



Unsupervised anomaly detection with a stacked transformer diffusion reconstruction framework ^{*}

Minjie Du ^{ID a}, Hengyu Xu ^{ID a}, Fulin Shang ^{ID a}, Zheng Wang ^{ID a}, Yining Hu ^{ID a,*},
Lizhe Xie ^{ID b,c,*}

^a School of Cyber Science and Engineering, Southeast University, Nanjing, 211189, Jiangsu, China

^b State Key Laboratory Cultivation Base of Research, Prevention and Treatment for Oral Diseases, Nanjing Medical University, Nanjing, 210000, Jiangsu, China

^c Jiangsu Province Engineering Research Center of Stomatological Translational Medicine, Nanjing Medical University, Gulou District, Nanjing, 210000, Jiangsu, China

ARTICLE INFO

Keywords:

Anomaly detection
Reconstruction-based
Diffusion
Isomorphic structural knowledge guidance

ABSTRACT

Reconstruction-based anomaly detection remains appealing for industrial inspection due to its ability to handle unknown defect types, yet diffusion models still suffer from semantic drift and geometric misalignment when reconstructing anomalous inputs. This paper presents STD RAD (Unsupervised Anomaly Detection with a Stacked Transformer Diffusion Reconstruction Framework), a fully transformer-based diffusion reconstruction model for unsupervised multi-class anomaly detection. Built upon the DiT paradigm, STD RAD replaces the conventional U-Net with a scalable stacked-transformer denoiser that provides global context modeling. A lightweight Adapter module is integrated into DiT blocks to improve feature modulation and facilitate faster optimization during training. To mitigate misalignment in latent diffusion, we introduce an ISKG (Isomorphic Structural Guidance) module that injects structurally compatible cues into the denoising trajectory. A multi-scale feature alignment strategy further enhances reconstruction fidelity across both textures and large structural regions. For anomaly scoring, multi-level representations extracted by a pretrained ResNet are compared between the input and its reconstruction, and their feature discrepancies are aggregated to produce anomaly maps. Extensive experiments on the MVTec AD and VisA benchmarks show that STD RAD achieves 95.8%/96.0% image/pixel-level AUROC on MVTec AD and 86.9%/97.2% on VisA, and qualitative results confirm cleaner reconstructions and more reliable suppression of anomalous patterns across diverse industrial scenarios.

1. Introduction

Anomaly detection is a fundamental problem in computer vision, with critical applications in industrial inspection, medical imaging, and safety-sensitive monitoring (Czimmermann et al., 2020; Liu et al., 2024; Tao et al., 2022). In these scenarios, anomalies often correspond to defects, diseases, or irregularities that may seriously compromise safety, product quality, or operational efficiency. The primary goal of anomaly detection is to determine whether an input image contains abnormal patterns and, if so, to precisely localize the affected regions. Compared with conventional recognition tasks where abundant annotated data define clear class boundaries, anomaly detection is uniquely challenging because anomalies are inherently rare, diverse, and often unknown during training. Models must therefore learn the distribution of normal data

and identify deviations without relying on exhaustive abnormal samples.

A variety of strategies have been proposed to address this challenge, which can be broadly categorized into three main paradigms: reconstruction-based methods (Tang et al., 2020; Zavrtnik et al., 2021b,c), feature based methods (Wan et al., 2021), and synthesis-based methods (Li et al., 2021; Xu et al., 2025).

Reconstruction-based methods rely on generative models trained solely on normal samples to capture the distribution of anomaly-free data. During inference, abnormal inputs are projected into the learned distribution, and reconstructed outputs are compared against the original inputs to reveal anomalies. These methods can be further divided into unsupervised (Yan et al., 2021) and supervised (or semi-supervised) (Akçay et al., 2018) approaches. In the unsupervised setting, models

^{*} This work was supported in part by the National Key Research and Development Program of China under Grant (2023YFC3010302), National Natural Science Foundation of China (Grant No. 82571165), and Key R&D Program of Jiangsu Province (BE2023836).

^{*} Corresponding authors.

E-mail addresses: duminjie@seu.edu.cn (M. Du), 220255610@seu.edu.cn (H. Xu), 220255590@seu.edu.cn (F. Shang), fiki@seu.edu.cn (Z. Wang), hyn.list@seu.edu.cn (Y. Hu), xielizhe@njmu.edu.cn (L. Xie).

<https://doi.org/10.1016/j.eswa.2026.131764>

Received 15 December 2025; Received in revised form 4 February 2026; Accepted 17 February 2026

Available online 20 February 2026

0957-4174/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

such as autoencoders (Chung et al., 2020; Liu & Wang, 2024; Liu et al., 2021), variational autoencoders, or GAN (Generative adversarial network) (Duan et al., 2023; Fang et al., 2025) are optimized with only normal samples. They are expected to fail when reconstructing unseen anomalies, leading to detectable residual errors. While appealing due to their annotation-free nature, these approaches struggle when defects cover large regions. They also tend to fail when anomalous patterns closely resemble normal structures. In the supervised or semi-supervised setting (Cao et al., 2025; Wang et al., 2025), a small number of annotated anomalies are incorporated during training. This additional guidance helps models learn to explicitly suppress or replace defective regions with plausible normal content. Although this improves reconstruction fidelity and localization accuracy, it requires costly labeling and may generalize poorly to unseen categories of anomalies.

Feature-based methods operate at the representation level. Instead of reconstructing images, they extract embeddings from pre-trained networks (He et al., 2016; Simonyan & Zisserman, 2014), or teacher-student architectures (Bergmann et al., 2020; Salehi et al., 2021; Wang et al., 2021), and anomalies are detected by measuring deviations from normal distributions in feature space. These methods often achieve high accuracy in image-level detection tasks. However, their reliance on pre-trained networks which are typically optimized on natural images, introduces a domain gap in industrial scenarios. As a result, they may fail to capture fine-grained structural defects and often produce inaccurate anomaly localization. Beyond conventional feature-based approaches, recent studies have explored interpretable and hybrid feature representations by combining deep neural networks with structured descriptors, such as quaternion- and moment-based features, for vision-related recognition and diagnosis tasks (El Ogri et al., 2024, 2026, 2021).

Synthesis-based methods attempt to address the scarcity of anomaly data by generating pseudo anomalies during training. They employ strategies such as cut-and-paste operations, texture perturbations, or external datasets to simulate defective regions, so that models can learn a decision boundary between normal and abnormal patterns (Hu et al., 2024; Liu et al., 2019, 2023). However, synthetic anomalies cannot fully capture the diversity of real defects, and the gap between synthetic and real anomalies may limit generalization.

Beyond these paradigms, reconstruction-based anomaly detection methods remain the most general and practical solution for real-world applications. Since they rely solely on normal samples to learn the intrinsic distribution of defect-free data, these models can naturally adapt to unseen or unknown anomaly types without requiring prior knowledge or manual labeling. This property makes them highly scalable and robust for complex industrial environments, where the types and appearances of defects are unpredictable and continuously evolving. By reconstructing each input toward its normal counterpart, reconstruction-based frameworks provide a unified and interpretable mechanism for detecting diverse anomalies across various categories, production lines, and domains, thus exhibiting strong generalization and deployment potential in real-world manufacturing scenarios.

Despite the progress achieved by these paradigms, existing methods still struggle to achieve robust anomaly localization in complex, multi-class industrial settings. Reconstruction methods often fail to maintain semantic or geometric consistency between the input and reconstruction. In real industrial inspection scenarios, anomalies exhibit diverse characteristics in both form and scale. Structural anomalies, such as geometric deformation, orientation changes, or missing components, are particularly challenging for reconstruction-based methods because even small semantic or geometric deviations can invalidate pixel-wise comparison. Textural anomalies and small-scale defects, including scratches, stains, or subtle surface irregularities, require high-fidelity reconstruction and sensitive localization. STDRAD is designed to address these challenges by combining structure-aware diffusion reconstruction with explicit guidance for geometric consistency and multi-scale feature comparison, enabling robust detection across both structural and textural anomaly types. In the context of diffusion based reconstruction, we use

the term semantic drift to describe cases where the reconstructed image deviates from the semantic identity of the input (e.g., category ambiguity or incorrect object orientation), even though the overall appearance remains plausible. We use geometric misalignment to denote inconsistencies in spatial structure, such as shifts in object position, rotation, or deformation between the input image and its reconstruction. These phenomena are particularly problematic for anomaly detection, as they undermine reliable pixel-wise comparison between the input and reconstructed images. In this paper, we use the term structural fidelity to refer to the preservation of global geometry and spatial layout between the input image and its reconstruction, including object shape, orientation, and relative part configuration. We use semantic alignment to denote the consistency of object identity and high-level semantic attributes during reconstruction, ensuring that the reconstructed image remains semantically compatible with the input rather than exhibiting category drift or geometric reorientation.

Feature-based methods are hindered by domain gaps and limited transferability. Synthesis-based methods, though useful, cannot fully represent the unpredictability of anomalies in practice. These limitations underscore the need for a framework that combines the high-fidelity reconstruction power of generative models with explicit mechanisms for semantic alignment and anomaly suppression.

Recently, diffusion models have emerged as a promising generative framework (Ho et al., 2020; Song et al., 2020). Their iterative denoising formulation allows for photo-realistic and semantically coherent reconstructions across diverse domains. Unlike traditional generative models that directly map latent codes to images, diffusion models gradually corrupt data with noise and learn to reverse this process step by step. This progressive refinement makes them highly effective at modeling complex structures and fine details. Intuitively, this property makes diffusion models suitable for anomaly detection, where the task is to reconstruct normal content while suppressing anomalies.

However, applying diffusion models directly to anomaly detection is problematic. When reconstructing inputs containing defects, conventional diffusion models frequently produce results with semantic misalignment or geometric inconsistency, such as rotations, reorientations, or texture shifts. These inconsistencies undermine anomaly detection, since meaningful comparison requires strict alignment between input and reconstruction. Consequently, naive diffusion approaches cannot be directly adopted for anomaly localization, particularly in multi-class industrial contexts.

To address this, guidance mechanisms are essential (He et al., 2024; Li et al., 2025). By constraining the denoising trajectory, guidance ensures that reconstructed images remain consistent with the input in both semantics and geometry, while abnormal regions are replaced with normal counterparts. This enables reconstructions to serve as reliable references for residual-based anomaly detection.

At the same time, the architectural foundation of diffusion models is evolving. Most existing diffusion approaches employ U-Net backbones, sometimes augmented with self-attention or lightweight transformer blocks, but still fundamentally organized in an encoder-decoder convolutional structure. In contrast, the Diffusion Transformer (DiT) (Peebles & Xie, 2023) paradigm represents a departure from this design. Instead of combining convolution with attention, DiT replaces the entire U-Net (Perez et al., 2018; Ronneberger et al., 2015; Van den Oord et al., 2016) with a stacked sequence of transformer blocks operating on patchified latent tokens. This fully transformer-based backbone provides a principled and scalable architecture, offering stronger global context modeling, consistent training dynamics, and favorable scaling properties, where increased model capacity directly improves generative quality.

While discriminative approaches (e.g., feature embedding methods like Dinomaly Guo et al., 2025) have recently achieved near-saturated performance on standard benchmarks, they often function as 'black boxes,' outputting anomaly scores without providing a visual explanation of the normal state. In contrast, reconstruction-based methods

offer intrinsic explainability by generating a defect-free reference. However, existing reconstruction models struggle to balance fidelity with structural consistency. Our work prioritizes this structural fidelity, arguing that in high-stakes industrial applications, the ability to generate a geometrically accurate reference image is as critical as the binary detection score itself.

Building upon these observations, we introduce Unsupervised Anomaly Detection with a Stacked Transformer Diffusion Reconstruction Framework (STD RAD), a guided latent diffusion architecture built entirely upon the DiT paradigm. Reconstruction-based methods remain highly general for industrial inspection, as they do not rely on predefined anomaly categories and can naturally accommodate a wide variety of defect types by restoring each input toward its normal counterpart. Motivated by this generality, STD RAD adopts a fully stacked transformer denoiser, replacing the conventional U-Net with a scalable architecture capable of global context modeling and structurally coherent latent refinement.

To address the semantic drift and geometric misalignment commonly observed in diffusion reconstruction, we introduce an Isomorphic Structural Knowledge Guidance (ISKG) module that is topologically isomorphic to the main denoiser. This structural homogeneity makes the guidance features compatible with the backbone in token organization and feature scale, enabling stable injection and improved semantic/geometric consistency along the reverse trajectory. In addition, a multi-scale feature alignment strategy further enhances reconstruction fidelity across both fine-grained textures and large structural defects, enabling reliable residual-based anomaly localization.

We adopt the single-model multi-class unsupervised anomaly detection setting, where a single model is jointly trained across all categories using only normal data. Category identifiers are used solely to distinguish different normal data distributions within the unified model and do not introduce semantic or anomaly-related supervision.

To facilitate understanding, the remainder of this paper is organized as follows. Section 2 reviews diffusion models and related anomaly detection paradigms. Section 3 details the proposed STD RAD framework, including the DiT-based latent denoiser, adapter mechanisms, Isomorphic Structural Knowledge Guidance, and the feature-space scoring module. Section 4 presents implementation details, datasets, evaluation metrics, and comprehensive experiments including ablation studies. Section 5 discusses the limitations of the proposed method and analyzes representative failure cases. Section 6 concludes the paper.

2. Related work

2.1. DDPM and DDIM

Recent advances in diffusion models demonstrate strong image synthesis across diverse domains. The Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) formulates generation as a two-stage process with a Markovian (Chung, 1967) forward corruption and a learned Gaussian reverse denoising. At each step, Gaussian noise is gradually injected into the clean data through a Markov chain, and a neural network is trained to invert this corruption by predicting and removing the noise.

While this iterative refinement yields high-fidelity samples, it typically requires hundreds or even thousands of denoising steps, leading to slow inference.

To alleviate this issue, the Denoising Diffusion Implicit Model (DDIM) (Song et al., 2020) reformulates the reverse process in a non-Markovian manner.

By introducing a deterministic trajectory that directly maps the noisy sample at step t to its predecessor at step $t-1$ using the same noise-prediction network, DDIM eliminates the stochastic sampling term of DDPM. This modification significantly reduces the number of required steps while maintaining perceptual quality, thus making diffusion mod-

els more practical for downstream tasks such as image restoration, reconstruction, and anomaly detection.

The following presents the formal expression of the DDPM and DDIM method. Let $\{\beta_t\}_{t=1}^T$ be a variance schedule, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$.

DDPM: forward (noising).

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1}, \beta_t I), \quad t = 1, \dots, T, \quad (1)$$

with the closed-form marginal

$$\begin{aligned} q(x_t | x_0) &= \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) I) \\ \Leftrightarrow x_t &= \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \end{aligned} \quad (2)$$

DDPM: reverse (denoising). The learned reverse transition is Gaussian with mean expressed via a noise predictor ϵ_θ :

$$\begin{aligned} p_\theta(x_{t-1} | x_t) &= \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I), \\ \mu_\theta(x_t, t) &= \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right), \end{aligned} \quad (3)$$

$$\sigma_t^2 = \tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t.$$

A widely used training objective is the simplified noise-matching loss:

$$\begin{aligned} \mathcal{L}_{\text{simple}}(\theta) &= \mathbb{E}_{t, x_0, \epsilon \sim \mathcal{N}(0, I)} \left[\left\| \epsilon - \epsilon_\theta(x_t, t) \right\|_2^2 \right], \\ \text{where } x_t &= \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon. \end{aligned} \quad (4)$$

DDIM: non-Markovian reverse with consistent parameterization. DDIM keeps the *same* forward process (1)–(2) and the *same* noise-prediction network ϵ_θ , but replaces the Markovian reverse kernel (3) by a non-Markovian update driven by the predicted clean image

$$\hat{x}_0(x_t, t) = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t, t) \right). \quad (5)$$

The DDIM step writes

$$\begin{aligned} x_{t-1} &= \sqrt{\bar{\alpha}_{t-1}} \hat{x}_0(x_t, t) \\ &\quad + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \epsilon_\theta(x_t, t) \\ &\quad + \sigma_t \xi, \end{aligned} \quad (6)$$

$$\text{where } \xi \sim \mathcal{N}(0, I), \quad \sigma_t = \eta \sqrt{\tilde{\beta}_t}, \quad \eta \geq 0.$$

where the posterior variance $\tilde{\beta}_t$ is the *same* as in DDPM (3). Setting $\eta = 0$ yields a deterministic sampler (no stochastic term) that preserves the DDPM marginals; choosing $\eta = 1$ recovers the DDPM sampling variance, thus DDIM is an acceleration that modifies only the reverse dynamics while keeping all symbols/schedules consistent with DDPM.

Notation. $x_0 \in \mathbb{R}^{H \times W \times C}$ is the clean image; x_t its noisy version at step t ; T the number of steps; $\beta_t \in (0, 1)$, $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$; I identity; $\epsilon \sim \mathcal{N}(0, I)$; $\epsilon_\theta(\cdot, t)$ the noise-prediction network; $\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$; η controls stochasticity in DDIM.

2.2. Latent diffusion models

Latent Diffusion Models (LDM) (Blattmann et al., 2022; Rombach et al., 2022) were proposed to address the high computational cost of performing diffusion directly in pixel space. The key idea is to employ a variational autoencoder (VAE) (Kingma & Welling, 2013) to compress images into a lower-dimensional latent representation, where both the forward noising and reverse denoising processes are conducted. In practice, LDM adopt the *same* DDPM/DDIM parameterization as in pixel space, but execute it entirely within the compact latent domain. After denoising, the latent variables are decoded back into the pixel

domain through the VAE decoder, which significantly reduces memory and computation requirements while preserving perceptual quality. In addition, LDM often incorporate conditioning mechanisms such as cross-attention, which enable flexible control with class labels or text prompts during denoising. By shifting the diffusion process into latent space, LDM achieve substantial acceleration and make large-scale, high-resolution generative modeling feasible.

The following presents the formal expression of the LDM method. Let \mathcal{E}_φ and D_φ denote the encoder and decoder.

Latent mapping and reconstruction.

$$z_0 = \mathcal{E}_\varphi(x_0), \quad \hat{x}_0 = D_\varphi(z_0) \approx x_0, \quad (7)$$

where $z_0 \in \mathbb{R}^d$ is the latent. Diffusion then proceeds in latent space with the *same* schedule $\{\beta_t, \alpha_t, \bar{\alpha}_t\}$.

LDM: forward and marginal in latent space.

$$\begin{aligned} q(z_t | z_{t-1}) &= \mathcal{N}(z_t; \sqrt{\alpha_t} z_{t-1}, \beta_t I), \\ q(z_t | z_0) &= \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t} z_0, (1 - \bar{\alpha}_t) I). \end{aligned} \quad (8)$$

LDM: reverse in latent space(DDPM-form).

$$\begin{aligned} p_\theta(z_{t-1} | z_t, c) &= \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t, c), \sigma_t^2 I), \\ \mu_\theta(z_t, t, c) &= \frac{1}{\sqrt{\alpha_t}} \left(z_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(z_t, t, c) \right), \end{aligned} \quad (9)$$

$$\sigma_t^2 = \tilde{\beta}_t.$$

where c is an optional condition injected via cross-attention.

LDM: reverse in latent space (DDIM-form). Using the same \hat{x}_0 -style reconstruction in latent space,

$$\hat{z}_0(z_t, t, c) = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(z_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(z_t, t, c) \right), \quad (10)$$

the DDIM update becomes

$$\begin{aligned} z_{t-1} &= \sqrt{\bar{\alpha}_{t-1}} \hat{z}_0(z_t, t, c) \\ &\quad + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \epsilon_\theta(z_t, t, c) \\ &\quad + \sigma_t \xi, \end{aligned} \quad (11)$$

$$\text{where } \xi \sim \mathcal{N}(0, I), \quad \sigma_t = \eta \sqrt{\tilde{\beta}_t}, \quad \eta \geq 0.$$

After denoising to z_0 , decode to pixels:

$$\hat{x}_0 = D_\varphi(z_0). \quad (12)$$

Notation. z_t is the latent at step t ; $\epsilon_\theta(\cdot, t, c)$ is the *same-form* noise predictor as in pixel space but operating on latents; $\{\beta_t, \alpha_t, \bar{\alpha}_t, \tilde{\beta}_t\}$ are *identical* schedules/definitions to those in DDPM/DDIM; η plays the same role controlling stochasticity.

2.3. Anomaly detection with reconstruction models

Reconstruction-based approaches remain one of the most widely used strategies for anomaly detection. The central idea is to train generative models only on normal samples, such that anomalous regions cannot be faithfully reconstructed and become detectable by comparing input and reconstruction. Representative works include RIAD (Zavrtanik et al., 2021c), which employs iterative inpainting to progressively restore occluded regions and highlight anomalies, and GaNomaly (Akcay et al., 2018), which integrates adversarial training with an encoder-decoder structure to capture distributional deviations between normal and abnormal inputs. UTRAD (Chen et al., 2022) further extends this

paradigm by introducing a two-stage reconstruction pipeline that leverages high-fidelity generation to mitigate reconstruction errors on complex textures. These approaches demonstrate strong performance in unsupervised settings, but they often fail when anomalies occupy large regions or share close resemblance to normal structures.

In this context, diffusion based frameworks have also been explored for anomaly detection. For instance, AnoDDPM (Wyatt et al., 2022) was the first to introduce diffusion models into medical anomaly detection, while subsequent methods such as DiffusionAD (Zhang et al., 2025) and DDAD (Mousakhan et al., 2024) combined synthetic anomalies or pre-trained score models for industrial applications. More recently, DiAD (He et al., 2024) proposed to incorporate semantic guidance into the denoising process of diffusion models, enabling the reconstruction of anomalous regions while preserving the original semantic information. Beyond such hybrid U-Net transformer backbones, there is considerable potential in adopting fully stacked Transformer architectures for diffusion denoising, as exemplified by the DiT paradigm. Compared with U-Net-style designs, DiT replaces the encoder-decoder hierarchy with a sequence of Transformer blocks that operate directly on latent tokens. This structure provides stronger global context modeling, more stable training dynamics, and favorable scalability, since increasing model depth or width consistently improves generative quality. These architectural advantages suggest that DiT-based denoising networks, although still relatively underexplored in anomaly detection, offer a promising direction for advancing reconstruction fidelity and robustness in multi-class industrial scenarios.

3. Method

We propose a DiT-based latent diffusion framework for multi-class anomaly detection. Our pipeline consists of four components: (1) a latent diffusion model operating on VAE latents with a fully stacked DiT denoiser, (2) an adapter augmented denoising backbone for parameter efficient fine tuning, (3) an ISKG interface for enforcing semantic and geometric consistency during denoising, and (4) a feature space anomaly scoring module that compares the input and the reconstruction using a frozen pretrained ResNet. An overview is given in Fig. 1.

Motivation. Reconstruction-based anomaly detection is founded on the principle that models trained only on normal data should fail to faithfully reproduce abnormal patterns. However, conventional reconstruction-based methods such as autoencoders or GAN-based models typically learn a single-pass mapping from an input image to its reconstruction. This direct reconstruction paradigm tends to preserve input-specific details and may undesirably reproduce anomalous regions, especially when anomalies are visually plausible or occupy large spatial areas, which weakens anomaly suppression.

Diffusion-based reconstruction follows a fundamentally different paradigm based on a noising–denoising process. During the forward diffusion stage, the input is progressively corrupted by noise, which systematically removes fine-grained and input-specific details, including abnormal patterns that are not supported by the normal-data distribution. The reverse denoising process is then learned exclusively from normal samples to predict the denoising direction that restores signals consistent with the normal manifold. As a result, diffusion reconstruction can be interpreted as a gradual projection toward the normal data manifold. This behavior naturally aligns with the objective of unsupervised anomaly detection, where anomalies should be suppressed rather than directly memorized.

At the same time, effective anomaly detection requires the reconstruction process to remain semantically and geometrically consistent with the input, so that deviations can be reliably attributed to anomalies rather than reconstruction artifacts. Stacked Transformer denoisers, as adopted in the DiT architecture, operate on a unified latent token space

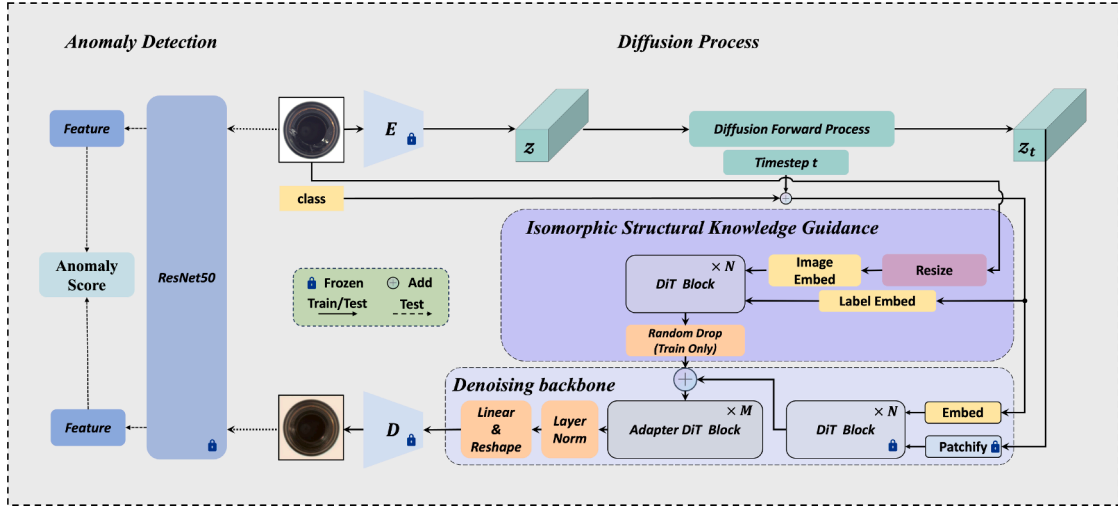


Fig. 1. Overview of the proposed STDRAD framework. The input RGB image x is encoded into a latent representation z , to which Gaussian noise perturbation is applied, yielding the noised latent representation z_t through a forward diffusion process. The corrupted latent representation is denoised by a linearly stacked Transformer-based DiT architecture. Simultaneously, a lightweight ISKG module extracts guiding features from x , which are injected after the N -th DiT block to facilitate the denoising process. The recovered latent representation \hat{z}_t is decoded back to the image space to produce the reconstructed image \hat{x} . Both x and \hat{x} are fed into a frozen ResNet-50 network to extract multi-scale features, and their differences are used to compute the final anomaly scores.

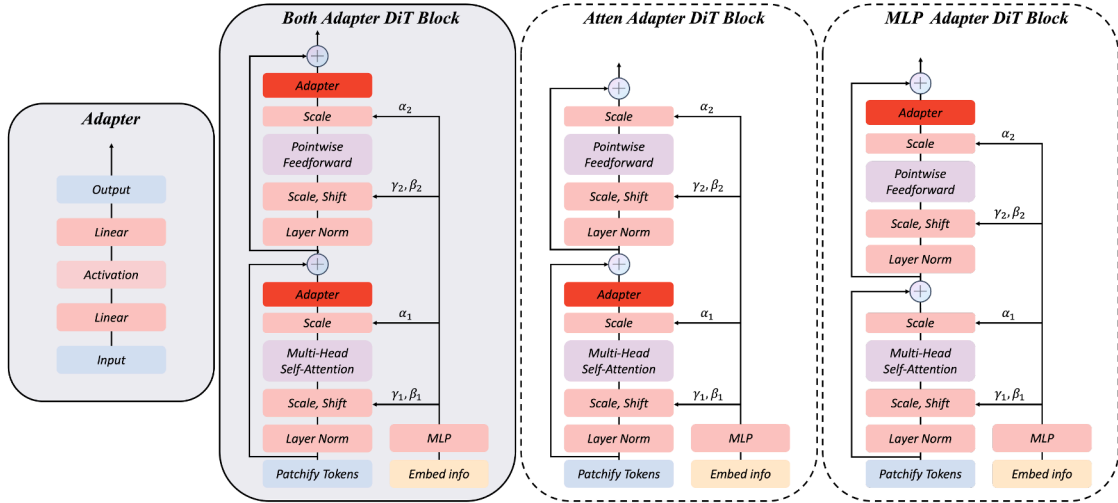


Fig. 2. Illustration of three different Adapter insertion configurations. From right to left: (1) MLP Adapter DiT Block - Adapter inserted only after the MLP module; (2) Atten Adapter DiT Block - Adapter inserted only after the attention module; (3) Both Adapter DiT Block - Adapters inserted after both the MLP and attention modules. The detailed architecture of the Adapter module is provided on the left.

and perform denoising through repeated self-attention-based refinement. This design allows structural information to be propagated consistently across diffusion steps and layers, reducing unpredictable geometric drift during reconstruction. By combining the anomaly-suppressing property of diffusion with the stable and consistent latent refinement enabled by stacked Transformer denoisers, the proposed framework provides a principled reconstruction mechanism that is well suited for unsupervised anomaly detection.

3.1. Denoising setup

Given an input image $x_0 \in \mathbb{R}^{H \times W \times C}$, we use a VAE encoder \mathcal{E}_ϕ to obtain a latent $z_0 = \mathcal{E}_\phi(x_0)$ and perform diffusion in latent space (see Section 2). Let $\{\beta_t\}_{t=1}^T$, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ denote the variance schedule. As in Latent Diffusion Models (LDM), we define the forward noising $q(z_t | z_{t-1}) = \mathcal{N}(\sqrt{\alpha_t} z_{t-1}, \beta_t I)$ and the marginal $q(z_t | z_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} z_0, (1 - \bar{\alpha}_t) I)$. During reverse denoising, our DiT predicts the noise $\epsilon_\theta(z_t, t, c)$, optionally conditioned on c (class labels).

3.2. DiT-based latent denoiser

Backbone. The denoising backbone adopts the standard Diffusion Transformer architecture and replaces the hierarchical convolutional structure of U-Net with a transformer-based design in latent space. The latent tensor is partitioned into non-overlapping patches. Each patch is projected into a token. The token sequence is equipped with deterministic positional embeddings to preserve spatial order. The diffusion timestep is encoded into a vector. The input category is encoded into another vector. The conditioning vector is obtained by combining the timestep encoding and the category encoding through linear transformations. The conditioning vector is injected into every transformer block through adaptive layer normalization.

The backbone contains multiple transformer blocks. Each block includes a multi-head self-attention module and a feed-forward network. The self-attention module constructs global interactions among all latent tokens. The feed-forward network refines token representations through nonlinear transformations. The repeated application of these

blocks yields progressive refinement of the noisy latent representation. The refinement process provides global coherence and local consistency without a multi-resolution pathway.

The stacked transformer backbone predicts the denoising direction at each reverse diffusion step. The prediction is fully determined by the conditioning vector and the latent tokens. The backbone operates in latent space and provides the representational capacity required for high-fidelity diffusion reconstruction.

Training objective. We adopt the simplified noise-matching loss in latent space:

$$\mathcal{L}_{\text{diff}}(\theta) = \mathbb{E}_{t, x_0, \epsilon \sim \mathcal{N}(0, I)} \left[\left\| \epsilon - \epsilon_\theta(z_t, t, c) \right\|_2^2 \right], \quad (13)$$

where $z_t = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t} \epsilon$, $z_0 = \mathcal{E}_\varphi(x_0)$.

with $z_0 = \mathcal{E}_\varphi(x_0)$ from normal training images only. The VAE ($\mathcal{E}_\varphi, D_\varphi$) is frozen.

The proposed ISKG-based alignment does not introduce any additional alignment loss term. The denoiser is optimized only by $\mathcal{L}_{\text{diff}}$ in latent space; alignment is implemented as latent-feature guidance injection inside the denoising network (Section 3.4), while the training objective remains unchanged.

3.3. Adapter DiT block

Although our objective is for the model to generate defect-free versions of input samples, this task remains a downstream image generation problem, where the Adapter mechanism can be leveraged for efficient finetuning. We therefore introduce a lightweight Adapter module consisting of a two-layer MLP with GELU activation. This module can accelerate the convergence of the diffusion process. It projects the input features to a reduced hidden dimension, applies a nonlinearity, and then projects them back to the original dimension. The Adapter structure can greatly accelerate model convergence and reduce training time. It is important to note that the Adapter structure described here is distinct from the adaptive layer normalization (adaLN) used within the DiT blocks. By incorporating the Adapter mechanism, our model can effectively reuse the pretrained DiT weights on the ImageNet dataset, enabling knowledge transfer from large-scale natural image distributions to the industrial anomaly reconstruction task. We design three different Adapter insertion positions. The structure of the Adapter is shown in Fig. 2.

3.4. ISKG for alignment

The ISKG network serves as a dedicated guidance extractor to provide structurally consistent cues for the denoising backbone. Alignment in our paper refers to feature-scale alignment in the latent token space. Specifically, ISKG is designed to be topologically compatible with the main DiT denoiser (same tokenization and block-style operations), so that its output guidance features match the token resolution and feature dimensionality of the backbone activations.

In practice, the input image x_0 is resized to $32 \times 32 \times 3$ to match the token scale used by the DiT backbone. It is then patchified into tokens to produce guidance features, which are additively injected into the corresponding DiT blocks during denoising. ISKG encodes multi-scale contextual representations that preserve normal structural priors while suppressing the influence of potential anomalies in the input.

The extracted guidance features are injected into the denoising backbone at corresponding layers through additive modulation. This mechanism ensures that the diffusion process restores images along semantically and geometrically consistent directions, allowing the network to reconstruct the input toward its normal appearance while attenuating abnormal regions.

We term our guidance module ‘Isomorphic’ because it mirrors the topological structure of the primary denoiser. This design choice is

grounded in the principle of feature alignment. By ensuring that the guidance features share the same tokenization, dimensionality, and attention dynamics as the latent space features, we minimize the information loss typically associated with cross-architecture fusion (e.g., CNN guiding ViT). This allows for the seamless injection of structural priors, effectively treating the input image’s geometry as an immutable knowledge constraint during the diffusion process. In essence, ISKG does not generate explicit noise predictions but instead enhances the denoiser’s capacity to recover correct orientations and shapes, thereby improving the fidelity and robustness of diffusion-based reconstruction. The primary role of ISKG is to constrain the diffusion reconstruction to follow the same global direction and orientation as the input image, rather than to introduce new semantic conditions or alter the content being generated.

To further improve representational compatibility in latent space, we initialize ISKG by copying the weights from the first N DiT blocks of the pretrained backbone. At the beginning of finetuning, these first N backbone blocks are frozen, and we ablate different N values in Table 7. This protocol is used to stabilize the aligned feature scales between ISKG and the backbone without changing the diffusion loss.

3.5. Anomaly detection

After denoising to \hat{z}_0 , we decode a reconstruction $\hat{x}_0 = D_\varphi(\hat{z}_0)$, where x_0 denotes the input image and \hat{x}_0 denotes its reconstructed counterpart obtained from the predicted clean latent. To robustly quantify abnormality, we compare representations of the input and the reconstruction, extracted by a frozen, pretrained ResNet \mathcal{F} . This reconstruction comparison mechanism also provides an intrinsic form of interpretability, as the resulting anomaly maps explicitly indicate which regions contribute to the anomaly decision.

Multi-layer features. Let $\{\mathcal{E}_k\}_{k=1}^K$ denote K intermediate layers (e.g., res1, res2, res3) of \mathcal{F} . We compute feature maps $F_k = \mathcal{F}^{(\mathcal{E}_k)}(x_0)$ and $\hat{F}_k = \mathcal{F}^{(\mathcal{E}_k)}(\hat{x}_0)$, each of shape $H_k \times W_k \times C_k$. We then define patch-wise distances $D_k \in \mathbb{R}^{H_k \times W_k}$:

$$D_k(i, j) = 1 - \frac{\langle F_k(i, j, :), \hat{F}_k(i, j, :) \rangle}{\|F_k(i, j, :)\|_2 \|\hat{F}_k(i, j, :)\|_2}. \quad (14)$$

Pixel-level evaluation. For quantitative assessment at the pixel level, the Area Under the Receiver Operating Characteristic curve (AUC) is employed as the primary metric. For each test image, the predicted anomaly map and its corresponding binary ground-truth mask are flattened into one-dimensional vectors. All pixel-wise predictions and labels are concatenated across the entire test set to form a unified evaluation pool. The false positive rate and true positive rate are then computed under varying thresholds to construct the ROC curve, and the final AUC value is reported as the pixel-level anomaly score. Pixel-level AUROC is computed following the standard ROC formulation. When the computed AUROC is below 0.5, this indicates that the predicted anomaly scores are globally inverted with respect to the ground-truth labels. In this case, we apply a symmetric adjustment, which corresponds to a category-level polarity correction. This operation is applied uniformly to all predictions of a category and preserves the relative ordering of anomaly scores.

Image-level evaluation. For quantitative assessment at the image level, we utilize a global-maximum aggregation strategy consistent with the implementation of EvalImageMax. For image-level evaluation, we adopt the EvalImageMax protocol. Each pixel-wise anomaly map is smoothed by applying 8 iterations of average pooling with a kernel size of 8×8 and stride 1. After smoothing, global max pooling is applied to obtain the final image-level anomaly score. These parameters are fixed for all datasets and experiments. This strategy provides a robust and consistent measure of image-level abnormality across different datasets and defect types.

Table 1

Comparison with state-of-the-art methods on MVTec AD dataset, with all metrics reported as $AUROC_{cls}/AUROC_{seg}$.

	Category	DiT	UTRAD	UniAD	DRAEM	DiAD	STDRAD
Objects	Bottle	65.1/72.9	98.2/95.6	99.5/97.8	97.5/87.6	99.8/98.3	100.0/98.4
	Cable	54.9/82.2	90.3/93.5	77.5/93.3	57.8/71.3	90.5/95.7	87.8/94.3
	Capsule	56.0/82.4	68.7/93.5	66.9/96.7	65.3/50.5	78.3/96.5	88.6/97.7
	Hazelnut	81.6/93.9	91.8/ 98.4	99.3/97.4	93.7/96.9	97.3/97.9	99.5/98.1
	Metal-nut	53.0/66.9	58.5/71.6	95.4/91.1	72.8/62.2	97.0/ 96.7	97.8/96.0
	Pill	57.3/63.1	78.2/95.2	73.1/90.5	82.2/94.4	86.8/93.9	91.9/96.1
	Screw	60.6/84.3	80.1/89.5	55.1/93.9	92.0/95.5	84.5/98.0	90.4/ 98.2
	Toothbrush	52.5/81.5	55.0/28.9	90.0/97.7	90.6/97.7	95.6/ 98.7	99.2/98.6
	Transistor	67.7/81.9	74.0/77.5	90.3/ 95.7	74.8/64.5	99.6/93.1	94.2/86.6
	Zipper	62.7/63.7	95.2/96.9	90.0/94.2	98.8/98.3	95.9/95.5	94.4/95.2
Textures	Carpet	90.4/91.7	95.3/96.9	100.0/98.7	98.0/98.6	98.6/ 98.9	98.3/ 98.9
	Grid	64.2/51.5	96.8/97.4	96.7/94.6	99.3/98.7	91.0/94.6	97.7/97.0
	Leather	88.2/95.3	99.4/99.0	100.0/99.1	98.7/97.3	96.8/96.2	98.7/98.7
	Tile	85.6/84.2	94.3/92.9	97.4/90.2	99.8/98.0	98.6/93.2	99.4/93.0
	Wood	93.6/86.2	98.2/91.4	97.3/93.4	99.8/96.0	98.9/92.3	99.1/92.7
	Mean	68.9/78.8	84.9/87.9	88.6/95.0	88.1/87.2	93.9/ 96.0	95.8/96.0

4. Experiment

4.1. Implementation details

All experiments are conducted on RGB input images with a spatial resolution of $256 \times 256 \times 3$. A frozen VAE is used to map between the pixel and latent domains, and the vae-ft-mse version released by HuggingFace is adopted without finetuning. The denoising backbone follows the standard DiT-XL/2 configuration with official pretrained weights, where each DiT block consists of multi-head self-attention and a two-layer feed-forward network with GELU activation. No dropout is applied inside the DiT attention or feed-forward layers, consistent with the original DiT design. Lightweight Adapter modules are implemented as two-layer MLPs with GELU activation and are inserted into the DiT blocks as described in Section 3.2.

The ISKG module shares the same transformer-style building blocks as the denoising backbone. During training, a dropout operation with a rate of 0.1 is applied to the guidance features before they are injected into the denoising backbone, randomly discarding 10% of the guidance information. This dropout is used only during training to prevent over-reliance on guidance signals and is disabled during inference.

For MVTec AD and VisA, the model is fine-tuned for 400 epochs with a fixed learning rate of 1×10^{-5} . No adaptive optimizer is used. The training batch size is set to 64. No data augmentation is applied during training, as the model is trained solely on normal samples. All experiments are conducted with a fixed random seed set to 22 to ensure deterministic behavior.

The diffusion process follows the standard deterministic DDIM formulation with $T = 10$ sampling steps and a linear noise schedule. The ISKG module contains five DiT blocks ($N = 5$). Training is performed jointly across all categories within a single unified model. For anomaly detection, a pretrained ResNet-50 model trained on ImageNet is used as the feature extractor. Feature maps from the `res2`, `res3`, and `res4` stages are selected to compute anomaly scores, and cosine distance is used as the similarity metric. For anomaly map post-processing, iterative average pooling with fixed settings is applied to improve spatial consistency, followed by bilinear interpolation to upsample anomaly maps to the input resolution. All diffusion-related hyperparameters follow standard settings commonly adopted in latent diffusion and DiT-based models, rather than being tuned for specific datasets. Architectural hyperparameters are fixed across experiments and selected based on the ablation studies in Section 4.4. All dataset preprocessing steps follow the official protocols of the MVTec AD and VisA benchmarks. All experiments are conducted on a single NVIDIA H100 GPU.

Table 2

Comparison of model size and computational cost with representative reconstruction-based anomaly detection methods. Trainable parameters are reported in millions (M). FLOPs are reported in GFLOPs under input resolution 256×256 with batch size fixed to 12 for all methods.

Method	Params (M)	FLOPs (G)
DiT-XL/2	675	148.2
CDAD	1099	128.1
DiAD	1330	622.6
STDRAD	831	185.6

Computational cost analysis. Since the proposed method adopts a DiT-based latent diffusion backbone, we further analyze its computational cost and model size to provide a fair comparison with existing approaches. We focus on representative reconstruction-based anomaly detection methods, which share similar inference pipelines and generative reconstruction objectives.

Table 2 reports the number of trainable parameters and FLOPs for DiT (Peebles & Xie, 2023), CDAD (Li et al., 2025), DiAD (He et al., 2024), and the proposed STDRAD. All FLOPs are measured under a unified setting with input resolution fixed to 256×256 and batch size fixed to 12 for all methods. This unified configuration ensures consistent inference-time conditions and avoids bias caused by different evaluation setups.

As shown in Table 2, although STDRAD is built upon a strong DiT (XL/2) backbone, its computational cost remains comparable to other diffusion-based reconstruction methods. In particular, STDRAD requires substantially fewer FLOPs than DiAD, while maintaining a similar order of trainable parameters and achieving superior detection performance. This comparison indicates that the performance gains of STDRAD are not merely due to increased computational budget, but arise from the proposed stacked transformer denoising architecture and isomorphic structural guidance design.

4.2. Datasets & evaluation metrics

MVTec-AD. The MVTec AD dataset (Bergmann et al., 2019) is a widely adopted industrial benchmark comprising 3629 defect-free training images and 1725 test images across 15 object and texture categories. Its diverse and realistic anomalies provide a rigorous testbed for evaluating unsupervised anomaly detection and localization, aligning well with the objectives of this study.

Table 3

Comparison with state-of-the-art methods on VisA dataset, with all metrics reported as $AUROC_{cls}/AUROC_{seg}$.

Category	DiT	UTRAD	UniAD	CDAD	DiAD	STDRAD
pcb1	69.0/87.0	67.6/76.2	92.8/93.3	86.7/97.9	85.5/ 99.2	86.4/97.4
pcb2	56.7/88.8	80.5/82.4	87.8/93.9	92.1/95.7	90.8/96.0	93.6/98.6
pcb3	67.6/92.3	82.4/92.3	78.6/97.3	80.7/97.0	85.4/96.9	84.0/ 97.9
pcb4	80.0/89.7	90.0/84.2	98.8/94.9	99.2/90.8	99.4/96.6	99.2/94.8
macaroni1	63.8/80.9	68.0/81.2	79.9/97.4	56.3/95.9	77.1/91.7	85.8/98.4
macaroni2	50.6/84.5	53.0/74.3	71.6/ 95.2	78.3/89.5	53.9/91.7	58.4/94.2
capsules	55.8/70.8	73.4/98.0	55.6/88.7	79.8/96.8	50.6/96.0	84.1/99.4
candle	69.4/95.1	89.3/96.2	94.1/98.5	80.0/95.2	88.4/97.2	90.0/96.5
cashew	59.3/93.7	66.8/53.1	92.8/98.6	90.2/93.7	92.0/84.4	92.1/98.3
chewinggum	60.2/88.6	78.1/87.6	96.3/98.8	89.9/96.3	90.1/94.5	92.5/98.5
fryum	75.4/94.7	92.0/91.9	83.0/95.9	80.9/95.1	85.9/96.0	88.9/94.4
pipe-fryum	62.1/98.3	80.8/88.8	94.7/98.9	91.9/96.8	96.2/98.2	88.3/97.9
Mean	64.0/88.7	76.8/83.9	85.5/95.9	83.8/95.1	82.9/94.9	86.9/97.2

VisA. The VisA dataset (Zou et al., 2022) is a large-scale industrial anomaly detection benchmark comprising 10,821 high-resolution images across 12 object categories. It includes 9,621 normal samples and 1,200 anomalous samples, each annotated with pixel-level defect masks. The dataset encompasses a wide range of industrial items with diverse geometric structures, surface textures, and material properties. Compared to MVTec AD, VisA features higher intra-class diversity and more complex real world scenes, offering a more challenging benchmark for evaluating unsupervised and multi-class anomaly detection methods.

Evaluation metrics. We follow standard practice and use two metrics: image-level $AUROC_{cls}$, assessing the ability to distinguish normal from anomalous samples, and pixel-level $AUROC_{seg}$, measuring localization accuracy. Both range from 0 to 1, with higher values indicating better performance. Following standard practice in reconstruction based anomaly detection. All results are reported under the fixed train/test splits of the benchmarks with a fixed random seed, without repeated runs or cross-validation.

4.3. Comparison with SOTAs

We conduct a comprehensive comparison between the proposed method and state-of-the-art reconstruction-based approaches on the MVTec AD and VisA datasets, including UniAD (You et al., 2022), UTRAD (Chen et al., 2022), CDAD (Li et al., 2025), DRAEM (Zavrtanik et al., 2021a), DiAD (He et al., 2024), and the baseline diffusion transformer DiT (Peebles & Xie, 2023). The DiT baseline represents an unguided Transformer based diffusion reconstruction model, where denoising is performed without any structural guidance or alignment mechanism.

The quantitative comparisons on the MVTec AD dataset (Table 1) demonstrate the consistent advantages of our framework across both object and texture categories. On MVTec AD, our method achieves the highest mean performance, with $AUROC_{cls}$ reaching 95.8% and $AUROC_{seg}$ reaching 96.0%, indicating accurate image-level discrimination and reliable pixel-level localization. On VisA (Table 3), STDRAD also attains the best mean $AUROC_{cls}/AUROC_{seg}$ of 86.9%/97.2%, substantially surpassing the DiT baseline and other competitors, which confirms that the proposed framework generalizes well to more diverse and challenging industrial scenarios. The inferior performance of the DiT baseline indicates that a strong Transformer backbone alone is insufficient for anomaly detection, and that effective guidance and alignment are critical for suppressing anomalous patterns in diffusion-based reconstruction.

Complementary qualitative results in Figs. 4 and 5 further show that our model reconstructs normal structures faithfully while effectively suppressing anomalous regions, enabling precise detection even under

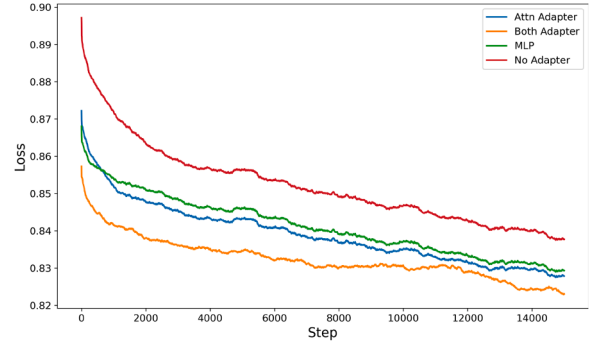


Fig. 3. Training loss curves for different adapter insertion strategies, including three insertion positions and the baseline DiT Block without any adapter.

challenging defect patterns. It is worth noting that the DiT backbone without any guidance mechanism corresponds to a vanilla diffusion reconstruction model in our experiments. In contrast, the DiT baseline without ISKG (denoted as Without Guide) exhibits noticeable geometric misalignment in these qualitative comparisons, such as orientation changes and structural shifts relative to the input image.

4.4. Ablation study

We perform ablation studies on key components of STDRAD, including the Adapter block, the ISKG module, its input forms, as well as the injection positions and injection methods of ISKG.

Effect of adapter DiT blocks. We design three types of Adapter insertion strategies based on the DiT block: MLP Adapter, which inserts the adapter after the MLP; Atten Adapter, inserted after the cross-attention; and Both Adapter, inserted after both the MLP and the cross-attention.

The training curves of these three variants, along with STDRAD without any adapter, are presented in Fig. 3. For clarity, each point on the curve represents the average loss over 760 optimization steps. Due to the nature of diffusion generative models, the training loss reflects the noise-prediction objective and is primarily analyzed at the level of optimization steps rather than epochs, and it does not directly indicate the final quality of generated images. As shown, the Both Adapter configuration exhibits a faster convergence trend in terms of step-wise diffusion training loss compared to the other variants. This observation indicates that inserting adapters into both the attention and MLP modules facilitates more efficient optimization under the diffusion noise-prediction objective. In addition, the proposed Adapter design enables the early-stage backbone blocks to remain frozen during training, reducing the number of trainable parameters of the denoising network by approximately 25%. This parameter reduction not only improves training efficiency but also stabilizes optimization by preventing unnecessary updates in structurally mature layers. The Adapter module is introduced for training efficiency and does not aim to change the final anomaly detection performance.

Effect of ISKG input type. We investigate the impact of different ISKG input types by comparing pixel-level images and latent representations as guidance signals. When feeding the ISKG with latent inputs that share the same representation space as the main denoising branch, the injected guidance inevitably carries the anomalous patterns embedded in the latent space. Consequently, the denoising process attempts to preserve these abnormal structures, resulting in reconstructed outputs that remain overly similar to the original inputs. Such behavior prevents the model from suppressing anomalies and fundamentally impairs its capability for anomaly detection.

In contrast, using the raw image at the pixel level provides a clean and structurally informative guidance signal that does not contain

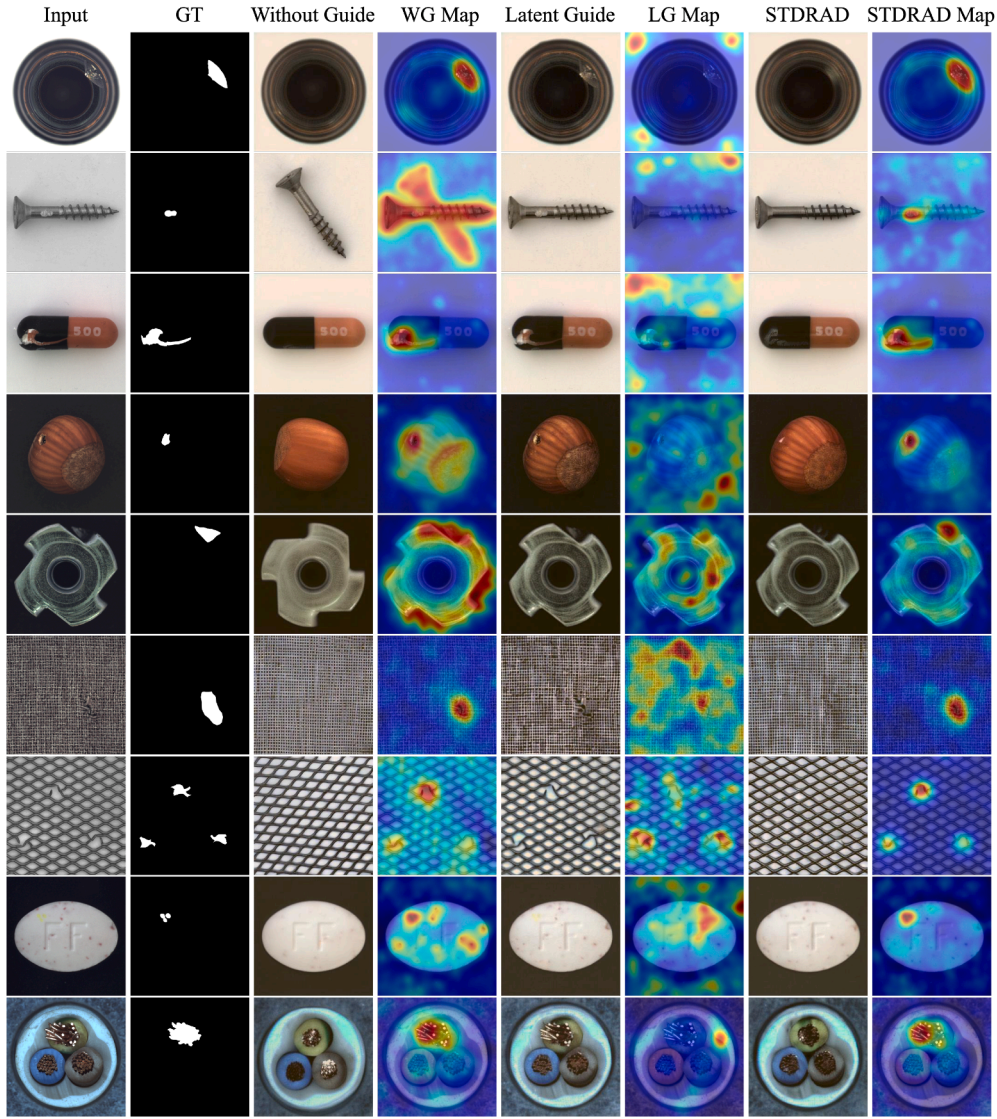


Fig. 4. Qualitative comparison on MVTec-AD. Each row shows an input image, its ground-truth anomaly mask (GT), and the outputs of several reconstruction variants: “Without Guide” and “WG Map” denotes reconstruction and heatmaps using the denoiser alone; “Latent Guide” and “LG Map” denote the reconstruction and heatmaps under Latent Guidance; “STD RAD” and “STD RAD Map” show the reconstruction and heatmaps of our proposed model. Heatmaps represent predicted anomaly responses. Our STD RAD delivers the cleanest normal-structure reconstruction and strongest suppression of anomalous regions.

Table 4
Ablation study on different ISKG input types.

ISKG Input Type	$AUROC_{cls}$	$AUROC_{seg}$
Latent	65.4	70.5
Image (ours)	95.8	96.0

Table 5
Ablation study on different ISKG backbone designs.

ISKG Backbone	$AUROC_{cls}$	$AUROC_{seg}$
DINOV2-based	83.3	91.2
Ours	95.8	96.0

latent-domain distortions. This pixel-level input effectively stabilizes the ISKG-to-Denoising interaction and guides the denoising process toward normal patterns. As a result, the reconstructed outputs exhibit proper anomaly removal and yield significantly improved detection and localization performance. The quantitative comparison is summarized in [Table 4](#).

Effect of alternative ISKG backbones. We evaluate an alternative configuration in which the ISKG module adopts a DINOv2 feature extractor. The variant preserves the same feature injection rule and the same interaction interface. Its performance is consistently lower than that of the DiT-based ISKG, as reported in [Table 5](#). The DINOv2 features present

weaker alignment with the latent token space of the main denoiser. The weaker alignment reduces the ability of the guidance module to provide reliable structural cues. As a result, the main denoiser must allocate more representational capacity to reconciling heterogeneous features. This additional burden decreases the fidelity of the reconstruction process. The proposed ISKG maintains structural homogeneity with the denoiser. The homogeneity ensures that the backbone receives features that follow consistent scales and attention patterns. This consistency allows the network to concentrate on the denoising trajectory rather than compensatory feature fusion.

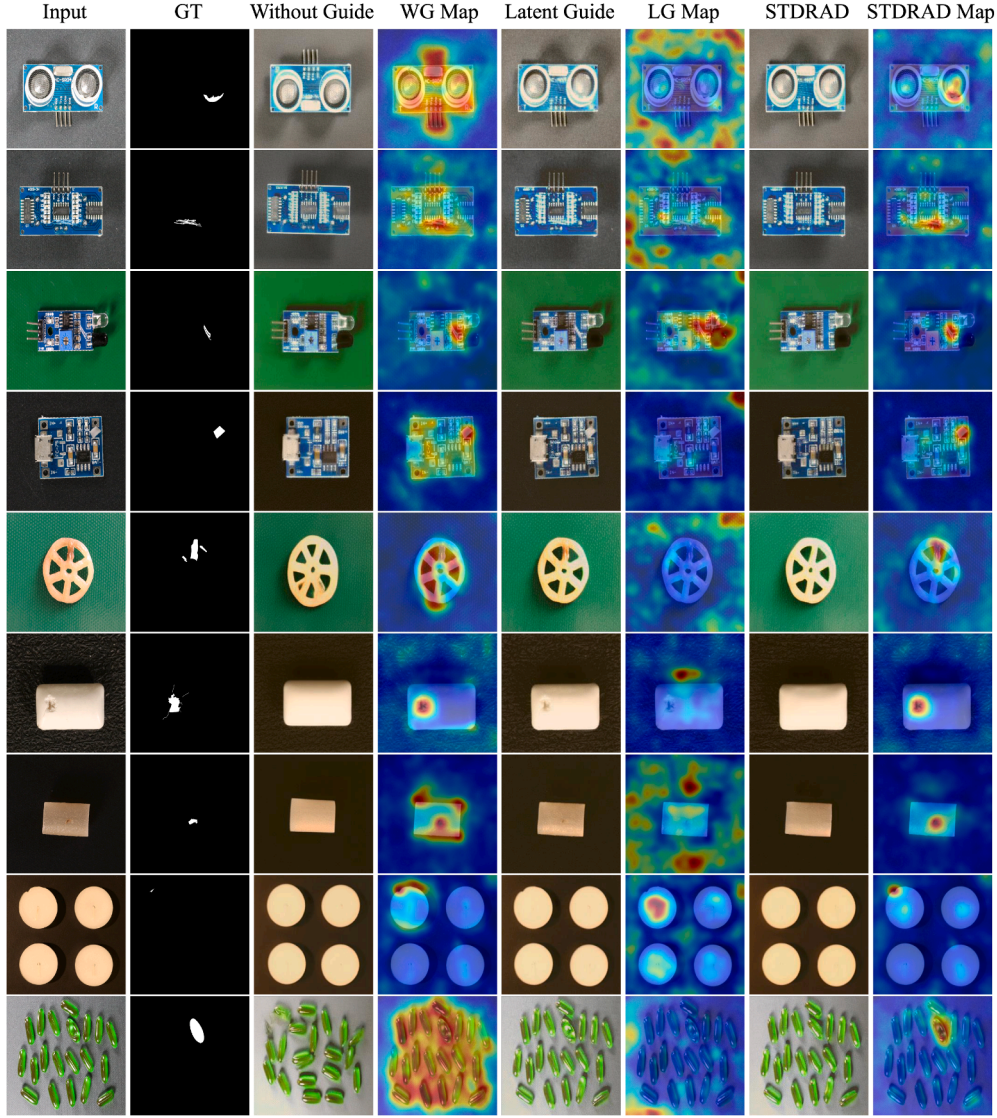


Fig. 5. Qualitative comparison on VisA. Each row shows an input image, its ground-truth anomaly mask (GT), and the outputs of several reconstruction variants: “Without Guide” and “WG Map” denotes reconstruction and heatmaps using the denoiser alone; “Latent Guide” and “LG Map” denote the reconstruction and heatmaps under Latent Guidance; “STD RAD” and “STD RAD Map” show the reconstruction and heatmaps of our proposed model. Heatmaps represent predicted anomaly responses. Our STD RAD delivers the cleanest normal-structure reconstruction and strongest suppression of anomalous regions.

Effect of ISKG fusion strategy. We compare two mechanisms for incorporating ISKG features into the backbone: cross-attention and direct additive fusion. As shown in Table 6, the additive strategy yields substantially higher performance. This difference arises from the distinct ways in which the two fusion mechanisms interact with the noise-prediction objective. In our setting, the optimization is driven solely by the noise-prediction loss, which provides limited supervisory pressure on attention weights. Under the cross-attention design, ISKG features are introduced through an auxiliary attention branch whose contribution can remain marginal throughout training. In contrast, additive fusion injects ISKG features directly into the backbone activations, ensuring that the guidance information participates in every forward update. This structural difference leads to a consistently stronger influence of ISKG features during denoising, resulting in improved reconstruction fidelity and anomaly suppression.

Effect of ISKG injection position. We explore various values of M and N to identify the most effective defect detection configuration. During initialization prior to training, the parameters of the ISKG are initialized by copying those of the first N DiT blocks from the backbone. These N

Table 6

Ablation study of ISKG fusion strategies. C-A denotes cross attention fusion and add denotes additive fusion.

ISKG Fusion	$AUROC_{cls}$	$AUROC_{seg}$
C-A	80.6	89.4
Add	95.8	96.0

blocks in the backbone are then frozen at the beginning of training. Given that the base model contains 28 DiT blocks, we fix $M + N = 28$ to maximize the benefits of the pre-trained DiT-XL/2 model.

Experimental results corresponding to different insertion positions are shown in Table 7. It is evident that the proposed ISKG provides strong guidance to the backbone denoising network. Even a minimal ISKG with $N = 1$ is sufficient to steer the model to generate samples aligned in direction and scale with the input. Increasing N further does not yield additional improvements in defect detection performance; on the contrary, it may lead to the model overfitting to excessive input

Table 7

Results of ISKG injected at different locations within the backbone network. M denotes the number of Adapter DiT Blocks, and N represents the number of preceding DiT blocks in the backbone.

$M - N$	27-1	23-5	18-10	13-15	8-20
$AUROC_{cls}$	93.4	95.8	91.9	91.6	87.3
$AUROC_{seg}$	94.6	96.0	93.0	93.6	92.8

Table 8

Ablation on ResNet-50 multi-layer selections for anomaly map computation. ✓ indicates inclusion of the corresponding feature layer. All metrics are reported as $AUROC_{cls}/AUROC_{seg}$.

f_1	f_2	f_3	f_4	f_5	AUROC
✓		✓			91.8 / 92.2
	✓	✓	✓		95.8 / 96.0
		✓	✓	✓	89.9 / 90.5
✓	✓	✓	✓		94.9 / 95.4
	✓	✓	✓	✓	95.3 / 95.5
✓	✓	✓	✓	✓	90.7 / 92.4

Table 9

Ablation on distance metrics for anomaly scoring. All results are reported as $AUROC_{cls}/AUROC_{seg}$.

Distance Metric	$AUROC_{cls}$	$AUROC_{seg}$
ℓ_1 distance	88.4	91.4
ℓ_2 distance	92.7	93.5
cosine distance	95.8	96.0

features, impairing its ability to suppress defects during generation. Empirically, the best performance is achieved when $M = 23$ and $N = 5$, indicating an optimal balance between guided features and backbone information fusion.

Effect of feature layer selection. We ablate different combinations of ResNet-50 feature layers for anomaly map computation. As summarized in Table 8, using intermediate layers (res2, res3, and res4) achieves the best overall performance. Lower-level features are overly sensitive to local texture noise, while higher-level features lose spatial precision. Therefore, res2/res3/res4 are adopted in the final model.

Effect of distance metric. We compare ℓ_1 , ℓ_2 , and cosine distance for feature-level anomaly scoring under identical settings. As shown in Table 9, cosine distance achieves the best performance. Accordingly, cosine distance is used in the final model.

5. Limitations

We clarify a representative limitation inherent to reconstruction based diffusion. When the input image is normal, the diffusion reconstruction introduces small but systematic background perturbations due to the iterative denoising process in latent space and the subsequent VAE decoding. These perturbations are visually subtle but spatially distributed across background regions. During anomaly scoring, the frozen CNN feature extractor responds deterministically to such perturbations, resulting in accumulated feature discrepancies even in the absence of true defects. Consequently, the anomaly map exhibits false-positive activations on normal samples, particularly in background areas with repetitive or low-contrast textures, as illustrated in Fig. 6.

In most cases, these activations remain low in magnitude and do not affect image-level anomaly classification. Only when the perturbations span a large area and produce consistently high anomaly responses do they lead to false positives at the image level.

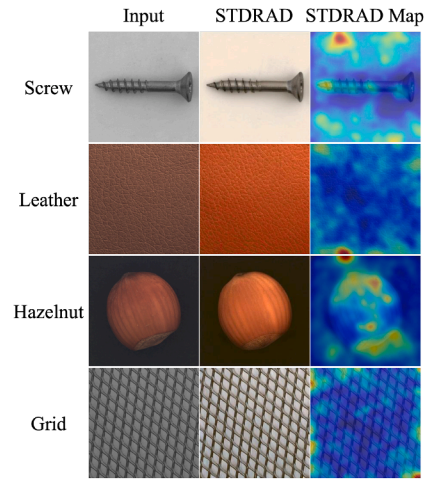


Fig. 6. Illustration of the identified limitation on normal samples. Although the input images contain no anomalies, mild background perturbations introduced during diffusion reconstruction lead to spurious activations in the anomaly maps.

6. Conclusion

In this work, we introduced STD RAD, a stacked transformer diffusion reconstruction framework for unsupervised multi-class anomaly detection. By employing a fully transformer-based denoiser, an Isomorphic Structural Knowledge Guidance mechanism, and a multi-scale feature alignment strategy, the framework achieves semantically aligned and geometrically consistent reconstruction, effectively addressing the misalignment issues of conventional U-Net based diffusion models and enabling reliable residual-based anomaly localization across diverse defect types. Importantly, STD RAD is not designed merely to pursue higher MVTEC AD or VisA leaderboard scores; rather, its primary goal is to explore how to construct a transformer-based, knowledge-guided, and parameter-efficient generative anomaly detection system suitable for real industrial scenarios. Moreover, because the architecture closely follows the design principles of DiT, advances in diffusion modeling and stacked-transformer research can be seamlessly integrated into our framework, providing strong compatibility and extensibility for future developments. Extensive experiments on industrial benchmarks demonstrate the effectiveness of STD RAD, highlighting its potential as a robust foundation for next-generation anomaly detection systems.

Looking forward, several concrete extensions of the proposed framework are worth exploring. One promising direction is online or continual adaptation, where the diffusion model incrementally updates its representation to accommodate distribution shifts in long-term industrial deployment. Another direction is extending STD RAD to video anomaly detection by incorporating temporal consistency into the diffusion reconstruction and guidance mechanisms, enabling the modeling of spatiotemporal anomalies. In addition, exploring more lightweight diffusion backbones or reducing denoising steps may further improve efficiency and facilitate deployment in real-time or resource-constrained industrial environments.

CRediT authorship contribution statement

Minjie Du: Writing – review & editing, Writing – original draft, Visualization, Investigation, Methodology, Software, Conceptualization; **Hengyu Xu:** Conceptualization; **Fulin Shang:** Conceptualization; **Zheng Wang:** Funding acquisition, Supervision, Conceptualization; **Yining Hu:** Funding acquisition, Supervision, Conceptualization; **Lizhe Xie:** Funding acquisition, Supervision, Conceptualization.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yining Hu reports financial support was provided by Ministry of Science and Technology of the People's Republic of China. Lizhe Xie reports financial support was provided by National Natural Science Foundation of China. Lizhe Xie reports financial support was provided by Jiangsu Provincial Department of science and technology. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Akay, S., Atapour-Abarghouei, A., & Breckon, T. P. (2018). GANomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision* (pp. 622–637). Springer.
- Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTEC ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9592–9600).
- Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2020). Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4183–4192).
- Blattmann, A., Rombach, R., Oktay, K., & Ommer, B. (2022). Retrieval-augmented diffusion models. <https://doi.org/10.48550/ARXIV.2204.11824>
- Cao, Y., Yao, H., Luo, W., & Shen, W. (2025). VarAD: Lightweight high-resolution image anomaly detection via visual autoregressive modeling. *IEEE Transactions on Industrial Informatics*, 21(4), 3246–3255.
- Chen, L., You, Z., Zhang, N., Xi, J., & Le, X. (2022). UTRAD: Anomaly detection and localization with u-transformer. *Neural Networks*, 147, 53–62.
- Chung, H., Park, J., Keum, J., Ki, H., & Kang, S. (2020). Unsupervised anomaly detection using style distillation. *IEEE Access*, 8, 221494–221502.
- Chung, K. L. (1967). Markov chains. Springer-Verlag, New York.
- Czimmermann, T., Ciuti, G., Milazzo, M., Chiuazzini, M., Roccella, S., Oddo, C. M., & Dario, P. (2020). Visual-based defect detection and classification approaches for industrial applications—a survey. *Sensors*, 20(5), 1459.
- Duan, Y., Hong, Y., Niu, L., & Zhang, L. (2023). Few-shot defect image generation via defect-aware feature manipulation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 571–578). (vol. 37).
- El Ogri, O., Jaouad, E.-M., Benslimane, M., & Hjouji, A. (2024). Automatic lip-reading classification using deep learning approaches and optimized quaternion meixner moments by GWO algorithm. *Knowledge-Based Systems*, 304, 112430.
- El Ogri, O., Jaouad, E.-M., & Hjouji, A. (2026). A computer-assisted medical diagnosis system for cancer diseases based on quaternion orthogonal rademacher-fourier moments and deep learning. *Biomedical Signal Processing and Control*, 112, 108744.
- El Ogri, O., Karmouni, H., Sayyouri, M., & Qjidaa, H. (2021). 3D image recognition using new set of fractional-order legendre moments and deep neural networks. *Signal Processing: Image Communication*, 98, 116410.
- Fang, Q., Su, Q., Lv, W., Xu, W., & Yu, J. (2025). Boosting fine-grained visual anomaly detection with coarse-knowledge-aware adversarial learning. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 16532–16540). (vol. 39).
- Guo, J., Lu, S., Zhang, W., Chen, F., Li, H., & Liao, H. (2025). Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *Proceedings of the computer vision and pattern recognition conference* (pp. 20405–20415).
- He, H., Zhang, J., Chen, H., Chen, X., Li, Z., Chen, X., Wang, Y., Wang, C., & Xie, L. (2024). A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8472–8480). (vol. 38).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
- Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840–6851.
- Hu, T., Zhang, J., Yi, R., Du, Y., Chen, X., Liu, L., Wang, Y., & Wang, C. (2024). Anomaly-diffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8526–8534). (vol. 38).
- Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. [arXiv:1312.6114](https://arxiv.org/abs/1312.6114).
- Li, C.-L., Sohn, K., Yoon, J., & Pfister, T. (2021). CutPaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9664–9674).
- Li, X., Tan, X., Chen, Z., Zhang, Z., Zhang, R., Guo, R., Jiang, G., Chen, Y., Qu, Y., Ma, L. et al. (2025). One-for-more: Continual diffusion model for anomaly detection. In *Proceedings of the computer vision and pattern recognition conference* (pp. 4766–4775).
- Liu, J., Wang, C., Su, H., Du, B., & Tao, D. (2019). Multistage GAN for fabric defect detection. *IEEE Transactions on Image Processing*, 29, 3388–3400.
- Liu, J., & Wang, F. (2024). Mixed-attention auto encoder for multi-class industrial anomaly detection. In *Icassp 2024-2024 IEEE international conference on acoustics, speech and signal processing (icassp)* (pp. 4120–4124). IEEE.
- Liu, J., Xie, G., Wang, J., Li, S., Wang, C., Zheng, F., & Jin, Y. (2024). Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1), 104–135.
- Liu, Y., Zhuang, C., & Lu, F. (2021). Unsupervised two-stage anomaly detection. [arXiv:2103.11671](https://arxiv.org/abs/2103.11671).
- Liu, Z., Zhou, Y., Xu, Y., & Wang, Z. (2023). SimpleNet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20402–20411).
- Mousakhan, A., Brox, T., & Tayyub, J. (2024). Anomaly detection with conditioned denoising diffusion models. In *Dagm german conference on pattern recognition* (pp. 181–195). Springer.
- Peebles, W., & Xie, S. (2023). Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4195–4205).
- Perez, E., Strub, F., De Vries, H., Dumoulin, V., & Courville, A. (2018). FILM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*. (vol. 32).
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684–10695).
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.
- Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M. H., & Rabiee, H. R. (2021). Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14902–14912).
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).
- Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. [arXiv:2010.02502](https://arxiv.org/abs/2010.02502).
- Tang, Y., Zhao, L., Zhang, S., Gong, C., Li, G., & Yang, J. (2020). Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129, 123–130.
- Tao, X., Gong, X., Zhang, X., Yan, S., & Adak, C. (2022). Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement*, 71, 1–21.
- Van den Oord, A., Kalchbrenner, N., Espeholt, L., Kavukcuoglu, K., Vinyals, O., Graves, A. et al. (2016). Conditional image generation with pixelcnn decoders. *Advances in Neural Information Processing Systems*, 29, 4790–4798.
- Wan, Q., Gao, L., Li, X., & Wen, L. (2021). Industrial image anomaly localization based on gaussian clustering of pretrained feature. *IEEE Transactions on Industrial Electronics*, 69(6), 6182–6192.
- Wang, F., Zhang, T., Wang, Y., Qiu, Y., Liu, X., Guo, X., & Cui, Z. (2025). Distribution prototype diffusion learning for open-set supervised anomaly detection. In *Proceedings of the computer vision and pattern recognition conference* (pp. 20416–20426).
- Wang, G., Han, S., Ding, E., & Huang, D. (2021). Student-teacher feature pyramid matching for anomaly detection. [arXiv:2103.04257](https://arxiv.org/abs/2103.04257).
- Wyatt, J., Leach, A., Schmon, S. M., & Willcocks, C. G. (2022). AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 650–656).
- Xu, X., Wang, Y., Huang, Y., Liu, J., Lei, X., Xie, G., Jiang, G., & Lu, Z. (2025). A survey on industrial anomalies synthesis. [arXiv:2502.16412](https://arxiv.org/abs/2502.16412).
- Yan, X., Zhang, H., Xu, X., Hu, X., & Heng, P.-A. (2021). Learning semantic context from normal samples for unsupervised anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 3110–3118). (vol. 35).
- You, Z., Cui, L., Shen, Y., Yang, K., Lu, X., Zheng, Y., & Le, X. (2022). A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35, 4571–4584.
- Zavrtanik, V., Kristan, M., & Skocaj, D. (2021a). Draem - a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 8330–8339).
- Zavrtanik, V., Kristan, M., & Skocaj, D. (2021b). Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8330–8339).
- Zavrtanik, V., Kristan, M., & Skocaj, D. (2021c). Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112, 107706.
- Zhang, H., Wang, Z., Zeng, D., Wu, Z., & Jiang, Y.-G. (2025). DiffusionAD: Norm-guided one-step denoising diffusion for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8), 7140–7152.
- Zou, Y., Jeong, J., Pemula, L., Zhang, D., & Dabeer, O. (2022). Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision* (pp. 392–408). Springer.