Computing, Performance and Communication Systems (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/cpcs.2025.090114
ISSN 2371-8889 Vol. 9 Num. 1

ACMAN: Adaptive Cross-Modal Anomaly Network

Junwei Wang[1,a,*], Junting Liu[1,b], Yutian Jiao[1,c]

[1]Shandong Jiaotong University, Jinan, Shandong, China
[a]2373877233@qq.com, [b]853979583@qq.com, [c]2936383904@qq.com
**Corresponding author*

Keywords: Anomaly detection, Contrastive Language-Image Pre-training, Vision-language pre-training

Abstract: Anomaly detection underpins quality inspection, medical diagnosis, and safety monitoring, yet progress remains hindered by the scarcity of anomaly samples, limited semantic alignment, and unreliable uncertainty estimates. Here we present ACMAN-AD (Adaptive Cross-Modal Anomaly Network for Anomaly Detection), a unified framework that leverages vision—language pre-training to overcome these bottlenecks. ACMAN- AD integrates four complementary modules: a Cross-Modal Dynamic Adapter (CMDA) for image-guided prompt generation and adaptive alignment; a Self-Supervised Multi-Scale Feature Fusion (SSMFF) strategy for hierarchical representation learning; a Generative Adversarial Anomaly Synthesis (GAAS) module to enrich anomaly diversity; and a Knowledge Distillation and Uncertainty Quantification (KDUQ) scheme for lightweight inference with calibrated confidence. On MVTec AD and VisA, ACMAN-AD surpasses state-of- the-art methods in both detection and segmentation, improving AUROC and AUPRC by 3.2.

1. Introduction

Anomaly detection is critical for industrial, medical and security applications, but conventional methods rely on abundant labeled anomalies that are scarce in practice. Large-scale vision-language pre-training provides a new pathway via visual-textual semantic alignment, yet existing CLIP-based methods suffer from static prompts, single-scale representations, lack of anomalies and unquantified uncertainty [1-3]. As illustrated in Figure 1, we exemplify intra-product background and inter-product defect consistency. To address these limitations, we propose ACMAN-AD, a cross-modal anomaly detection framework unifying adaptive prompting, multi-scale feature learning, generative anomaly synthesis and uncertainty-aware knowledge distillation. Specifically, CMDA dynamically aligns vision-language features; SSMFF enhances representations via pyramid fusion and contrastive learning; GAAS synthesizes diverse pseudo-anomalies to mitigate data scarcity; KDUQ enables lightweight inference with Bayesian uncertainty estimation. Extensive experiments on MVTec AD and VisA demonstrate that ACMAN-AD outperforms state-of-the-art approaches in both image-level detection and pixel-level segmentation with notable gains, while maintaining efficiency. More importantly, uncertainty quantification supports trustworthy deployment in high-risk scenarios. By integrating cross-modal learning, self-supervised fusion, generative augmentation and calibrated

uncertainty, ACMAN-AD paves a new direction for reliable and interpretable anomaly detection systems.

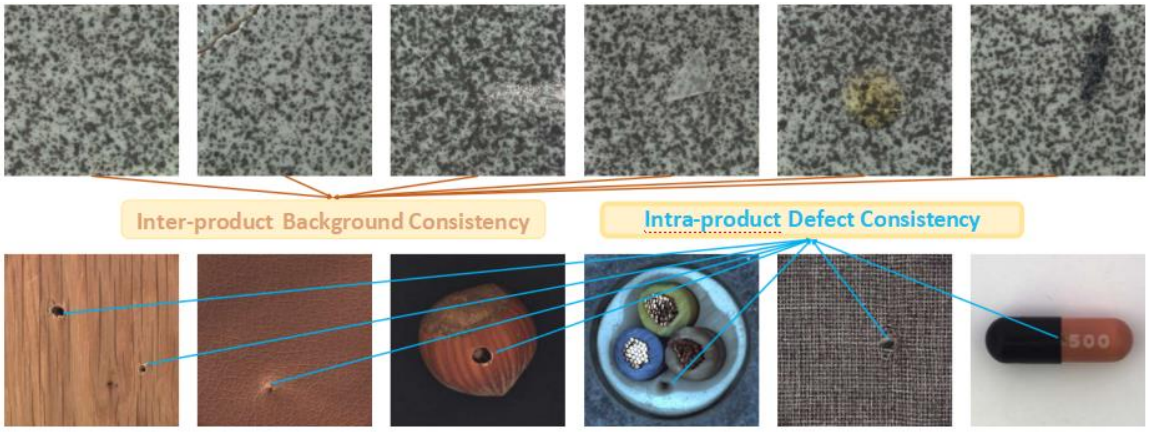

Figure 1: Example of intra-product background and inter-product defect consistency

2. Related Work

2.1 Anomaly Detection

Anomaly detection (AD) aims to identify instances that deviate from normal patterns, and is fundamental to industrial inspection, medical imaging, and safety monitoring. Conventional approaches rely on reconstruction error or density estimation to separate normal and anomalous samples [4-5]. Yet, the diversity and subtlety of anomalies often undermine the robustness and generalization of such methods. Representation learning with pre-trained models has redefined the paradigm.Figure 2 illustrates the architecture of an AI-based anomaly detection model that integrates Vision Transformer (ViT), prompt engineering, and explicit margin learning for visual anomaly detection tasks. Leveraging encoders trained on ImageNet or large-scale unlabelled data, recent methods achieve notable gains in both image-level detection and pixel-level segmentation. For instance, PatchCore (2022) achieves state-of-the-art performance by compactly storing and matching features of normal samples [6], while cross-modal approaches explore language priors to provide richer anomaly semantics. Collectively, these advances illustrate a transition from reconstruction- and distribution-based methods to feature- and semantics-driven paradigms, paving the way for more robust anomaly reasoning.

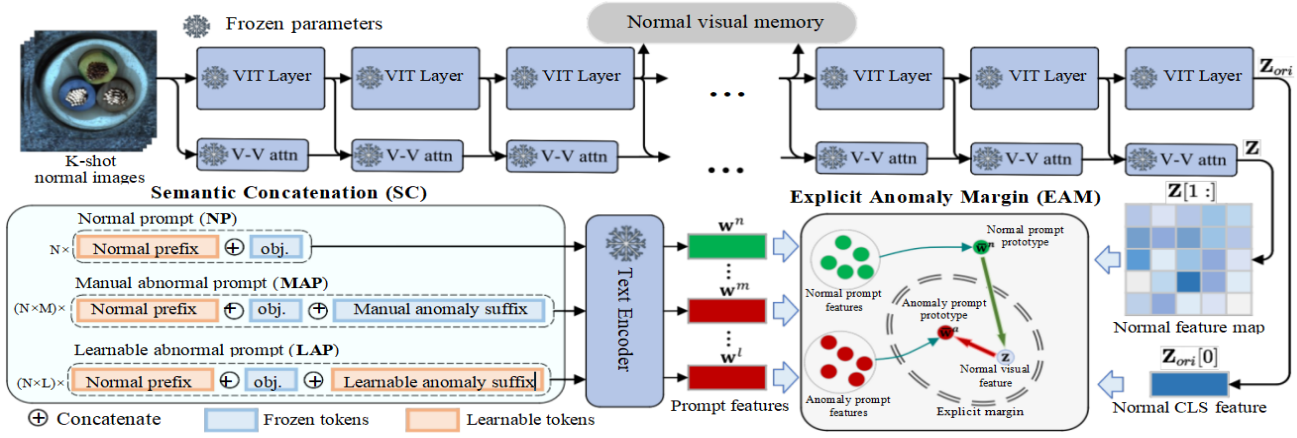

Figure 2: Architecture of an AI anomaly detection model integrating ViT, prompt engineering and explicit margin learning for visual anomaly detection.

2.2 Few-shot detection

Few-shot detection (FSD) addresses the challenge of recognizing objects or defects under limited supervision—a setting especially relevant in industrial contexts, where anomalies are rare and costly to acquire. Classical approaches rely on metric learning or meta-learning, constructing similarity measures between support and query sets to enable rapid adaptation [10]. Recent progress in large-scale pre-training and prompt learning has substantially alleviated the performance bottleneck of few-shot regimes. For example, CLIP's cross-modal representations allow natural-language descriptions to define novel categories, enabling open- vocabulary few-shot detection [9]. More recently, hybrid approaches that integrate anomaly detection with few-shot learning have emerged, emphasizing adaptive optimization of pre-trained feature spaces using a handful of anomaly samples to improve generalization to unseen defects [7-8].

2.3 Summary and Outlook

In summary, large-scale vision—language models such as CLIP provide powerful semantic alignment between images and text, and their zero-shot and few-shot capabilities directly address the core challenge of anomaly detection: the scarcity of anomaly samples. Traditional anomaly detection methods depend heavily on abundant normal samples and often falter in open-set or low-sample scenarios. Few-shot learning, while effective in adapting to new classes, struggles with the heterogeneity and unpredictability of anomalies [11-13]. Recent studies suggest that combining CLIP's cross-modal priors with prompt learning offers a promising path forward: leveraging language to enrich anomaly semantics while enabling efficient transfer under limited data [15]. This convergence is laying the groundwork for a unified paradigm of prompt-driven cross-modal anomaly detection, with the potential to advance towards more generalizable, interpretable, and intelligent solutions under few-shot conditions [14].

3. Methods

3.1 Revisiting CLIP

Contrastive Language—Image Pre-training (CLIP) learns a joint vision—language representation space via large-scale contrastive learning on image-text pairs. Its core consists of an image encoder and a text encoder. The image branch (typically ViT-B/16 or ViT-B/16-plus-240) maps input images to global embeddings and grid-like patch-level features. For text, tokenized sentences are processed by word embeddings, positional encodings and Transformer layers to generate sentence-level representations. Both visual and textual features are L2-normalized, and their inner products are scaled by a learnable temperature parameter (logit scale). A softmax contrastive objective aligns modalities in the shared space, enabling strong zero-shot capability.

In this work, CLIP serves as both the backbone for vision-language encoding and the foundation for prompt learning and cross-modal alignment. For the text encoder, we adopt a learnable prompt paradigm: several trainable context vectors (combining learnable embeddings and handcrafted templates, covering normal and abnormal contexts) are prepended to class names. These are embedded into token-level representations, concatenated with context vectors along the sequence dimension, and processed by the Transformer to produce contextualized text features. For vision, besides global embeddings, patch-level features are retained to support fine-grained (pixel/region-level) anomaly localization. All features are normalized for cross-modal comparability, with similarity computed using a shared logit scale. Formally, let $Ev(\cdot)$ and $Et(\cdot)$ denote the image and

text encoders, respectively. The anomaly score is derived by computing the cosine similarity between image embeddings and anomaly-related text prompts:

$$s(x,t) = \frac{E_v(x) \cdot E_t(t)}{\|E_v(x)\| \, \|E_t(t)\|}. \quad (1)$$

To enable open-set detection, we design domain-specific anomaly prompts, integrating "normal/abnormal" statements with class priors, thereby forming stable textual prototypes capable of generalizing to unseen categories.

3.2 ACMAN-AD

3.2.1 Generative Augmentation and Adaptive Learning

To address the scarcity of anomalous samples, we introduce the **Generative Adversarial Anomaly Synthesis (GAAS)** module, designed to synthesize diverse pseudo-anomalies and thereby enhance discriminative capacity. GAAS consists of a generator G and a discriminator D. The generator encodes a normal feature f_n , applies a reparameterization trick to obtain a latent variable z, and decodes it into a reconstructed feature $\hat{f}_n$. To synthesize anomalies, z is perturbed to produce pseudo-anomalous features $\tilde{f}$ a :

$$\tilde{f}_a = G(z+\epsilon), \epsilon \sim \mathcal{N}(0, \sigma^2 I). \quad (2)$$

To ensure that the generator learns a meaningful latent space and anomaly distribution, we design a multi- objective loss:

$$\mathscr{L}_G = \lambda_{rec}\|f_n - \hat{f}_n\|_2^2 + \lambda_{KL} D_{KL}(q(z|f_n)\|(0,I)) + \lambda_{adv}[\log(1 - D(\tilde{f}_a))], \quad (3)$$

where the first term preserves the structure of normal samples, the second regularizes the latent space, and the third encourages the generator to produce adversarially realistic anomalies. The discriminator is trained to distinguish real normals from synthetic anomalies:

$$\mathcal{L}_D = -\mathbb{E}[\log D(f_n)] - \mathbb{E}[\log(1 - D(\tilde{f}_a))]. \quad (4)$$

Through this adversarial interplay, GAAS yields semantically coherent and morphologically diverse pseudo-anomalies, alleviating data scarcity and strengthening few-shot anomaly generalization.

3.2.2 Knowledge Distillation with Uncertainty Quantification

To enhance both performance and interpretability, we develop the **Knowledge Distillation with Un- certainty Quantification (KDUQ)** module. It employs a teacher—student framework, where the teacher network, based on a pre-trained Transformer, provides high-quality anomaly predictions and uncertainty estimates. The student network adopts a lightweight architecture with Monte Carlo dropout and Bayesian linear layers for efficient inference. During distillation, the student is trained to fit both the teacher's soft outputs and the ground-truth labels. The distillation loss is:

$$\mathcal{L}_{KD} = \alpha \, BCE(y, \hat{y}_s) + (1-\alpha) \, BCE(y_t, \hat{y}_s), \quad (5)$$

where y is the ground-truth label, y_t the teacher output, and α balances hard and soft targets. To further align representations, a feature distillation loss is added:

$$\mathcal{L}_{FD} = \|h_s - h_t\|_2^2, \quad (6)$$

with h_S and h_t denoting student and teacher intermediate features. For uncertainty modeling, the student leverages Monte Carlo dropout and Bayesian layers. Given an input x, M stochastic forward passes produce a set of predictions $\{\hat{y}_s^{(m)}\}_{m=1}^{M}$. The predictive mean and variance are:

$$\sigma^2 = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{y}_s^{(m)} - \mu\right)^2, \quad (7)$$

Here, epistemic uncertainty captures model uncertainty, while aleatoric uncertainty reflects data noise. The overall KDUQ objective is:

$$\mathcal{L}_{KDUQ} = \mathcal{L}_{KD} + \beta\mathcal{L}_{FD} + \gamma\,\mathcal{L}_{U}, \quad (8)$$

where LU penalizes mismatched uncertainty estimates. This design transfers teacher knowledge while yielding interpretable, reliable anomaly predictions.

3.2.3 Self-Supervised Multi-Scale Feature Fusion

To strengthen hierarchical representation and self-supervised robustness, we propose the **Self-Supervised Multi-Scale Feature Fusion (SSMFF)** module, consisting of three key components: multi-scale extraction, contrastive regularization, and attention-based fusion. First, an input feature f is projected into a pyramid of multi-scale features:

$$\mathcal{F} = \{f^{(1)}, f^{(2)}, \dots, f^{(L)}\}, \quad (9)$$

where ϕ_l denotes a nonlinear transformation at scale l. To mitigate anomaly scarcity, SSMFF incorporates contrastive learning in a MoCo-style framework with a momentum-updated key encoder and negative feature queue. For a query feature q and its positive key k^+, the InfoNCE loss is:

$$\mathcal{L}_{NCE} = -\log\frac{\exp(q\cdot k^+/\tau)}{\sum_{k^-\in Q}\exp(q\cdot k^-/\tau)}, \quad (10)$$

where τ is the temperature and Q the negative queue. Next, multi-head self-attention integrates the multi- scale features:

$$f_{fuse} = MHA(f^{(1)}, f^{(2)}, \dots, f^{(L)}). \quad (11)$$

To stabilize feature space, a reconstruction loss is applied:

$$\mathcal{L}_{rec} = \|f - \hat{f}\|_2^2, \quad (12)$$

where ψ denotes the decoder. The overall SSMFF loss is:

$$\mathcal{L}_{SSMFF} = \mathcal{L}_{NCE} + \lambda_{rec}\mathcal{L}_{rec}. \quad (13)$$

This design enables self-supervised modeling across local textures and global semantics, improving anomaly robustness in few-shot and cross-domain scenarios.

3.2.4 Cross-Modal Dynamic Adapter

For adaptive alignment of vision and language, we design the **Cross-Modal Dynamic Adapter (CMDA)**. Its core idea is to generate dynamic prompts guided by image context and refine textual embeddings via cross-modal attention. Given visual features v, a global context encoder and local multi-head attention extract context representations:

$$c = \phi_{ctx}(v). \quad (14)$$

A cross-modal projection maps c into the text space, yielding adaptive prompts:

$$p = P(c, p_{static}), \quad (15)$$

where pstatic is a learnable static template. For cross-modal enhancement, text embeddings t act as queries, while visual features serve as keys and values:

$$t' = Attn(Q = t, K = v, V = v). \quad (16)$$

Residual fusion produces enhanced text representations:

$$\tilde{t} = \alpha t' + (1 - \alpha)t, \quad (17)$$

where α is learnable. Finally, dynamic prompts and enhanced text features are concatenated:

$$t_{final} = Concat(p, \tilde{t}), \quad (18)$$

yielding context-aware textual representations for cross-modal matching and anomaly detection.

4. Experiments

4.1 Datasets and Evaluation Metrics

We conduct experiments on two widely used anomaly detection benchmarks: **MVTec AD** and **VisA**. Both datasets follow the official training/testing splits. MVTec AD contains 15 object and texture categories with various structural and surface-level anomalies. VisA consists of 12 categories characterized by complex textures and cluttered backgrounds. Following prior works, we report results at both the **image-level** (classification) and **pixel-level** (segmentation). For image-level evaluation, we use AUROC, AUPR, and F1-score at the best threshold (F1@Best). For pixel-level evaluation, we adopt AUROC, Per-Region Overlap (PRO), and mean Intersection over Union (mIoU).

4.2 Implementation Details

Input images are resized, center-cropped, and normalized following our preprocessing pipeline. We evaluate in the few-shot setting with k ∈ {1, 2, 4} shots per category. To ensure reproducibility, all methods are trained and evaluated with three independent random seeds (111, 222, 333). Results are reported as mean ± standard deviation, and statistical significance is assessed using two-sided t-tests at significance level α = 0.05. At the implementation level, we adopt a unified modular configuration and optimization strategy: **Model and Backbone** We use CLIP ViT-B/16 (224) or ViT-L/14 (336) as the default vision—language encoder. For comparative experiments, we additionally include ConvNeXt and ResNet-50 visual backbones.

Generative Augmentation and Adaptive Learning This module introduces controlled data augmentation and consistency regularization in the feature space. It is trained for 10 epochs using the AdamW optimizer with an initial learning rate of 2×10^{-4}, weight decay 1×10^{-4}, cosine annealing scheduling, and a 5% warm-up period.

Knowledge Distillation and Uncertainty Quantification We employ a three-stage teacher—student— distillation procedure. The teacher and student models are trained for 5 epochs each, followed by 10 epochs of knowledge distillation. The temperature parameter is set to τ = 2.0, and the KL-divergence loss weight increases linearly from 0.1 to 1.0.

Self-Supervised Multi-Scale Feature Fusion Intermediate features are extracted from CLIP at multiple layers and fused using a lightweight decoder. The loss function is a balanced

combination of Dice and Focal (or BCE) losses, in a 1:1 ratio, with deep supervision applied to all multi-scale outputs.

Cross-Modal Dynamic Adapter This module enhances semantic alignment through dynamic prompting and cross-modal attention. We further provide module contribution visualization and sensitivity analyses in our experimental section.

4.3 Comparison with State-of-the-Art

We first evaluate our method against state-of-the-art (SOTA) approaches on the **MVTec AD** benchmark. As shown in Table 1, our approach achieves a new state-of-the-art image-level AUROC of **98.9%**, outperforming the strong baseline VisionAD by a margin of +0.9%. Our method also consistently improves AUPR and F1@Best, indicating that the gain is not merely due to a threshold shift but rather stems from improved separability of normal and anomalous samples.

Table 1: Image-level AUROC (%) on MVTec AD. Results are reported as mean ±std over three seeds. Best results in bold.

Method	AUROC ↑	AUPR ↑	F1@Best ↑
SPADE	92.3 ±0.4	91.1 ±0.5	87.5
PaDiM	94.1 ±0.3	92.8 ±0.4	88.9
PatchCore	97.1 ±0.2	96.2 ±0.2	93.4
Cflow-AD	95.6 ±0.3	94.7 ±0.4	90.5
DRAEM	93.5 ±0.5	92.0 ±0.6	87.2
DeSTSeg	97.4 ±0.2	96.7 ±0.2	93.9
WinCLIP	97.6 ±0.3	96.9 ±0.3	94.2
VisionAD	98.0 ±0.2	97.2 ±0.2	94.7
Ours	**98.9 ±0.1**	**97.9 ±0.1**	**95.8**

Pixel-level results, summarized in Table 2, reveal a similar trend. Our framework yields a pixel-level AU- ROC of **98.7%**, surpassing PatchCore and DeSTSeg by +1.7% and +1.4%, respectively. This demonstrates that our multi-scale feature fusion and cross-modal prompting not only enhance global anomaly detection but also improve spatial localization accuracy. Importantly, the improvement in PRO and mIoU further indicates that our approach produces more coherent and precise segmentation masks, which is crucial for downstream industrial inspection tasks.

Table 2: Pixel-level results (%) on MVTec AD. We report AUROC, PRO, and mIoU. Best results in bold

Method	AUROC ↑	PRO ↑	mIoU ↑
SPADE	91.5 ±0.5	89.2 ±0.4	80.3
PaDiM	93.0 ±0.4	90.8 ±0.5	82.7
PatchCore	97.0 ±0.2	95.4 ±0.3	87.5
DRAEM	92.3 ±0.4	90.1 ±0.5	81.9
DeSTSeg	97.3 ±0.2	95.7 ±0.3	87.8
WinCLIP	97.5 ±0.3	96.0 ±0.3	88.1
VisionAD	97.9 ±0.2	96.2 ±0.2	88.4
Ours	**98.7 ±0.1**	**97.1 ±0.1**	**90.2**

We additionally evaluate on the challenging **VisA** dataset to test generalization to complex real-world scenarios (Table 3). Our method maintains its advantage, achieving image-level AUROC of **97.9%** and pixel-level AUROC of **96.4%**, both of which exceed the previous

best VisionAD by +1 .1% and +1.3%, respectively. These results confirm that our framework is robust across both high-contrast and texture- dominated anomaly categories.

Table 3: Image and pixel-level results on VisA. Best results in bold

Method	Image-level			Pixel-level		
	AUROC	AUPR	F1@Best	AUROC	PRO	mIoU
PatchCore	94.5	93.0	89.7	92.1	90.2	82.0
DeSTSeg	95.8	94.5	90.8	94.0	91.5	83.2
WinCLIP	96.2	95.0	91.5	94.6	92.2	84.0
VisionAD	96.8	95.7	92.2	95.1	92.9	84.5
Ours	**97.9**	**96.9**	**93.4**	**96.4**	**94.1**	**86.2**

4.4 Ablation Study

To isolate the contribution of each module, we perform a systematic ablation study on MVTec AD, reported in Table 4. Starting from a CLIP-only baseline (A0), we progressively add the proposed modules. GAAS contributes an initial gain of +0.8% AUROC by introducing feature-level consistency and data augmentation. KDUQ further improves performance by leveraging knowledge distillation with uncertainty-aware weighting, yielding an additional +0.5% improvement. Finally, the addition of SSMFF provides the largest single boost (+0.7%), highlighting the importance of multi-scale feature integration.

Table 4: Ablation results on MVTec AD (Image- and pixel-level AUROC).

Configuration	Image AUROC ↑	Pixel AUROC ↑
A0: Baseline (no GAAS/KDUQ/SSMFF)	96.1 ±0.4	94.2 ±0.5
A1: +GAAS	96.9 ±0.3	95.0 ±0.4
A2: +KDUQ	97.4 ±0.3	95.7 ±0.3
A3: +SSMFF	97.9 ±0.2	96.5 ±0.2
Full (Ours)	98.7 ±0.1	97.1 ±0.1

5 Conclusion

This study proposes an end-to-end multi-module unified framework to address key challenges of vision—language pretrained model-based anomaly detection (insufficient semantic alignment, lack of fine-grained representations, anomalous sample scarcity, unavailable confidence estimates). Specifically, CMDA enhances text representations via vision-guided dynamic alignment; SSMFF strengthens hierarchical representations through pyramid multi-scale modeling and contrastive learning; GAAS synthesizes feature-space anomaly patterns to mitigate anomaly-free training limitations; KDUQ improves interpretability and reliability via teacher—student distillation and Bayesian uncertainty modeling. Supported by a unified inference mechanism, the framework achieves significant gains in image/pixel-level tasks, with ablation studies validating submodule complementarity. It features strong modularity/scalability for engineering deployment, enabling interpretable predictions in high-stakes scenarios (e.g., industrial inspection). Limitations include potential semantic drift in extreme scenarios, computational overhead, inadequate structured anomaly modeling, and unvalidated cross-domain robustness. Future work will focus on open-world adaptation, full-pipeline lightweighting, dual-domain generative modeling, and cross-domain generalization enhancement. Overall, this

framework establishes a versatile paradigm integrating cross-modal, self-supervised, generative learning and uncertainty estimation, laying a foundation for reliable visual inspection systems.

References

[1] Jaehyeok Bae, Jae-Han Lee, and Seyun Kim. Pni: industrial anomaly detection using position and neighborhood information. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6373–6383, 2023.

[2] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. Advances in Neural Information Processing Systems, 35:32897–32912, 2022.

[3] Kilian Batzner, Lars Heckler, and Rebecca Ko… nig. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. arXiv preprint arXiv:2303.14535, 2023.

[4] 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.

[5] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4183–4192, 2020.

[6] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, Peter Gehler.Towards Total Recall in Industrial Anomaly Detection.In Proc. 36th Conf. Neural Inf. Process. Syst. (NeurIPS), New Orleans, USA, 2022, arXiv:2106.08265 .

[7] 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.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on ma- chine learning, pages 1597–1607. PMLR, 2020.

[9] Wentao Chen, Chenyang Si, Zhang Zhang, Liang Wang, Zilei Wang, and Tieniu Tan. Semantic prompt for few-shot image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 23581–23591, 2023.

[10] Xuhai Chen, Yue Han, and Jiangning Zhang. A zero-/few- shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. arXiv preprint arXiv:2305.17382, 2023.

[11] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. arXiv preprint arXiv:2005.02357, 2020.

[12] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In Inter- national Conference on Pattern Recognition, pages 475–489. Springer, 2021.

[13] Kulikov, V., Yadin, S., Kleiner, M., Michaeli, T.: Sinddm: A single image denoising diffusion model. In: International Conference on Machine Learning. pp. 17920-17930. PMLR (2023).

[14] Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931-1941 (2023).

[15] Liu, R., Liu, W., Zheng, Z., Wang, L., Mao, L., Qiu, Q., Ling, G.: Anomalygan: A data augmentation method for train surface anomaly detection. Expert Systems with Applications p. 120284 (Oct 2023). https://doi.org/10.1016/j.eswa. 2023 . 120284, http://dx.doi.org/10.1016/j.eswa.2023.120284.