Free Lunch of Image-mask Alignment for Anomaly Image Generation and Segmentation

Xiangyue Li[1], **Xiaoyang Wang**[2], **Zhibin Wan**[1], **Quan Zhang**[2], **Yupei Wu**[3], **Tao Deng**[1], **Mingjie Sun**[1*]
[1]School of Computer Science & Technology, Soochow University
[2]Xi'an Jiaotong-Liverpool University
[3]Aqrose Technology

Abstract

This paper aims at generating anomalous images and their segmentation labels to address the lack of real-world anomaly samples and privacy issues. Departing from conventional approaches that use masks solely to guide the generation of anomaly images, we propose a dual-branch training strategy for the generative model. This strategy enables the simultaneous production of anomaly images and masks, with an Alignment Regularization loss that ensures the coherence between the generated images and their masks. During inference, only the image-generation branch is activated to produce synthetic samples for training the downstream segmentation model. Furthermore, we propose to integrate well-trained generative model into the training of segmentation models, utilizing a Generative Feedback loss to refine the segmentation model's performance. Experiments show our method's IoU metrics exceed previous methods by 5.03%, 5.68% and 16.63% on Real-IAD (industrial), polyp (medical) and Floor Dirty (indoor) datasets. The code is publicly accessible at *https://github.com/huanyin/anomaly-alignment*.

1 Introduction

The objective of anomaly image generation and segmentation is to produce anomalous images, such as industrial product defects or diseased areas in medical imagery, along with their corresponding mask labels in a single process. This task addresses the scarcity of real-world anomaly samples and the associated privacy issues. It is a crucial task in computer vision, characterized by the simultaneous creation of images and masks, which distinguishes it from traditional image generation tasks. By creating these synthetic samples, we can overcome the shortage of real ones and improve the performance and generalization ability of the downstream segmentation models to better handle various anomaly situations.

In previous research on anomaly image generation and segmentation, prevalent approaches tend to employed a pipeline where the guidance mask was utilized as a condition for the generative model, specifying the location of the anomaly within the synthetic image. Once the synthetic image was produced, the mask was treated as the segmentation label. However, misalignment happens when anomaly regions in the generated images fail to accurately align with the conditional mask. This phenomenon, known as the mask label drift issue, negatively impacts the performance of the segmentation models by introducing false training signals.

This drift issue may stem from the generative model's incomplete understanding of the image-mask correspondence during training. To enhance this understanding, we propose a dual-branch training approach for the generative model. Specifically, different from previous methods that focused solely on the generation of anomaly images (Figure. 1(a)), we

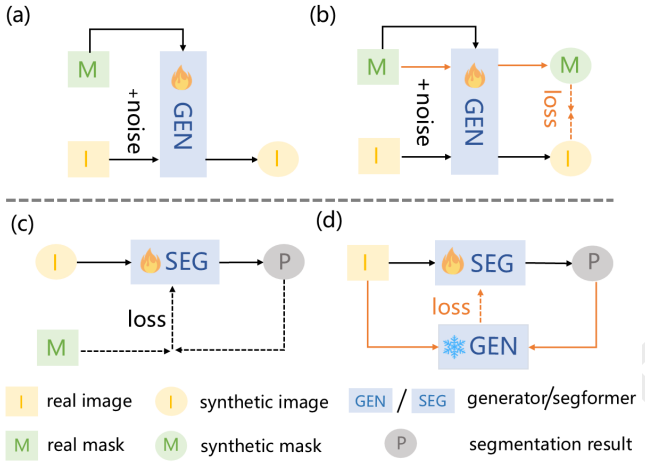

Figure 1: Comparison between previous methods and ours. (a) Previous mask-guided image generation suffers from label drift issues where generated images are not well aligned with the mask. (b) Our dual-branch training strategy introduces an Alignment Regularization by incorporating the mask into the noise-adding and denoising process. It effectively improves the image-mask corresponding of the generative model and alleviates the label drift problem. (c) Previous methods only adopt conventional segmentation training on real and synthetic samples and fail to fully exploit the image-mask corresponding within the well-trained diffusion model. (d) We propose to leverage the diffusion model to further improve the segmentation model by directly evaluating its predictions with the generation objective, which we term as the Generative Feedback.

*Corresponding author (mjsun@suda.edu.cn).

introduce an innovative step where the binary mask is also been generated during the training of the shared generative model (Figure. 1(b)). This dual-branch framework allows us to assess the discrepancy between the noise predicted for the image and that for the mask by the shared generative model. This discrepancy is termed as the Alignment Regularization loss to improve the generative model's image-mask alignment during training. It is worth noting that only the image generation branch is employed during inference to generate samples to train downstream segmentation models.

We attribute such improvement to the fact that denoising both binary mask and image creates an explicit link between their features, mitigating their distribution gap and establishing a robust internal representation of their relationship at the pixel level. Consequently, the contour characteristics of anomalies can be more rapidly learned through the mask generation branch. These contour characteristics are then transferred to the image generation branch via the Alignment Regularization loss during training, ultimately enhancing the image-mask corresponding of the generative model.

Then, we further delve deeper into the "free lunch" of the aforementioned image-mask correspondence within the well-trained generative model: when the condition mask closely matches the anomaly area within the input noisy image, the generative model can easily reconstruct the image from its noisy counterpart, resulting in a smaller standard denoise loss; conversely, if there is a mismatch between the condition mask and the anomaly area in the input noisy image, the corresponding denoise loss increases substantially as well.

Building on this observation, we propose to integrate the generative model into the training of the segmentation model, to complement the traditional training strategy (Figure. 1(c)). Here is how the process unfolds for a given image: first, this image is fed into the segmentation model. The segmentation model then predicts a logit mask, which serves as the condition for the generative model. Subsequently, we introduce noise to the image, treating the noisy image as the input for the generative model, as seen in Figure. 1(d). The standard denoising loss of the generative model is harnessed to finetune the segmentation model to improve its performance.

The main contributions of this paper are as follows:

- We introduce a dual-branch training approach that enables the simultaneous generation of anomaly images and their masks. In Addition, we incorporate an Alignment Regularization loss between the generated images and masks to enhance their correspondence. Only the image-generation branch is activated when synthesizing samples, with the image-mask drift issue alleviated.
- We propose to integrate the well-trained generative model into the training of downstream segmentation models via Generative Feedback loss. This loss, derived from the original denoising loss of the generative model, can serve as an indicator of the segmentation model's accuracy in identifying anomaly areas for its finetuning.
- Experiments show our method outperforms previous approaches in IoU metrics by 5.03% on the Real-IAD dataset (industrial), 5.68% on the polyp dataset (medical), and 16.63% on the Floor Dirty dataset (indoor).

2 Related Work

2.1 Anomaly Image Generation

Since anomaly data is extremely scarce in the real world, researchers want to expand the anomaly dataset in various ways. Previous methods [Zavrtanik *et al.*, 2021; Lin *et al.*, 2021] transferred the existing anomalies or textures to normal images, but they lacked realism and diversity. Then researchers used the GAN-based [Goodfellow *et al.*, 2020] model SDGAN [Niu *et al.*, 2020] and Defect-GAN [Zhang *et al.*, 2021] to generate anomalies on normal samples, but the shortage of anomaly samples limits training. Diffusion models [Ho *et al.*, 2020; Yao *et al.*, 2023; Wu *et al.*, 2025; Qiu *et al.*, 2025] have gradually gained widespread application due to their high-quality generation performance. Recently, AnomalyDiffusion [Hu *et al.*, 2024] has made progress in few-shot anomaly generation, while requiring separate models to generate different types of anomalies.

2.2 Anomaly Image Segmentation

Anomaly Image Segmentation is a crucial computer vision task of localization and segmentation [Sun *et al.*, 2020b; Sun *et al.*, 2021b; Sun *et al.*, 2021a; Sun *et al.*, 2020a; Sun *et al.*, 2024] that aims to identify and segment the abnormal regions within an image. Previous researchers often adopt reconstruction-based methods [Schlegl *et al.*, 2019; Cao *et al.*, 2023], which discover defects by analyzing the residual before and after reconstructing the images. Embedding-based methods [Lee *et al.*, 2022; Cao *et al.*, 2022] can be used, where pre-trained encoder networks are used to extract features, followed by using clustering methods to detect defects.

3 Method

3.1 Overview

The core of this work is the image-mask Alignment Regularization strategy to address the drift issue in the Stable Diffusion (SD) [Rombach *et al.*, 2022] model. The well-learned alignment is then used to assist the segmentation training by high-quality training data augmentation and a novel Generative Feedback mechanism.

Figure 2 shows the framework of our method. For generative training, we finetune an SD model for mask-conditioned image generation as shown in Figure 2 (a). Here, our proposed Alignment Regularization is applied to enforce the model to learn precise image-mask alignment at the pixel level. For segmentation training, we first use the well-tuned SD model to generate highly aligned training samples to augment the training data for segmentation models, as shown in Figure 2 (b) and (c). Then, we transfer the image-mask correspondence from SD model into a Generative Feedback supervision. This feedback mechanism, based on the SD model's noise prediction objective, guides the segmentation model to produce masks that accurately correspond to the image.

3.2 Generative Model Training

Standard Finetuning for Conditional Image Generation

Our primary objective is to generate high-quality image-mask pairs to augment the training data for our segmentation model. To achieve this, we first finetune an SD model

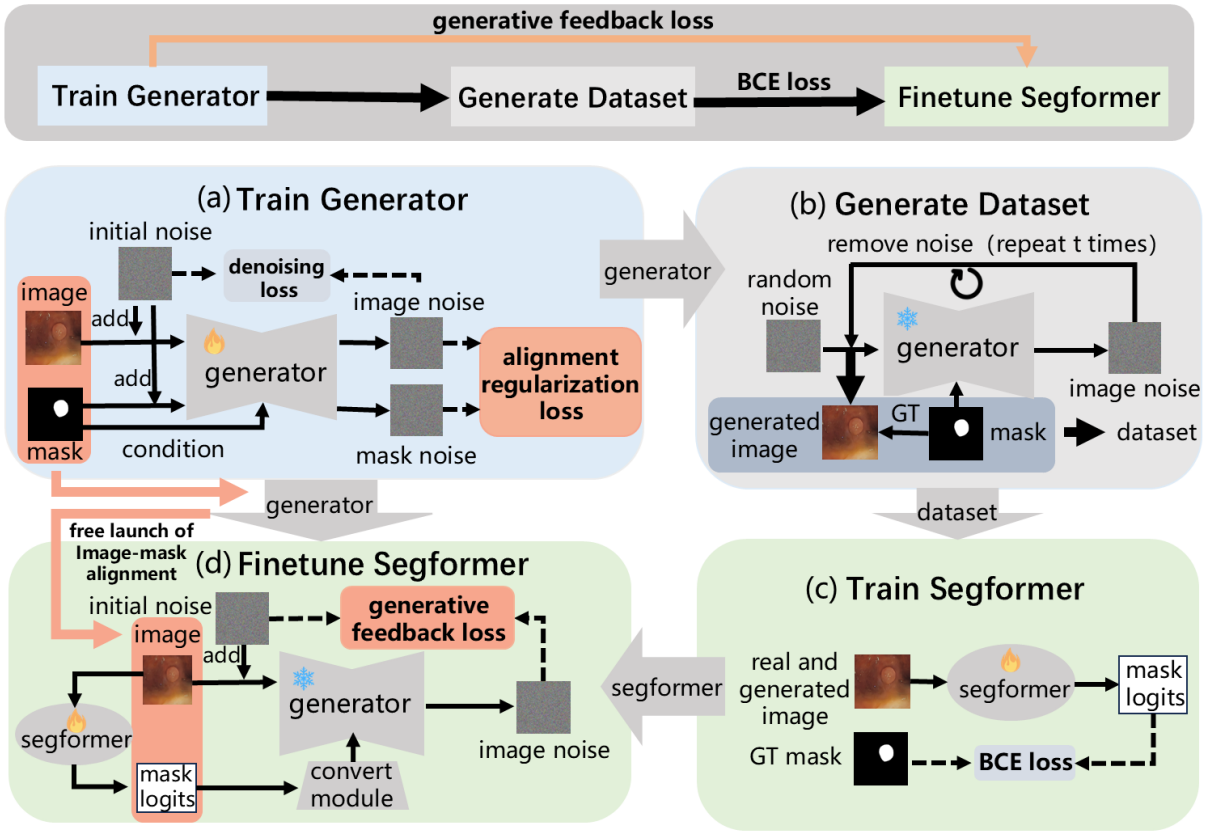

Figure 2: The overview of our framework. (a) We propose a dual-branch training strategy for the diffusion model, incorporating an Alignment Regularization loss to mitigate the drift problem in diffusion. This enables the generative model to robustly learn precise image-mask alignment. (b) With the well-learned alignment, the diffusion model is then used to generate a series of synthetic images based on ground truth masks, which form our generative dataset. (c) We train the segmentation model using both the real dataset and the generated dataset, enhancing its ability to generalize across diverse data. (d) The segmentation model is further improved by the diffusion model. By leveraging the image-mask alignment learned by the diffusion model, we introduce a Generative Feedback mechanism to iteratively refine the segmentation outputs with the generative objective of the diffusion model.

on the training dataset for conditional image generation. The SD model comprises four main components: a Variational Autoencoder (VAE) [Kingma and Welling, 2013], a CLIP Text Encoder [Radford *et al.*, 2021], a Scheduler, and a U-Net [Ronneberger *et al.*, 2015]. The model is conditioned on both a binary mask m, which specifies the background and the foreground anomalous region, as well as a textual description p, which provides additional contextual details for the content to be generated. Ultimately, m is leveraged as the ground truth label for the generated anomalous image.

During training, the process begins with the VAE encoder $\mathcal{E}$ compressing the input image x_0 into a latent feature z_0. Concurrently, the CLIP Text Encoder τ converts the tokenized prompt p into text embedding vectors $\tau(p)$. In the forward diffusion process, the Scheduler progressively adds Gaussian noise ϵ to the latent feature z_0 over t timesteps, resulting in z_t, which approaches pure Gaussian noise as t increases. During the reverse (denoising) process, the U-Net is trained to predict the added noise ϵ based on the noised latent feature z_t, the timestep t, and the text embedding vector $\tau(p)$. The finetuning loss $\mathcal{L}_{sd}$ is calculated as the Mean Squared Error (MSE) between the actual noise ϵ and the predicted noise ϵ_θ, as defined in Equation 1, as follows:

$$\mathcal{L}_{sd} = \mathbb{E}_{\mathcal{E}(x),p,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta(p)\right)\|_2^2\right]. \quad (1)$$

After finetuning, the SD model becomes capable of generating anomalous images from randomly sampled noise, conditioned on a given binary mask. However, a limitation arises due to the model's coarse understanding of the mask, which results in a label drift issue. Specifically, the generated image-mask pairs exhibit misalignment between the binary mask and the corresponding foreground and background regions in the image. The misaligned training pairs can introduce inconsistencies in the training data, potentially producing false supervision signals during subsequent segmentation training.

Enhancing Image-Mask Alignment with Regularization

To address the drift issue observed in conventional finetuning, we propose a novel strategy termed Alignment Regularization to enhance correspondence between generated images and their conditioning masks. Our hypothesis identifies feature misalignment between masks and images as the root

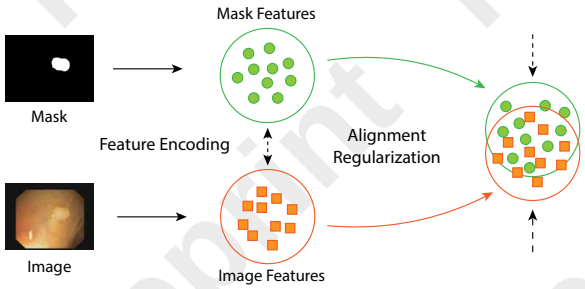

Figure 3: Illustration of the effect of our Alignment Regularization mechanism. By enforcing consistency on the noise predictions on both mask and images, their feature distribution gap can be mitigated, thus enhancing the guidance of the mask.

cause of this drift phenomenon. Due to their inherent visual differences, a distribution gap can emerge between their encoded features, hindering the model's ability to learn effective mask-to-image correlations and ultimately leading to undesired foreground drift in the generated samples.

To overcome this challenge, we introduce an Alignment Regularization mechanism that treats masks not only as conditions but also as denoising targets. Specifically, we apply identical noise to both images and masks, then enforce consistency between their respective noise predictions through our regularization term. As illustrated in Figure 3, this approach enables the model to adapt to both image and mask features simultaneously, thereby bridging their distribution gap. Furthermore, this mechanism establishes explicit links between the representations of the image and mask, strengthening the mask's effectiveness as a conditioning signal.

We introduce a dual-head training strategy that simultaneously processes the image and mask. Specifically, during the diffusion process, the framework employs the VAE encoder $\mathcal{E}$ to compress the mask m into a latent feature h_0. The same noise ϵ, added to the latent image, is applied to the latent feature h_0 over t timesteps to produce noisy latent feature h_t:

$$h_0 = \mathcal{E}(m), \quad h_t = \sqrt{\bar{\alpha}_t} h_0 + \sqrt{1-\bar{\alpha}_t}\epsilon. \tag{2}$$

In the denoising stage of the generative process, the model conditions on both the mask m and the text embedding $\tau(p)$. Using the noise-added feature h_t as input to the U-Net model, the framework predicts the noise ϵ_h for the mask region as:

$$\epsilon_h = \epsilon_\theta\left(h_t, t, \tau_\theta(p), m\right). \tag{3}$$

To enforce structural alignment, the Alignment Regularization term $\mathcal{L}_{al}$ computes the mean squared error (MSE) between the noise ϵ_z predicted by the input image head and the noise ϵ_h predicted by the mask head, as shown below:

$$\mathcal{L}_{al} = \mathbb{E}_{\mathcal{E}(x),\mathcal{E}(m),p,\epsilon\sim\mathcal{N}(0,1),t,m}\left[\|\epsilon_z - \epsilon_h\|_2^2\right]. \tag{4}$$

The overall training objective for generation $\mathcal{L}_{gen}$ combines the conventional finetuning loss $\mathcal{L}_{sd}$ and the proposed structural alignment loss $\mathcal{L}_{al}$, then weighted by a factor α:

$$\mathcal{L}_{gen} = \mathcal{L}_{sd} + \alpha\mathcal{L}_{al}. \tag{5}$$

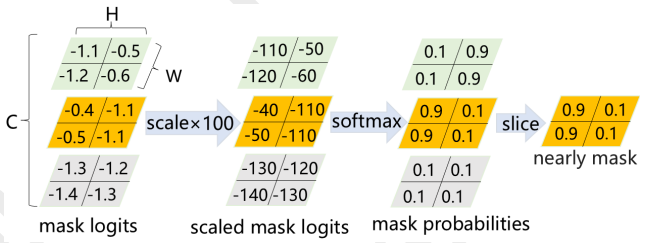

Figure 4: Illustration of the module to convert logits to masks by amplifying values, applying softmax along the channel dimension, and extracting the second channel's value as the approximate mask.

3.3 Segmentation Model Training

Segmentation Training on Augmented Dataset

Building upon our novel finetuning strategy, the SD model is now capable of generating high-quality image-mask pairs. We first leverage this capability to augment our training dataset. Specifically, we generate synthetic training images conditioned on binary masks and combine these with real data pairs. The segmentation model is then trained on this combined dataset. The training process optimizes a cross-entropy loss, $\mathcal{L}_{ce}$, which minimizes the discrepancy between the model-predicted logits and the ground truth masks.

Given real images x_{real} and synthetic images x_{syn}, the segmentation model f produces corresponding logits l_{real} and l_{syn}, respectively. The loss is then computed as:

$$\mathcal{L}_{ce} = \frac{1}{N_{real}}\sum_{i=0}^{N_{real}}\mathcal{H}(l_{real}^i, y^i) + \frac{1}{N_{syn}}\sum_{j=0}^{N_{syn}}\mathcal{H}(l_{syn}^j, y^j), \tag{6}$$

where y represents the binary masks, and $\mathcal{H}$ denotes the cross-entropy calculation. By training on these high-quality and diverse augmented data, we enable the segmentation model to develop more robust representations compared to training solely on real data, yielding better performance.

Improving Segmentation with Generative Feedback

To further leverage the image-mask alignment learned in the SD model, we introduce a novel Generative Feedback loss, $\mathcal{L}_{fb}$, which utilizes the well-trained SD model to evaluate the output of the segmentation model, providing guidance and iterative refinement. During training, the segmentation model f initially predicts logits l for a given input image x. These logits are then processed through a conversion module $\mathcal{T}$ to obtain an approximate mask $\tilde{m}$, which can be defined as:

$$l = f(x), \quad \tilde{m} = \mathcal{T}(l). \tag{7}$$

The logits l have dimensions of $N \times C \times H \times W$, where the channel dimension $C = 1$ corresponds to the prediction scores for the background and foreground. They are then converted by module $\mathcal{T}$ defined as in following equations:

$$\tilde{m} = \text{softmax}(\frac{l}{\lambda}), \tag{8}$$

$$\tilde{m} = \tilde{m}[:, 1, :, :], \tag{9}$$

It is also illustrated in Figure 4. First, the logits are rescaled by a temperature parameter $\lambda \in (0, 1]$, followed by a softmax

Method	Training Data	ClinicDB		ETIS		EndoScene		Kvasir		ColonDB		AVE	
		mDice	**mIoU**	**mDice**	**mIoU**	**mDice**	**mIoU**	**mDice**	**mIoU**	**mDice**	**mIoU**	**mDice**	**mIoU**
PVT	train-real(1450)	**93.7**	**88.9**	78.7	70.6	90.0	83.3	91.7	86.4	80.8	72.7	86.98	80.38
	+ train-real(1450)	93.4	88.7	77.2	69.5	89.5	82.9	**92.1**	**87.1**	80.9	72.8	86.62	80.20
	+ train-ArSDM(1450)	92.2	87.5	80.6	72.9	88.2	81.2	91.5	86.3	**81.7**	**73.8**	86.84	80.34
SAnet	train-real(1450)	91.6	85.9	75.0	65.4	88.8	81.5	90.4	84	75.3	67.0	84.22	76.76
	+ train-real(1450)	92.2	87.2	78.3	69.8	88.6	82.2	90.4	84.8	75.4	68.2	84.98	78.44
	+ train-ArSDM(1450)	91.4	86.1	78.0	69.5	90.2	83.2	91.1	85.6	77.7	70.0	85.68	78.88
Segformer	train-real(100)	84.6	76.8	71.2	64.5	87.3	80.2	88.6	81.0	74.4	65.8	81.20	73.65
	+ train-ArSDM(1450)	88.7	82.0	73.9	66.7	85.7	78.2	90.3	83.7	76.3	67.8	82.96	75.66
	+ train-ours(1450)	93.2	88.3	**81.8**	**74.4**	**92.0**	**86.6**	90.6	84.2	80.4	73.3	**87.59**	**81.34**

Table 1: The comparison with other SOTA methods for the polyp segmentation, using the metrics of mIoU (%) and mDice (%), where higher values mean more accurate segmentation. The result with the highest score is highlighted in **bold** and the second highest is underlined. "+" denotes the combination of "train-real" dataset and another dataset and dataset(n) indicates this dataset consists of n samples.

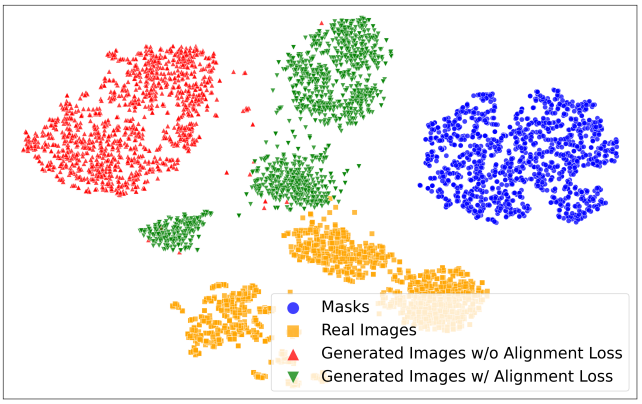

Figure 5: VAE-encoded distribution via t-SNE on polyp dataset.

normalization along the channel dimension. The temperature parameter λ sharpens the softmax output, emphasizing the foreground scores. As shown in Equation 9, we extract the foreground predicted probabilities by slicing the second channel of the predictions to approximate a binary mask, also shown by the yellow probability map in Figure 4.

The approximate mask $\tilde{m}$ is then used as a condition for the SD model, along with a text embedding vector $\tau(p)$. The SD model then processes the noise-added latent image feature z_t to generate predicted noise ϵ_{fb}. Our feedback loss, $\mathcal{L}_{fb}$, is then defined as the Mean Squared Error (MSE) between the real noise ϵ added during the forward diffusion process and the noise ϵ_{fb} predicted by the SD model, as follows:

$$\mathcal{L}_{fb} = \mathbb{E}_{\mathcal{E}(\tilde{m}),p,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_{fb}\left(z_t, t, \tau(p), \tilde{m}\right)\|_2^2\right]. \tag{10}$$

The final segmentation training loss, $\mathcal{L}_{seg}$, is a combination of the cross-entropy loss and the Generative Feedback loss:

$$\mathcal{L}_{seg} = \mathcal{L}_{ce} + \mathcal{L}_{fb}. \tag{11}$$

This dual loss mechanism allows the learned alignment expertise of the generative model to guide and refine the segmentation process. Specifically, if the predicted mask aligns closely with the real mask, the SD model can effectively reconstruct the image, resulting in a predicted noise close to the real noise and, therefore, reducing the $\mathcal{L}_{fb}$. Conversely, significant deviations between the predicted and actual masks will hinder the reconstruction, leading to less accurate noise predictions and an increase in $\mathcal{L}_{fb}$. By integrating $\mathcal{L}_{fb}$ into the overall segmentation loss, the discrepancy between the approximate mask and the actual mask is effectively reduced. By doing so, the segmentation model is encouraged to generate higher-quality mask prediction close to the ground truth.

Method	FID↓	IS ↑	LPIPS ↓
ArSDM	361.55	1.98	0.869
Ours	162.01	3.25	0.698

Table 2: Quantitative comparison about the generated image quality of different generative models on the polyp dataset. Both ArSDM and our method use 100 real image-mask pairs for training.

4 Experiment

4.1 Dataset and Metrics

The evaluation experiments are conducted on multiple scenarios, including medical polyp dataset (ETIS [Silva *et al.*, 2014], CVC-ClinicDB/CVC-612 [Bernal, 2015], CVC-ColonDB [Tajbakhsh *et al.*, 2015], En-doScene [Vázquez *et al.*, 2017], Kvasir [Jha *et al.*, 2020]), industrial dataset

Dataset	Method	Category	mDice ↑	mIoU ↑
MVTec-AD	AnomalyDiffusion	-	77.60	71.84
	Ours	-	**78.62**	**72.91**
Real-IAD	AnomalyDiffusion	easy	70.98	66.35
		hard	67.31	63.91
		average	69.15	65.13
	Ours	easy	**74.68**	**72.39**
		hard	**69.32**	**67.94**
		average	**72.00**	**70.16**
Floor Dirt	AnomalyDiffusion	stains	35.61	35.23
		faeces	34.98	34.51
		average	35.30	34.87
	Ours	stains	**49.52**	**48.94**
		faeces	**57.60**	**54.06**
		average	**53.56**	**51.50**

Table 3: Anomaly segmentation comparison on MVTec-AD, Real-IAD and Floor Dirt dataset. The highest score is in **bold**.

(Real-IAD [Wang *et al.*, 2024], MVTec-AD[Bergmann *et al.*, 2019]), and Floor Dirty dataset. Frechet Inception Distance (FID) [Heusel *et al.*, 2017], Inception Score (IS) [Salimans *et al.*, 2016] and Learned Perceptual Image Patch Similarity (LPIPS) [Zhang *et al.*, 2018] are adopted to evaluate synthetic image quality. Intersection over Union (IoU) and Dice coefficient are utilized to evaluate the accuracy of downstream segmentation models.

4.2 Implementation Details

Using LoRA [Hu *et al.*, 2022] to add image conditions for Stable Diffusion is adopted as the baseline generative model. Segformer [Xie *et al.*, 2021] is adopted as the baseline segmentation model. AdamW is adopted as the optimizer. Input and output images are constrained to 512 × 512. The learning rate is set as 10^{-5}. The batch size is set as 4. For the inference of the diffusion model, the classifier free guidance scale is set as 7. We set the factor α for $\mathcal{L}_{al}$ to 0.7. For each dataset, we first finetune SD via real samples. Then, real masks are utilized to guide SD to generate synthetic samples. Ultimately, real samples are combined with augmented synthetic samples to train downstream segmentation models (see details in complementary materials).

4.3 Comparisons with SOTA Methods

Synthetic Data Quality Evaluation

In order to explore the matching degree between the generated image and the mask, we carry out relevant experiments. Figure 5 displays the data distribution of real masks, real images, generated images without alignment loss, and generated images with alignment loss after encoding by the VAE. The figure shows that without the alignment loss, there is a significant distribution gap between the generated images and both the real masks and real images. However, after adding the alignment loss, the distribution gap between the generated images and the real masks is noticeably reduced, and the gap with the real images also becomes relatively smaller.

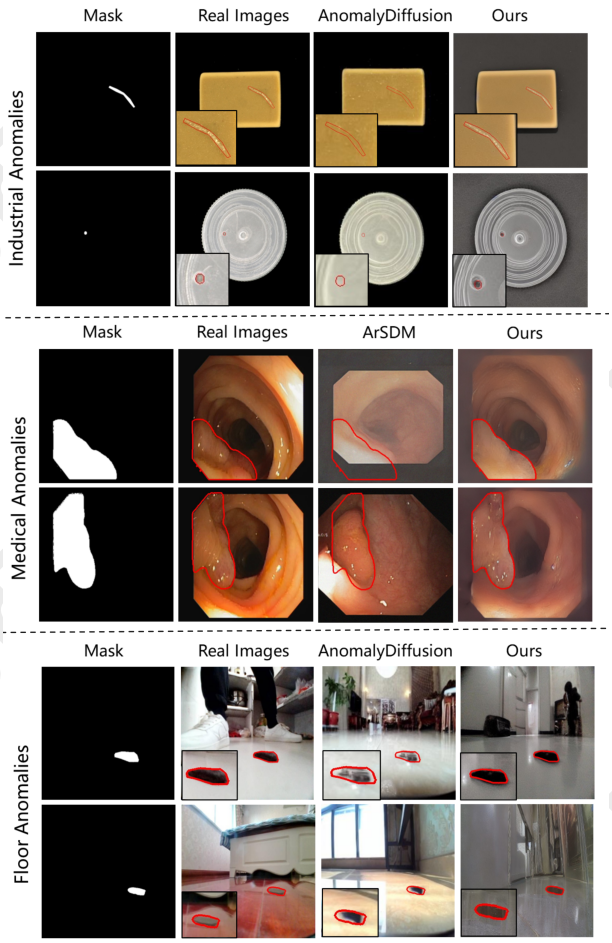

Figure 6: Qualitative comparison of different image generative methods. The red region marks the boundary position of GT mask.

For qualitative synthetic data assessment, we compare images generated by our approach to those from AnomalyDiffusion [Hu *et al.*, 2024] and ArSDM [Du *et al.*, 2023]. Figure 6 shows the results. In medical anomaly datasets with larger Ground-truth (GT) masks, our model's images closely match the GT, while ArSDM's images show misalignment and irregular shapes. For industrial and floor anomaly datasets with smaller GT masks, our model maintains high consistency between generated and real masks, unlike AnomalyDiffusion's images, which have significant shape inconsistencies.

In terms of the quantitative synthetic data evaluation, we compare our method to previous models using FID, IS, and LPIPS metrics on the medical polyp dataset. Results in Table 2 show our method yields lower FID and LPIPS scores, indicating greater similarity to real images, and a higher IS score, suggesting high-quality and diverse image generation.

Segmentation Accuracy Evaluation

As shown in Table 1, in the medical anomalies scenario, we primarily compare segmentation performance with data gen-

No.	$\mathcal{L}_{al}$	$\mathcal{L}_{fb}$	Segmentation Metrics	
			mDice ↑	mIoU↑
1			81.21	73.65
2	✓		87.11	80.46
3		✓	87.28	80.51
4	✓	✓	**87.59**	**81.34**

Table 4: Ablation studies about the proposed loss functions. The segmentation metrics are the average results obtained from five polyp datasets. The result with the highest score is marked in **bold**.

No.	Sample Numbers	Segmentation Metrics	
		mDice ↑	mIoU↑
1	20	58.74	47.27
2	50	85.64	78.80
3	100	87.59	81.34
4	150	**88.80**	**82.22**

Table 5: Ablation studies about different numbers of samples used for training the generative model on the polyp dataset. The segmentation metrics are the average results obtained from five polyp datasets. The result with the highest score is marked in **bold**.

erated by ArSDM on the polyp dataset. We train our generative model on a smaller dataset with only 100 image-mask pairs to simulate the scarcity of labelled data in medical scenarios. Compared to using 100 samples to train ArSDM, our method yields much better segmentation performance when training segformer with generated images. ArSDM struggles with limited samples, and even with 1,450 samples, training PVT [Dong *et al.*, 2021] or SANet [Wei *et al.*, 2021] doesn't match our results. Our method excels in datasets like ETIS and EndoScene, which greatly differ from the training set, by better capturing image-mask relationships. ArSDM, however, only generates images similar to the training set and fails to adapt to significant differences. Thus, our improvements are notable on these datasets. On other test sets similar to the training set, where segmentation performance is near saturation, the proposed method shows no significant gains.

Table 3 shows industrial anomaly segmentation performance comparisons on the MvTec-AD and Real-IAD datasets. The Real-IAD dataset, split into easy (10 objects) and hard (20 objects) based on generation difficulty, reveals that segformer trained with our synthetic samples achieves 1.07% and 5.03% higher average mIoU on MvTec-AD and Real-IAD, respectively, than segformer trained on the samples generated by AnomalyDiffusion. Similar accuracy improvement can be observed on the Floor Dirty dataset.

4.4 Ablation Study

Contribution of Main Components

The main components: Alignment Regularization loss $\mathcal{L}_{al}$ and Generative Feedback loss $\mathcal{L}_{fb}$, play a crucial role in enhancing the performance and stability of the model. Table 4 shows the individual and combined impact of these loss functions. Using $\mathcal{L}_{al}$ and $\mathcal{L}_{fb}$ separately brings gains of 5.9% and 6.07% on mDice and 6.81% and 6.86% on mIoU over baseline, respectively. Combining these two loss functions raises the performance to a new level, leading to improvements of 6.38% on mDice and 7.69% on mIoU. This indicates the effectiveness of the two main loss functions. It is worth noting that the Alignment Regularization impacts the segmentation model indirectly by enhancing its training data, while Generative Feedback mechanism is directly involved in its optimization process. Importantly, these two loss functions have a synergistic effect in improving the segmentation model.

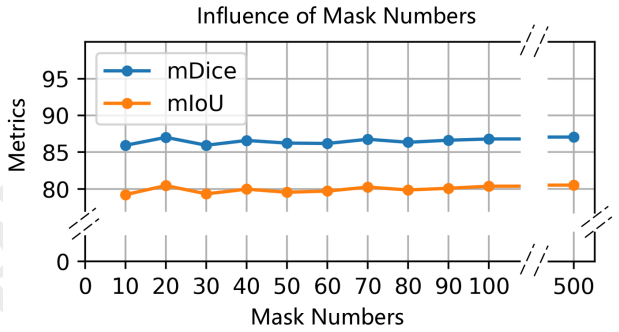

Figure 7: Ablation studies about different numbers of real masks used for the inference of the generative model. The segmentation metrics are the average results obtained from five polyp datasets.

Sample Size on Generative Model Training

Table 5 shows the impact of sample size on finetuning the generative model. As samples increase from 50 to 150, the segmentation model's mIoU improves from 78.80% to 82.22%. With only 20 samples, mIoU drops to 47.27%, likely due to insufficient tuning of the diffusion model, causing misalignments in synthetic data and misleading segmentation model optimization. This highlights the importance of adequate training samples for diffusion model finetuning.

Real Mask Numbers for Generative Model Inference

This study examines how the number of real masks guiding synthetic sample generation affects downstream segmentation performance. Figure 7 shows a non-linear relationship between real mask number and segmentation performance. mIoU fluctuates minimally from 10 to 60 masks, peaking at 80.47% with 20 masks and bottoming at 79.24% with 10 masks. From 60 to 100 masks, mIoU slightly increases to 80.38%. The high performance with few real masks indicates our method's robustness and low sensitivity to mask quality.

5 Conclusion

Our method introduces Alignment Regularization loss during generative model training to improve diffusion's understanding of image-mask relationships and output quality. It also integrates Generative Feedback loss into segmentation model training to optimize performance and accuracy. Extensive experiments validate the proposed approach's effectiveness, supporting their application and value in downstream tasks.

Acknowledgments

This work was supported by Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62302328), Jiangsu Province Foundation for Young Scientists (Grant No. BK20230482), Suzhou Key Laboratory Open Project (Grant No. 25SZZD07) and Jiangsu Manufacturing Strong Province Construction Special Fund Project (Grant Name: Research and Development and Industrialization of Intelligent Service Robots Integrating Large Model and Multimodal Technology).

Contribution Statement

For the co-authors of this paper, Xiangyue Li and Xiaoyang Wang contributed equally to this work.

References

[Bergmann *et al.*, 2019] 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*, 2019.

[Bernal, 2015] Sánchez F. J. Fernández-Esparrach G. Gil D. Rodríguez C. Vilariño F Bernal, J. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

[Cao *et al.*, 2022] Yunkang Cao, Qian Wan, Weiming Shen, and Liang Gao. Informative knowledge distillation for image anomaly segmentation. *Knowledge-Based Systems*, 248:108846, 2022.

[Cao *et al.*, 2023] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023.

[Dong *et al.*, 2021] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021.

[Du *et al.*, 2023] Yuhao Du, Yuncheng Jiang, Shuangyi Tan, Xusheng Wu, Qi Dou, Zhen Li, Guanbin Li, and Xiang Wan. Arsdm: colonoscopy images synthesis with adaptive refinement semantic diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023.

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

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[Hu *et al.*, 2022] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[Hu *et al.*, 2024] Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[Jha *et al.*, 2020] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, 2020.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Lee *et al.*, 2022] Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10:78446–78454, 2022.

[Lin *et al.*, 2021] Dongyun Lin, Yanpeng Cao, Wenbin Zhu, and Yiqun Li. Few-shot defect segmentation leveraging abundant defect-free training samples through normal background regularization and crop-and-paste operation. In *2021 IEEE International Conference on Multimedia and Expo*, 2021.

[Niu *et al.*, 2020] 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.

[Qiu *et al.*, 2025] Kunpeng Qiu, Zhiqiang Gao, Zhiying Zhou, Mingjie Sun, and Yongxin Guo. Noise-consistent siamese-diffusion for medical image synthesis and segmentation. *arXiv preprint arXiv:2505.06068*, 2025.

[Radford *et al.*, 2021] 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*, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015.

[Salimans *et al.*, 2016] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2016.

[Schlegl *et al.*, 2019] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, 2019.

[Silva *et al.*, 2014] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery*, 9:283–293, 2014.

[Sun *et al.*, 2020a] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Yanchun Xie, and Jiashi Feng. Adaptive roi generation for video object segmentation using reinforcement learning. *Pattern Recognition*, 106:107465, 2020.

[Sun *et al.*, 2020b] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Bingfeng Zhang, and Yao Zhao. Fast template matching and update for video object tracking and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[Sun *et al.*, 2021a] Mingjie Sun, Jimin Xiao, and Eng Gee Lim. Iterative shrinking for referring expression grounding using deep reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, 2021.

[Sun *et al.*, 2021b] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Y. Goulermas. Discriminative triad matching and reconstruction for weakly referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4189–4195, 2021.

[Sun *et al.*, 2024] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Cairong Zhao, and Yao Zhao. Unified multi-modality video object segmentation using reinforcement learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):6722–6734, 2024.

[Tajbakhsh *et al.*, 2015] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging*, 35(2):630–644, 2015.

[Vázquez *et al.*, 2017] David Vázquez, Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Antonio M. López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, 2017(1):4037190, 2017.

[Wang *et al.*, 2024] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[Wei *et al.*, 2021] Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S Kevin Zhou, and Shuguang Cui. Shallow attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention*, 2021.

[Wu *et al.*, 2025] Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, 2021.

[Yao *et al.*, 2023] Siyue Yao, Mingjie Sun, Bingliang Li, Fengyu Yang, Junle Wang, and Ruimao Zhang. Dance with you: The diversity controllable dancer generation via diffusion models. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

[Zavrtanik *et al.*, 2021] 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 and Pattern Recognition*, 2021.

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

[Zhang *et al.*, 2021] Gongjie Zhang, Kaiwen Cui, Tzu-Yi Hung, and Shijian Lu. Defect-gan: High-fidelity defect synthesis for automated defect inspection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision and Pattern Recognition*, 2021.