ASBench: Image Anomalies Synthesis Benchmark for Anomaly Detection

Qunyi Zhang, Songan Zhang, Jiaqi Liu, Jinbao Wang, *Member, IEEE*, Xiaoning Lei, Guoyang Xie, *Member, IEEE*, Guannan Jiang, *Member, IEEE*, Zhichao Lu, *Member, IEEE*

***Abstract*—Anomaly detection plays a pivotal role in manufacturing quality control, yet its application is constrained by limited abnormal samples and high manual annotation costs. While anomaly synthesis offers a promising solution, existing studies predominantly treat anomaly synthesis as an auxiliary component within anomaly detection frameworks, lacking systematic evaluation of anomaly synthesis algorithms. Current research also overlooks crucial factors specific to anomaly synthesis, such as decoupling its impact from detection, quantitative analysis of synthetic data and adaptability across different scenarios. To address these limitations, we propose ASBench, the first comprehensive benchmarking framework dedicated to evaluating anomaly synthesis methods. Our framework introduces four critical evaluation dimensions: (i) the generalization performance across different datasets and pipelines (ii) the ratio of synthetic to real data (iii) the correlation between intrinsic metrics of synthesis images and anomaly detection performance metrics, and (iv) strategies for hybrid anomaly synthesis methods. Through extensive experiments, ASBench not only reveals limitations in current anomaly synthesis methods but also provides actionable insights for future research directions in anomaly synthesis. Code is available at https://github.com/M-3LAB/ASBench.**

***Impact Statement*—Industrial image anomaly detection is a highly popular field in both academia and industry. However, academic research on industrial image anomaly detection primarily focuses on unsupervised anomaly detection, whereas industrial applications often directly employ supervised training methods to develop the required models. Anomaly synthesis can effectively bridge the gap between academia and industry by transforming unsupervised anomaly detection into supervised model training. Our paper presents the first comprehensive benchmark for analyzing anomaly synthesis in industrial images. Through a detailed examination of various anomaly synthesis methods, we aim to identify the currently optimal synthesis method and narrow the gap between industrial practices and academic research.**

***Index Terms*—Anomaly detection, Defect detection, Unsupervised learning**

Received 10 October 2025; revised 9 January 2026; accepted 27 March 2026. This work was supported by the National Natural Science Foundation of China under Grant 62576218. (Qunyi Zhang, Songan Zhang and Jiaqi Liu contributed equally to this work. Corresponding authors: Guoyang Xie; Guannan Jiang.)

Qunyi Zhang and Songan Zhang are with Global Institute of Future Technology, Shanghai Jiao Tong University. (Email: zqyeleven@sjtu.edu.cn, songanz@sjtu.edu.cn)

Jinbao Wang is with School of Artificial Intelligence and also with the National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, China. (Email: wangjb@szu.edu.cn)

Xiaoning Lei, Guoyang Xie and Guannan Jiang are with Contemporary Amperex Technology Co.,Ltd. (Email: leixn01@outlook.com, guoyang.xie@ieee.org, jianggn@catl.com)

Jiaqi Liu and Zhichao Lu are with Department of Computer Science, City University of HongKong. (Email: liu_jiaqi_@outlook.com, zhichao.lu@cityu.edu.hk)

I. Introduction

ANOMALY detection has emerged as a critical technology in modern manufacturing quality control and healthcare monitoring, playing an important role in ensuring product reliability and enhancing diagnostic accuracy [7] [32]. Despite its significance, the practical deployment of anomaly detection systems is often constrained by two major challenges: the scarcity of abnormal samples and the prohibitively high costs associated with manual annotation. These limitations significantly impede the performance and scalability of the detection models. To address these issues, anomaly synthesis has gained traction as a promising solution, offering the capability to generate representative abnormal samples. This approach not only alleviates the data scarcity problem but also substantially reduces the reliance on costly manual annotation efforts. Jan and Christian (2022) [9] summarize the anomaly synthesis methods. The significant potential has been demonstrated in various domains, including surface defect detection and medical image analysis. Anomaly synthesis can be effectively integrated with real anomaly data or directly applied to enhance detection systems.

While anomaly synthesis holds transformative potential for advancing anomaly detection capabilities, current anomaly synthesis methods remain largely confinded to auxiliary roles within anomaly detection frameworks. This paradigm neglects rigorous benchmarking and detailed exploration of the methodological nuances and performance characteristics unique to anomaly synthesis algorithms. The absence of systematic evaluation frameworks creates critical gaps in understanding algorithmic strengths, operational boundaries and domain-specific adaptability-limitations. To address this critical gap, our reasearch focuses specifically on anomaly synthesis algorithms, with the primary objectives of establishing a systematic evaluation framework and conducting rigorous performance analysis. The proposed framework aims to provide scientific foundations for anomaly detection tasks, offering theoretical guidance for the selection and optimization of synthesis methods in practical applications.

Existing benchmark studies in this field primarily focus on anomaly detection tasks, yet fail to address critical challenges specific to anomaly synthesis. The current research landscape exhibits the following limitations:

1) There has been no systematic evaluation of different anomaly synthesis methods across various datasets, thereby lacking guidance in practical applications.
2) Since the detection and generation components of the

model are not decoupled, it is impossible to assess the magnitude of the impact of anomaly synthesis on the performance of anomaly detection models.
3) Instead of direclty conducting quantitative analysis on synthesized abnormal images, the effectiveness of anomaly synthesis methods is only indirectly reflected through the metrics of anomaly detection models
4) An analysis of how the data proportion of synthetic anomaly samples influences anomaly detection tasks is absent.
5) Current anomaly detection frameworks predominantly employ single anomaly synthesis algorithms, neglecting potential benefits from hybrid application of multiple synthesis strategies.

To address these limitation, we propose ASBench, a comprehensive benchmark to evaluate the anomaly synthesis from four perspectives:

1) **Cross Dataset and Detection Methods Comparison**: Cross-dataset generalization capability and detection model compatibility evaluation of individual synthesis methods.
2) **Data Ratio Impact**: Investigate the impact of different sample proportions of synthesis anomalies on the performance of anomaly detection models.
3) **Metric Correlation**: Correlation analysis between intrinsic quality metrics of the anomaly images generated and downstream detection performance metrics.
4) **Hybrid Strategies**: Evaluate the effects of mixing abnormal data samples generated by multiple anomaly synthesis algorithms.

As shown in Table I, compared to other benchmark studies in the field of anomaly detection, our ASBench stands out for its unique features in anomaly synthesis decoupling, adjustable anomaly sample ratios, and diverse pipeline configurations. The overall workflow of ASBench, along with the chapter structure detailing analyzes from various perspectives, is illustrated in Fig. 1.

Key Takeaways: Through extensive experimentation, we have identified several key insights:

1) No single anomaly synthesis method universally dominate across all datasets and algorithms.
2) The sample proportions of generated anomalies exert no significant influence on the performance of anomaly detection models.
3) There is no correlation between the intrinsic metrics of generated images and detection performance.
4) Combined usage of multiple anomaly synthesis can improve the accuracy of anomaly detection.

Our main contributions are listed below.

1) **First Comprehensive Benchmark for Anomaly Synthesis**: To address the research gap in evaluation standards for anomaly generation, we propose ASBench, a standardized and unified benchmark designed for comprehensive investigation and experimental assessment of anomaly generation methodologies.
2) **Open-IAS Framework for Unified Evaluation**: We propose Open-IAS, a flexible framework integrating 12 anomaly synthesis methods, 4 detection pipelines, and 5 industrial datasets, resulting in a total of 19,680 data. providing standardized protocols and reproducible baselines to accelerate future research in anomaly synthesis.
3) **Critical Analysis and Future Directions**: Through ASBench, we perform granular comparisons across anomaly proportions, generation strategies, dataset adaptability, and evaluation metrics. This reveals limitations in existing methods and presents actionable research directions.

The remainder of this paper is structured as follows. Section II reviews related literature on anomaly detection pipelines and anomaly synthesis methods. Section III introduces the experimental settings of our proposed benchmark, ASBench. In Section IV, a comprehensive experimental evaluation is presented along with a discussion of the results from four perspectives. Finally, Section V concludes the paper.

II. Related Work

A. Anomaly Detection

With the release of the MVTec AD dataset [2], the development of industrial image anomaly detection has shifted from a supervised paradigm to an unsupervised one [12], [22]. In the unsupervised paradigm, the training set consists solely of normal images, while the test set includes both normal and labeled anomalous images. Research in unsupervised industrial image anomaly detection has gradually evolved into two main categories: feature embedding-based methods and reconstruction-based methods.

Feature embedding-based methods can be further divided into four subcategories. Specifically, (i) teacher-student distillation models typically comprise a pretrained teacher network and a trainable student network. During training, the knowledge of normal sample features extracted by the teacher network is distilled into the student network. During inference, discrepancies between features extracted by the student network for anomalous samples and those extracted by the teacher network facilitate anomaly detection. State-of-the-art results in this category have been achieved by methods such as RD4AD and RD++. (ii) One-class classification methods generate anomalies either at the image level or the feature level and then learn to classify anomalous images or features. Representative methods in this subcategory include CutPaste [19] and SimpleNet [24]. (iii) Mapping-based methods utilize pretrained models to extract image features, which are then mapped to a desired distribution using a feature mapping module. During testing, if the sample's features deviate from the expected distribution, the sample is deemed anomalous. Techniques in this category frequently employ normalizing flow modules to map features to a multivariate Gaussian distribution. (iv) Memory-based methods use pretrained networks to extract features from training samples, which are subsequently sampled and stored in a memory bank. During testing, anomaly scores are computed by measuring the distance between the test sample's features and the features stored in the memory bank.

Reconstruction-based methods share a similar overall architecture. These methods typically involve self-supervised

Off-the-shelf Work & Ours	IM-IAD [35]	ADer [42]	MMAD [18]	MulSen-AD [21]	RAD [5]	Real-IAD [33]	PAD [45]	ASBench
Contains Anomaly Synthesis Method	✓	✗	✗	✗	✗	✓	✓	✓
Decoupling of Synthesis and Detection	✗	✗	✗	✗	✗	✗	✗	✓
Multiple Datasets	7	11	4	1	1	1	1	5
Different Abnormal Data Ratio	✗	✗	✗	✗	✗	✗	✗	✓
Different Pipeline Combination	✗	✗	✗	✗	✗	✗	✗	✓

TABLE I: Comparison of different anomaly detection benchmarks. Our ASBench is the first to provide a decoupled analysis of the individual components in anomaly synthesis methods.

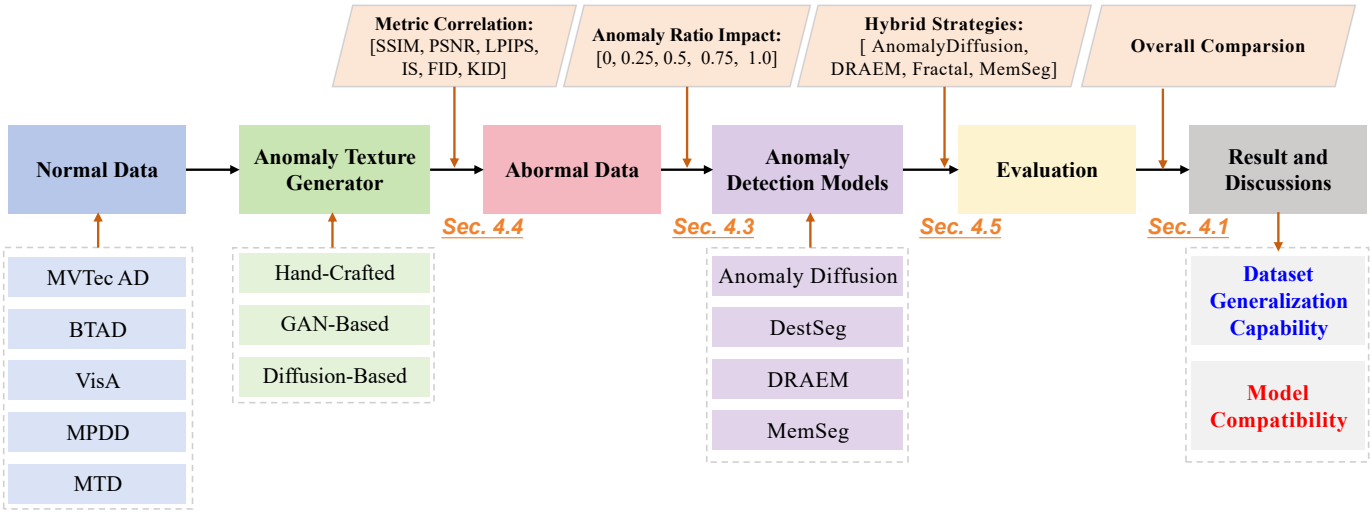


Fig. 1: The overall workflow of ASBench. We disentangle different stages in the anomaly synthesis and anomaly detection pipeline, and discuss the impact of variables at each stage on anomaly detection in separate subsections of Section IV.

training, wherein normal images and artificially generated anomalous images are reconstructed into normal images. Anomaly localization is achieved during testing by comparing the differences between the reconstructed and original images. Autoencoders are the most commonly used reconstruction networks for anomaly detection, as seen in DRAEM [37], DSR [39], and NSA [28]. While methods leveraging generative adversarial networks (GANs) as reconstruction networks are fewer, they have demonstrated outstanding performance, indicating untapped potential in this area. Recently, transformers, as a foundational model in computer vision, have also been employed as reconstruction networks and have shown impressive results in anomaly detection. Diffusion models, which are among the most popular generative models currently, have set new SOTA benchmarks in anomaly detection tasks with methods such as DDAD [26] and DiffusionAD [41].

The field of anomaly detection has seen excellent development in available datasets. MVTec AD [2], as the first comprehensive dataset, covering multiple object and defect types, has become the widely recognized benchmark in the field. MTD [16] is specifically designed for saliency detection in tile images, focusing on the identification of anomalies at a micro-scale. BTAD [25] is a dataset from real industrial scenarios where images have relatively simple backgrounds, making the detection task more straightforward. MPPD [17] focuses on defects in metal painted parts during the manufacturing process, emphasizing detection capability under diverse viewing angles and complex conditions. The innovation of the VisA dataset [47] lies in its introduction of a multi-instance sample setting. These datasets align with the objectives of the anomaly detection tasks targeted by ASBench from various perspectives.

PAD [46] and Real-IAD [33] share a key characteristic with MPPD by focusing on multi-viewpoint imaging to closely approximate real industrial conditions. Comprehensive 3D datasets such as MVTec 3D-AD [4] and Real3D-AD [23] are primarily oriented towards anomaly detection and localization in 3D data, which does not fully align with ASBench's current focus on 2D image detection tasks.

B. Anomaly Synthesis

The development of anomaly generation techniques has evolved from simple image-level manipulations to complex generative approaches. Hand-crafted method is the mainstream research direction in the early stages of anomaly synthesis research. Initial methods primarily operated at the patch level for augmentation. Techniques such as CutOut [8] and RIAD [38] simulated anomalies by blacking out certain patches, while methods like CutPaste [19] further replaced patches with textures from other regions to mimic anomalies. However, the shapes of anomalies generated through these approaches often differed significantly from real anomalous shapes. Consequently, subsequent research gradually shifted toward using random noise to simulate anomalous regions. Representative methods include DRAEM [37], which employs Perlin noise [13] and external texture datasets to generate anomalous textures, and FractalAD [34], which utilizes fractal noise. Building upon these approaches, MemSeg[36]

constrained anomalies to appear only in foreground regions, whereas NSA [28] filled anomalous areas with the object's own texture via Poisson image editing.

With the advancement of generative models, methods based on generative models have gradually become mainstream in anomaly synthesis, as they produce textures that are closer to real anomalies and offer greater diversity. Early methods such as DFMGAN [11] and AdaBLDM [20] utilized GANs [27] and latent diffusion models to generate defective images, thereby improving anomaly generation capabilities under few-shot conditions. In recent years, diffusion models [14] have been introduced to the anomaly synthesis task. Methods such as AnomalyDiffusion [15] enable precise control over the spatial distribution and visual appearance of anomalies, capable of generating more realistic anomalous images. RealNet [43] combined adaptive diffusion models with feature selection modules to enhance anomaly detection performance.

Based on the distinct characteristics of existing anomaly synthesis methods, we have selected patch augmentation-based and random noise-based methods from the Hand-crafted synthesis category, as well as GANs-based and diffusion-based methods from the generation-model-based category for our benchmark.

III. ASBench

A. Problem Definition

The proposed ASBench systematically addresses three critical dimensions in various anomaly synthesis evaluation: cross-dataset adaptability, detection algorithm compatibility, and anomaly ratio sensitivity. The focus of ASBench is on existing anomaly detection models and related anomaly synthesis techniques. This research first evaluates the generated images of anomaly synthesis methods independently by decoupling the anomaly detection and synthesis components. Subsequently, it combines existing anomaly synthesis methods with different anomaly detection models, summarizing the performance across various industrial datasets. Additionally, the interactions between multiple anomaly synthesis methods are evaluated to identify the most effective and realistic anomaly generation strategies, in order to explore the optimal approach for anomaly detection tasks in industrial environments.

The experimental framework of this study can be formally summarized by the following relationship:

$$Trained\ Model = A \otimes (B, C, D),$$

where A represents the anomaly detection algorithm, B denotes the anomaly synthesis method, C corresponds to the industrial dataset and D specifies the proportion of synthetic abnormal samples in the training data. The $\otimes$ operator represents composability between components. This formulation establishes a systematic approach to evaluate the interdependencies between critical components.

B. Implementation Details

The research conducts systematic comparisons across 5 industrial anomaly detection datasets by integrating 12 anomaly synthesis approaches with 4 detection models. The overall workflow is shown in Fig. 1, the anomaly synthesis and detection processes are decoupled into different stages.

Table II lists 12 anomaly synthesis methods used in ASBench (marked in purple). The criteria for selecting methods to be implemented for ASBench are that different shapes of the abnormal region and whether model training is required.

Regarding the selection of anomaly detection pipelines, our research incorporates four representative architectures: DRÆM, DestSeg, MemSeg, and AnomalyDiffusion. DRÆM employs pixel-level composition and segmentation for anomaly identification, while DestSeg leverages feature distillation and contrastive learning to effectively capture subtle discrepancies in complex textures. In contrast to DRÆM's pixel-level fusion strategy, MemSeg operates through feature space segmentation, demonstrating particular efficacy in scenarios requiring distinct anomaly separation within feature representations. AnomalyDiffusion adopts a novel paradigm by decoupling abnormal appearance and structural information, directing the model's attention to minute anomalous patterns. These algorithms exhibit complementary strengths in segmentation mechanisms, anomaly saliency processing, and feature comparison strategies, contributing to the development of a robust detection system capable of addressing diverse anomaly characteristics.

For conducting thorough ablation studies, we utilize five publicly available datasets, MVTec AD [2], BTAD [25], VisA [47], MPDD [17], and MTD [16]. Table III offers a detailed summary of these datasets, encompassing the number of samples (covering both training and test sets, with normal and abnormal samples), the total number of classes, image resolution, and the primary characteristics of each dataset. All datasets include pixel-level annotations.

Regarding evaluation metrics, we employ Area Under the Receiver Operating Characteristics (AU-ROC/AUC), Area Under Precision-Recall (AUPR/AP), and Per-Region Overlap (PRO) [1] to evaluate the abilities of anomaly localization. These five metrics collectively address both anomaly detection and segmentation performance at image and pixel granularities, ensuring comprehensive and precise evaluation.

For experimental configurations, all parameter settings strictly adhere to the default specifications outlined in the detection models' publicly available source codes, maintaining methodological consistency and reproducibility.

IV. Result and Discussions

This section explores current anomaly synthesis methods and discusses the essential components of the proposed uniform settings. Each subsection outlines the experimental methodologies employed, presents a detailed analysis of the results, and identifies unresolved challenges along with potential avenues for future research.

A. Overall Comparison

1) Comparative Experimental Framework: We conducted a comprehensive comparison across 5 datasets, 12 synthesis methods, and 4 detection algorithms, and the detailed results

Paradigm		Methods
Hand-Crafted	Patch Augmentation	**CutOut** [8], **CutPaste**, **CutPaste_Scar** [19], **FPI** [30], PII [31], RIAD [38]
	Random Shape Noise	**DestSeg** [44], **DRAEM** [37], **FractalAD** [34], **MemSeg** [36], **NSA** [28]
Generative-Model-Based	GAN-Based	**DFMGAN** [11], Con-GAN [10], Defect-GAN [40]
	Diffusion-Model-Based	**RealNet** [43], **AnomalyDiffusion** [15], DefectDiffu [29], AdaBLDM [20]

TABLE II: Representative anomaly synthesis algorithms for ASBench. The bolded ones indicate our re-implemented method.

Datasets	Sample Number			Classes	Image Resolution		Main Feature
	Train Set	Test Set			Min	Max	
	Normal	Normal	Abnormal				
MVTec AD [2]	3629	467	1258	15	700	1024	Basic benchmark standard
BTAD [25]	1799	451	290	3	600	1600	Real-world manufacturing
VisA [47]	8659	962	1200	12	960	1562	Multi-instance IAD
MPDD [17]	888	176	282	6	1024	1024	Multi-view Anomalies
MTD [16]	902	50	392	1	113	491	Micro-scale anomalies

TABLE III: Statistics of the selected datasets with diverse features used in ASBench.

Dataset	MVTEC		BTAD		VisA		MPDD		MTD	
Metric	AUC Image	AP Pixel	AUC Image	AP Pixel	AUC Image	AP Pixel	AUC Image	AP Pixel	AUC Image	AP Pixel
AnomalyDiffusion	0.9796(a)	0.7599(d)	0.9481(s)	**0.6748(m)**	0.8813(s)	0.4179(s)	**0.9883(s)**	**0.4640(m)**	**0.9732(d)**	**0.6801(s)**
CutOut	0.8799(d)	0.2644(s)	0.8796(m)	0.1630(s)	0.7959(a)	0.1832(s)	0.9218(s)	0.2018(s)	0.7449(d)	0.1150(s)
CutPaste	0.9539(s)	0.5659(s)	**0.9643(d)**	0.4159(m)	0.8558(s)	0.3576(s)	0.9181(s)	0.1688(a)	0.8870(s)	0.4892(d)
CutPaste Scar	0.9417(s)	0.3851(s)	0.8848(s)	0.2078(s)	0.8529(s)	0.2542(s)	0.8918(d)	0.1969(a)	0.8428(s)	0.1128(s)
DestSeg	**0.9892(s)**	**0.7948(s)**	0.9445(d)	0.5010(s)	**0.9508(d)**	0.3727(a)	0.9749(s)	0.3401(d)	0.9476(s)	0.3754(s)
DFMGAN	0.9393(s)	0.6027(s)	0.9397(s)	0.3355(m)	0.8208(d)	0.2879(a)	0.8832(m)	0.2696(s)	0.6902(m)	0.0807(a)
DRAEM	0.9889(s)	0.7810(s)	0.9500(s)	0.5049(s)	0.9362(d)	**0.4933(s)**	0.9819(s)	0.3774(d)	0.9470(s)	0.3565(s)
FPI	0.9557(d)	0.6527(s)	0.9366(d)	0.3508(m)	0.9338(d)	0.4030(s)	0.9300(d)	0.2487(a)	0.8405(d)	0.2200(s)
Fractal	0.9858(s)	0.7327(s)	0.9442(d)	0.5621(s)	0.9348(s)	0.4866(s)	0.9598(s)	0.4193(m)	0.9090(s)	0.2710(s)
MemSeg	0.9841(s)	0.7515(s)	0.9469(s)	0.5836(s)	0.9025(m)	0.2568(s)	0.9257(m)	0.3070(s)	0.9322(s)	0.3129(s)
NSA	0.9813(d)	0.6957(d)	0.9541(m)	0.5175(m)	0.9210(d)	0.4822(s)	0.9625(d)	0.2703(a)	0.8674(d)	0.1927(s)
RealNet	0.9815(s)	0.7130(s)	0.9235(m)	0.2788(m)	0.8601(m)	0.2358(s)	0.9395(s)	0.2939(s)	0.9191(s)	0.2207(s)

TABLE IV: Optimal algorithm performance of 12 synthesis methods on 5 datasets. Bolded data represents the optimal results among the twelve methods. Different suffixes represent different detection methods. (s)=DestSeg, (d)=DRAEM, (m)=MemSeg, (a)=AnomalyDiffusion.

are presented in Table IV. To visualize the overall performance, we generated radar charts (Fig. 2) depicting AUC Image and AP Pixel aggregated from all datasets. Notably, for the overall evaluation, we adopted a weighted averaging approach using the number of subclasses within each dataset as weighting coefficients to account for inter-dataset categorical imbalances.

The weighted average formula is as follows:

$$\overline{M} = \sum_{i=1}^{5} w_i \cdot M_i, \quad w_i = \frac{N_i}{\sum_{j=1}^{5} N_j}, \tag{1}$$

where M_i denotes the metric value on dataset i, N_i is the number of subclasses in dataset i, and weights w_i normalize subclass counts to mitigate categorical imbalance.

Additionally, Fig. 3 and 4 present separate radar charts that illustrate AUROC Image and AUPR Pixel metrics for individual datasets, providing granular insights into the performance of methods across different data environments.

2) Analysis of Results: The experimental results indicate that no single anomaly synthesis method demonstrates universal superiority. Our investigation reveals that current anomaly synthesis methods and detection algorithms, predominantly validated on the MVTec AD dataset. Consequently, optimal performance on MVTec AD is achieved by employing the synthesis method originally proposed within the respective detection algorithms. However, when these approaches are extended to alternative datasets and integrated with different processing algorithms, the previously superior detection algorithms and synthesis methods fail to maintain their performance advantages, demonstrating significant performance degradation in cross-dataset applications.

As evidenced by Fig. 3 and Table IV, significant performance variations are observed across different combinations of detection algorithm and synthesis method on various datasets. Regarding image-level anomaly classification, the four synthesis methods, DestSeg, DRAEM, Fractal, and AnomalyDiffusion exhibit generally stable performance across all four detection algorithms. With notable exceptions: CutPaste combined with DRAEM's algorithm achieves state-of-the-art results on BTAD, where textures are relatively simple and homogeneous. In contrast, the VisA dataset contains multi-instance objects, making it more suitable for methods like FPI and NSA that employ seamless image editing and blending, or Fractal-based techniques that generate diversified anomalies through synthetic fractal patterns. For datasets focusing on multi-view, fine-grained details (e.g., MPDD, MTD), generative models such as AnomalyDiffusion and RealNet perform better, as they excel at simulating detailed anomalies. For pixel-level anomaly localization, the AnomalyDiffusion detection pipelines demonstrates robust compatibility with most synthesis methods on

This article has been accepted for publication in IEEE Transactions on Artificial Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAI.2026.3680823

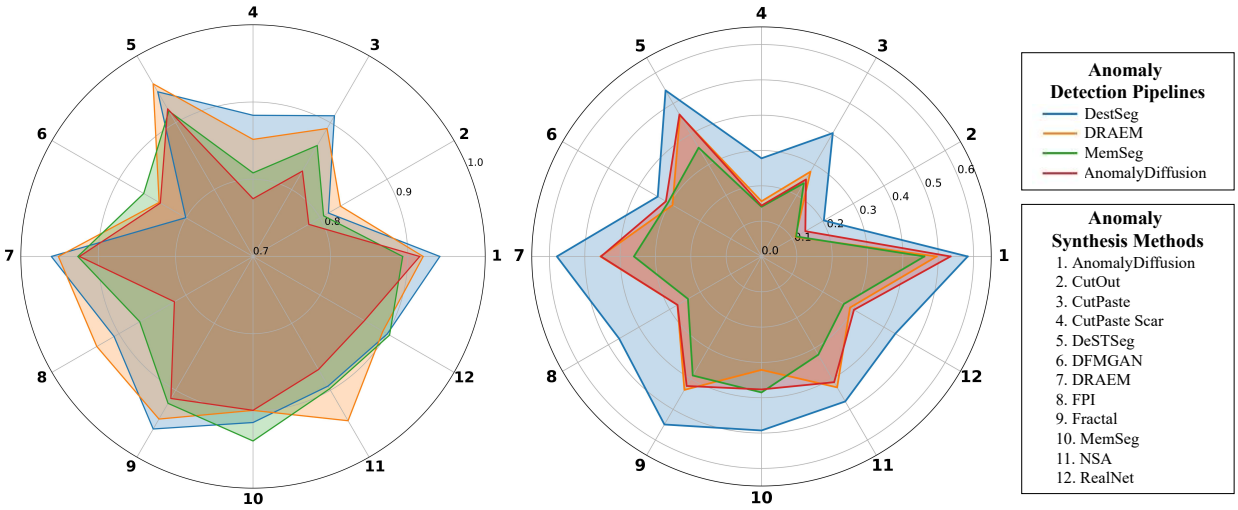
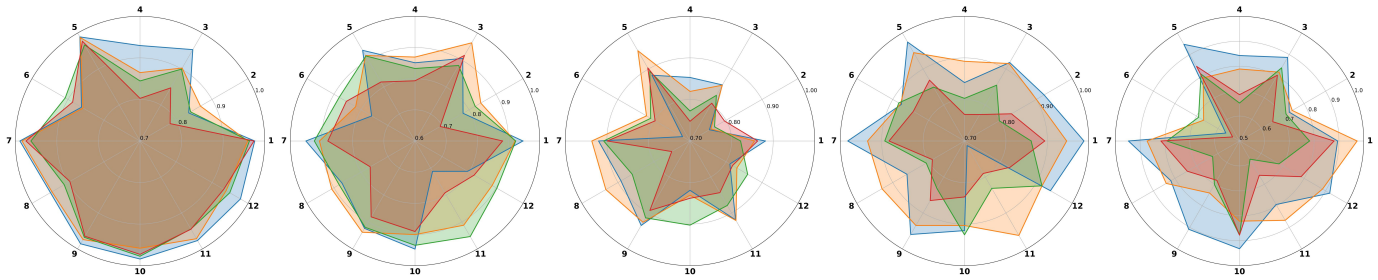
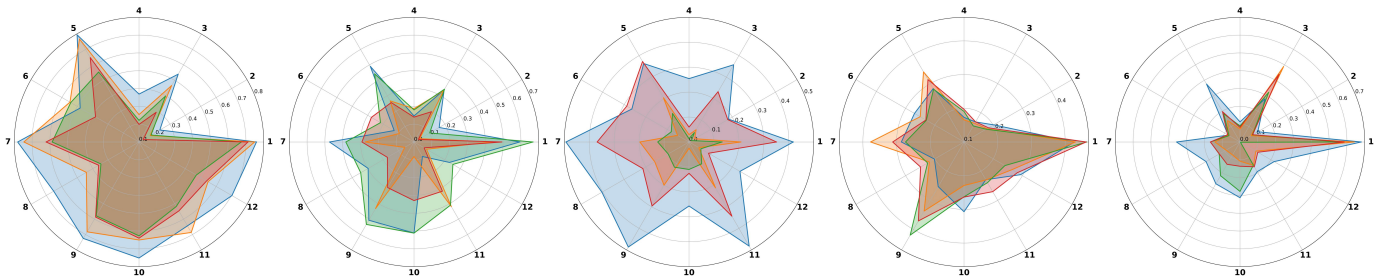


Fig. 2: Performance comparison of different anomaly synthesis methods across various detection pipelines and all datasets. For datasets, results are computed via weighted averaging, where weights correspond to the number of subclasses per dataset. Axes represent the anomaly synthesis methods, lines correspond to the detection pipelines, and vertices quantify the performance of these synthesis methods across different detection pipelines.



(a) MVTec AD (b) BTAD (c) VisA (d) MPDD (e) MTD

Fig. 3: Image-level AUROC performance comparison of different anomaly synthesis methods across various detection pipelines and different datasets.



(a) MVTec AD (b) BTAD (c) VisA (d) MPDD (e) MTD

Fig. 4: Pixel-level AUPR performance comparison of different anomaly synthesis methods across various detection pipelines and different datasets.

VisA, MPDD, BTAD, and MTD datasets. However, on the MVTec AD dataset, synthesis strategies derived from DRAEM and DestSeg yield superior localization performance compared to AnomalyDiffusion.

3) Implications and Future Challenges: The experimental findings underscore necessity of strategic selection between detection pipelines and synthesis methods based on task-specific requirements. Optimal performance requires dual adaptation: first, to dataset-specific characteristics (e.g., background complexity and positional variance), and second, to task priorities (classification versus localization). This empirical evidence motivates pursuit of a universal anomaly synthesis methods capable of autonomously adapting to heterogeneous data distributions while maintaining cross-algorithm compatibility. This is a critical direction for advancing generalizable anomaly detection systems.

B. Analysis of Anomaly Synthesis Methods Across Datasets and Detection Models

1) Comparative Experimental Framework: Based on the experiments conducted across five datasets, twelve synthesis methods, and four detection models, we further analyze the following aspects. First, the cross dataset and detection model performance analysis is performed. We compute the average performance metrics for each synthesis method across detection pipelines. We construct the heatmap (Fig. 5) with synthesis methods versus datasets to visually present performance variations across different dataset-method combinations. By averaging the performance metrics across all five datasets, we generate another heatmap (Fig. 5), which takes synthesis methods and detection models as axes. Fig. 5 primarily highlights the performance fluctuations of identical synthesis methods across different detection models.

Then, to analyze model robustness, we calculate the mean and variance of performance metrics for each synthesis method across different datasets and under different detection models. And we construct Fig. 6 to quantify the discrepancy in robustness among the methods.

2) Analysis of Results: As shown in the Fig. 2 and Fig. 5, through comparative analysis of synthesis methods, we observe that approaches capable of generating complex and diverse anomalous regions significantly outperform handcrafted designs or those relying on real anomaly samples, and the potential advantages of real anomalies remaining underutilized. Handcrafted methods like such as CutPaste, CutOut, and FPI demonstrate inferior performance. Approaches leveraging real anomalies, including DFMGAN, RealNet, and AnomalyDiffusion, also do not exhibit significant advantages. The more effective methods, DRAEM, DestSeg, Fractal, and NSA, employ complex and diverse masks to increase anomaly diversity and mitigate overfitting risks. Defining anomalous regions through Fractal patterns or Perlin Noise proves more effective than random cropping or simulating real anomalies, as these techniques introduce greater stochasticity that improves model generalization capabilities.

Notably, MemSeg synthesis method restricts anomalies to object foregrounds, unlike DRAEM, to reduce background noise interference. However, this strategy yields no significant performance gains; in some cases, MemSeg underperforms DRAEM. The excessive foreground concentration limits anomaly diversity, failing to cover potential background noise patterns and increasing false positives. Since background noise is typically subtle, models focusing solely on foregrounds overlook such anomalies, compromising overall detection effectiveness. Moreover, MemSeg's overly strict foreground constraints require dataset-specific parameter tuning for optimal results, diminishing its generalization capability and further impairing detection performance.

Fig. 5 reveals that DestSeg and DRAEM achieve the highest average performance among the evaluated approaches, with DRAEM exhibiting particularly low variance. DestSeg's strength stems from its two-stage training strategy: In the first stage, it aligns the student model's features with those of the teacher model, adopting the distillation-based anomaly detection framework from prior unsupervised methods [3], [6]. This allows knowledge transfer without requiring anomaly samples. The second stage introduces a segmentation model for anomaly localization while keeping the pre-trained teacher-student models fixed. Building on this foundation of first stage, DestSeg delivers robust performance across diverse synthetic anomalies. In addition, DRAEM demonstrates minimal sensitivity to anomaly data ratios and synthesis method variations. Its reconstruction-based architecture enables training under extreme conditions—even in anomaly-free scenarios, the model can localize anomalies by detecting discrepancies between input images and their reconstructions.

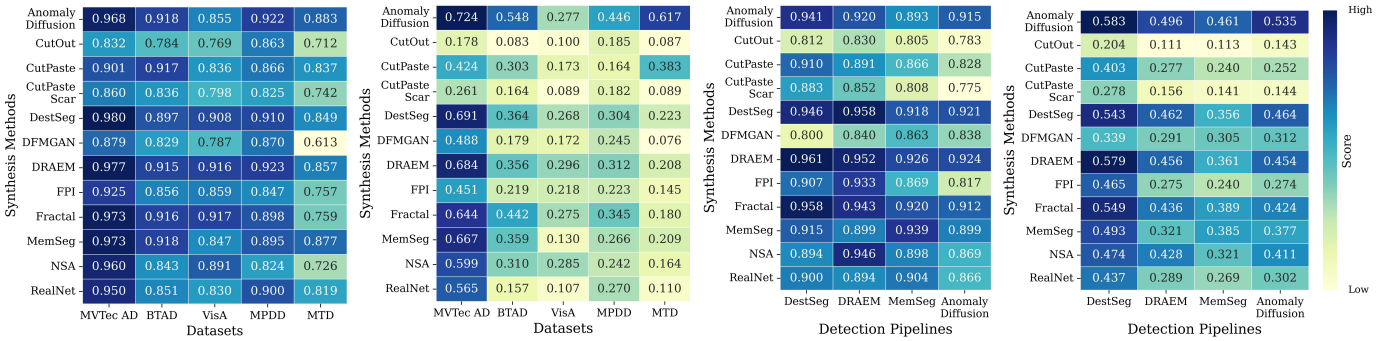
3) Implications and Future Challenges: Based on the preceding analysis, the optimal synthesis methods are influenced by detection models and dataset characteristics. These collective observations underscore that no single synthesis method is universally optimal across all evaluation scenarios, necessitating careful calibration among datasets, evaluation metrics, and downstream detection pipelines. The core challenge lies in designing highly generalizable anomaly synthesis approaches that balance diversity and authenticity in generated anomalies. Future work should explore more efficient utilization of real data while developing models robust to background noise and cross-dataset discrepancies.

As shown in Fig. 7, the comparison between MemSeg and DRAEM synthesis methods reveals that anomaly synthesis methods must strike a balance between anomaly diversity and background complexity, as overemphasizing foreground objects can yield detrimental effects. Harmonizing anomaly diversity with background noise management emerges as the pivotal challenge for advancing anomaly synthesis and detection performance. For instance, adjust masks while preserving fidelity to real anomalies; or synergistically integrate authentic anomaly-derived masks with procedurally generated masks (e.g., Perlin/fractal noise) to enhance structural diversity.

Furthermore, regarding downstream detection pipelines, current stable pipelines universally incorporate U-Net for segmentation, underscoring its essential role in ensuring robustness. Future work should optimize U-Net for unsupervised industrial scenarios, particularly for long-tail distributions where anomalies are sparse, to fundamentally enhance detection

Dataset	MVTec AD	BTAD	VisA	MPDD	MTD
DestSeg($\mathcal{D}$)	DestSeg	DRAEM	Fractal	AnomalyDiffusion	DestSeg
DRAEM($\mathcal{D}$)	DestSeg	CutPaste	DestSeg	NSA	AnomalyDiffusion
MemSeg($\mathcal{D}$)	MemSeg	NSA	Fractal	MemSeg	MemSeg
AnomalyDiffusion($\mathcal{D}$)	AnomalyDiffusion	CutPaste	DestSeg	AnomalyDiffusion	AnomalyDiffusion

TABLE V: Optimal anomaly synthesis methods across different algorithms and datasets based on image-level AUROC. The $\mathcal{D}$ represents detection pipeline.



(a) Image-level AUROC (b) Pixel-level AUPR (c) Image-level AUROC (d) Pixel-level AUPR

Fig. 5: Heatmaps of image-level AUROC and pixel-level AUPR performance for anomaly synthesis methods. The images (a) and (b) represent performance across different datasets. The images (c) and (d) represent performance across different detection pipelines. AnomalyDiffusion, DRAEM, and DestSeg establish a clear advantage in anomaly synthesis effects over other methods.

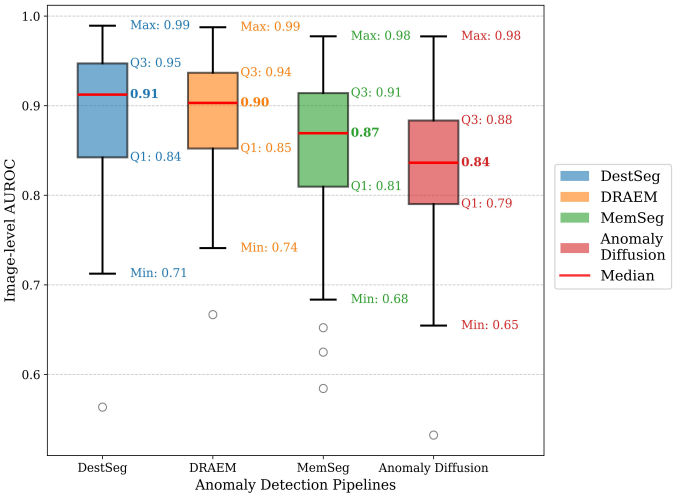


Fig. 6: Comparison of detection pipelines: (a) DestSeg, (b) DRAEM, (c) MemSeg, and (d) Anomaly Diffusion.

generalization capabilities.

C. Effect of Different Anomaly Ratios

1) Comparative Experimental Framework: Considering the original implementations, the default anomaly-to-total ratios (i.e., the proportion of anomalous samples in the training set) were set to 0.5 and 1.0 across the four detection algorithms. For our experiments, we systematically varied this ratio in the training sets (0.25, 0.5, 0.75, and 1.0) on the MVTec AD and BTAD datasets.

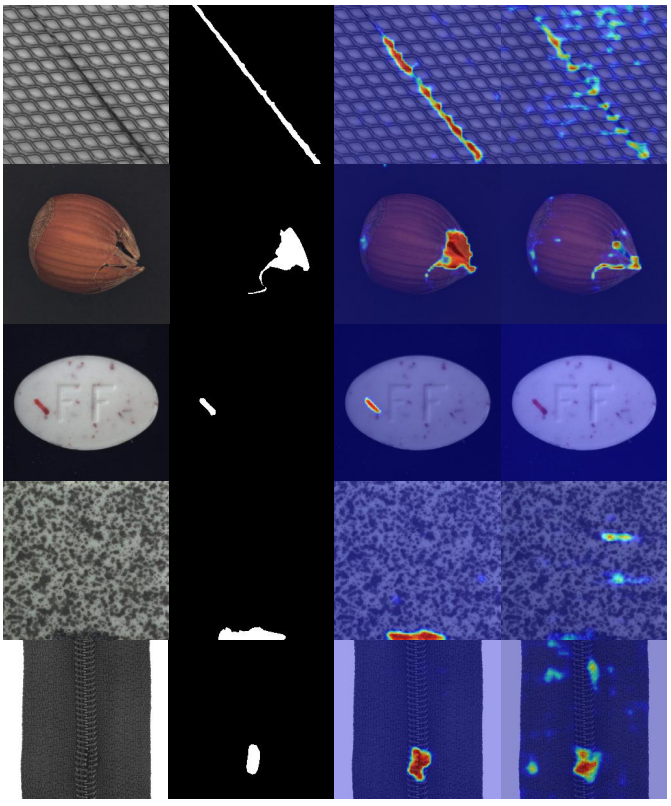


Fig. 7: Visualization comparison between synthesis methods: DRAEM and MemSeg.

This configuration enables dual analytical perspectives: (a) averaging across four detection models to evaluate how data ratios affect performance under different synthesis strategies, and (b) averaging across synthesis methods to examine how anomaly ratios impact individual model performance. On the other hand, we investigate the relationship between synthesis method stability and performance by analyzing how the performance differential (between optimal and worst-case scenarios) varies with anomaly sample ratios. The model-averaged performance results and the performance differentials are recorded in Table VII and VIII. Based on tabulated mean values and performance gaps, we generate scatter plots and calculate the related correlation coefficients to analyze.

2) Analysis of Results: As shown in Table VI and Fig. 8, the experimental results under anomaly-free conditions demonstrate significantly inferior detection performance compared to scenarios containing anomalous samples. The significant difference observed demonstrate the efficacy of anomaly synthesis in enhancing detection capabilities across these algorithms. However, Fig. 10 illustrates that, after introducing synthetic anomalies, different ratios exhibit minimal performance variations, and no statistically significant optimal ratio point has been identified through current comparative analyzes. Notably, peak performance rarely occurs at higher anomaly concentrations. A particularly illustrative case is DFMGAN, which exhibits marked performance degradation at ratio=1.0. The visualization comparison is shown in Fig. 9. This phenomenon stems from DFMGAN's design paradigm that leverages some real anomalies. Excessive synthetic anomalies during training induce overfitting to specific real anomaly patterns, consequently impairing generalization capacity on unseen anomaly types. In particular, According to the Fig. 10, similar degradation is observed in AnomalyDiffusion detection algorithms at ratio = 1.0, likely attributable to their intrinsic anomaly-focused feature encoding mechanism that amplifies the risks of synthetic pattern overfitting.

For synthesis methods utilizing authentic anomalous samples, excessive incorporation of synthesized anomalous samples may increase the possibility of overfitting. In this context, while approaches such as DFMGAN and AnomalyDiffusion enhance the authenticity of generated anomalous samples, they do not exhibit superior performance in anomaly detection tasks. Although the generated anomalies more closely approximate the distribution of real-world anomalies, an overabundance of synthetic samples in the training set may cause the model to over-rely on specific features of these samples, thereby inducing overfitting and compromising the model's generalization capability on unseen data.

In particular, regarding the visual effect, we anticipate that the samples generated by AnomalyDiffusion will exhibit superior quality compared to DFMGAN, demonstrating greater authenticity and diversity. AnomalyDiffusion achieves better anomaly detection performance despite its synthetic samples' higher fidelity. This observation implies that AnomalyDiffusion achieves a more effective equilibrium between authenticity and diversity in anomaly generation, thereby optimizing detection efficacy. In contrast, DFMGAN's anomaly synthesis process may overemphasize domain-specific anomalous features in certain scenarios, rendering detection models more susceptible to overfitting and impeding their ability to generalize across heterogeneous anomaly types.

Our investigation of ratio-dependent detection performance revealed that the seven top-performing methods exhibited greater robustness against ratio variations. Pearson's coefficients obtained from two independent datasets (-0.902409 and -0.701109) approached -1, demonstrating statistically significant strong negative linear relationships. These p-values ($p = 0.000059$ and 0.011074, respectively), being substantially below the 0.05 significance threshold, provide robust evidence that the observed correlations between ratio stability and detection efficacy are not attributable to random variation.

3) Implications and Future Challenges: To better utilize limited anomalous samples and prevent overfitting, technical improvements in anomaly synthesis methods are crucial. Excess synthetic samples fail not only to effectively enhance model performance but may induce overfitting to training data, resulting in suboptimal performance on unseen data. Therefore, it is essential to achieve an optimal equilibrium between the quantity and quality of generated samples while avoiding redundant or informationally redundant output. Adaptive generation strategies should be developed to more accurately simulate characteristics of scarce anomalous samples, enhancing diversity and authenticity through controlled stochastic processes. This optimization prevents overspecialization on specific anomaly features during generation, thereby improving model robustness and generalization capacity.

D. Correlation between Data Metrics and Performance

1) Comparative Experimental Framework: Our comprehensive analysis across five datasets evaluates the correlation between detection performance and synthetic image quality metrics. We compare generated anomalies with real defective samples using six established image generation metrics: Structural Similarity Index (SSIM) for structural similarity, Peak Signal-to-Noise Ratio (PSNR) for pixel-level differences, Learned Perceptual Image Patch Similarity (LPIPS) for perceptual quality, Inception Score (IS) for diversity and quality assessment, with Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) quantifying distributional discrepancies between synthetic and real data. These metrics collectively provide a multidimensional evaluation of image generation fidelity, reconstruction accuracy, and perceptual characteristics.

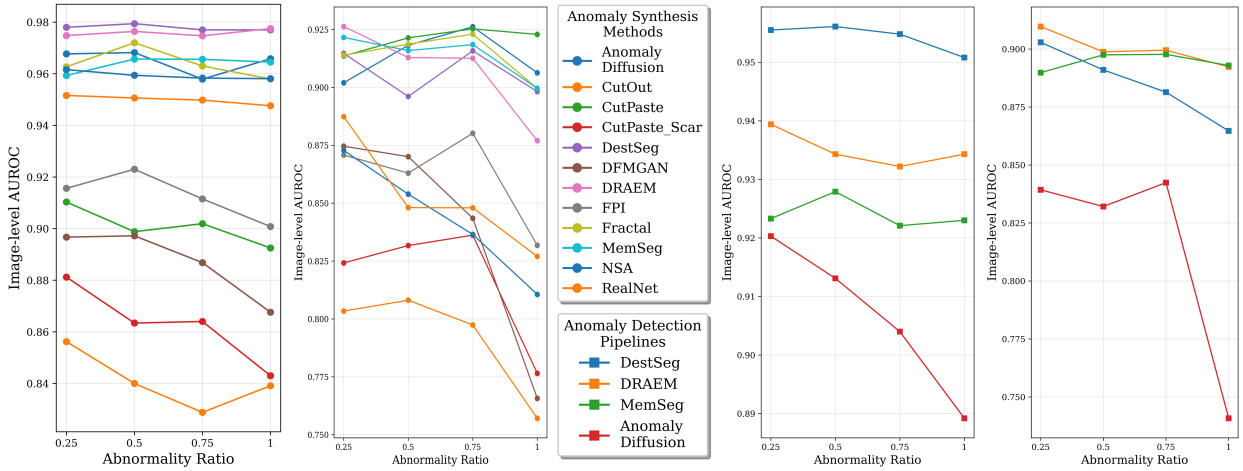
The analytical framework employs two-dimensional scatter plots, Fig. 11 visualizing relationships between generation metrics and detection performance indicators, accompanied by Pearson correlation coefficients with statistical significance levels (p-values), as detailed in Table IX .

2) Analysis of Results: The comprehensive analysis reveals weak correlations between anomaly detection performance metrics and image generation quality indicators. As evidenced by the tabular data and scatter plots of metric relationships, conventional image synthesis measures, including PSNR, SSIM, LPIPS, KID, FID, and IS demonstrate limited predictive capability for detection effectiveness.

The computed Pearson coefficients and corresponding p-values predominantly exhibit two patterns: (1) Near-zero cor-

Dataset	DestSeg	DRAEM	MemSeg	AnomalyDiffusion
Metric	AUROC / AUPR / PRO	AUROC / AUPR / PRO	AUROC / AUPR / PRO	AUROC / AUPR / PRO
MVTec AD	0.5850 / 0.0362 / 0.1536	0.7425 / 0.0531 / 0.2324	0.7968 / 0.1440 / 0.5065	0.7985 / 0.0730 / 0.3416
BTAD	0.5421 / 0.0330 / 0.1572	0.7578 / 0.0572 / 0.2436	0.6277 / 0.0442 / 0.4859	0.8028 / 0.0361 / 0.2822
VisA	0.5276 / 0.0071 / 0.1850	0.7462 / 0.0046 / 0.3265	0.6852 / 0.0082 / 0.4862	0.5666 / 0.0500 / 0.6029
MPDD	0.3735 / 0.0345 / 0.1874	0.7808 / 0.0756 / 0.2718	0.7832 / 0.1411 / 0.4326	0.7642 / 0.1676 / 0.6418
MTD	0.6345 / 0.0820 / 0.1646	0.7401 / 0.0742 / 0.2150	0.5620 / 0.0743 / 0.4753	0.4814 / 0.0776 / 0.2921

TABLE VI: Performance comparison (AUROC/AUPR/PRO) of various methods under anomaly-free conditions on different datasets.



(a)Mvtec_AD (b)BTAD (c)MVTec AD (d)BTAD

Fig. 8: Impact of anomaly ratio on anomaly synthesis methods performance: (a&b) image-level AUROC of different anomaly synthesis methods on MVTec AD and BTAD. (c&d) image-level AUROC of different anomaly detection pipelines on MVTec AD and BTAD. A performance degradation is often observed when the abnormality ratio is set to 1.

Method	Anomaly Diffusion	CutOut	CutPaste	CutPaste Scar	DestSeg	DFMGAN	DRAEM	FPI	Fractal	MemSeg	NSA	RealNet
Mean	0.9649	0.8410	0.9009	0.8629	0.9778	0.8871	0.9759	0.9127	0.9639	0.9638	0.9593	0.9499
Variance	2.27E-05	1.28E-04	5.44E-05	2.44E-04	1.25E-06	1.92E-04	1.84E-06	8.61E-05	3.45E-05	9.19E-06	2.63E-06	2.90E-06
Max	0.9682	0.8562	0.9103	0.8812	0.9794	0.8972	0.9775	0.9230	0.9720	0.9657	0.9616	0.9516
Min	0.9579	0.8288	0.8925	0.8430	0.9770	0.8676	0.9747	0.9008	0.9580	0.9593	0.9581	0.9476
Range (Max - Min)	0.0102	0.0274	0.0178	0.0382	0.0024	0.0296	0.0028	0.0223	0.0141	0.0064	0.0035	0.0040

TABLE VII: Performance of anomaly synthesis methods on different proportions based on image AUROC and MVTec AD dataset.

Method	Anomaly Diffusion	CutOut	CutPaste	CutPaste Scar	DestSeg	DFMGAN	DRAEM	FPI	Fractal	MemSeg	NSA	RealNet
Mean	0.9132	0.7915	0.9208	0.8172	0.9062	0.8385	0.9072	0.8614	0.9138	0.9139	0.8435	0.8526
Variance	1.22E-04	5.45E-04	2.56E-05	7.59E-04	1.10E-04	2.54E-03	4.46E-04	4.40E-04	1.03E-04	9.63E-05	6.99E-04	6.36E-04
Max	0.9262	0.8081	0.9253	0.8362	0.9158	0.8746	0.9263	0.8802	0.9230	0.9216	0.8728	0.8874
Min	0.9020	0.7571	0.9136	0.7765	0.8961	0.7657	0.8770	0.8318	0.8996	0.8996	0.8106	0.8270
Range (Max - Min)	0.0242	0.0510	0.0117	0.0597	0.0197	0.1089	0.0493	0.0484	0.0234	0.0220	0.0622	0.0604

TABLE VIII: Performance of anomaly synthesis methods on different proportions based on image AUROC and BTAD dataset.

relation coefficients coupled with statistically insignificant p-values , supporting the null hypothesis of no linear association; (2) Moderately elevated correlation coefficients accompanied by non-significant p-values, indicating insufficient statistical evidence for meaningful correlations.

These dual patterns collectively confirm the absence of statistically robust linear relationships between synthetic image quality metrics and anomaly detection performance across the evaluated frameworks. The weak correlation can be mainly attributed to two factors. Firstly, this discrepancy originates from the fundamentally divergent optimization objectives between conventional quality metrics and detection metrics. Image quality metrics predominantly assess reconstruction fidelity and visual similarity through pixel-wise comparisons and structural/textural analyzes. The metrics, typically computed at the image domain or holistic level, inadequately capture localized anomalies that characterize most synthetic anomalies. In contrast, detection-oriented metrics prioritize localized anomaly characteristics by design, systematically evaluating spatial precision to identify deviant patterns at the pixel or region level.

Secondly, existing generative models exhibit limitations in simulating the intrinsic complexity of real-world anomalies. When synthesized anomalies fail to precisely replicate the

This article has been accepted for publication in IEEE Transactions on Artificial Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAI.2026.3680823

Dataset	MVTec-AD		BTAD		VisA		MPDD		MTD	
Comparison	Cor	P-value	Cor	P-value	Cor	P-value	Cor	P-value	Cor	P-value
SSIM vs AUROC	0.0832	0.7972	0.1220	0.7056	0.1958	0.7056	-0.1624	0.6141	0.2360	0.4603
PSNR vs AUROC	-0.0472	0.8843	0.1314	0.6840	0.0336	0.6840	-0.1813	0.5729	-0.4960	0.1010
LPIPS vs AUROC	0.1068	0.7412	-0.2138	0.5046	0.2278	0.5046	0.2053	0.5221	-0.2194	0.4933
IS vs AUROC	0.6023	0.0382	-0.0482	0.8817	0.6882	0.8817	0.5332	0.0742	0.0701	0.8286
FID vs AUROC	0.0358	0.9121	-0.4318	0.1610	-0.0163	0.1610	0.1645	0.6094	-0.2800	0.3782
KID vs AUROC	-0.4152	0.1796	-0.3319	0.2919	-0.3915	0.2919	-0.0339	0.9167	-0.3698	0.2368

TABLE IX: Correlation and P-values between six metrics and image-level AUROC across datasets.

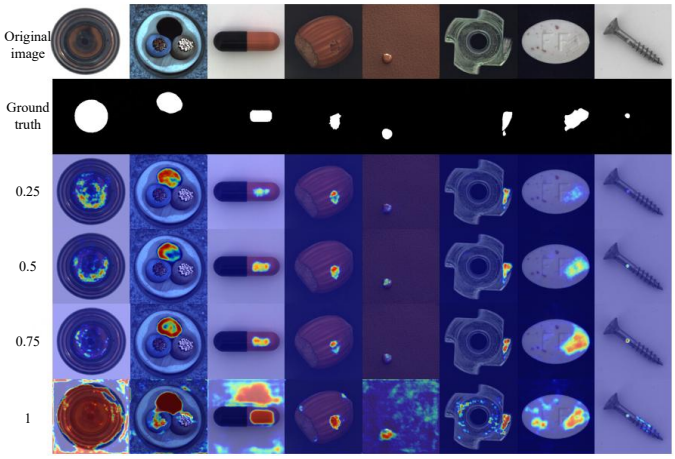


Fig. 9: Visualization comparison of the synthesis method DFMGAN between different anomaly ratios.

subtle morphological and contextual variations of authentic defects, significant discrepancies emerge between superficial image quality and practical detection utility. For example, certain synthetically generated anomalies may achieve high scores on traditional quality metrics while lacking discriminative features essential for effective detection, ultimately leading to suboptimal detection performance.

3) Implications and Future Challenges: The limited predictive value for detection tasks of current evaluation requires the development of specialized assessment criteria for synthetic abnormal samples. An enhanced evaluation paradigm should concurrently operate at pixel-level and image-level dimensions, rather than rely solely on holistic image quality metrics. Such a dual-scale framework would enable more reliable quality assessment of synthetic anomalies while permitting prediction when utilizing data for detector training, thereby conserving human and material resources typically expended in repetitive experimental iterations of detection tasks.

This challenge parallels difficulties encountered in evaluation of natural image generation. Some methods assess generation quality through combined global-local analyzes. For example, portrait synthesis validation examines both overall reasonability and localized correctness. Drawing inspiration from these approaches, synthetic anomaly evaluation should incorporate Global consistency and Local fidelity. The establishment of such multidimensional evaluation standards represents a critical research direction for improving synthetic anomaly generation and application efficacy.

E. Performance of Mixed Synthesis methods

1) Analysis of Results:

2) Comparative Experimental Framework: This section is based on four synthesis algorithms—Fractal, DRAEM, AnomalyDiffusion, and MemSeg—to construct and evaluate eleven hybrid combinations through training and testing, aiming to assess their synergistic effects in anomaly synthesis tasks. The experiments use the weighted average of performance metrics across the dataset as a reference. The results of the hybrid methods are compared with those of each individual method and theoretical combined means. The comparison of visualization is shown in Fig. 12. And bar charts are plotted for different detection pipelines, as shown in Fig. 13, to analyze the performance differences among various combinations.

From Fig. 13, it can be observed that most hybrid methods outperform the best results achieved by individual methods. A few hybrid methods show slightly lower performance than the optimal single method, but still surpass the weaker individual methods. Additionally, by comparing the theoretical and actual performance of hybrid methods, we find that the actual performance of hybrid methods generally exceeds the theoretical mean, indicating that the combination of multiple methods indeed produces synergistic effects.

This improvement primarily stems from the diversity, complementarity, and robustness of the different algorithms. Each algorithm has distinct strengths and characteristics when handling anomalies. Combining these methods can cover a broader range of anomaly patterns, thereby enhancing overall detection performance. Furthermore, hybrid methods, by integrating the results of multiple algorithms, reduce the sensitivity of individual methods to specific noise or data distributions, thus improving overall robustness.

Notably, mixed methods incorporating AnomalyDiffusion show significant improvement in the pixel-level metric compared to others. AnomalyDiffusion generates abnormal samples with high quality and diversity at pixel-level, better simulating real-world anomaly data. This allows the model to learn more refined anomaly features during training, while other methods primarily focus on image-level.

3) Implications and Future Challenges: Although the diversity, complementarity, and robustness of different methods are advantages of hybrid approaches, selecting optimal combination of methods and adjusting their weights remains a complex optimization problem. The significant improvement in pixel-level performance when AnomalyDiffusion is incorporated highlights necessity of including high-quality pixel-level samples. In other words, it is worth exploring anomaly synthesis methods that operate at both image and pixel level.

This article has been accepted for publication in IEEE Transactions on Artificial Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAI.2026.3680823

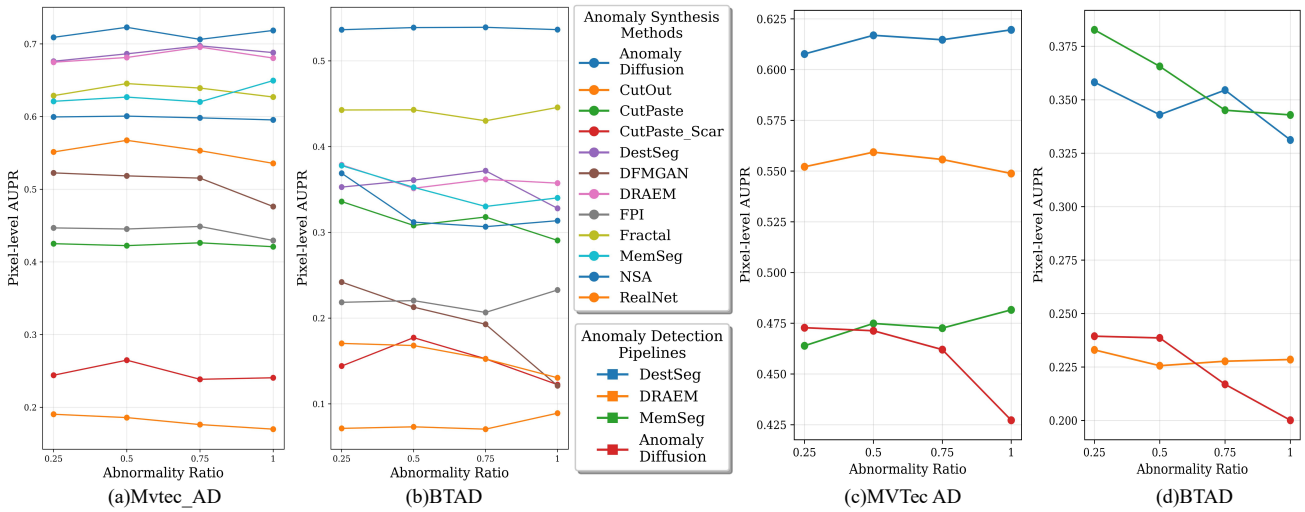


Fig. 10: Impact of Anomaly Ratio on Anomaly Detection Pipelines Performance: (a&b) pixel-level AUPR of different anomaly synthesis methods on MVTec AD and BTAD. (c&d) pixel-level AUPR of different anomaly detection pipelines on MVTec AD and BTAD. A performance degradation is often observed when the abnormality ratio is set to 1.

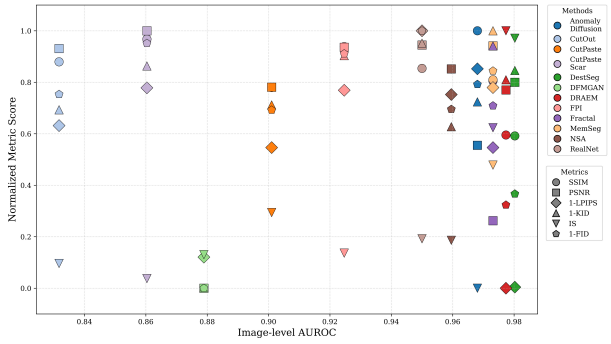


Fig. 11: Correlation analysis between generated image quality metrics (SSIM, PSNR, LPIPS, IS, KID, FID) and anomaly detection performance (image-level AUROC) on MVTec-AD.

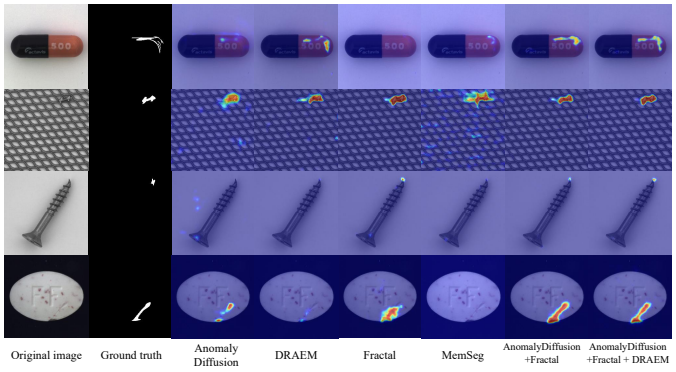


Fig. 12: Visualization comparison between single synthesis method and mixed synthesis method.

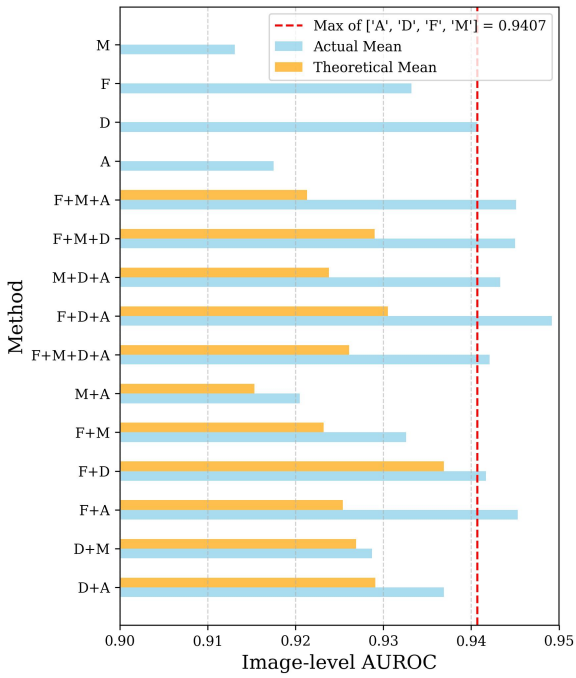


Fig. 13: Comparison of performance (image-level AUROC) using mixed synthesis augmentation methods versus individual approaches. **Blue part**: Actual performance of hybrid synthesis methods. **Yellow part**: Theoretical performance computed from individual methods' results. **Red line**: Maximum AUROC achieved by any single synthesis method.

V. Conclusion

We introduce ASBench in this paper, the first comprehensive benchmark for image anomaly synthesis, featuring 12 anomaly synthesis methods, 4 anomaly detection pipelines, and 5 industrial datasets across 4 key evaluation dimensions. Through extensive experiments, we have gained crucial insights into the performance of anomaly synthesis, including the discovery that no single method is universally optimal; the non-linear influence of synthetic data ratios, where higher proportions do not guarantee better performance; the weak correlation between conventional image quality metrics and downstream detection efficacy; and the notable performance boost from hybridizing complementary synthesis methods. On

top of these findings, we present several intriguing future lines for image anomaly synthesis. For example, future work could focus on (1) designing anomaly synthesis methods with greater generalizability to balance diversity and realism; (2) developing adaptive synthesis strategies to optimize the trade-off between the quantity and quality of synthetic samples; (3) constructing novel task-oriented evaluation metrics that can reliably predict detection performance; and (4) systematically optimizing the combination and weighting mechanisms when hybridizing multiple synthesis approaches.

REFERENCES

[1] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021.

[2] 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 (CVPR)*, June 2019.

[3] 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.

[4] Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. *arXiv preprint arXiv:2112.09045*, 2021.

[5] Yuqi Cheng, Yunkang Cao, Rui Chen, and Weiming Shen. Rad: A comprehensive dataset for benchmarking the robustness of image anomaly detection. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, pages 2123–2128. IEEE, 2024.

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

[7] Huilin Deng, Hongchen Luo, Wei Zhai, Yanming Guo, Yang Cao, and Yu Kang. Prioritized local matching network for cross-category few-shot anomaly detection. *IEEE Transactions on Artificial Intelligence*, 5(9):4550–4561, 2024.

[8] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout, 2017.

[9] Jan Diers and Christian Pigorsch. A survey of methods for automated quality control based on images. *International Journal of Computer Vision*, 131(10):2553–2581, 2023.

[10] Zongwei Du, Liang Gao, and Xinyu Li. A new contrastive gan with data augmentation for surface defect recognition under limited data. *IEEE Transactions on Instrumentation and Measurement*, 72:1–13, 2022.

[11] Yuxuan Duan, Yan Hong, Li Niu, and Liqing Zhang. Few-shot defect image generation via defect-aware feature manipulation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 571–578, 2023.

[12] Clement Fung, Chen Qiu, Aodong Li, and Maja Rudolph. Model selection of anomaly detectors in the absence of labeled validation data. *IEEE Transactions on Artificial Intelligence*, pages 1–10, 2025.

[13] John C Hart. Perlin noise pixel shaders. In *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS workshop on Graphics hardware*, pages 87–94, 2001.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[15] 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*, volume 38, pages 8526–8534, 2024.

[16] Yibin Huang, Congying Qiu, and Kui Yuan. Surface defect saliency of magnetic tile. *The Visual Computer*, 36:85 – 96, 2018.

[17] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pages 66–71, 2021.

[18] Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. In *The Thirteenth International Conference on Learning Representations*.

[19] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021.

[20] Hanxi Li, Zhengxun Zhang, Hao Chen, Lin Wu, Bo Li, Deyin Liu, and Mingwen Wang. A novel approach to industrial defect generation through blended latent diffusion model with online adaptation. *arXiv preprint arXiv:2402.19330*, 2024.

[21] Wenqiao Li, Bozhong Zheng, Xiaohao Xu, Jinye Gan, Fading Lu, Xiang Li, Na Ni, Zheng Tian, Xiaonan Huang, Shenghua Gao, et al. Multi-sensor object anomaly detection: Unifying appearance, geometry, and internal properties. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9984–9993, 2025.

[22] Jing-Xiao Liao, Bo-Jian Hou, Hang-Cheng Dong, Hao Zhang, Xiaoge Zhang, Jinwei Sun, Shiping Zhang, and Feng-Lei Fan. Quadratic neuron-empowered heterogeneous autoencoder for unsupervised anomaly detection. *IEEE Transactions on Artificial Intelligence*, 5(9):4723–4737, 2024.

[23] Jiaqi Liu, Guoyang Xie, Ruitao Chen, Xinpeng Li, Jinbao Wang, Yong Liu, Chengjie Wang, and Feng Zheng. Real3d-ad: A dataset of point cloud anomaly detection. *Advances in Neural Information Processing Systems*, 36:30402–30415, 2023.

[24] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20402–20411, 2023.

[25] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, page 01–06. IEEE, June 2021.

[26] Arian Mousakhan, Thomas Brox, and Jawad Tayyub. Anomaly detection with conditioned denoising diffusion models. In *DAGM German Conference on Pattern Recognition*, pages 181–195. Springer, 2024.

[27] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[28] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *European Conference on Computer Vision*, pages 474–489. Springer, 2022.

[29] Qingfeng Shi, Jing Wei, Fei Shen, and Zhengtao Zhang. Few-shot defect image generation based on consistency modeling. In *European Conference on Computer Vision*, pages 360–376. Springer, 2024.

[30] Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, and Bernhard Kainz. Detecting outliers with foreign patch interpolation. *Machine Learning for Biomedical Imaging*, 1(April 2022):1–27, April 2022.

[31] Jeremy Tan, Benjamin Hou, Thomas Day, John Simpson, Daniel Rueckert, and Bernhard Kainz. Detecting outliers with poisson image interpolation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 581–591. Springer, 2021.

[32] Xian Tao, Shaohua Yan, Xinyi Gong, and Chandranath Adak. Learning multiresolution features for unsupervised anomaly localization on industrial textured surfaces. *IEEE Transactions on Artificial Intelligence*, 5(1):127–139, 2024.

[33] 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*, pages 22883–22892, 2024.

[34] Xuan Xia, Weijie Lv, Xing He, Nan Li, Chuanqi Liu, and Ning Ding. Fractalad: A simple industrial anomaly detection method using fractal anomaly generation and backbone knowledge distillation. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2024.

[35] Guoyang Xie, Jinbao Wang, Jiaqi Liu, Jiayi Lyu, Yong Liu, Chengjie Wang, Feng Zheng, and Yaochu Jin. Im-iad: Industrial image anomaly detection benchmark in manufacturing. *IEEE Transactions on Cybernetics*, 2024.

[36] Minghui Yang, Peng Wu, and Hui Feng. Memseg: A semi-supervised method for image surface defect detection using differences and

commonalities. *Engineering Applications of Artificial Intelligence*, 119:105835, 2023.

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

[38] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, 2021.

[39] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dsr–a dual subspace re-projection network for surface anomaly detection. In *European conference on computer vision*, pages 539–554. Springer, 2022.

[40] 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*, pages 2524–2534, 2021.

[41] Hui Zhang, Zheng Wang, Dan Zeng, Zuxuan Wu, and Yu-Gang Jiang. Diffusionad: Norm-guided one-step denoising diffusion for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[42] Jiangning Zhang, Haoyang He, Zhenye Gan, Qingdong He, Yuxuan Cai, Zhucun Xue, Yabiao Wang, Chengjie Wang, Lei Xie, and Yong Liu. Ader: A comprehensive benchmark for multi-class visual anomaly detection. *arXiv preprint arXiv:2406.03262*, 2024.

[43] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16699–16708, 2024.

[44] Xuan Zhang, Shiyu Li, Xi Li, Ping-Chia Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3914–3923, 2022.

[45] Qiang Zhou, Weize Li, Lihan Jiang, Guoliang Wang, Guyue Zhou, Shanghang Zhang, and Hao Zhao. Pad: A dataset and benchmark for pose-agnostic anomaly detection. *Advances in Neural Information Processing Systems*, 36:44558–44571, 2023.

[46] Qiang Zhou, Weize Li, Lihan Jiang, Guoliang Wang, Guyue Zhou, Shanghang Zhang, and Hao Zhao. Pad: A dataset and benchmark for pose-agnostic anomaly detection. *Advances in Neural Information Processing Systems*, 36:44558–44571, 2023.

[47] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision*, pages 392–408. Springer, 2022.

Qunyi Zhang received her B.S. degree in Electrical and Computer Engineering & M.S. degree in Electronic Information from Shanghai Jiao Tong University, China, in 2023 and 2026 respectively. Her research interest focuses on anomaly detection.

Songan Zhang received the B.S. and M.S. degrees in automotive engineering from Tsinghua University, Beijing, China, in 2013 and 2016, respectively. In 2021, she earned the Ph.D. degree in mechanical engineering from the University of Michigan, Ann Arbor, MI, USA. Upon graduation, she joined Ford Motor Company as a Research Scientist where she made contributions to pioneering innovations in smart manufacturing and advanced driver assist systems. She is currently an Assistant Professor with the Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai, China. Her research interests include accelerated and safety evaluation of autonomous vehicles; verification methods for autonomous driving systems; model-based, trustworthy, and data-efficient reinforcement learning and meta-learning; and the application of foundation models for decision-making in autonomous vehicles and intelligent transportation systems.

Jiaqi Liu received his B.S. degree from Dalian University of Technology in 2019 and his M.S. degree from the Southern University of Science and Technology, China, in 2024. His research interests include anomaly detection and multimodal large language model training.

Jinbao Wang (Member, IEEE) received his Ph.D. degree from the University of Chinese Academy of Sciences in 2019. He is currently an Assistant Professor with the School of Artificial Intelligence, Shenzhen University, Shenzhen, China. His research interests include digital human modeling and driving, image anomaly detection, computer vision, and machine learning.

Xiaoning Lei received a Master of Science degree from Central South University. He is currently working in CATL. His research interests include computer vision and digital humans. He has published more than 10 papers in top-tier conferences(NeurIPS, InterSpeech, ICASSP).

Guoyang Xie (Member, IEEE) received his Ph.D. degree in Computer Science from University of Surrey, UK, in 2023. He is currently an Senior Algorithm Manager in Department of Intelligent Manufacturing at CATL. His current research includes AI for manufacturing and industrial image anomaly detection. He is the Associate Editor of IEEE Data Descriptions. He has published more than 28 papers in peer-reviewed top-tier conferences and journals.

Guannan Jiang (Member, IEEE) received his Ph.D. degree in Robotics & Computer Vision at University of New South Wales, Australia, in 2016. He is currently a Senior Algorithm Manager in Department of Intelligent Manufacturing at CATL. His current research includes computer vision, pattern recognition, and automation. He owns over 30 authorized patents and has authored over 40 papers in referred journals and conferences, including ACM Multimedia, IEEE CVPR, NeurIPS and AAAI.

Zhichao Lu (Member, IEEE) received his Ph.D. degree in Electrical and Computer Engineering from Michigan State University, USA, in 2020. He is currently an Assistant Professor in Department of Computer Science at City University of Hong Kong. His current research focuses on the intersections of evolutionary computation, learning, and optimization, notably on developing efficient, reliable, and automated machine learning algorithms and systems. He received the GECCO-2019 best paper award and the 2024 IEEE-CCF Cloud Computing Best Paper Award.