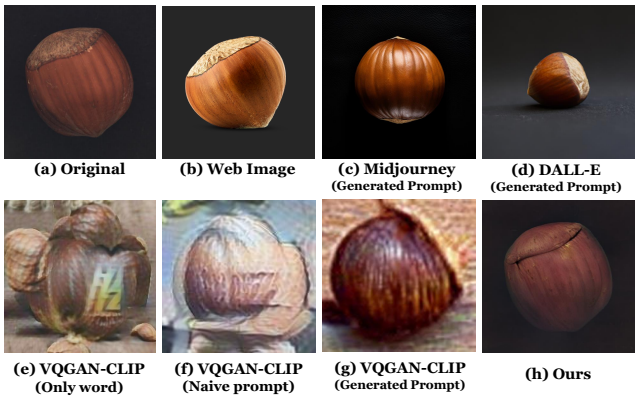


Figure 3. **Generated images in the preliminary experiment.** (a) The original image of *hazelnut* in MVtecAD dataset. (b) The image retrieved by a keyword of 'hazelnut' from the web. (c),(d) The images generated using Midjourney and DALL-E, respectively, using a captioning of the original image as a prompt. (e),(f),(g) The images generated using the VQGAN-CLIP model based on 'hazelnut', 'A photo of a hazelnut', and the captioning of the original image, respectively. (h) The image generated by our method.

The results of preliminary experiments are represented in Table 1, showing that performance is affected depending on the generated images. Interestingly, preliminary results show that good image quality does not necessarily help improve performance as presented in the low performance of DALL-E. Meanwhile, given the performance of web images is relatively high, it can be assumed that performance can be improved if elements such as outlines and angles of the original image are preserved in the additional images.

Therefore, while simply generated images cannot be utilized for anomaly detection models, we need to consider several additional factors for generated images to train non-defective distribution effectively. First, generated images should be similar to the appearance of the provided non-defective images while preserving their visual variance. Second, it is important to find the best prompt to generate visually well-structured images. Lastly, based on the above two information contents, we should create images with a small semantic gap even when an insufficient number of non-defective images is given.

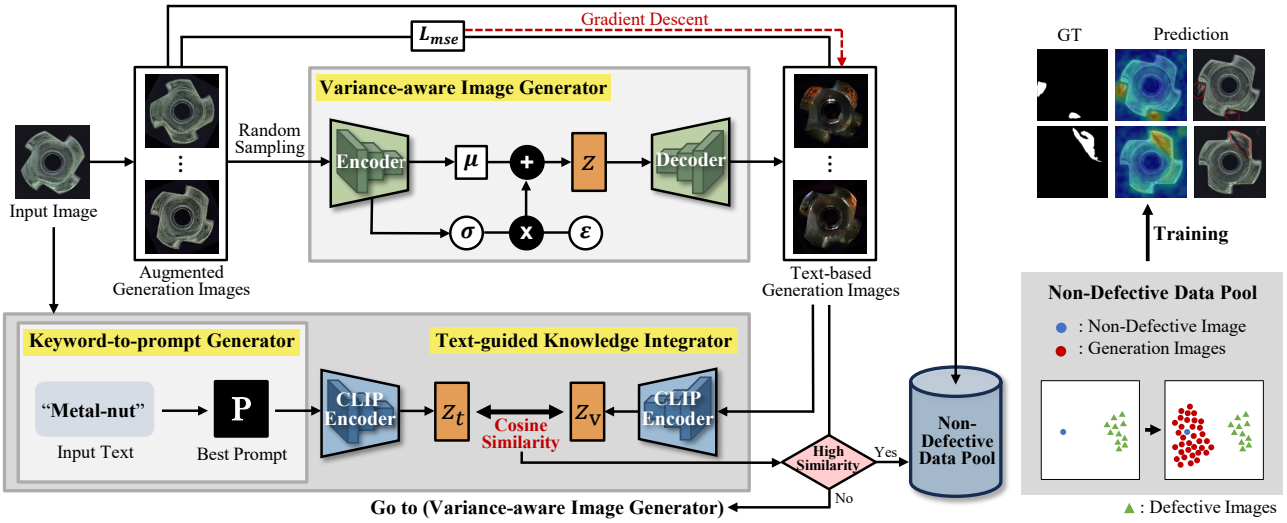


Figure 4. **Overview of our framework.** Our framework comprises a keyword-to-prompt generator, a variance-aware image generator, and a text-guided knowledge integrator. The keyword-to-prompts generator creates prompts from key input words and selects the best one that matches an input image. A variance-aware image generator creates non-defective images, encoding their visual features into a normal distribution to maintain variance. Our process updates through iteration, and a text-guided knowledge integrator selects the optimal images by comparing the similarity of their latent distribution to the text prompts.

4. Methodology

As shown in Fig. 4, our framework comprises the keyword-to-prompt generator, variance-aware image generator, and text-guided knowledge integrator. The keyword-to-prompt generator produces a group of prompts consisting of different word combinations based on the essential words presented in the input text, then selects the best prompt that contains comparable information to the input image among the generated ones. The variance-aware image generator generates a set of non-defective images, encoding the visual features of the non-defective images into the normal distribution to preserve their variance. The generation process is iterated with the generator update, and the text-guided knowledge integrator determines the optimal set of generated images by assessing the latent distribution similarity between the image set and the text prompts generated by the keyword-to-prompt generator. Finally, the optimal set of generated images is utilized as the additional training set for the baseline anomaly detection model.

4.1. Generation Modules

The image generation module comprises a keyword-to-prompt generator and a variance-aware image generator. The keyword-to-prompt generator integrates the target object name W^o with the set of predefined status words ($\{W_1, W_2, \ldots, W_T\}$) to generate multiple candidate prompts ($\{S_1, S_2, \ldots, S_T\}$). Then, the keyword-to-prompt generator selects the best prompt from the candidate prompts according to their latent distances with the latent feature of the original images I^o. Then, we convert the input image $\{I^o\}$ into multiple augmented images $I_1, I_2, \ldots, I_N$, which are fed into the variance-aware image generator, which encodes the image I_i into the corresponding latent distribution $\mathbf{z}_i$. Finally, we decode $\mathbf{z}_i$ by sampling to generate a new multiple image set $\{I_1^+, I_2^+, \ldots, I_M^+\}$.

4.1.1 Keyword-to-Prompt Generator

The keyword-to-prompt generator produces the most appropriate prompt P based on the given object name W^o and the original training images I^o. With WordNet [11], we construct a set of T-different words $\{W_1, W_2, \ldots, W_T\}$ to obtain the candidate prompts $(S_1, S_2, \ldots, S_T)$ where S_t is "a $\{W^o\}$ with $\{W_t\}$" replaced by a corresponding word of $\{W_1, W_2, \ldots, W_T\}$. Using WordNet [11], we can collect a large database containing synonyms, hypernyms, hyponyms, and part-whole relationships with high semantic relevance of W^o.

We find the optimal prompts from the candidate prompts of $(S_1, S_2, \ldots, S_T)$ with two stages: distance-based outlier removal and embedding similarity estimation. First, we remove the outlier prompts by using the L2 distance between the latent features of the candidate prompt and the original image. We define the positive prompt set where the outlier prompts were removed by S^p, which is determined by $S^p = S_j | d(S_j, I^o) > 0.5$, where $d(S_i, I^o) = ||f(I^o) - G(S_i)||_2$ and $f(I)$ and $G(S)$ are the image embedding vector of the image I and text embedding vector of the prompt S, respectively. We utilized CLIP encoders [19] to obtain the image and text embedding vectors. Then, from the positive prompt

set S^p, we determine the best prompt P by using the cosine similarity as follows:

$$P = \arg\max_{S_j \in S^p} \cos\big(f(I^o), G(S_j)\big), \tag{1}$$

where $\cos\big(f(I^o), G(S_j)\big)$ means the cosine similarity between the two embedding vectors of $f(I^o)$ and $G(S_j)$.

4.1.2 Variance-aware Image Generator

As a baseline architecture of the variance-aware image generator, we employ the VQGAN model [10]. This model integrates a vector quantization process, which maps the continuous latent space to a discrete codebook. Due to this characteristic, the encoder's output specifically follows the approximate posterior distribution of the latent vector, which enables to effectively reconstruct the given input image via the decoder. The vector quantization process plays a pivotal role, as it effectively represents the image within the latent space by mapping continuous representations to a fixed codebook. Additionally, the VQGAN model ensures that the distribution of the latent variable closely approximates the standard normal distribution.

The architecture of VQGAN can be represented as:

$$\begin{aligned} I' &= p^o\big(E(I)\big) \\ E(I) &= [e\big(q_1(I)\big), e\big(q_2(I)\big), \ldots, e\big(q_D(I)\big)] \\ e\big(q_d(I)\big) &= \arg\min_{\mathbf{e}_k \in \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}} ||q_d(I) - \mathbf{e}_k||_2, \end{aligned} \tag{2}$$

where $p^o\big(E(I)\big)$ is a decoder to generate an entire image by integrating the patches each generated from the d-th latent vector sampled from a normal distribution of $\mathcal{N}\{e\big(q_d(I)\big), 1\}$, $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K\}$ are the codebook vectors embedded in VQGAN, and K is the codebook size.

However, existing VQGANs uniformly utilize the variance σ of the latent vector by 1, ignoring the possible diversity in the patch-wise appearance of the target object. Especially, the appearance diversity must be considered for anomaly detection models where a trained distribution of non-defective images determines defects. To address the issue, we extend the VQGAN architecture to predict the variance of latent variables. The variance-aware image generator can be represented as:

$$I_m^+ = p\Big(\mathbb{E}_{\forall i}[E(I_i)], \Sigma_{\forall i}\big(E(I_i)\big)\Big), \tag{3}$$

where $p(E, \Sigma)$ represents the extended decoder of VQGAN sampling d-th patch's latent vector from the distribution of $\mathcal{N}\{E_d, \Sigma_d\}$, E_d and Σ_d represent the d-th column of E and Σ, respectively, and $\Sigma_{\forall i}\big(E(I_i)\big)$ is a function estimating the column-wise variance vectors of $E(I_i)$ for all possible i. We should remind that I_m^+ and I_i are the m-th generated image and the i-th augmented one, respectively. Thus, our variance-aware image generator extracts the latent variables according to the variance value estimated from the given image and the target patch. Due to the sampling process in the VQGAN, the generated images I_m^+ differ at each sampling iteration m.

4.2. Text-guided Knowledge Integrator

In this procedure, we generate the non-defective images aligned well with the best prompt P, adding them to the non-defective data pool for the anomaly detection model. From P selected from the keyword-to-prompt generator, we extract a textual clip feature z_t through a clip text encoder as $z_t = G(P)$. At the same time, we generate the images $\{I_1^+, I_2^+, \ldots, I_M^+\}$ by using the variance-aware image generator, and the visual clip feature is estimated by averaging multiple image features as:

$$z_v = \mathbb{E}_{\forall m}\big[f(I_m^+)\big]. \tag{4}$$

The text-guided knowledge integrator repeats the generation of the image set, scoring the image set by the cosine similarity between the image feature expectation z_v and the best prompt feature z_t. Here, we define the l-th generated image set as $\{I_1^{l+}, I_2^{l+}, \ldots, I_M^{l+}\}$, and its corresponding visual clip feature as z_v^l. Based on the cosine similarity between z_v^l and z_t, we select the best set of generated images, which are used to train the following anomaly detection model. We define the best set of generated images as:

$$\begin{aligned} \{\widehat{I}_1^+, \widehat{I}_2^+, \ldots, \widehat{I}_M^+\} &\equiv \{I_1^{\alpha+}, I_2^{\alpha+}, \ldots, I_M^{\alpha+}\}, \\ \alpha &= \arg\max_{l \in \{1, \ldots, L\}} \cos(z_t, z_v^l), \end{aligned} \tag{5}$$

where L is the number of iteration to generate the image.

At every iteration to generate the new set of images, we simultaneously update the variance-aware image generator. This update is necessary for the image generator to generate images similar to the input of non-defective images, and second, we can enhance the variety of generated image sets. We update all parameters of the variance-aware image generator, including the encoder, decoder, and codebook vector parameters. In addition to the original loss of VQGAN, we also utilize the Mean Squared Error (MSE) loss to generate images similar to the input images. The MSE loss is derived as follows:

$$\mathcal{L}_{mse}(I^o, \{I_1^+, \ldots, I_M^+\}) = \frac{1}{n}\sum_{m=1}^{M} ||I^o - I_m^+||_2^2. \tag{6}$$

Consequently, the total training loss can be represented as follows:

$$\mathcal{L} = \mathcal{L}_{mse}(I^o, \{I_1^+, \ldots, I_M^+\}) + \lambda \mathcal{L}_{vq}\big(\{I_1, \ldots, I_N\}\big), \tag{7}$$

where $\mathcal{L}_{vq}\big(\{I_1, \ldots, I_N\}\big)$ is the conventional loss for VQGAN [10] with the input image set of $\{I_1, \ldots, I_N\}$ and λ is a user-defined hyperparameter. We utilize the Adam optimizer to update the estimated gradients across the variance-aware image generator.

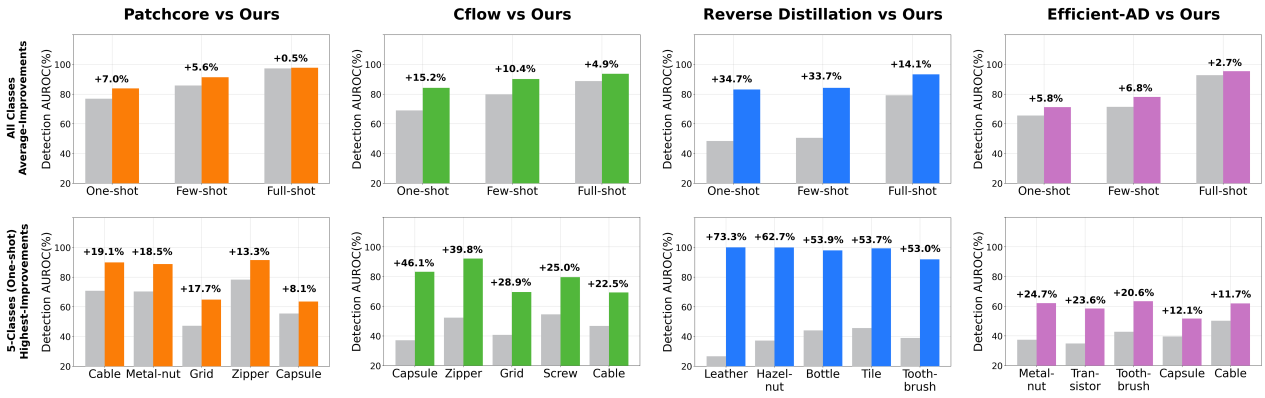


Figure 5. **Generalization test for anomaly detection in MVTecAD dataset.** The first row shows the average improving score across different baselines and varying numbers of non-defective images. The second rows present the average score for the highest-improving five classes.

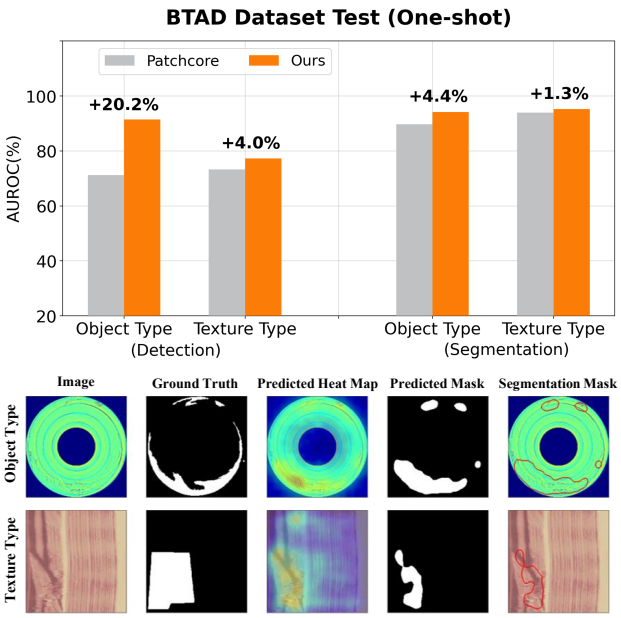


Figure 6. **Generalization test in BTAD dataset.** We have conducted experiments on detection and segmentation performance on both object and texture types of BTAD dataset [18] using Patchcore as baseline in one-shot setting. Our method improves the overall performance, particularly for the object type.

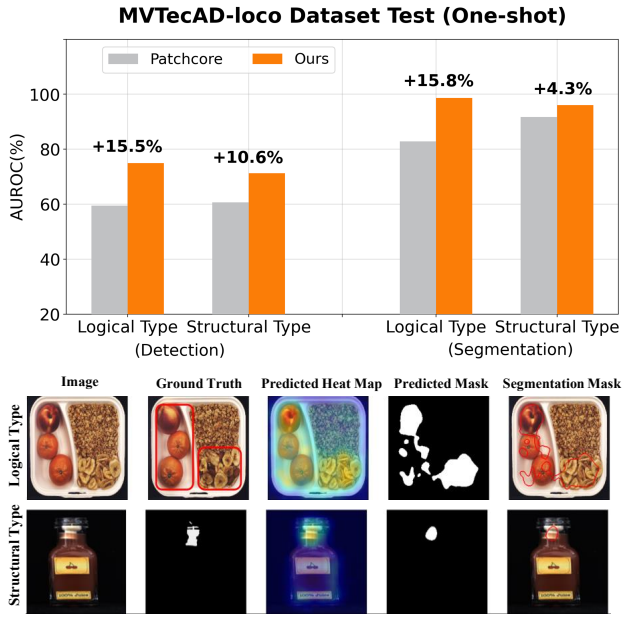


Figure 7. **Generalization test in MVTec-LOCO AD dataset.** We have conducted experiments on detection and segmentation performance in MVTec-LOCO AD dataset [4] using Patchcore as baseline in one-shot setting. Our method improves the performance for overall cases.

5. Experimental Result

5.1. Quantitative Results

We have conducted a comparative analysis of the generalization performance of our newly designed framework for one-shot, few-shot (5 images), and full-shot (entire training images) compared to the baselines. Firstly, the integrated average performance of the classes is derived by applying the state-of-the-art model [2] and the latest baselines [9, 14, 21] to the MVTecAD dataset [3]. Also, we have analyzed the five classes with the highest improvement to show the strengths of our method. Secondly, we have conducted experiments using the BTAD dataset [18] consisting of two types of objects and textures to find the object types in which our model performs well. Finally, we have validated the model's performance by assigning tasks

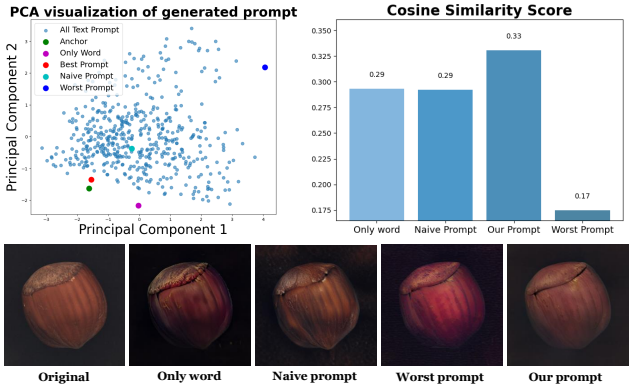


Figure 8. **(Top-Left) Similarity comparison between input text and each prompt, (Top-Right) Similarity comparison between input image and each prompt, (Bottom) Qualitative results of generated images.** The prompt form by text generator is used as follows: "hazelnut" for Only word, "a photo of a hazelnut" for Naïve prompt, "a {hazelnut} with {cobnut}" for Our prompt, and "a {hazelnut} with {decantherous}" for Worst prompt

with complicated object types of the MVTec-LOCO AD dataset [4].

In Fig. 5, the experimental results on generalized detection using MVTecAD dataset indicate enhanced performance across all baselines, thus verifying our model's ability to generalize. The results show an average increase of 15.7% in one-shot, 14.1% in few-shot, and 5.6% in full-shot, confirming that our framework can successfully achieve impressive performance even with a few training images. Additionally, we have analyzed the performance of the five highest improving classes in one-shot tasks, as shown in the second rows of Fig. 5. The results show that object types, such as the cable and metal-nut classes from the Patchcore model, show impressive improvements of 19.1% and 18.5%, respectively. Other baselines show a similar trend in object types, and the reverse distillation model exhibits a significant increase of 73.3% even in texture types such as leather.

As shown in Fig. 6, we have conducted additional experiments on the BTAD dataset to revalidate which defects our model works well. As in the graph of Fig. 6 (Top), the detection AUROC shows improvements of 20.2% for object type (1st class) and 4.0% for texture type (2nd class), compared to baseline. Similarly, the segmentation AUROC has also shown more significant improvements for the object type, with 4.4% for object types and 1.3% for texture types. In the qualitative results of Fig. 6 (Bottom), the object type has resulted in better outcomes for locating the anomaly segments close to the ground truth. As a result, this experiment confirms that our model is more robust on the object types.

However, failures happen in more diverse and complicated situations in the real world; thus, it is necessary to verify the robustness of our model in more complex settings. MVTec-LOCO AD dataset is suitable for testing complex tasks since it consists of complicated classes, such as logically structured arrangement (logical type) and surface defect recognition (structural type). Fig. 7 shows the experimental results on the MVTec-LOCO AD dataset. For the logical type, our method shows improvements of 15.5% for the detection AUROC and 15.8% for segmentation AUROC. Furthermore, for the structural type, our method also shows improvements for both cases, with 10.6% for detection AUROC and 4.3% for segmentation AUROC. In Fig. 7(Bottom), the qualitative results also show that our method can effectively detect and classify the changes in the placement structure of a breakfast box and the subtle differences in a juice bottle.

Table 2. **Evaluation of image quality.** The image generated by our prompt show the highest diversity and quality based on Inception Score (IS), Structural Similarity Index (SSIM), Peak Signal to Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS).

Method	Prompt	IS(↑)	SSIM(↑)	PSNR(↑)	LPIPS(↓)
Original	N/A	1.00	1.00	-	0.00
Only word	"a hazelnut"	1.62	0.50	19.01	984.15
Naïve prompt	"a photo of hazelnut"	1.34	0.78	23.46	764.48
Worst prompt	"a {hazelnut} with {decantherous}"	4.56	0.79	23.57	782.75
Our prompt	"a {hazelnut} with {cobnut}"	**22.59**	**0.90**	**28.39**	**701.84**

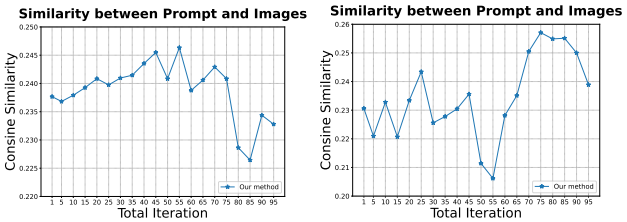


Figure 9. **The trend of similarity score between the generated image and the best prompt according to the number of iterations.** (Left), The trend of similarity score in hazelnut class, (Right) The trend of similarity score in leather class in MVTec-AD dataset.

5.2. Experimental Analysis

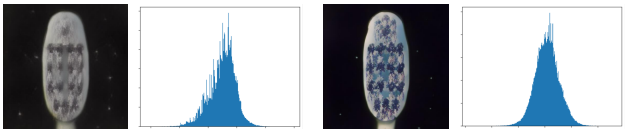
5.2.1 Optimizing results between our prompt and generated images

As shown in Fig. 8, we visualize the results of the Keyword-to-prompt Generator successfully finding the best prompt, and the images generated by Variance-aware Image Generator and Text-guided Knowledge Integrator based on the best prompt.

Fig. 8 (Top-Left) is a graph comparing the distance based on the similarity between input text and prompts based on the cosine similarity value. The worst prompt (blue dot) and

Table 3. **Ablation studies using Patchcore. [21]** 1) The performance is enhanced by the text prompt created with the text generator. 2) Initially, performance rises as the number of generated images increases, but it eventually reaches a saturation point over time. 3) In addition, the image generation module uses the variance method to guarantee a uniform distribution of the latent vector. This technique directly affects the quality and performance of the resulting images.

Component	Image-Level (%)	Gain (%)	Pixel-Level (%)	Gain (%)
Baseline	**74.5(±1.6)**	-	**97.0(±0.1)**	-
1) Text prompt				
w/o Text	77.6(±0.6)	(+3.1%)	**97.1(±0.0)**	**(+0.1%)**
w/ Text (ours)	**79.1(±0.6)**	**(+4.6%)**	96.9(±0.0)	(-0.1%)
2) N-Generated images				
1-copy	78.6(±0.4)	(+4.1%)	97.0(±0.0)	(+0.0%)
10-copy	77.9(±0.4)	(+3.4%)	96.9(±0.0)	(-0.1%)
20-copy	78.4(±0.5)	(+3.9%)	**97.0(±0.0)**	**(+0.0%)**
30-copy	**79.0(±0.7)**	**(+4.5%)**	97.0(±0.0)	(+0.0%)
50-copy	78.5(±0.3)	(+4.0%)	96.9(±0.1)	(-0.1%)
100-copy	78.4(±0.5)	(+3.9%)	96.9(±0.0)	(-0.1%)
3) Variance-aware				
w/o Variance	74.3(±0.4)	(-0.2%)	97.0(±0.0)	(+0.0%)
w/ Variance (ours)	**78.9(±0.5)**	**(+4.3%)**	**97.3(±0.0)**	**(+0.3%)**



w/o Variance **w/Variance (ours)**

Figure 10. **Qualitative results by variance-aware method.** The variance-aware method affect the distribution of image latent vectors, it also affect to the quality of image generation.

our prompt (red dot) show a stark difference. Fig. 8 (Top-Right) is a graph comparing the similarity scores of input images and each prompt, and we can show that the prompt received the highest score (33%) and the worst prompt received the lowest score (17%). We find that our prompts have characteristics close to the non-defective image and are evidence that they can boost the performance of text-based knowledge integration modules. In Fig. 8 (Bottom), we show that the qualitative image result generated by the best prompt is most similar to the original image. In Table 2, we additionally compared the quantitative results based on the qualitative result images in Fig. 8. We compared IS, SSIM, PSNR, and LPIPS scores and showed that our method had the highest diversity and high visual similarity to the original compared to other prompts.

In Fig. 9, we analyze the trend of similarity between the best prompt and the generated image for each iteration. Looking at the results, the generated image and the number of iterations are not unconditionally proportional, but the performance is optimized at a certain point and then decreases. We prove our model is looking for optimal images while generating images with different properties.

5.2.2 Ablation Studies

Table 3 shows the impact of the respective component of our method. The results with our prompt generator show 4.6% improvement compared to the baseline.

We also analyze the effect of the quantity of the text-based generation images on the performance of the text generator. The model's performance continuously improves as the number of text-based generation images rises and tends to reach a saturation point. The evidence is that excessive accumulation of non-defective data can act as critical noise in data expression. We found a performance difference of up to 1.1% depending on the number of images.

Lastly, we compare the results of using the variance-aware method. The variance method had an impact of approximately 4.3%. Additionally, In Fig. 10, we compared qualitative results for the Variance-aware method. The Variance-aware method made latent vector from the distribution of $\mathcal{N}\{E_d, \Sigma_d\}$ value in a form close to the normal distribution, and the generation quality is also improved.

6. Conclusion

To solve the issues in data shortages of the anomaly detection task in large-scale industrial manufacturing, our study proposes a new framework composed of variance-aware image generator, keyword-to-prompt generator, and text-guided knowledge integrator. Based on diverse real-world industrial situations and data shortage scenarios, we have conducted extensive experiments using four baselines and three datasets for comparative analysis. The experimental results show the impressive performance of our framework in all of the assumed scenarios, especially with a single non-defective image. We have found that the uniform distribution of the latent vector helps improve performance by preserving the characteristics of non-defective images during regeneration. With a text-based multimodal model, our study shows our framework's potential to effectively perform anomaly detection and segmentation, even in industrial environments with insufficient data. Furthermore, our method shows impressive performance even with a limited number of non-defective images, which enables us to avoid the problem of intertwining the defective images while effectively gathering the large-scale non-defective image set.

Acknowledgements: This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2021-0-01341, Artificial Intelligence Graduate School Program(Chung-Ang University) and 2021-0-02067, Next Generation AI for Multi-purpose Video Search and 2014-3-00123, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis) and a grant (22193MFDS471) from the Ministry of Food and Drug Safety in 2024.

References

[1] Jaehyeok Bae, Jae-Han Lee, and Seyun Kim. Image anomaly detection and localization with position and neighborhood information. *arXiv*, 2022. 2

[2] Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. *arXiv*, 2023. 2, 6, 12, 13, 15

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 1, 2, 3, 6, 12, 14, 15

[4] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130 (4):947–969, 2022. 6, 7, 12, 16

[5] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *Computer Vision–ECCV 2022*, pages 88–105. Springer, 2022. 2

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019. 18

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 18

[8] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*, pages 475–489. Springer, 2021. 1

[9] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022. 6, 12, 13, 15

[10] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020. 2, 5, 11

[11] Christiane Fellbaum. Wordnet and wordnets. In *Encyclopedia of Language and Linguistics*, pages 665–670. Elsevier, 2005. 4, 11

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[13] Google. Google image search. https://www.google.com/imghp?hl=ko&ogbl, 2023. Accessed on 2023. 3

[14] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 98–107, 2022. 1, 2, 6, 12, 13, 14

[15] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023. 2

[16] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429, 2022. 2

[17] MidJourney. Midjourney home page. https://www.midjourney.com/home, 2023. Accessed on 2023. 3

[18] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics*, pages 01–06. IEEE, 2021. 6, 12, 16

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. Proceedings of Machine Learning Research, 2021. 2, 4, 11

[20] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2, 3

[21] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection, 2021. 1, 2, 3, 6, 8, 12, 13, 14, 16

[22] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1907–1916, 2021. 1

[23] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1088–1097, 2022. 1

[24] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. Proceedings of Machine Learning Research, 2018. 1

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

[26] Shelly Sheynin, Sagie Benaim, and Lior Wolf. A hierarchical transformation-discriminating generative model for few shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8495–8504, 2021. 1

[27] Minghui Yang, Peng Wu, and Hui Feng. Memseg: A semi-supervised method for image surface defect detection using differences and commonalities. *Engineering Applications of Artificial Intelligence*, 119:105835, 2023. 2

[28] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1

[29] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv*, 2021. 1, 2

[30] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021. 2