Journal of Multimedia Information System VOL. 12, NO. 3, September 2025 (pp. 87-94): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2025.12.3.87

VLM-Guided Inpainting for Anomaly Detection

Jungyeon Seo[1], Kibeom Hong[1*]

Abstract

Anomaly detection (AD) aims to identify regions in an image that deviate from the expected distribution of normal visual data, a task critical for applications such as industrial inspection. Recent CLIP-based approaches have enabled zero-shot anomaly detection by comparing image features with text-derived embeddings, leveraging pretrained vision-language alignment. While effective in general scenarios, these methods struggle to capture domain-specific normality and often fail to accurately localize subtle anomalies. We introduce a novel framework that integrates CLIP-guided mask inference with a diffusion-based generative inpainting module trained on normal data. To improve semantic consistency and reconstruction fidelity, we incorporate score distillation sampling (SDS) loss, which aligns the inpainted output with the distribution of normal images in the embedding space. Our method is model-agnostic and can be integrated into existing CLIP-based detectors without requiring anomaly annotations. Experiments on datasets from industrial and medical domains demonstrate consistent improvements when integrated with various backbones in both image-level and pixel-level detection tasks. Qualitative results show improved reconstruction and precise localization of fine-grained anomalies.

Key Words: Anomaly Detection, Vision-Language Models, Diffusion Models, Score Distillation Sampling.

I. INTRODUCTION

Anomaly Detection is the task of identifying anomalies in the form of defects in the objects or textures. In computer vision, this typically involves detecting anomalous regions that deviate from the expected distribution of natural image data [1]. Although several works [2-3] have employed deep neural networks for anomaly detection, they often suffer from limited representation capability and poor generalization to diverse or complex datasets.

Recent advancements in vision-language models (VLMs), such as CLIP, have significantly enhanced anomaly detection tasks by leveraging powerful semantic representations learned from large-scale vision-language data. Consequently, several CLIP-based approaches [4-6] typically identify anomalies based on discrepancies between visual features extracted from input data and normal reference embeddings derived from pretrained VLMs by leveraging their zero-shot capability.

While VLM-based methods have evolved to include patch-level attention and finer-grained analysis [7], they fundamentally rely on pretrained CLIP encoders that are not optimized for capturing the specific data distribution of a given anomaly detection dataset. This reliance limits their ability to model nuanced variations in normality specific to industrial or domain-specific datasets, ultimately constraining anomaly localization performance.

To address these shortcomings, in this paper, we have employed the strengths of generative models, which inherently excel at reconstructing and generating detailed visual content. Crucially, these generative models [8-9] trained on normal data distributions naturally excel at inpainting tasks effectively reconstructing missing or masked regions of an image. Since anomalies inherently represent deviations from the normal data distribution, reconstruction errors by generative models can directly indicate anomalous regions.

To this end, we propose a novel anomaly detection framework that synergistically combines the semantic generalization capabilities of VLMs with the detailed reconstruction strengths of generative models. Specifically, we integrate an explicit semantic-guided inpainting mechanism within a CLIP-based anomaly detection pipeline, enhanced by using a score distillation sampling (SDS) loss [10]. Our key hypothesis is that accurate reconstruction of missing regions guided by VLM semantics, reinforced through SDS loss, inherently indicates precise anomaly localization capabilities.

Our main contributions are as follows:

Manuscript received June 03, 2025; Revised June 16, 2025; Accepted June 19, 2025. (ID No. JMIS-25M-06-016)
Corresponding Author (*): Kibeom Hong, +82-10-8979-2065, kb.hong@sookmyung.ac.kr
[1]Department of Computer Science, Sookmyung Women's University, Yongsan, Seoul, Korea, seocindy2002@sookmyung.ac.kr, kb.hong@sookmyung.ac.kr

1) We propose a new anomaly detection paradigm that explicitly connects anomaly localization with generative inpainting capabilities guided by semantic embeddings from VLMs.

2) We develop a unified framework integrating vision-language semantics and generative reconstruction, significantly enhancing the precision and granularity of anomaly localization.

3) Extensive experiments demonstrate our approach's superior anomaly detection and localization performance compared to existing state-of-the-art VLM-based methods across multiple benchmark datasets, particularly emphasizing performance improvements on industrially relevant datasets such as MVTec-AD.

II. RELATED WORKS

2.1. CLIP-Based Anomaly Detection

Recently, vision-language models (VLMs), particularly CLIP, have been explored for anomaly detection for their powerful multimodal representation capabilities and ability to generalize to unseen concepts without requiring task-specific full-supervision.

WinCLIP [4] pioneers a novel approach to zero- and few-shot anomaly detection as a retrieval task in the CLIP embedding space by comparing patch-level visual features with a compositional ensemble of text prompts that represent normal concepts. It incorporates a diverse prompt ensemble and multi-scale spatial feature aggregation to align visual and textual representations, demonstrating strong performance in language-guided anomaly classification and localization.

AnomalyCLIP [5] further advances zero-shot anomaly detection by addressing the limitations of existing prompt engineering-based approaches. Unlike prior methods which rely on extensive manual construction of object-specific prompt ensembles, AnomalyCLIP introduces an object-agnostic prompt learning framework that captures the semantics of generic normality and abnormality loss, which combines global and local contextual information, to learn two universal text prompts representing normal and abnormal states from auxiliary data. This formulation allows for simplified and transferable prompt design, making the model highly adaptable across domains without requiring prompt modification.

More recently, AA-CLIP [6] tackles CLIP's limited anomaly awareness by enhancing its ability to distinguish between normal and abnormal features in both visual and textual embedding spaces. It employs a two-stage strategy: first generating anomaly-aware text anchors, then aligning patch-level visual features with them for precise localization. Using residual adapters for lightweight fine-tuning, AA-CLIP effectively learns anomaly-specific representations while maintaining CLIP's generalization, achieving strong zero-shot performance even in data-scarce settings.

2.2. Generative Anomaly Detection

Generative methods mainly focus on modeling the distribution of normal data and detect anomalies based on reconstruction-based methods. They mainly use generative models such as variational autoencoders (VAE) [11], generative adversarial networks (GANs) [12], and diffusion [13] models to attempt to reconstruct normal images, where anomalies are assumed to produce large reconstruction errors that manifest as noticeable deviations between the input and the generated output.

f-AnoGAN [14] improves the efficiency of GAN-based anomaly detection by introducing an encoder that directly maps input images to the latent space, removing the need for iterative optimization. Anomalies are scored using both pixel-level reconstruction error and feature residuals from the discriminator. Trained only on normal data, f-AnoGAN achieves strong performance, especially in medical imaging.

AnomalySD [9] adapts Stable Diffusion [15] for few-shot multi-class anomaly detection by introducing hierarchical text prompts and a foreground mask mechanism. During inference, it employs multi-scale and prototype-guided masking strategies to localize and inpaint anomalous regions as normal, with anomaly scores computed from the inpainting results.

III. PROPOSED METHODS

3.1. Preliminary

3.1.1. CLIP-Based Anomaly Detection

Contrastive language-image pretraining (CLIP) models have demonstrated strong generalization capabilities across a wide range of vision-language tasks by learning a joint embedding space for images and natural language descriptions. CLIP is trained using a contrastive loss [16], where image and text pairs are pulled together in the embedding space while mismatched pairs are pushed apart. Specifically, given a batch of image-text pairs, the model maximizes the cosine similarity [17] of correct pairs while minimizing it for incorrect ones across all combinations, effectively aligning semantically related image and text representations. The loss function is defined as a symmetric cross-entropy over cosine similarities (equation (1)):

$$L_{CLIP} = \frac{1}{2N}\sum_{i=1}^{N}\left(log\frac{\exp(\text{sim}(x_i,y_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(x_i,y_j)/\tau)} + log\frac{\exp(\text{sim}(y_i,x_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(y_i,x_j)/\tau)}\right), \quad (1)$$

where x_i and y_i denote the image and text embeddings of the i-th matched pair in a batch of size N, $\mathrm{sim}(\cdot,\cdot)$ is the cosine similarity, and τ is a temperature scaling parameter. This symmetric formulation encourages both image-to-text and text-to-image alignment.

In the context of anomaly detection, CLIP enables image regions to be semantically aligned with predefined textual concepts such as "normal" or "abnormal." Recent methods utilize this capability by computing the similarity between visual features and prompt-engineered textual embeddings to perform zero-shot anomaly classification or segmentation. This formulation allows for detection of semantically meaningful anomalies even without access to anomaly samples during training.

3.1.2. Diffusion-Based Inpainting

Diffusion models have emerged as powerful generative models capable of producing high-fidelity images through a denoising process that iteratively transforms noise into structured content. In anomaly detection, these models leveraged for image inpainting, where anomalous regions are masked and reconstructed if they were normal. The difference between the original and reconstructed images is then used to infer the presence and severity of anomalies. This reconstruction-based paradigm, particularly when guided by semantic conditions, enables precise localization and detection of subtle anomalies.

3.2. Proposed Methods

We propose a novel reconstructive anomaly detection framework that integrates vision-language models and diffusion-based inpainting, as shown in Fig. 1. The objective is to precisely localize anomalous regions by leveraging the complementary strengths of CLIP for semantic conditioning and a text-guided generative model for high-fidelity reconstruction. Our framework operates in a zero-shot setting without requiring any task-specific anomaly training data.

The core intuition behind the proposed method is that anomalies, by definition, deviate from the learned distribution of normality. A generative model trained on normal data will therefore struggle to reconstruct anomalous regions, leading to high reconstruction error. By integrating vision-language alignment through CLIP and using text-conditioned inpainting with diffusion, we can guide this reconstruction process to reflect what is semantically expected.

3.2.1. CLIP-Based Embedding and Mask Inference

Firstly, we employ a pre-trained CLIP-based anomaly detection model to extract both image and text embeddings. Given an input image x and a normality prompst y, CLIP provides the image embedding $f_{img}(x)$ and text embedding $f_{txt}(y)$. These features serve as semantic guidance throughout the pipeline. After that, an irregular binary mask M is generated to simulate occluded or uncertain regions within the input image. This mask is applied elementwise to the input image to produce a masked image $x_M = M \odot x$, where $\odot$ denotes element-wise multiplication.

3.2.2. Reconstructive Anomaly Detection

The masked image x_M and the text prompt y are then passed to a diffusion-based generative model, which aims to reconstruct the full image $\hat{x}$ by inpainting the masked

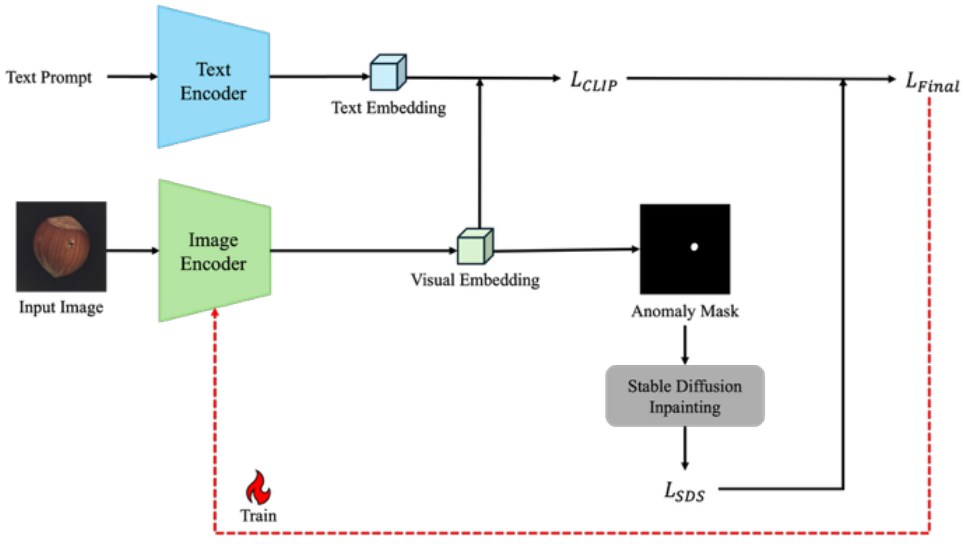

Fig. 1. Representative examples from the CQAD dataset. It generates an anomaly mask from CLIP visual embeddings and feeds it into a Stable Diffusion based inpainting model. Then the gradients derived from the resulting SDS loss is combined with the CLIP loss to form the final training objective. By jointly optimizing the image adapter using this loss, the adapter is optimized to improve spatial localization of anomalies.

regions while preserving semantic consistency with the provided prompt. To ensure that the model can effectively represent the distribution of normal images within the target domain, we fine-tune the generative model using only clean normal data. Specifically, we fine-tune only the decoder layers of the UNet to enhance its ability to represent the distribution of normal images while preserving the general generative capacity of the pretrained backbone. This selective tuning enables the model to reconstruct high-frequency details and textures characteristic of normal samples, improving its ability to detect anomalies as deviations from this learned normality.

To validate the effectiveness of this selective fine-tuning strategy, we compare it with full fine-tuning of the entire UNet architecture. As shown in Table 1, the partial tuning configuration achieves lower FID scores across classes-for instance, reducing the score from 22.7 to 20.8 in the Screw class-indicating better preservation of the normal image distribution and sharper reconstruction quality. This result confirms that partial tuning strikes a better balance between domain adaptation and generative generalization, which is critical for precise anomaly localization.

3.2.3. Optimization via Score Distillation Sampling (SDS)

To further refine the quality and semantic consistency of the reconstructed image, we apply score distillation sampling (SDS) as the core optimization objective. SDS enables CLIP-based mask inference to be guided by the generation process so that the output image better reflects the distribution of normal data. This facilitates more accurate detection by the CLIP anomaly detector, as the generated image aligns with the learned normality in both appearance and semantics. In our framework, SDS optimizes the generative model to align its output with the embedding of the original input image rather than a text prompt. This allows the model to directly capture the normality distribution represented in real image space.

Table 1. Class-wise FID score comparison on the MVTec-AD dataset between over architecture (full UNet tuning) and partial architecture (decoder-only tuning).

Class	FID score	
	Over architecture	Partial architecture
Bottle	112.9	**80.5**
Cable	62.0	64.9
Capsule	41.0	**29.4**
Carpet	46.5	**36.4**
Grid	63.4	**60.8**
Hazelnut	46.6	**36.9**
Leather	99.8	**80.1**
Metal nut	91.5	**86.3**
Pill	59.9	**50.7**
Screw	22.7	**20.8**
Tile	84.5	**54.1**
Transistor	63.8	80.2
Toothbrush	75.7	**60.0**
Wood	102.1	**80.4**
Zipper	110.7	**56.5**

The SDS loss is computed by minimizing the distance between the inpainted results within CLIP-detected regions and the distribution of real normal images (equation (2)):

$$\nabla_\theta L_{SDS}(\phi, x = g(\theta)) \triangleq E\{t,\varepsilon\}\left[w(t)\left(\hat{\varepsilon}_\phi(z_t;\, y, t) - \varepsilon\right) \partial x/\partial\theta \right], \tag{2}$$

where $\hat{\varepsilon}_\phi$ is the predicted noise at timestep t with the masked noisy latent z_t. In the context of our work, we integrate the SDS loss into CLIP-based anomaly detection frameworks to enhance anomaly localization capabilities. By incorporating semantic guidance from vision-language models and leveraging the reconstruction strengths of generative models, our approach aims to achieve precise and semantically consistent anomaly detection. Consequently, our final objective is to be formulated as follow (equation (3)):

$$L_{final} = L_{CLIP} + \lambda \cdot L_{SDS}, \tag{3}$$

where λ controls the influence of the SDS guidance during optimization. Unlike conventional loss functions, the SDS loss is defined through the gradient of a guidance objective with respect to the input image. Specifically, it measures how the diffusion model's output, guided by a target text prompt, changes the synthesized image, and uses this as a directional signal to update the anomaly detection model. Although this involves a gradient operator within the loss definition, this is consistent with the optimization formulation proposed in [10].

While our method introduces additional components such as SDS loss and a fine-tuned diffusion model during training, it maintains the same computational complexity as the CLIP-based baseline during inference. Therefore, our method does not incur any inference-time trade-off, which ensures practicality for deployment. To quantify training overhead, we measured peak GPU memory usage. AA-CLIP showed a 12.9% increase (from 19,262MB to 21,755MB), while AnomalyCLIP exhibited a larger rise of 389.8% (from 3,585MB to 17,558MB). Notably, this memory overhead is strictly confined to the training phase and is considered acceptable given the substantial performance improvements achieved by our method.

IV. EXPERIMENTS

4.1. Experiment Setups

To evaluate the effectiveness of our proposed anomaly detection framework, we conducted experiments on three widely recognized benchmarks covering both industrial and medical domains. For the industrial domain, we utilized the MVTec-AD [1] and VisA [18] datasets. MVTec-AD contains over 5,000 high-resolution images across 15 categories with pixel-precise annotations for all anomalous regions. VisA comprises images from 12 industrial object categories and provides pixel-level annotations as well. For the medical domain, we employed the CVC-ClinicDB [19] dataset, a publicly available colonoscopy dataset consisting of 612 frames extracted from 31 video sequences. Each frame is annotated with expert-labeled polyp masks, allowing for pixel-level anomaly segmentation evaluation in a clinical setting.

For quantitative evaluation, we employed four standard metrics: Image-level AUROC (Image-AUROC) and image-level average precision (Image-AP) assess the model's ability to distinguish between normal and anomalous images, evaluating the precision-recall trade-off at the image level. Pixel-level AUROC (Pixel-AUROC) and Pixel-level Average Precision (Pixel-AP) measure the capability of the model to localize anomalies at the pixel level, providing a comprehensive assessment of both detection and localization performance.

We selected AA-CLIP and Anomaly CLIP as our primary baselines because they represent the current state-of-the-art in CLIP-based anomaly detection. Both methods build directly on the CLIP encoder without introducing fundamentally different architectural components, making them structurally comparable to our approach. This shared foundation enables a fair evaluation of our method's effectiveness. By integrating with these strong baselines, we aim to demonstrate performance improvements within the same architectural paradigm.

4.2. Quantitative Results

As shown in Table 2 and Table 3, we report quantitative comparisons with previous studies [5-6] on both industrial domains (MVTec-AD and VisA datasets) and medical domains (CVC-ClinicDB dataset). Our approach demonstrates consistently strong performance compared to prior CLIP-based methods across four evaluation metrics. On the various datasets, our method achieves the highest image-level scores and competitive pixel-level results in most cases, demonstrating effective detection and localization capabilities.

Notably, our method can be applied on top of existing CLIP-based anomaly detection pipelines, yielding consistent improvements regardless of the underlying backbone. For example, on the MVTec-AD dataset, we observe a notable increase in Image-AUROC for the AA-CLIP baseline, increasing from 99.2 to 99.5. This result highlights the benefit of SDS-guided inpainting in modeling fine-grained normality. Although our proposed method based on AA-CLIP exhibits slightly lower pixel-level performance than the baseline in a few cases on MVTec-AD and VisA,

Table 2. Quantitative comparisons of AA-CLIP and our framework on industrial domain datasets (MVTec-AD and VisA) and medical domain datasets (Clinic DB).

Method	MVTec-AD			
	Image-level		Pixel-level	
	AUROC	AP	AUROC	AP
AA-CLIP [6]	99.2	99.6	**99.6**	86.3
AA-CLIP+ our framework	**99.5**	**99.7**	99.5	**86.5**
	VisA			
	Image- level		Pixel-level	
	AUROC	AP	AUROC	AP
AA-CLIP [6]	97.0	97.8	**97.6**	**56.0**
AA-CLIP+ our framework	**97.2**	**98.0**	97.3	55.6
	CVC-ClinicDB			
	Pixel-level			
	AUROC		AP	
AA-CLIP [6]	99.72		97.83	
AA-CLIP+ our framework	**99.75**		**98.06**	

Table 3. Quantitative comparisons of AnomalyCLIP and our framework on industrial domain datasets (MVTec-AD and VisA) and medical domain datasets (Clinic DB).

Method	MVTec-AD			
	Image-level		Pixel-level	
	AUROC	AP	AUROC	AP
AnomalyCLIP [5]	91.6	96.4	91.1	81.4
AnomalyCLIP+ our framework	**94.6**	**97.4**	**94.3**	**87.5**
	VisA			
	Image- level		Pixel-level	
	AUROC	AP	AUROC	AP
AnomalyCLIP [5]	81.9	85.4	95.5	**86.8**
AnomalyCLIP+ our framework	**89.1**	**90.3**	**95.8**	78.4
	CVC-ClinicDB			
	Pixel-level			
	AUROC		AP	
AnomalyCLIP [5]	82.9		68.1	
AnomalyCLIP+ our framework	**84.9**		**68.3**	

it consistently improves image-level metrics and performs robustly across domains, including the medical setting.

4.3. Qualitative Results

Fig. 2 illustrates qualitative comparisons of anomaly localization results between our method and baseline approaches on representative samples from MVTec-AD and VisA datasets. For each example, we present the input image, the ground-truth anomaly mask, and the predicted anomaly maps from both the baseline (AA-CLIP) and our proposed framework. Specifically, we include a Screw class sample from the MVTec-AD dataset and a Cashew class sample from the VisA dataset. In both cases, the anomaly maps produced by our method more strongly highlight the anomalous regions compared to the baseline.

These qualitative results illustrate how our method produces more accurate and spatially coherent anomaly maps, effectively highlighting fine-grained defects that are often missed or mislocalized by prior methods. Notably, our framework better captures subtle texture anomalies and small structural defects, which are common in real-world industrial settings. In the Screw sample, the anomaly- a small deformation near the tip- is subtle and localized. While AA-CLIP roughly identifies the region, it lacks spatial precision. Our method produces a sharper response that aligns more closely with the ground-truth, demonstrating improved localization of small structural defects. In the *Cashew* sample, a fine surface scratch is better captured by our method, with a focused anomaly map and reduced false positives. This highlights the effectiveness of diffusion-based reconstruction in preserving texture consistency, guided by CLIP semantics.

The improvements stem from the integration of semantic guidance through CLIP and the precise inpainting ability of diffusion models, reinforced via SDS optimization. This allows the reconstructed image to closely resemble the normal data distribution, thereby amplifying the discrepancy in anomalous regions.

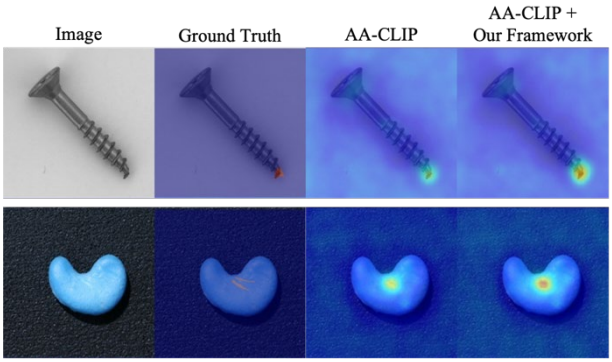

Fig. 2. Visualization of anomaly localization result of AA-CLIP and our model on MVTec-AD and VisA.

Table 4. Ablation study results on the effect of SDS loss weight (λ) on anomaly detection performance of our model on the MVTec-AD dataset.

SDS loss weight (λ)	MVTec-AD			
	Image-level		Pixel-level	
	AUROC	AP	AUROC	AP
0.1	99.39	99.68	99.55	86.58
0.5	99.41	99.70	99.35	86.44
1	99.39	99.68	99.46	86.42
5	**99.51**	**99.71**	**99.55**	**86.51**
10	99.49	99.77	99.37	86.47

4.4. Ablation Study on SDS Weight λ

To investigate the influence of the SDS loss weight λ in our final objective, we conduct an ablation study by varying $\lambda \in \{0.1, 0.5, 1, 5, 10\}$. As shown in Table 4, performance improves steadily as λ increases, peaking at λ=5, after which further increases result in diminishing returns or slight performance drops. This trend indicates that a moderate SDS guidance strength is essential: small weights underutilize the semantic alignment benefits of SDS, while excessively large weights may dominate the optimization, leading to overfitting or instability in generation. The best results across both MVTec-AD and VisA are achieved when λ=5, which we adopt in all subsequent experiments.

V. CONCLUSION

In this paper, we introduce a novel anomaly detection framework that bridges the strengths of vision-language models and generative models. By integrating a semantic-guided inpainting mechanism enhanced with score distillation sampling (SDS) loss, our method achieves fine-grained anomaly localization grounded in both semantic understanding and visual reconstruction. Unlike prior CLIP-based approaches that rely heavily on global embeddings, our method adapts to the underlying data distribution through generative reconstruction, enabling superior precision in localizing subtle anomalies. Extensive experiments conducted on standard benchmarks, including the industrially focused MVTec-AD dataset, demonstrate that our approach not only improves localization accuracy but also enhances overall detection performance when integrated with existing CLIP-based pipelines. Our findings highlight the value of bridging semantic reasoning with generative modeling in advancing the state-of-the-art for vision-language anomaly detection.

ACKNOWLEDGEMENT

This research was supported by Sookmyung Women's University Research Grants (1-2403-2034) and the Seoul Business Agency (SBA) through the 2024 Artificial Intelligence Technology Commercialization Support Program (SY240213).

REFERENCES

[1] P. Bergmann, M. Fauser, D. Sattlegger, and C. 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*, pp. 9592-9600, 2019.

[2] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318-14328, 2022.

[3] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: a patch distribution modeling framework for anomaly detection and localization," in *International Conference on Pattern Recognition*, pp. 475-489, 2021.

[4] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "Winclip: Zero-/few-shot anomaly classification and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19606-19616, 2023.

[5] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection," *arXiv Preprint arXiv:2310.18961*, 2023.

[6] W. Ma, X. Zhang, Q. Yao, F. Tang, C. Wu, and Y. Li, et al., "AA-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip," *arXiv Preprint arXiv:2503.06661*, 2025.

[7] H. Deng, Z. Zhang, J. Bao, and X. Li, "Bootstrap fine-grained vision-language alignment for unified zero-shot anomaly localization," *arXiv Preprint arXiv:2308.15939*, 2023.

[8] J. Pirnay and K. Chai, "Inpainting Transformer for Anomaly Detection," in: S. Sclaroff, C. Distante, M. Leo, G. M. Farinella, F. Tombari, (eds), *Image Analysis and Processing – ICIAP 2022*, *Lecture Notes in Computer Science*, vol. 13232, 2022.

[9] Z. Yan, Q. Fang, W. Lv, and Q. Su, "Anomalysd: Few-shot multiclass anomaly detection with stable diffusion model," *arXiv Preprint arXiv:2408.01960*, 2024.

[10] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

[11] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv Preprint arXiv:1312.6114*, 2013.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley and S. Ozair, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672-2680, 2014.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840-6851, 2020.

[14] T. Schlegl, P. Seebock, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "F-anogan: fast unsupervised anomaly detection with generative adversarial networks," *Medical Image Analysis*, vol. 54, pp. 30-44, 2019.

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684-10695, Jun. 2022.

[16] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *in NeurIPS*, 2018.

[17] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," in *Advances in Neural Information Processing Systems*, vol. 31, pp. 1-12, 2018.

[18] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pretraining for anomaly detection and segmentation," in *European Conference on Computer Vision*, pp. 392-408, 2022.

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

AUTHORS

Jungyeon Seo is currently pursuing her B.S. degree in the Department of Software at Sookmyung Women's University, Korea, from 2021. Her research interests include computer vision, ML/DL, and artificial intelligence.

Kibeom Hong is currently an assistant professor in the Department of Software at Sookmyung Women's University, Korea, since 2024. He received B.S. and the Ph.D. degrees in computer science from Yonsei University, Seoul, Korea in 2023. His research interests include generative models, neural style transfer and domain generalization.