electronics

MDPI

Article

Semantically Grounded Reverse Distillation for Unsupervised Anomaly Detection

Chunlei Li [1], Xin Yuan [1], Junpu Wang [1], Ruimin Yang [1] and Hongling Zhao [2,*]

[1] School of Information and Communication Engineering, Zhongyuan University of Technology, Zhengzhou 450007, China; lichunlei1979@zut.edu.cn (C.L.); 2024119038@zut.edu.cn (X.Y.); 6984@zut.edu.cn (J.W.); 6132@zut.edu.cn (R.Y.)
[2] College of Distance Learning, Zhengzhou University, Zhengzhou 450001, China
* Correspondence: hlzhao@zzu.edu.cn

Abstract

Knowledge distillation-based unsupervised anomaly detection has achieved strong performance in industrial inspection. However, existing Reverse Distillation (RD) methods suffer from two fundamental limitations: the bottleneck compresses semantically entangled CNN features into a one-class representation that inadequately characterizes normal structural regularity, and the pairwise distillation objective leaves the co-learned teacher-student geometry semantically underconstrained. To address these issues, we propose Semantically Grounded Reverse Distillation (SGRD), which integrates a frozen vision foundation model (DINOv2 in this work) as a fixed semantic reference into the RD framework. Specifically, a foundation-aware semantic bottleneck conditions the bottleneck encoding pathway with foundation-model patch semantics to produce a structurally coherent one-class code. Furthermore, a foundation-semantic representation anchoring module jointly constrains both teacher and student representations toward the foundation-model semantic subspace, discouraging semantically arbitrary representation geometry. Experiments on MVTec AD achieve 98.5% image-level AUROC and 98.3% pixel-level AUROC, while experiments on VisA achieve 99.8% image-level AUROC and 94.2% pixel-level AUPRO, with the proposed method attaining the strongest overall average performance among the compared methods on both benchmarks.

Keywords: anomaly detection; knowledge distillation; vision foundation model; reverse distillation; semantic grounding

1. Introduction

Anomaly detection (AD) aims to identify and localize abnormal regions that deviate from normal visual patterns. It plays an important role in industrial inspection [1–4], medical imaging [5,6], and video surveillance [7,8]. Since anomalous samples are scarce and expensive to annotate, unsupervised AD methods trained solely on normal images have been extensively studied.

Existing unsupervised AD approaches can be broadly categorized into reconstruction-based and feature embedding-based methods. Reconstruction-based methods [9–11] detect anomalies by measuring discrepancies between inputs and their reconstructions. However, the strong generalization capability of deep models often leads to faithful reconstruction of anomalous regions, causing false negatives [12,13]. Feature embedding-based methods [14–16] model the distribution of normal features and detect anomalies by measur-

Check for updates

Received: 22 September 2026
Revised: 27 October 2026
Accepted: 4 November 2026
Published: 7 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

ing deviations from the learned distribution but typically rely on complex density estimation or large-scale feature matching that limits practical scalability.

Knowledge distillation (KD)-based methods have emerged as a compelling alternative, leveraging representational discrepancies between teacher and student networks to identify anomalies [1,17]. Since the student is trained only on normal samples, anomalous patterns cannot be faithfully reconstructed during inference, producing measurable discrepancies that indicate defects. Early KD-based methods [18] apply multi-scale feature distillation across feature pyramid levels to aggregate layer-wise differences, achieving competitive performance. However, without anomaly supervision, the student tends to over-generalize, thus blurring anomaly boundaries and degrading detection accuracy. Reverse Distillation (RD) [19] mitigates this by adopting an asymmetric encoder-decoder architecture with a compact bottleneck and pairwise distillation objective under reverse data flow. Subsequent improvements [20,21] have further explored anomaly synthesis for explicit denoising supervision and guided information injection to improve detail reconstruction, achieving strong performance. However, these approaches do not address two fundamental limitations intrinsic to the RD paradigm. First, its bottleneck structure compresses CNN features in a semantically agnostic manner, conflating structurally meaningful normal patterns with anomaly-correlated perturbations that are statistically inseparable in CNN activation space. Second, its pairwise distillation objective leaves the co-learned teacher-student manifold geometrically underconstrained, admitting semantically arbitrary solutions that satisfy pairwise proximity yet lack meaningful structural organization. Consequently, the teacher-student discrepancy becomes an unreliable anomaly indicator for defects that deviate at the structural or semantic level rather than at the local texture level, as illustrated in Figure 1. These two limitations call for a source of structural information that does not originate from the CNN features themselves. The patch tokens of a self-supervised vision foundation model such as DINOv2 [22] predominantly encode object-centric structural semantics at a coarse spatial granularity, which allows them to serve as an external structural reference. We therefore use DINOv2 as a frozen reference and retain the CNN teacher-student pair as the detector, so that the multi-scale discrepancy continues to carry the anomaly signal.

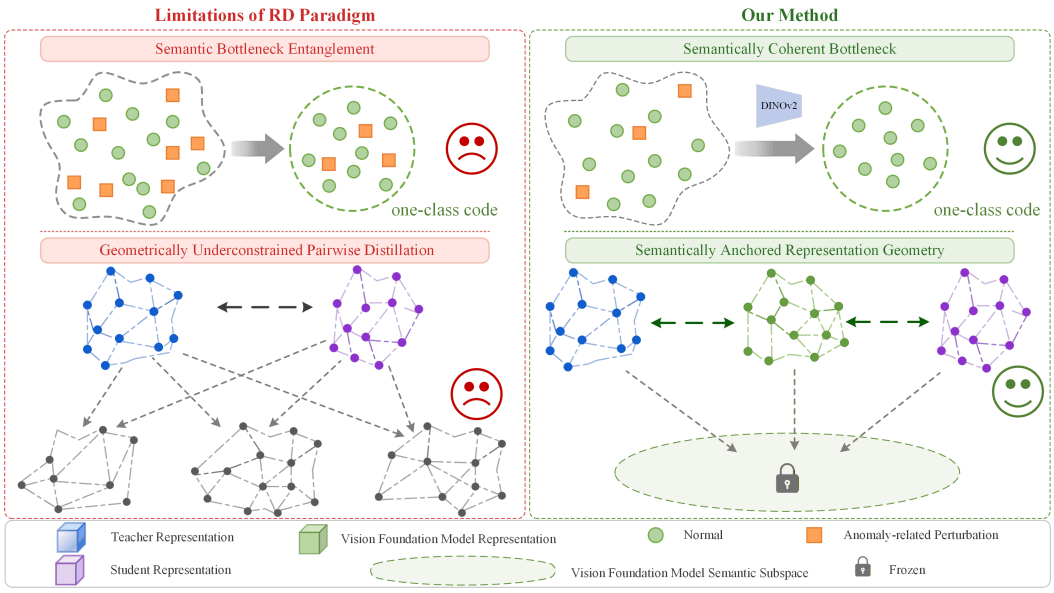


Figure 1. Motivation of the proposed SGRD. Existing RD methods suffer from semantic bottleneck entanglement and geometrically underconstrained pairwise distillation. By grounding reverse distillation in a frozen foundation-model semantic subspace, SGRD produces semantically coherent bottleneck representations and semantically anchored representation geometry.

To address both limitations, we propose Semantically Grounded Reverse Distillation (SGRD) for unsupervised anomaly detection, which integrates a frozen vision foundation model (FM), instantiated as DINOv2 in this work, as a fixed semantic authority into the RD pipeline. Specifically, a foundation-aware semantic bottleneck is proposed to resolve the semantic deficiency of the bottleneck by conditioning its encoding pathway with FM patch-token semantics, producing a one-class code that faithfully reflects the structural distribution of normal objects. Furthermore, a foundation-semantic representation anchoring module is proposed to address the geometric underdetermination of pairwise distillation by simultaneously anchoring both sides of the teacher-student pair to the FM's semantic subspace, thus constraining the co-learned manifold toward a pre-established semantic coordinate system.

The main contributions of this paper are summarized as follows:

(1) We identify two fundamental limitations intrinsic to the RD paradigm, namely, semantic bottleneck entanglement and geometrically underconstrained pairwise distillation, and propose SGRD to address both through foundation-model semantic grounding.
(2) We propose a foundation-aware semantic bottleneck, which conditions bottleneck encoding with FM patch-token semantics to produce a semantically coherent one-class representation of normal structures.
(3) We propose foundation-semantic representation anchoring, which anchors both teacher and student representations to the FM semantic subspace, constraining the co-learned representation geometry with semantic structure. Its teacher-side term acts on the semantically modulated teacher feature, coupling the two modules through the gradient path during training.

2. Related Work

2.1. Unsupervised Anomaly Detection

Unsupervised anomaly detection is usually trained only on normal images, and a test sample is flagged when it departs from the learned notion of normality. This setting differs from learning with noisy labels in image classification, for example, UCRT [23], which selects a class-balanced clean subset from mislabeled training data and then extends it through semi-supervised training of a classifier. A large share of existing anomaly detection work follows a reconstruction-based strategy, where a network learns to reproduce normal inputs or normal features and the reconstruction error at test time becomes the anomaly signal. DSR [11] projects images into a dual subspace and detects surface defects from the discrepancy between the two reconstructions. DiAD [9] instead relies on a diffusion model to recover the normal appearance of an input image, and UniAD [10] introduces a neighbor-masked attention module to keep a transformer-based reconstruction network from simply copying anomalous regions. DMAD [13] pursues a related goal by modeling the diversity of plausible normal reconstructions rather than committing to a single deterministic one. Even with these adjustments, an expressive decoder can still generalize somewhat beyond the normal training distribution, and the reconstruction discrepancy does not always separate anomalies reliably on its own.

Feature embedding-based methods avoid this pixel-level reconstruction step. They represent inputs in the feature space of a frozen, pretrained backbone and identify anomalies from how a test feature relates to the distribution of normal features, whether through an explicit distance measure, a density estimate, or a learned decision boundary. PatchCore [15] stores a coreset of representative normal patch features in a memory bank and scores each test patch by its distance to the nearest stored exemplar. SimpleNet [14] instead perturbs normal features with synthetic noise and trains a lightweight discriminator to separate perturbed features from clean ones, while SuperSimpleNet [24] extends this idea to a setting

that can also draw on a handful of labeled anomalies when they happen to be available. MSFlow [25] takes yet another route and models the density of normal features with a multi-scale normalizing flow, and MB-PFM [26] instead maps features from one pretrained space into a second one, flagging anomalies from the resulting mapping discrepancy. These methods are often effective, though the memory bank or the density model still has to be built and queried per category, and this adds storage and inference overhead as the number of categories grows.

2.2. Knowledge Distillation and Reverse Distillation for Anomaly Detection

Knowledge distillation offers an alternative view of the same problem. A fixed teacher network and a trainable student are compared under normal-only supervision, and their discrepancy at test time serves as the anomaly indicator, since the student is never exposed to abnormal patterns during training and tends to reproduce the teacher's response poorly on anomalous regions. Pull & Push [18] distills teacher features into a student across a shared feature-pyramid data flow and aggregates the layer-wise differences. EfficientAD [1] keeps the same teacher-student principle but favors a lighter patch description network, prioritizing millisecond-level inference over architectural complexity. DeSTSeg [17] instead pairs a denoising student, trained on synthetically corrupted inputs, with a separate segmentation network that fuses the multi-level teacher-student discrepancy into the final anomaly map. AST [27] targets a different concern, the tendency of a same-architecture teacher-student pair to over-generalize, by attaching a normalizing-flow-based student that is structurally asymmetric to the teacher. Without this kind of asymmetry, a student trained only under a pairwise proximity objective can still learn to mimic the teacher too broadly, and the resulting anomaly map tends to blur rather than isolate defect boundaries.

Reverse Distillation addresses the same over-generalization tendency through a different asymmetric design. A frozen, pretrained encoder first compresses the input into a compact one-class bottleneck embedding, and a separate decoder, structurally mirrored relative to the encoder, reconstructs the encoder's multi-scale features from this embedding alone under a reversed data flow. The decoder never receives the raw input directly, so it has fewer opportunities to leak anomalous detail through the bottleneck, and the resulting discrepancy tends to stay more spatially localized. RD++ [21] builds directly on this design and synthesizes pseudo-anomalies during training to provide explicit denoising supervision, while Liu et al. [20] revisit the bottleneck's information flow and add an expert network to sharpen fine-grained detail reconstruction. DTGD [28] adapts the same bottleneck idea for a differently structured student. Its student is a single denoising encoder-decoder with the bottleneck placed inside, between its own encoder and decoder, and it is guided by two frozen teachers: one trained on normal images, the other exposed to synthetic anomalies. DBKD [29] follows a related but distinct route, pairing a conventional student that shares the teacher's structure with a second, bottleneck-based branch that reconstructs the teacher's multi-scale features through an asymmetric decoder; the discrepancy maps from both branches are then aggregated. And RDMS [30] instead trains three independent student decoders, each dedicated to reconstructing a different teacher layer, coupled with a multi-level feature fusion module that replaces the original one-class bottleneck.

Most of these extensions adjust the training signal or the decoder's reconstruction target, while two aspects of the original bottleneck design have drawn comparatively less attention. The bottleneck compression is shaped mainly by CNN activation statistics computed within the target category, and little in the training objective pushes the resulting one-class code toward structural regularity rather than domain-specific texture statistics. The pairwise distillation objective raises a related concern, since it constrains only the relative proximity between teacher and student, so the absolute geometry of their shared representation

space depends largely on initialization and optimization dynamics rather than on semantic structure. These two aspects appear to have received limited direct treatment in the RD literature so far, and the resulting discrepancy may become a less dependable cue for defects that diverge from normal samples at the structural level while remaining close to them in local texture.

2.3. Foundation Models in Anomaly Detection

A separate line of work builds the anomaly detector directly on a large, pretrained vision or vision-language foundation model rather than a task-specific CNN. WinCLIP [31] and AnomalyCLIP [32] adapt CLIP's joint image-text embedding through prompt design for zero- and few-shot anomaly classification and segmentation, while APRIL-GAN [33] instead trains a small set of adapter layers on top of a frozen CLIP encoder for the same task. A second direction replaces the backbone itself with a self-supervised vision transformer such as DINOv2: Dinomaly [12] trains a single, unified reconstruction model on a frozen DINOv2 backbone and reports state-of-the-art results on MVTec AD and VisA under both unified multi-class and class-separated settings, while AnomalyDINO [34] and SuperAD [35] skip training altogether, building a training-free memory bank directly from DINOv2 patch tokens for few-shot or zero-shot segmentation.

These methods differ from RD and its descendants along a more basic axis than the shot regime: the foundation model constitutes the entire detection backbone and operates end-to-end at its native transformer resolution, rather than serving as an auxiliary reference alongside a separate, lightweight feature extractor. This distinction holds regardless of whether the foundation-model backbone is evaluated under a zero-shot, few-shot, unified multi-class, or per-category configuration. A related idea appears outside the industrial domain: the Q-Former Autoencoder [36] keeps a frozen foundation-model encoder for medical images and steers a single reconstruction pathway with a perceptual loss computed from a separate pretrained masked autoencoder. DINOv2 remains frozen throughout training in the present work and stays outside the detection backbone; it is introduced instead as a fixed external reference that conditions the CNN bottleneck and anchors both sides of the teacher-student representation, while the discrepancy between the CNN teacher and student continues to carry the anomaly signal.

3. Method

3.1. Model Architecture

The proposed SGRD framework builds on the asymmetric encoder-decoder organization of reverse distillation and introduces a frozen vision foundation model as a semantic reference. As illustrated in Figure 2, the framework consists of a frozen teacher encoder; a trainable student decoder; a frozen DINOv2 foundation model; and two semantic grounding modules, namely, foundation-aware semantic bottleneck (FASB) and foundation-semantic representation anchoring (FSRA). The teacher encoder adopts an ImageNet-pretrained WideResNet50 and extracts multi-scale representations from the input image. These representations are processed through the bottleneck pathway to obtain a compact one-class representation, which serves as the sole input to the student decoder. The student reconstructs the corresponding teacher representations from this code under the reverse distillation objective.

DINOv2 provides fixed patch-token features that are used to introduce semantic guidance at two stages of the framework. FASB uses the foundation-model features to modulate the deepest teacher representation before multi-scale bottleneck aggregation. The semantic response is incorporated through a residual formulation, so that the original CNN representation remains available while additional structural information is introduced

at the highest feature level. The lower-level teacher representations are retained during aggregation to preserve fine-grained spatial information. In parallel, FSRA operates during training on the FASB-modulated teacher representation, the deepest student representation, and the corresponding DINOv2 features. These representations are mapped into a shared semantic comparison space, where FM-referenced alignment constraints are imposed on both the teacher-side and student-side features.

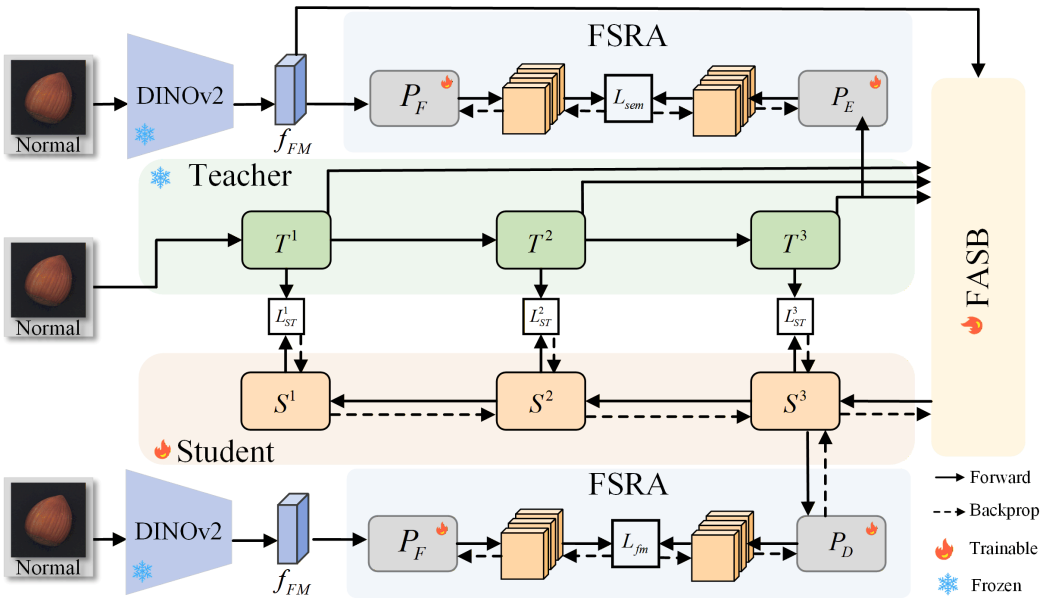


Figure 2. Overview of the proposed SGRD framework. A frozen DINOv2 semantically guides bottleneck encoding through FASB and anchors the teacher-student feature manifold through FSRA. During inference, anomaly localization is obtained from multi-scale feature discrepancies between teacher and student representations.

FASB and FSRA therefore provide complementary semantic guidance within the reverse distillation pipeline. FASB acts on the representation encoding pathway before bottleneck compression, whereas FSRA introduces additional representation-level constraints during optimization. The overall training objective combines the reverse distillation loss with the two semantic alignment terms of FSRA. During inference, the FASB pathway remains active and DINOv2 continues to provide the semantic features required for bottleneck modulation, while the projection heads introduced by FSRA are discarded. The student then reconstructs the multi-scale teacher representations from the semantically modulated bottleneck code, and the anomaly map is computed from the accumulated discrepancies between the corresponding teacher and student features. The final anomaly signal remains within the teacher-student discrepancy, while foundation-model semantics influence the representation encoding and training process.

In SGRD, semantic grounding refers to integrating DINOv2 visual patch-token features into two complementary parts of the reverse-distillation pipeline. FASB uses these input-dependent features to condition the deepest CNN teacher feature before bottleneck aggregation, so the one-class code is conditioned on the FM features of each input image during both training and inference. FSRA uses the same frozen FM features as a shared reference for training-time constraints on the FASB-modulated teacher representation and the deepest student representation. In the complete model, the teacher-side term $\mathcal{L}_{sem}$ is computed from the FASB-modulated feature, allowing its gradient to reach FASB's trainable attention projections while the FM parameters remain frozen. Unlike prompt-based

CLIP approaches [31,32], SGRD uses DINOv2 visual patch tokens without text prompts or image-text similarity.

3.2. Foundation-Aware Semantic Bottleneck

The RD bottleneck compresses multi-scale CNN features in a semantically entangled space where structurally meaningful patterns and anomaly-correlated perturbations are inseparable by convolutional statistics alone. This semantic blindness causes the one-class code to inadequately characterize the normal structural manifold, weakening the teacher-student discrepancy for anomalies that deviate at the semantic rather than texture level. To address this, we propose FASB, which conditions the bottleneck encoding pathway on FM patch-token semantics. Since DINOv2 patch tokens primarily encode object-centric structural semantics rather than low-level texture statistics, FASB selectively enhances semantically coherent CNN activations while suppressing semantically irrelevant responses. This disentangles semantic structure from texture statistics before bottleneck compression, yielding a one-class code that better reflects the structural distribution of normal objects.

Figure 3 illustrates the overall architecture of FASB. Since FM patch-token representations predominantly encode object-centric structural semantics at a coarse spatial granularity, semantic grounding is architecturally meaningful only at the highest CNN feature level, where representational abstraction becomes commensurate with FM patch-token semantics. Formally, let $f_E^{(3)} \in \mathbb{R}^{C_E \times H_E \times W_E}$ denote the deepest feature map of the teacher encoder, carrying the dominant semantic signal that governs the direction of the one-class code. Let $f_{FM} \in \mathbb{R}^{C_F \times H_F \times W_F}$ denote the patch-token feature map produced by the frozen FM backbone. Both are resampled to a shared low spatial resolution $L \times L$, at which the attention captures global structural correspondences rather than local pixel correlations:

$$\hat{f}_E^L = \text{Resample}(f_E^{(3)}, L), \quad \hat{f}_{FM}^L = \text{sg}(\text{Resample}(f_{FM}, L)), \tag{1}$$

where $\text{sg}(\cdot)$ denotes stop-gradient, preventing any gradient from entering the frozen FM. The cross-attention projections are:

$$Q = W_q \hat{f}_{FM}^L, \quad K = W_k \hat{f}_E^L, \quad V = W_v \hat{f}_E^L, \tag{2}$$

where W_q, W_k, W_v are learnable 1×1 convolutional projections with output dimension d. FM patch-token semantics are used as queries, while CNN features serve as keys and values, enabling the FM to identify semantically coherent CNN feature patterns. In this assignment, each FM token queries the CNN positions through the affinity matrix A, and the aggregated content is drawn from the CNN values, so the response injected in Equation (4) is composed of projections of the teacher's own features, weighted under the guidance of the FM. This assignment is compared with alternative fusion forms in Section 4.4 multi-head attention with temperature scaling then produces a semantically filtered response:

$$A = \text{Softmax}\left(\frac{\hat{Q}\hat{K}^\top}{\sqrt{d/n_h}\,\tau}\right), \quad \hat{O} = A\hat{V}, \tag{3}$$

where $\hat{Q}, \hat{K}, \hat{V} \in \mathbb{R}^{N \times d}$ are sequence-form projections with $N = L^2$, n_h the number of attention heads, and τ a temperature parameter. The affinity matrix A quantifies the semantic alignment between FM patch tokens and CNN spatial positions; the aggregated output $\hat{O}$ concentrates on CNN features that are semantically congruent with FM's structural understanding.

The response is upsampled and injected residually:

$$\hat{f}_E^{(3)} = f_E^{(3)} + \lambda_s \cdot \text{Upsample}(\text{Conv}(\hat{O}), H_E, W_E), \tag{4}$$

where λ_s is a residual scaling factor that governs the injection strength; $\hat{O}$ is first reshaped from $\mathbb{R}^{N\times d}$ to $\mathbb{R}^{d\times L\times L}$, after which $\mathrm{Conv}(\cdot)$ denotes a learnable 1×1 convolutional projection mapping the attended output from dimension d to C_E; and $\mathrm{Upsample}(\cdot)$ denotes bilinear interpolation to the original spatial resolution (H_E, W_E). The residual formulation is critical: it preserves the CNN's texture-discriminative cues, which are essential for fine-grained anomaly localization, while injecting semantic structural guidance into the representation.

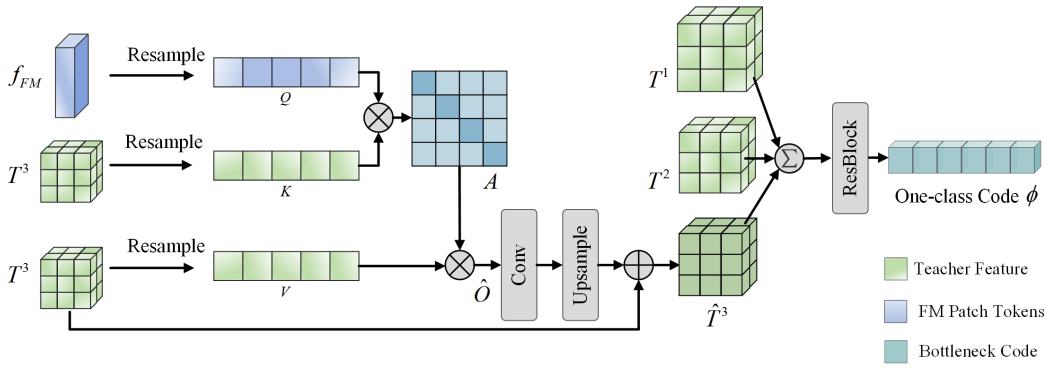


Figure 3. Architecture of the proposed foundation-aware semantic bottleneck. FM patch-token semantics guide bottleneck encoding to produce a semantically grounded one-class representation.

The semantically modulated $\hat{f}_E^{(3)}$ replaces $f_E^{(3)}$ as the highest-level input to FASB's multi-scale aggregation stage. Together with the lower-level features $f_E^{(1)}$ and $f_E^{(2)}$, which preserve fine-grained spatial detail, the aggregated representation is compressed into the one-class code $\boldsymbol{\phi}$, the sole input to the student decoder.

By semantically grounding the bottleneck encoding pathway, FASB produces a compact one-class representation aligned with the FM's object-centric structural semantics. Applying FM constraints to lower CNN levels would instead introduce semantically mismatched supervision on texture-dominated representations, degrading fine-grained localization sensitivity, as validated in Table 9.

3.3. Foundation-Semantic Representation Anchoring

FASB resolves the semantic deficiency of the bottleneck at the structural encoding level. However, a complementary limitation remains in the representation geometry co-learned by the teacher-student pair. The distillation objective constrains only their relative proximity while leaving the absolute organization of the learned manifold underconstrained. Consequently, multiple semantically arbitrary solutions can satisfy the pairwise distillation objective, causing the learned representation geometry to depend primarily on initialization and optimization dynamics rather than semantic structure. This weakens discriminability for anomalies that deviate structurally or semantically while remaining locally texture-consistent.

To address this, we propose FSRA, which introduces the frozen FM as a universal semantic reference frame that simultaneously constrains both sides of the teacher-student pair. Rather than relying solely on pairwise alignment, FSRA enforces a tripartite co-grounding in which both the teacher-side bottleneck representation and the student-side decoder representation are jointly drawn toward the FM semantic subspace, anchoring the co-learned manifold to a semantically structured representation space that preserves object-centric structural regularity and discourages convergence to semantically arbitrary solutions.

Formally, FSRA introduces learnable projection heads P_E, P_D, and a shared FM head P_F, each implemented as a single 1×1 convolutional layer, projecting heterogeneous features into

a shared d_p-dimensional semantic comparison space. On the encoder side, FSRA operates on the FASB-modulated feature $\hat{f}_E^{(3)} \in \mathbb{R}^{C_E \times H_E \times W_E}$ obtained from Equation (4), instead of the raw frozen teacher activation $f_E^{(3)}$. Because the teacher-side anchoring loss $\mathcal{L}_{sem}$ is computed from the FASB-modulated feature $\hat{f}_E^{(3)}$, its gradient can update the projection heads and propagate through FASB's trainable parameters, including the attention projections W_q, W_k, and W_v, while the DINOv2 backbone remains frozen. Since $\hat{f}_E^{(3)}$ also enters the multi-scale aggregation that produces the one-class code, from which the student reconstructs the teacher features under $\mathcal{L}_{RD}$, these parameters receive gradient contributions from both the reverse-distillation objective and the teacher-side anchoring objective, which couples the training of the two proposed modules rather than optimizing them in isolation. Let $f_{dec} \in \mathbb{R}^{C_D \times H_\phi \times W_\phi}$ denote the student decoder's deepest output (semantically richest on the decoder side). The FM feature is spatially aligned to the resolution of $\hat{f}_E^{(3)}$ and treated as a fixed semantic target via stop-gradient:

$$\bar{f}_{FM} = \mathrm{sg}(\mathrm{Upsample}(f_{FM}, H_E, W_E)). \tag{5}$$

The three projected representations are:

$$\tilde{f}_E = P_E(\hat{f}_E^{(3)}), \quad \tilde{f}_{dec} = P_D(f_{dec}), \quad \tilde{f}_F = P_F(\bar{f}_{FM}), \tag{6}$$

where all projections share output shape $\mathbb{R}^{d_p \times H_E \times W_E}$, with d_p denoting the projection dimension. The two anchoring terms measure per-position cosine distances in the projected space:

$$\mathcal{L}_{sem} = \frac{1}{H_E W_E} \sum_{h,w} \left(1 - \frac{\tilde{f}_E(h,w)^\top \tilde{f}_F(h,w)}{\|\tilde{f}_E(h,w)\| \cdot \|\tilde{f}_F(h,w)\|} \right), \tag{7}$$

$$\mathcal{L}_{fm} = \frac{1}{H_E W_E} \sum_{h,w} \left(1 - \frac{\tilde{f}_{dec}(h,w)^\top \tilde{f}_F(h,w)}{\|\tilde{f}_{dec}(h,w)\| \cdot \|\tilde{f}_F(h,w)\|} \right). \tag{8}$$

The roles of the two anchoring terms are complementary. $\mathcal{L}_{sem}$ draws the FASB-modulated teacher representation toward the FM's semantic subspace, guiding the encoding pathway that feeds the one-class code to reflect the FM's structural understanding of normal objects. $\mathcal{L}_{fm}$ draws the student's deepest reconstruction toward the same subspace, constraining the student's representation toward the FM's semantic structure rather than an arbitrary nearby geometry. Together they anchor both sides of the teacher-student pair to a shared, pre-established semantic coordinate system. The combined FSRA objective is:

$$\mathcal{L}_{FSRA} = \lambda_1 \mathcal{L}_{sem} + \lambda_2 \mathcal{L}_{fm}, \tag{9}$$

where λ_1 and λ_2 are balancing hyperparameters.

The overall training objective combines the reverse distillation loss with the proposed semantic anchoring objective:

$$\mathcal{L}_{total} = \mathcal{L}_{RD} + \mathcal{L}_{FSRA}, \tag{10}$$

where $\mathcal{L}_{RD}$ denotes the multi-scale reverse distillation loss defined in the RD framework. FSRA is applied only at the highest semantic level, where CNN representations are most compatible with FM patch-token semantics. Extending FM supervision to lower texture-dominated CNN levels introduces semantically mismatched constraints that degrade localization sensitivity.

3.4. Inference

At inference, the FSRA projection heads are discarded. Given an input image $\mathbf{x}$, the FM first extracts patch-token features f_{FM}. FASB takes f_{FM} and $\{f_E^k\}$ as input to semantically

modulate $f_E^{(3)}$ and aggregate multi-scale features into the semantically enriched one-class code $\boldsymbol{\phi}$, which is subsequently decoded by the student decoder D to reconstruct $\{f_D^k\}_{k=1}^3$.

The pixel-level anomaly localization map is computed by accumulating the cosine distances between teacher and student features across all scales, followed by upsampling to input resolution and Gaussian smoothing:

$$S_{AL}(h,w) = \mathcal{G}_\sigma\left(\sum_{k=1}^{3} \Psi\left(1 - \frac{f_E^k(h,w)^\top f_D^k(h,w)}{\|f_E^k(h,w)\| \cdot \|f_D^k(h,w)\|}\right)\right), \tag{11}$$

where $\Psi(\cdot)$ denotes bilinear upsampling to the input image resolution $H \times W$, and $\mathcal{G}_\sigma$ denotes Gaussian filtering with standard deviation σ.

The image-level anomaly detection score is defined as the maximum anomaly response over all spatial positions, which avoids the dilution effect of mean-based aggregation on small-defect categories:

$$S_{AD} = \max_{h,w} S_{AL}(h,w). \tag{12}$$

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate the proposed method on two widely used unsupervised anomaly detection benchmarks: MVTec AD [37] and VisA [38]. MVTec AD contains 15 industrial categories (10 object and 5 texture classes) with pixel-level anomaly annotations, while VisA comprises 10,821 high-resolution images from 12 categories with diverse structural and surface defects. Following the standard protocol, only normal samples are used for training, whereas both normal and anomalous samples are included during testing. To further examine our method on more challenging data, we evaluate it on the MPDD [39] and BTAD [40] datasets. MPDD contains six categories of metal parts (888 normal training images, 176 normal and 282 anomalous test images) captured with varying orientations, positions, distances, and illuminations against non-uniform backgrounds. Similarly, BTAD comprises real-world images of three industrial products featuring anomalies such as body and surface defects (1799 normal training images, and 451 normal and 290 anomalous test images). Both datasets provide only normal images for training, and both normal and anomalous images for testing.

Evaluation Metrics. Image-level anomaly detection performance is evaluated using the Area Under the Receiver Operating Characteristic curve (AUROC). For anomaly localization, we report pixel-level AUROC together with the Area Under the Per-Region-Overlap curve (AUPRO) on MVTec AD, MPDD, and BTAD, and AUPRO on VisA, where AUPRO integrates the per-region overlap score up to an average per-image false-positive rate of 0.3, following the convention introduced by Bergmann et al. [41] and adopted by prior work on both benchmarks. In addition to the metrics above, we report Average Precision (AP), i.e., the area under the precision–recall curve, and F1-max, i.e., the maximum F1-score over all possible thresholds, at both the image level and the pixel level on MVTec AD. All four metrics are computed per category from the same anomaly scores as the AUROC and then averaged over the 15 categories; at the pixel level, all pixels of a category's test set are pooled. AUROC, AUPRO, and AP are threshold-independent: they are obtained by sweeping the threshold over the anomaly scores (up to a false-positive rate of 0.3 for AUPRO), so no decision threshold is selected or tuned for them. F1-max, in contrast, is obtained using the ground-truth labels of the evaluation set and therefore represents an upper bound rather than a deployable operating point. In practical deployments that require a binary decision, the threshold should be calibrated on held-out normal images only.

Implementation Details. A separate model is trained for each category. Input images are resized to 256×256, with a batch size of 16 for 200 epochs. The teacher encoder adopts a

frozen ImageNet-pretrained WideResNet50, and the student decoder is randomly initialized. All trainable modules are optimized using AdamW with learning rate and weight decay both set to 1×10^{-4}. Within FASB, the cross-attention is computed at a spatial resolution of $L = 16$ with $n_h = 4$ heads and a projection dimension of $d = 256$, under a softmax temperature of $\tau = 0.07$ and a residual coupling strength of $\lambda_s = 0.3$. FSRA projects features into a shared $d_p = 256$-dimensional space, with the two anchoring terms weighted by $\lambda_1 = 0.1$ and $\lambda_2 = 0.03$. At inference, the anomaly map is smoothed with a Gaussian filter of $\sigma = 4$. The vision foundation model is a frozen DINOv2-ViT-B/14, and experiments are conducted on a single NVIDIA RTX 4090 GPU. On MVTec AD, RD and SGRD are each trained with three random seeds (42, 222, and 1234) under identical settings, and the image-level AUROC, pixel-level AUROC, and AUPRO reported for the two methods are averaged over the three runs. On VisA, the results are obtained from a single run per category.

For the additional robustness analysis, we evaluate both RD and SGRD on MVTec AD under additive Gaussian noise with $\sigma_n \in \{0, 0.02, 0.05, 0.10\}$. Here, σ_n denotes the noise standard deviation in the original $[0, 1]$ image-intensity space. Specifically, the perturbed input is defined as $\tilde{x} = x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_n^2 I)$. Since the test images are first normalized using the ImageNet statistics, the corresponding perturbation is injected into the normalized inputs after channel-wise scaling by the inverse normalization standard deviation. Model parameters remain fixed throughout this evaluation. For each noise level, the same random seed is used for RD and SGRD, resulting in identical noise realizations for corresponding test images. All other evaluation settings, including anomaly-map smoothing and image-level score computation, remain unchanged from the main experiments.

4.2. Result Analysis

MVTec AD Dataset. Tables 1 and 2 report the image-level detection and pixel-level localization results on MVTec AD. As shown in Table 1, the proposed method achieves the best total average image-level AUROC of 98.5%, outperforming all compared methods. In particular, our method achieves perfect detection performance on six categories, including carpet, leather, tile, bottle, hazelnut, and metal nut, while maintaining consistently strong performance across both texture and object categories. None of the ten compared baselines reaches a perfect score on all six of these categories at once; several attain 100% on individual categories such as hazelnut or bottle, yet each still falls short on at least one of the remaining five. Compared with RD, the strongest baseline by total average image-level AUROC, the proposed method raises this score from 98.1% to 98.5% and surpasses RD++ by 1.6%. These results demonstrate that incorporating foundation-model semantics into reverse distillation yields more discriminative structural representations.

For pixel-level localization, Table 2 shows that the proposed method achieves the best total average AUROC and AUPRO of 98.3% and 94.6%, respectively, narrowly ahead of DBKD, the closest overall competitor, at 98.2%/94.3%. While some competing methods obtain higher scores on individual categories, they often exhibit substantial category-level performance fluctuations. For instance, RD achieves strong localization results on Screw and Zipper but performs notably worse on Wood and Metal nut, whereas RD++ performs competitively on Tile yet shows clear degradation on Cable and Transistor. In contrast, our method maintains consistently balanced localization performance across both texture and object categories, attaining the highest AUPRO among all compared methods on Carpet (97.6%) alongside a tied-best AUROC of 99.2% on the same category. Compared with RD, our method improves the total average pixel-level AUROC and AUPRO by 0.5% and 0.7%, respectively, while surpassing RD++ by 1.5% AUROC and 1.7% AUPRO. Broken down by category group, the advantage over DBKD comes mainly from the texture categories (97.9%/95.5% against 97.6%/95.1%), while the two methods are close on the object cat-

egories, at 98.5%/94.1% and 98.5%/93.9%, respectively. These results demonstrate that semantically grounding the reverse distillation process substantially improves the robustness and spatial precision of anomaly localization, with the proposed semantic grounding framework preserving texture-level sensitivity on most categories while strengthening structural discrimination.

Table 1. Image-level anomaly detection AUROC(%) on MVTec AD dataset.

	Category	UniAD [10]	Patch-Core [15]	RD [19]	MB-PFM [26]	RD++ [21]	Sim-pleNet [14]	DiAD [9]	MSFlow [25]	Efficien-tAD [1]	DBKD [29]	Ours
Textures	Carpet	99.8	97.0	98.7	100	98.0	98.2	99.4	95.5	96.0	97.1	**100**
	Grid	99.3	91.4	95.5	98.0	98.6	99.3	98.5	**99.7**	95.2	94.1	99.1
	Leather	100	100	100	100	100	99.9	99.8	99.8	98.0	100	**100**
	Tile	99.9	96.0	99.2	99.6	99.5	100	96.8	99.1	99.7	99.5	**100**
	Wood	98.9	93.8	98.7	99.5	99.0	**99.7**	**99.7**	99.5	99.3	99.1	99.3
	Texture Avg.	99.6	95.6	98.4	99.4	99.0	99.4	98.8	98.7	97.6	98.0	**99.7**
Objects	Bottle	100	100	99.8	100	99.6	99.9	99.7	100	97.9	100	**100**
	Cable	94.8	95.3	96.2	98.8	90.8	**99.1**	94.8	91.0	91.4	98.0	98.7
	Capsule	93.7	96.8	96.5	94.5	91.9	97.7	89.0	**98.0**	74.2	**98.0**	93.7
	Hazelnut	100	99.3	100	100	100	99.8	99.5	99.6	99.1	100	**100**
	Metal nut	98.3	99.1	100	100	100	95.4	99.1	99.1	95.6	100	**100**
	Pill	94.4	86.4	95.9	96.5	96.0	98.2	95.7	96.4	93.0	**98.3**	96.0
	Screw	95.3	94.2	**97.8**	91.8	88.8	93.8	90.7	89.0	86.4	95.5	96.8
	Toothbrush	89.7	**100**	99.2	88.6	**100**	90.7	99.7	96.1	96.1	92.5	99.7
	Transistor	99.8	98.9	95.7	97.8	92.1	99.1	99.8	96.4	99.6	**99.8**	98.3
	Zipper	98.6	97.1	98.5	97.4	98.5	92.8	95.1	**99.3**	90.6	98.6	96.5
	Object Avg.	96.5	96.7	98.0	96.5	95.8	96.7	96.3	96.5	92.4	**98.1**	98.0
Total Average		97.5	96.4	98.1	97.5	96.9	97.6	97.2	97.2	94.1	98.0	**98.5**

Table 2. Pixel-level anomaly localization AUROC/AUPRO (%) on MVTec AD dataset.

	Category	UniAD [10]	Patch-Core [15]	RD [19]	MB-PFM [26]	RD++ [21]	Sim-pleNet [14]	DiAD [9]	MSFlow [25]	Efficien-tAD [1]	DBKD [29]	Ours
Textures	Carpet	98.4/44.4	98.1/80.0	99.0/97.0	99.2/96.9	99.1/97.1	99.0/96.1	98.6/90.6	98.3/93.0	95.0/85.0	98.7/97.0	**99.2/97.6**
	Grid	97.7/92.9	98.4/71.4	98.9/96.2	98.8/96.0	98.7/**97.2**	98.5/94.3	96.6/86.8	99.1/93.7	91.2/87.3	98.1/95.1	**99.1**/97.1
	Leather	98.8/96.8	99.2/92.4	99.4/99.1	99.4/98.8	99.4/98.8	99.1/97.1	98.8/91.3	98.7/94.1	96.3/89.4	**99.4**/99.1	99.3/**99.2**
	Tile	92.3/78.4	90.3/85.7	95.7/90.8	96.2/88.7	96.7/91.6	**97.1**/**93.9**	92.4/90.7	95.1/85.6	93.7/37.3	96.5/92.9	95.3/91.0
	Wood	93.2/86.7	90.8/78.9	95.3/90.9	95.6/92.6	95.7/92.9	95.6/86.3	93.3/**97.5**	93.3/88.8	94.4/91.4	95.2/91.5	**96.4**/92.5
	Texture Avg.	96.1/89.8	95.4/81.7	97.7/94.8	97.8/94.6	97.9/95.5	97.9/93.5	95.9/91.4	96.9/91.0	94.1/78.1	97.6/95.1	**97.9/95.5**
Objects	Bottle	98.0/93.1	97.4/92.0	98.8/96.9	98.4/95.4	98.2/95.1	98.8/94.6	98.4/86.6	96.4/87.2	98.3/90.4	**99.0/97.1**	98.8/97.0
	Cable	96.9/86.1	93.6/91.0	97.2/90.6	96.7/**94.2**	86.9/82.4	**98.7**/93.2	96.8/80.5	96.2/85.6	97.7/87.1	98.0/92.6	98.1/93.5
	Capsule	98.8/92.1	98.0/90.9	98.7/95.7	98.3/91.7	98.6/95.6	**99.3**/94.2	97.1/87.2	99.0/96.2	96.4/88.3	98.9/**96.4**	98.5/94.8
	Hazelnut	98.0/94.1	97.6/86.7	98.9/95.6	99.1/**96.7**	99.1/96.0	98.9/93.7	98.3/91.5	98.4/95.8	98.3/90.2	**99.2**/95.5	99.1/95.7
	Metal nut	93.3/81.8	96.3/90.4	97.3/92.3	97.2/**94.6**	96.5/92.5	98.0/91.0	97.3/90.6	98.4/87.8	98.1/86.7	98.4/93.5	**98.7**/93.7
	Pill	96.1/95.3	90.6/92.3	98.2/**96.5**	97.2/96.1	98.1/96.1	**99.2**/95.7	95.7/95.7	97.8/94.1	95.7/95.1	98.0/96.3	97.9/95.6
	Screw	99.2/95.2	98.9/83.6	99.6/**98.2**	98.7/93.4	99.4/96.1	99.5/97.8	**99.7**/95.0	98.9/94.7	94.5/83.2	99.3/97.3	99.2/96.7
	Toothbrush	98.4/87.9	98.8/84.7	99.1/94.2	98.6/90.7	99.0/93.8	99.1/93.5	99.0/**95.0**	98.5/84.2	97.8/85.4	99.2/94.4	**99.7**/93.4
	Transistor	97.4/93.5	92.3/93.9	92.5/78.2	87.8/74.9	88.7/73.2	**99.0**/**94.9**	95.1/90.0	86.5/82.4	94.1/90.3	96.4/79.5	97.1/85.8
	Zipper	98.0/92.6	95.7/92.4	98.3/95.6	98.2/94.8	98.0/93.8	90.1/82.4	96.2/91.6	98.2/92.2	**98.9**/81.7	98.3/**96.4**	97.9/94.7
	Object Avg.	97.4/91.2	95.9/89.8	97.9/93.4	97.0/92.2	96.2/91.5	98.1/93.1	97.2/89.7	96.8/90.0	96.0/87.8	98.5/93.9	**98.5/94.1**
Total Average		97.0/90.7	95.7/87.1	97.8/93.9	97.3/93.0	96.8/92.9	98.0/93.2	96.8/90.7	96.8/90.4	95.4/84.6	98.2/94.3	**98.3/94.6**

Figure 4 presents qualitative localization results on representative MVTec AD categories. The proposed method accurately localizes diverse defect types while maintaining clean background suppression and precise structural boundaries. In structurally challenging categories such as grid and metal nut, the predicted anomaly maps remain compact and well aligned with the ground-truth regions, consistent with the pixel-level scores reported in Table 2, where the proposed method attains the highest AUROC among all compared methods on these two categories (99.1% and 98.7%, respectively). This agreement between the visual and quantitative results suggests that the semantic grounding introduced by FASB and FSRA helps keep the predicted anomaly region concentrated on the true defect area rather than dispersing across the background, an effect that appears most clearly in categories with a repeating global structure.

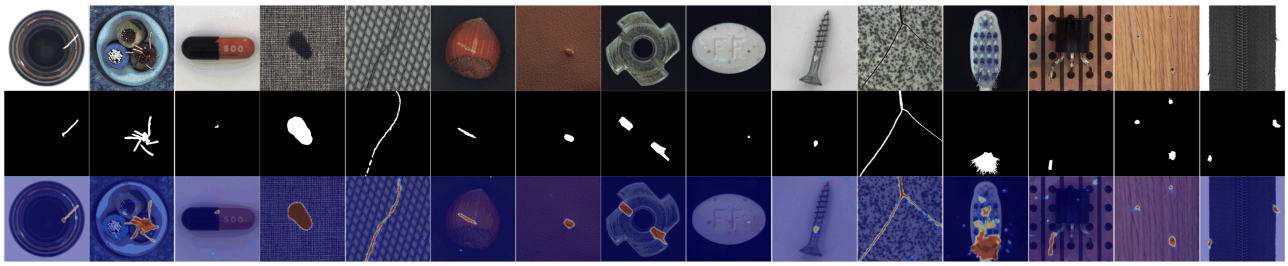


Figure 4. Qualitative anomaly localization results on the MVTec AD dataset. From top to bottom: input image, ground-truth mask, and predicted anomaly map.

Precision–Recall-Based Metrics. Table 3 reports image-level and pixel-level AP and F1-max on MVTec AD, averaged over the 15 categories. Among the methods in Table 3, SGRD obtains the highest average on all four metrics. Compared with RD, it improves image-level AP and F1-max by 0.4 percentage points each, and pixel-level AP and F1-max by 3.2 and 1.4 percentage points, respectively. The image-level differences are small, whereas the pixel-level gains are larger. Pixel-level AP and F1-max are much lower than pixel-level AUROC for all methods, because anomalous pixels account for only a small fraction of all pixels, so spurious responses on normal regions weigh heavily on both metrics. The larger pixel-level gain over RD under this stricter criterion is consistent with the weakened responses at structurally similar normal positions observed when FASB or FSRA is added to RD in the qualitative analyses in Section 4.3 .

Table 3. Image-level and pixel-level AP and F1-max (%) on MVTec AD, averaged over the 15 categories.

Method	I-AP	I-F1max	P-AP	P-F1max
UniAD [10]	98.8	96.2	43.4	49.5
DiAD [9]	99.0	96.5	52.6	55.5
RD [19]	99.0	97.5	56.3	59.6
Ours	**99.4**	**97.9**	**59.5**	**61.0**

Note: The results of UniAD and DiAD are taken from [9]. Bold numbers indicate the best results in the table.

Statistical Analysis Over Random Seeds. Table 4 reports RD and SGRD over three random seeds. Averaged over the 15 categories, SGRD improves on RD by 0.43 percentage points in image-level AUROC (95% CI [+0.23, +0.64], $p = 0.006$), by 0.50 in pixel-level AUROC ([+0.25, +0.75], $p = 0.013$), and by 0.63 in AUPRO ([+0.43, +0.84], $p = 0.002$). All three differences remain below 0.05 after Holm correction across the three metrics, and every SGRD run exceeds every RD run on all three metrics, indicating that the improvement does not depend on the particular random seed.

VisA Dataset. Table 5 reports the results on the VisA dataset, which contains more structurally diverse and fine-grained anomalies than MVTec AD. The proposed method achieves the best average image-level AUROC of 99.8%, together with a pixel-level AUPRO of 94.2% that matches the best score among the compared methods. Compared with EfficientAD [1], our method improves image-level AUROC by 1.7% and surpasses the RD baseline by 3.8%, demonstrating strong and consistent performance under challenging anomaly scenarios. Moreover, our method achieves the highest image-level AUROC on 11 out of 12 categories, including perfect scores on chewing gum, PCB3, and pipe fryum. The sole exception is cashew, where RD attains a marginally higher image-level AUROC (99.6% against 99.4%), although the proposed method still records the highest pixel-level AUPRO in this category (95.5%). For pixel-level localization, the proposed method maintains consistently reliable localization quality across heterogeneous structures and fine-grained defect patterns, particularly on complex PCB assemblies and multi-instance object categories,

and ties with DTGD [28] on the dataset average at 94.2%. These results demonstrate that semantically grounding reverse distillation with foundation-model priors improves both image-level discrimination and localization robustness on structurally diverse industrial anomalies, with the remaining gap on a handful of fine-grained categories pointing to pixel-level localization, rather than image-level detection, as the aspect most sensitive to further refinement. Figure 5 further confirms that semantically grounded reverse distillation produces spatially coherent anomaly localization across heterogeneous defect structures, highlighting its robustness under fine-grained industrial inspection scenarios.

Table 4. Multi-seed statistical comparison between RD and SGRD on MVTec AD (%).

Method	Seed	Image AUROC	Pixel AUROC	AUPRO
RD	42	98.1	97.8	93.9
	222	98.2	97.8	93.8
	1234	98.0	97.8	94.0
	Mean ± std	98.10 ± 0.10	97.80 ± 0.00	93.90 ± 0.10
SGRD	42	98.6	98.2	94.5
	222	98.5	98.3	94.5
	1234	98.5	98.4	94.6
	Mean ± std	**98.53 ± 0.06**	**98.30 ± 0.10**	**94.53 ± 0.06**
Δ [95% CI]		+0.43 [+0.23, +0.64]	+0.50 [+0.25, +0.75]	+0.63 [+0.43, +0.84]
p (Welch)		0.006	0.013	0.002

Note: Each entry is the average over the 15 categories. Std is the sample standard deviation over the three seeds. Δ denotes SGRD−RD (difference of the seed means), and its 95% confidence interval uses the Welch–Satterthwaite degrees of freedom. p is the two-sided p-value of Welch's t-test. Means are computed from the per-seed averages; the Total Average entries in Tables 1 and 2 are obtained from category values rounded to one decimal and may therefore differ in the last digit.

Table 5. Image-level anomaly detection AUROC/pixel-level anomaly localization AUPRO (%) on the VisA dataset.

Category	PatchCore [15]	RD [19]	AST [27]	DSR [11]	Efficien-tAD [1]	SimpleNet [14]	SuperSim-pleNet [24]	DMAD [13]	DTGD [28]	Ours
Candle	98.6/94.0	92.2/92.2	99.4/94.1	86.4/79.7	98.4/**95.7**	92.5/89.8	97.1/93.6	94.2/88.3	96.2/93.0	**99.6**/94.2
Capsules	81.6/85.5	90.1/56.9	85.4/68.2	93.4/74.5	93.5/**96.9**	78.9/84.8	81.5/80.1	88.7/90.7	90.5/96.7	**99.9**/95.1
Cashew	97.3/94.5	**99.6**/79.0	95.1/79.1	85.2/61.5	97.2/94.2	91.9/82.6	93.0/86.3	97.4/89.0	97.2/93.7	99.4/**95.5**
Chewing gum	99.1/84.6	99.7/**92.5**	100/78.5	97.2/58.2	99.9/83.1	99.0/84.1	99.3/87.7	98.3/85.9	99.3/86.4	**100**/91.0
Fryum	96.2/85.3	96.6/81.0	99.0/58.7	93.0/65.5	96.5/86.7	95.4/90.3	96.8/79.3	96.7/84.2	96.7/**92.1**	**99.7**/91.6
Macaroni1	97.5/95.4	98.4/71.3	93.9/87.2	91.7/57.7	99.4/**99.0**	94.2/97.3	93.1/95.9	94.7/94.6	98.4/97.7	**99.7**/94.0
Macaroni2	78.1/94.4	97.6/68.0	72.1/80.4	79.0/52.2	96.7/**98.8**	71.8/85.9	75.0/89.7	88.1/97.4	92.4/98.4	**99.9**/97.0
PCB1	98.5/94.3	97.6/43.2	99.2/89.3	89.1/61.3	98.5/**97.1**	92.5/88.7	96.9/92.9	95.6/94.0	97.5/96.6	**99.9**/95.5
PCB2	97.3/89.2	91.1/46.4	98.4/85.7	96.4/84.9	99.5/**95.0**	93.6/89.0	97.5/85.4	97.7/89.1	95.1/92.3	**99.8**/93.5
PCB3	97.9/90.9	95.5/80.3	97.4/87.7	97.0/79.5	98.9/94.0	92.6/90.3	94.4/83.0	98.7/80.9	97.5/**94.8**	**100**/93.7
PCB4	99.6/90.1	96.5/72.2	99.7/80.1	98.5/62.1	98.9/92.7	97.9/81.6	98.4/87.3	99.8/79.5	99.9/92.3	**99.9**/**93.4**
Pipe fryum	99.8/95.7	97.0/68.3	99.4/89.4	94.3/80.5	99.7/94.7	94.6/91.1	97.6/88.0	99.4/**96.6**	99.7/96.4	**100**/95.9
Average	95.1/91.2	96.0/70.9	94.9/81.5	91.8/68.1	98.1/94.0	91.2/88.0	93.4/87.4	95.8/89.2	96.7/94.2	**99.8**/**94.2**

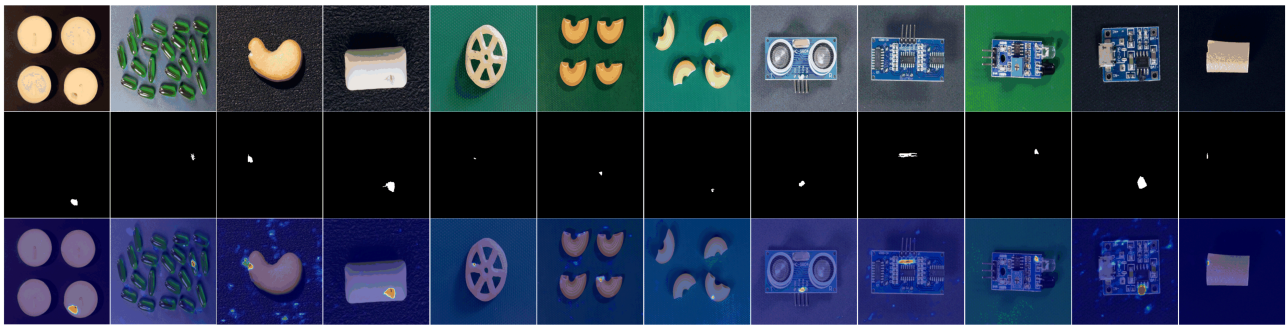


Figure 5. Qualitative anomaly localization results on VisA dataset. From top to bottom: input image, ground-truth mask, and predicted anomaly map.

Comparison with a Foundation-Model-Based Detector. Table 6 lists the results reported in the original paper of Dinomaly [12], which builds the whole detector on a DINOv2 ViT encoder-decoder and reports both a unified multi-class model and a class-separated variant. These numbers are quoted from [12] and have not been re-run under our setting; the detector is a DINOv2-based ViT encoder-decoder in Dinomaly and a WideResNet50 teacher-student pair in SGRD, and the input resolution (392×392 versus 256×256) and the training schedule also differ. Dinomaly obtains higher values on MVTec AD for all metrics in Table 6, most clearly in pixel-level AP and F1-max, while on VisA the two methods are within one percentage point on both metrics. SGRD retains the CNN teacher-student detector of reverse distillation and uses the frozen DINOv2 as a semantic reference; its contribution is evaluated against RD under the same protocol in Table 4. The superiority statements in this section refer to the methods compared in Tables 1, 2 and 5.

Table 6. Reference results of Dinomaly [12], quoted from the original paper, and SGRD (%). I/P: image/pixel level. "–": not reported.

	MVTec AD							VisA	
Method	**I-AUROC**	**I-AP**	**I-F1max**	**P-AUROC**	**P-AP**	**P-F1max**	**AUPRO**	**I-AUROC**	**AUPRO**
Dinomaly, unified multi-class	99.6	99.8	99.0	98.4	69.3	69.2	94.8	98.7	94.5
Dinomaly, class-separated	99.7	–	–	99.9	–	–	95.0	98.9	95.1
SGRD	98.5	99.4	97.9	98.3	59.5	61.0	94.6	99.8	94.2

4.3. Ablation Analysis

Effectiveness of Semantic Grounding Components. We first examine the contribution of FASB by incorporating it into the RD baseline while keeping the remaining components unchanged, as shown in Table 7. The RD baseline achieves 98.1% image-level AUROC, 97.8% pixel-level AUROC, and 93.9% AUPRO. With FASB introduced, these scores increase to 98.4%, 98.1%, and 94.5%, corresponding to gains of 0.3, 0.3, and 0.6 percentage points, respectively. As described in Section 3.2, FASB uses DINOv2 patch tokens as queries against the deepest teacher feature and injects the resulting semantically filtered response through a residual connection. The original CNN feature therefore remains in the pathway, while the deepest representation receives additional foundation-model guidance before multi-scale bottleneck aggregation. This design allows the bottleneck pathway to incorporate higher-level structural information while retaining the local visual information provided by the CNN encoder. The improvement across all three metrics suggests that the conditioned representation provides a more suitable basis for subsequent one-class encoding and student reconstruction. The 0.6 percentage-point increase in AUPRO further indicates that the benefit of the modified bottleneck representation extends to pixel-level localization, where fine-grained visual information remains important.

Table 7. Ablation studies of effectiveness of semantic grounding components on the MVTec AD dataset (%).

RD Baseline	**FASB**	**FSRA**	**Image AUROC**	**Pixel AUROC**	**AUPRO**
✓	–	–	98.1	97.8	93.9
✓	✓	–	98.4	98.1	94.5
✓	–	✓	98.2	98.0	94.5
✓	✓	✓	**98.5**	**98.3**	**94.6**

We next evaluate FSRA independently by adding it to the RD baseline without FASB, as shown in Table 7. Image-level AUROC, pixel-level AUROC, and AUPRO increase from 98.1%, 97.8%, and 93.9% to 98.2%, 98.0%, and 94.5%, respectively. In this configuration, the teacher encoder remains frozen and FASB is absent. Consequently, the encoder-side anchoring term

$\mathcal{L}_{sem}$ operates on the original teacher feature and can update the learnable projection head P_E, but it cannot modify the teacher representation or the bottleneck encoding pathway. The student-side term $\mathcal{L}_{fm}$ remains active with the same formulation as in the full model and constrains the deepest decoder representation toward the shared DINOv2 semantic reference through P_D. Thus, the FSRA-only setting retains the semantic anchoring objectives on both sides at the loss level, while only the student-side representation can be directly modified by these constraints in the absence of FASB. The resulting gains of 0.1, 0.2, and 0.6 percentage points in image-level AUROC, pixel-level AUROC, and AUPRO suggest that semantic alignment at the representation stage can improve anomaly detection and localization even without modifying the bottleneck encoding pathway. In the RD + FSRA setting, FASB is absent and the FSRA projection heads are discarded at inference, so no FM-conditioned bottleneck pathway is retained and the inference-time feature pathway and anomaly-scoring rule remain the same as in RD. When FASB and FSRA are combined, performance reaches 98.5% image-level AUROC, 98.3% pixel-level AUROC, and 94.6% AUPRO, exceeding either module applied alone on all three metrics. In the full configuration, $\mathcal{L}_{sem}$ is applied to the FASB-modulated teacher representation, allowing the semantic constraint to participate in the learning of the representation that enters the bottleneck, while $\mathcal{L}_{fm}$ constrains the corresponding student reconstruction toward the same semantic reference. The combined result is therefore consistent with FASB and FSRA contributing at different stages of the reverse distillation process, with FASB enriching the deepest teacher feature used for bottleneck aggregation and FSRA further constraining the semantic alignment between the teacher-side representation and the student reconstruction.

Qualitative Analysis of FASB. Figure 6 compares the RD baseline against RD equipped with FASB alone on six MVTec AD categories. Grid, screw, and metal nut are considered together here. Grid is grouped among the texture classes in the MVTec AD taxonomy, but its normal appearance, together with that of screw and metal nut, is built from a fairly regular, repeating spatial layout, a diamond mesh, a threaded shaft, and a tri-lobed body, which is the setting Section 3.2 identifies as most susceptible to bottleneck entanglement between structurally coherent variation and genuine defects. On grid, the RD baseline traces the diagonal crack only partially and additionally activates a few unrelated mesh intersections; once FASB is introduced, the trace becomes more continuous and most of the isolated activations recede. Metal nut shows a related pattern. The baseline correctly marks the scratch near one edge but also raises a secondary response near an unrelated cutout on the opposite side, and this secondary response is largely absent once FASB is added. Screw shows a smaller version of the same tendency, with the predicted region drawn somewhat more tightly around the annotated defect, though the change here is less pronounced than for the other two categories.

Carpet, tile, and wood are included as categories whose normal surface pattern is comparatively irregular and grain-like rather than governed by a repeating geometric layout, a setting where a bottleneck unable to separate semantic structure from texture statistics might be expected to let its response spread into visually busy but otherwise normal regions. Carpet and wood show a modest version of this effect; the region highlighted by FASB stays a little more confined to the annotated area than under the RD baseline, without a large change in overall shape. Tile shows comparatively little difference between the two settings in this instance; the baseline already produces a fairly compact response, and FASB does not visibly reshape it further. This range, from a clear correction on grid and metal nut to a barely visible one on tile, is broadly consistent with FASB contributing an average gain of 0.3 percentage points in image-level AUROC and 0.6 in AUPRO in Table 7, an improvement that need not appear with equal strength in every category or every image. The overall pattern supports the reading in Section 3.2 that FASB mainly acts by weakening

CNN activations at positions whose local statistics resemble a defect but whose relation to the object's semantic layout does not, an effect naturally more visible in categories where such structurally similar decoys are common.

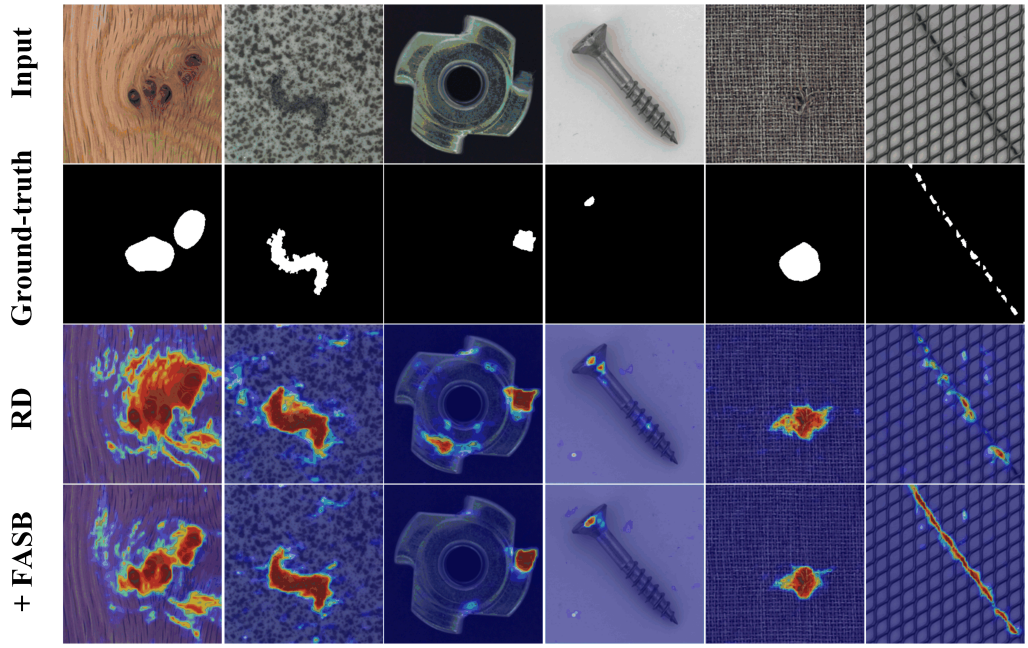


Figure 6. Qualitative anomaly localization results. From top to bottom: input image, ground-truth mask, anomaly map predicted by the RD baseline, and anomaly map predicted by RD + FASB.

Qualitative Analysis of FSRA. Figure 7 applies the same comparison to FSRA, isolated from FASB, on six categories whose typical defects rearrange the object's structure or spatial arrangement while leaving local texture largely unchanged. Transistor offers a clear illustration of this pattern. The ground truth marks two adjacent leads as bent, while a third lead beside them remains intact. The RD baseline concentrates its response mainly on one of the two bent leads and additionally raises a comparable response on the intact lead beside it, effectively treating a structurally similar but normal lead as anomalous while under-covering part of the true defect. With FSRA introduced, the response on the intact lead is largely suppressed, and the remaining response extends across both bent leads, aligning more closely with the full extent of the ground-truth mask. Cable shows a related pattern along a different axis. The ground truth marks a ring-shaped region corresponding to a missing wire core, and the RD baseline response only partially covers this ring while extending slightly into an adjacent, intact wire. With FSRA, the response covers the ring more completely and the spread into the neighboring wire is reduced. Zipper shows a similar tendency in a case where the defect occupies two separate regions along the fastener, the RD baseline responds mainly to one of them, leaving the other only weakly indicated, while the FSRA map covers both regions more evenly.

Capsule, pill, and toothbrush involve smaller, more localized marks, a fine scratch, a surface stain, and a cluster of damaged bristles, where the RD baseline already produces a reasonably compact response. FSRA brings a comparatively small refinement in these cases—a slight tightening or a marginal increase in contrast around the defect—rather than a visibly different map. Taken together, these six examples suggest that FSRA is most visible where a structurally similar but normal region would otherwise compete with the true defect for the strongest response, or where a defect spans more than one disconnected region; both situations are tied to the geometric underdetermination discussed in Section 3.3. This tendency is at least directionally consistent with the aggregate pattern reported above

in Table 7, where the gain from FSRA alone is more modest on image-level AUROC than on AUPRO relative to FASB, a pattern not incompatible with an effect that leans toward improving how completely a defect region is covered rather than raising the single peak score behind image-level detection, though the two examples discussed here are illustrative rather than a demonstration of this correspondence. Overall, in both comparisons in Table 7, adding FSRA raises all three metrics, and together with the qualitative examples in Figure 7, this is consistent with the FM-referenced anchoring of the teacher-side and student-side representations contributing to anomaly detection and localization.

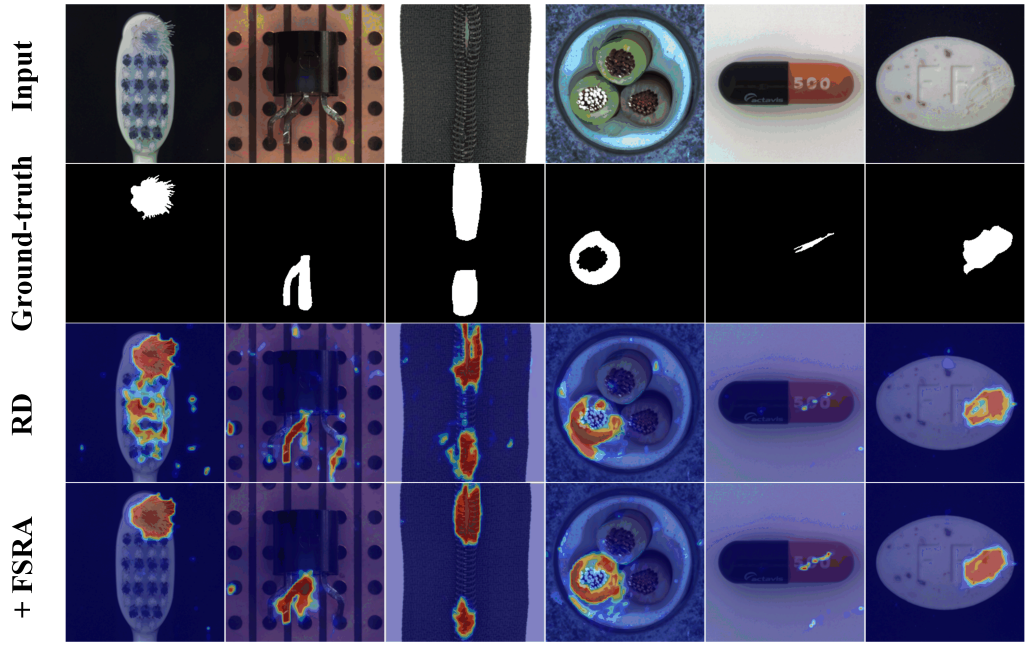


Figure 7. Qualitative anomaly localization results. From top to bottom: input image, ground-truth mask, anomaly map predicted by the RD baseline, and anomaly map predicted by RD + FSRA.

4.4. Discussion

Effect of FM Fusion Form. We examine how the FM features should enter the bottleneck pathway by replacing FASB in RD+FASB (without FSRA) with four alternatives acting on the same deepest teacher feature (Table 8): Reverse (CNN features as queries, FM tokens as keys and values), Concat (channel concatenation of the resampled FM feature and $f_E^{(3)}$ followed by a 1×1 convolution), ChannelAtt (a channel gate computed from the global descriptors of both features), and SelfAtt (self-attention on the CNN feature without FM). All variants use the same residual form with $\lambda_s = 0.3$ and warm-up schedule and the same initialization, and the attention-based variants share L, n_h, d, and τ with FASB. FASB obtains the highest value on all three metrics (98.31%, 97.97%, and 94.45%), with margins of 0.07–0.16, 0.04–0.12, and 0.22–0.28 points over the alternatives in image-level AUROC, pixel-level AUROC, and AUPRO, respectively. These results indicate that, with the same frozen DINOv2 features available, the highest scores are obtained with the query–key/value assignment adopted in FASB.

Table 8. Ablation studies of effect of FM fusion form on the MVTec AD dataset (%).

Fusion Form	Image AUROC	Pixel AUROC	AUPRO
RD	98.10	97.80	93.90
Reverse (Q: CNN, K/V: FM)	98.15	97.85	94.18
Concat	98.24	97.90	94.17
ChannelAtt	98.15	97.88	94.20
SelfAtt (w/o FM)	98.18	97.93	94.23
FASB (Q: FM, K/V: CNN)	**98.31**	**97.97**	**94.45**

Effect of Semantic Coupling Strength. We investigate the residual scaling factor λ_s in FASB to examine how the strength of foundation-model guidance affects anomaly detection performance, as shown in Figure 8. The factor λ_s controls the contribution of the semantically filtered response to the original deepest teacher feature in the residual formulation. When $\lambda_s = 0$, the residual semantic response is removed and the FASB path reduces to the original feature pathway. Since FSRA remains active in this experiment, the resulting configuration corresponds to the RD+FSRA setting in Table 7, yielding 98.2% image-level AUROC, 98.0% pixel-level AUROC, and 94.5% AUPRO. Raising λ_s to 0.1 brings a small gain of 0.1 percentage point on both AUROC metrics, with AUPRO unchanged at 94.5%. A further increase to 0.3 produces a clearer improvement, lifting the three metrics by 0.3, 0.3, and 0.1 percentage points over $\lambda_s = 0$, and this setting gives the best result among the four values tested. Pushing λ_s to 0.5 instead lowers all three metrics by 0.4, 0.4, and 0.3 percentage points relative to $\lambda_s = 0.3$, with image-level AUROC even dropping slightly below the $\lambda_s = 0$ setting. This shape of the curve, rising and then falling, fits the residual design described in Section 3.2, where the CNN activation is meant to stay the primary signal and the semantic term only supplies a correction on top of it. A moderate λ_s seems to let this correction tighten the structural coherence of the bottleneck code without pushing aside the fine-grained, texture-sensitive cues the encoder already carries, while a larger value lets the semantic term take over a larger share of the residual sum, at some apparent cost to the local sensitivity the original feature was meant to preserve. Taken together, these results point to a moderate coupling strength as the more reliable choice, with $\lambda_s = 0.3$ offering the best balance among the settings examined.

Effect of FM Supervision Granularity. We finally investigate the feature level at which DINOv2 supervision should be introduced. Table 9 compares three configurations in which FM guidance is applied to all three CNN levels, the highest two levels, or only the highest level. Applying FM supervision to all three levels yields 98.2% image-level AUROC, 97.9% pixel-level AUROC, and 94.3% AUPRO. Removing Level 1 supervision increases these values to 98.3%, 98.0%, and 94.4%, respectively. The best performance is obtained when supervision is applied only to Level 3, reaching 98.5% image-level AUROC, 98.3% pixel-level AUROC, and 94.6% AUPRO. Relative to supervision at all three levels, the Level 3-only configuration improves image-level AUROC, pixel-level AUROC, and AUPRO by 0.3, 0.4, and 0.3 percentage points, respectively. The observed trend is related to the different abstraction levels of the CNN representations. Lower-level features retain more local edge, texture, and spatial details, whereas the deepest feature contains more abstract structural information. DINOv2 patch tokens are obtained at a comparatively coarse spatial granularity and are therefore more compatible with the semantic abstraction of the deepest CNN feature. Applying the same foundation-model constraint to lower-level features may introduce a mismatch in representation granularity, where detailed local responses are directly compared with a reference that emphasizes higher-level structural information. The performance decrease observed for the all-level and two-highest-level configurations is consistent with this interpretation, particularly for pixel-level localization

and AUPRO. Restricting FM supervision to Level 3 allows the foundation model to guide the most semantically abstract CNN feature while leaving the lower-level features available to preserve the local information required for fine-grained anomaly localization.

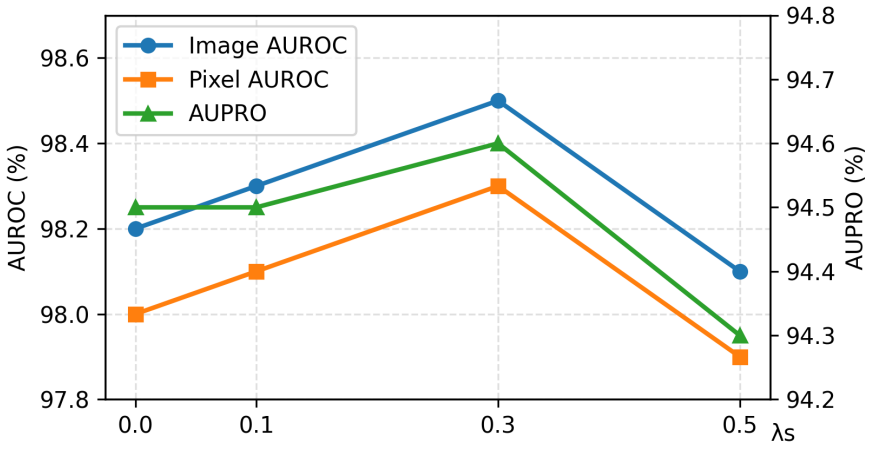


Figure 8. Ablation studies of semantic coupling strength in FASB on the MVTec AD dataset.

Table 9. Ablation studies of effect of FM supervision granularity on the MVTec AD dataset (%).

Level 1	Level 2	Level 3	Image AUROC	Pixel AUROC	AUPRO
✓	✓	✓	98.2	97.9	94.3
–	✓	✓	98.3	98.0	94.4
–	–	✓	**98.5**	**98.3**	**94.6**

Effect of Input Noise Perturbation. We further evaluate the robustness of detection and localization performance under additive Gaussian noise by varying the noise standard deviation σ_n over $\{0, 0.02, 0.05, 0.10\}$. As shown in Table 10, both RD and SGRD exhibit progressive performance degradation as the input noise increases, while SGRD generally shows smaller performance drops from the clean condition. The performance gap between SGRD and RD widens as the noise level increases and is most pronounced at $\sigma_n = 0.10$, particularly in AUPRO. On the full MVTec AD dataset, SGRD decreases from 98.5% to 76.6% in image AUROC, from 98.3% to 84.3% in pixel AUROC, and from 94.6% to 62.5% in AUPRO, corresponding to drops of 21.9, 14.0, and 32.1 percentage points, respectively. The corresponding decreases for RD are 24.9, 19.8, and 40.9 percentage points. A similar trend is observed in the texture categories, where the decreases in image AUROC, pixel AUROC, and AUPRO are 4.7, 4.1, and 6.9 percentage points for SGRD, compared with 7.1, 5.3, and 12.4 percentage points for RD. The object categories show the same direction of change, with SGRD exhibiting drops of 30.6, 18.9, and 44.7 percentage points, compared with 33.9, 27.1, and 55.2 percentage points for RD. These results suggest that SGRD is more robust to input degradation than RD. The observed trend is compatible with the design of FASB, which introduces foundation-model guidance at the deepest teacher representation through a residual connection while retaining lower-level CNN features in the bottleneck pathway. This design allows semantic guidance to be incorporated at the highest feature level while preserving access to local visual information. Since the noise experiment evaluates the fully trained SGRD model, the observed robustness reflects the combined effect of the FASB-modulated encoding pathway and the training-time semantic anchoring provided by FSRA, rather than the isolated contribution of either module.

Table 10. Robustness to additive Gaussian noise on the MVTec AD dataset (%). Each entry reports Image AUROC/Pixel AUROC/AUPRO, averaged over the corresponding category group. The Drop column denotes the absolute performance decrease from $\sigma_n = 0.00$ to $\sigma_n = 0.10$, reported in percentage points.

Group	Method	$\sigma_n = 0.00$			$\sigma_n = 0.02$			$\sigma_n = 0.05$			$\sigma_n = 0.10$			Drop (σ_n: 0 → 0.10)		
		I-AUC	P-AUC	PRO	I-AUC	P-AUC	PRO	I-AUC	P-AUC	PRO	I-AUC	P-AUC	PRO	I-AUC	P-AUC	PRO
Textures	RD	98.4	97.7	94.8	98.2	97.2	94.6	97.4	96.1	91.8	91.3	92.4	82.4	7.1	5.3	12.4
	Ours	**99.7**	**97.9**	**95.5**	**99.6**	**97.5**	**95.4**	**98.9**	**96.6**	**94.0**	**95.0**	**93.8**	**88.6**	**4.7**	**4.1**	**6.9**
Objects	RD	**98.0**	97.9	93.4	92.1	93.2	79.9	78.1	83.0	58.4	64.1	70.8	38.2	33.9	27.1	55.2
	Ours	**98.0**	**98.5**	**94.1**	**92.2**	**95.1**	**84.7**	**78.8**	**88.3**	**66.4**	**67.4**	**79.6**	**49.4**	**30.6**	**18.9**	**44.7**
Total	RD	98.1	97.8	93.9	94.1	94.5	84.8	84.5	87.4	69.5	73.2	78.0	53.0	24.9	19.8	40.9
	Ours	**98.5**	**98.3**	**94.6**	**94.7**	**95.9**	**88.3**	**85.5**	**91.0**	**75.6**	**76.6**	**84.3**	**62.5**	**21.9**	**14.0**	**32.1**

Effect of Additional Input Perturbations. Beyond additive Gaussian noise, we evaluate four further perturbations that are common in industrial inspection, namely, Gaussian blur, illumination change, random occlusion, and stripe noise, each with three severity levels given in the note of Table 11. The perturbations are generated in the original $[0, 1]$ intensity space and clamped to this range before the ImageNet normalization is applied. The illumination change scales the brightness and the contrast of each image by $1 \pm \delta$ with independently drawn signs, and stripe noise adds one offset drawn from $\mathcal{N}(0, \sigma_s^2)$ to every image row, shared by the three color channels. In SGRD, the perturbed image is also fed to the frozen DINOv2. Both methods degrade as the severity increases. SGRD scores higher than RD in all 36 comparisons of Table 11 (four perturbations, three levels, and three metrics) and loses fewer points on every metric under every perturbation. The margins at Level 1 are modest, but they widen with severity: at Level 3, SGRD exceeds RD by 0.8 to 1.8, 0.9 to 2.7, and 2.0 to 5.8 points in image AUROC, pixel AUROC, and AUPRO, respectively, and for every perturbation and metric this margin is larger than at Level 1. Stripe noise yields the largest AUPRO margin, where SGRD retains 67.9% at Level 3 against 62.1% for RD, and under Gaussian blur the AUPRO margin grows from 1.1 to 3.1 points between Level 1 and Level 3. Occlusion needs a separate reading. The image AUROC of both methods already falls to 57.1% (RD) and 58.1% (SGRD) with two blocks and stays between 55.0% and 58.1% as more blocks are added. Because the image score is the maximum response over all positions (Equation (12)), a gray block that does not occur in the normal training images can raise the score of normal and defective images alike, so the image-level differences should be interpreted with caution, and this setting is better regarded as a stress test. The pixel-level metrics still decrease with the covered area, and with eight blocks SGRD keeps 79.9% pixel AUROC and 38.2% AUPRO, compared with 78.5% and 35.2% for RD. These trends are compatible with the design of FASB, which injects DINOv2 semantics only into the deepest teacher feature through a residual path and leaves the lower-level features available for local detail, as well as with the semantic anchoring that FSRA provides during training. The absolute scores of both methods still fall sharply under occlusion and strong stripe noise, so SGRD reduces the degradation observed for RD without removing it.

Table 11. Robustness to additional test-time perturbations on the MVTec AD dataset (%). Each entry reports image AUROC/Pixel AUROC/AUPRO (I-AUC/P-AUC/PRO), averaged over all 15 categories. The Drop column denotes the absolute performance decrease from the clean condition to Level 3, reported in percentage points.

Perturbation	Method	Clean			Level 1			Level 2			Level 3			Drop (Clean → Level 3)		
		I-AUC	P-AUC	PRO	I-AUC	P-AUC	PRO	I-AUC	P-AUC	PRO	I-AUC	P-AUC	PRO	I-AUC	P-AUC	PRO
Gaussian blur	RD	98.1	97.8	93.9	97.3	97.1	92.0	90.3	94.6	85.3	82.1	91.2	76.3	16.0	6.6	17.6
	Ours	**98.5**	**98.3**	**94.6**	**97.8**	**97.7**	**93.1**	**91.8**	**95.6**	**86.6**	**83.7**	**93.1**	**79.4**	**14.8**	**5.2**	**15.2**
Illumination	RD	98.1	97.8	93.9	97.7	97.2	93.5	95.9	96.4	91.2	93.4	94.6	86.9	4.7	3.2	7.0
	Ours	**98.5**	**98.3**	**94.6**	**98.2**	**97.9**	**94.3**	**96.4**	**97.2**	**92.2**	**94.2**	**95.5**	**88.9**	**4.3**	**2.8**	**5.7**
Occlusion	RD	98.1	97.8	93.9	57.1	91.1	71.1	56.6	85.6	54.0	55.0	78.5	35.2	43.1	19.3	58.7
	Ours	**98.5**	**98.3**	**94.6**	**58.1**	**92.0**	**72.9**	**57.8**	**86.9**	**56.5**	**56.8**	**79.9**	**38.2**	**41.7**	**18.4**	**56.4**
Stripe noise	RD	98.1	97.8	93.9	92.5	91.3	79.3	85.3	86.4	70.7	78.0	82.1	62.1	20.1	15.7	31.8
	Ours	**98.5**	**98.3**	**94.6**	**93.3**	**93.4**	**82.7**	**86.6**	**88.8**	**75.0**	**79.7**	**84.8**	**67.9**	**18.8**	**13.5**	**26.7**

Note: Levels 1–3 correspond to Gaussian blur with $\sigma_b = 1, 2, 3$ pixels; illumination change with $\delta = 0.1, 0.2, 0.3$; random occlusion with 2, 4, and 8 gray 32×32 blocks; and stripe noise with $\sigma_s = 0.02, 0.05, 0.10$. Model parameters are fixed, and RD and SGRD are evaluated on identical perturbed images. Bold numbers indicate the better result of the two methods (a smaller value for the Drop columns).

Evaluation on More Challenging Datasets. To further examine SGRD under more demanding conditions, we evaluate it on MPDD and BTAD (Tables 12 and 13). On MPDD, SGRD obtains average scores of 91.1% image-level AUROC, 98.1% pixel-level AUROC, and 93.6% AUPRO, which are the highest averages among the compared methods on all three metrics. At the category level, SGRD achieves the highest AUPRO on Bracket White and Metal Plate and the highest image-level AUROC on Bracket Brown. On BTAD, SGRD obtains average scores of 94.0%, 97.5%, and 75.5%, where the average image-level and pixel-level AUROC are the highest among the compared methods and the average AUPRO is marginally higher than that of FastFlow (75.3%). On Class 02, which is the most challenging category for all methods at the image level, SGRD achieves the highest image-level AUROC (85.9%). These results indicate that SGRD performs competitively on datasets with more demanding conditions than MVTec AD and VisA.

Table 12. Image-level AUROC/pixel-level AUROC/AUPRO (%) on the MPDD dataset.

Category	CFLOW [42]	DRAEM [43]	FastFlow [44]	RD [19]	EfficientAD [1]	Ours
Bracket Black	**96.5**/**97.7**/89.5	65.4/95.9/**93.1**	62.8/94.4/79.7	50.7/96.4/89.1	85.5/97.2/90.0	83.3/96.9/88.2
Bracket Brown	49.7/95.0/83.6	70.8/76.1/58.3	47.9/92.8/85.6	68.7/**97.3**/94.0	85.3/92.5/81.8	**90.6**/97.0/**94.4**
Bracket White	79.9/96.7/78.7	84.0/86.6/69.1	50.8/90.2/49.7	83.7/98.1/84.7	**96.5**/97.7/93.3	88.6/**98.5**/**95.0**
Connector	**99.5**/97.4/91.3	83.3/90.0/72.1	52.4/92.7/77.0	93.1/98.6/95.3	50.0/**99.3**/**97.8**	95.2/99.0/94.4
Metal Plate	97.4/**98.6**/91.5	80.1/96.2/82.8	86.9/93.0/74.6	91.9/97.5/92.0	**100.0**/96.5/83.6	98.7/**98.6**/**94.6**
Tubes	**99.6**/**99.2**/**96.8**	61.3/76.9/44.8	89.4/98.4/94.9	89.7/99.1/96.0	95.3/99.0/96.3	90.4/98.4/94.9
Average	87.1/97.4/88.6	74.1/87.0/70.0	65.0/93.6/76.9	79.6/97.8/91.9	85.4/97.0/90.4	**91.1**/**98.1**/**93.6**

Table 13. Image-level AUROC/pixel-level AUROC/AUPRO (%) on the BTAD dataset.

Category	CFLOW [42]	DRAEM [43]	FastFlow [44]	RD [19]	EfficientAD [1]	Ours
Class 01	**98.6**/95.4/66.8	80.2/59.1/21.6	94.5/93.1/60.8	96.3/96.3/74.8	94.2/64.6/29.8	96.4/**96.4**/**75.6**
Class 02	82.1/94.8/54.3	65.2/69.3/27.6	84.3/96.2/**68.0**	83.0/96.4/58.9	75.4/85.7/44.4	**85.9**/**96.5**/65.5
Class 03	**100.0**/**99.6**/**98.2**	74.9/48.7/18.2	96.3/99.3/97.0	99.7/**99.6**/83.4	95.2/96.1/89.0	99.7/**99.6**/85.5
Average	93.6/96.6/73.1	73.4/59.0/22.5	91.7/96.2/75.3	93.0/97.4/72.4	88.3/82.1/54.4	**94.0**/**97.5**/**75.5**

Pretrained Knowledge of the Foundation Model. DINOv2 is pretrained with a self-supervised objective on LVD-142M, which comprises about 142 million images retrieved from a large pool of uncurated web images according to their visual similarity to images in

curated datasets such as ImageNet-22k, ImageNet-1k, and Google Landmarks [22]. Neither class labels nor text are used, and MVTec AD and VisA are not among the curated source datasets described in [22]. Since this collection is broad and diverse, DINOv2 may carry generic prior knowledge of object and texture appearance related to the categories of MVTec AD and VisA, and we do not regard the reference as free of such knowledge. Its weights remain frozen and are never updated on either benchmark, and it receives no anomalous samples or anomaly annotations during training. The model of normality for each category is learned from the normal training images of that category, and the anomaly map is still computed from the teacher-student discrepancy (Equation (11)), with DINOv2 acting through FASB and FSRA. Here, "unsupervised" follows the standard protocol of the two benchmarks, i.e., no anomalous samples or anomaly annotations are used for training. Like the compared methods built on ImageNet-pretrained backbones, SGRD relies on pretrained representations, and its results should be interpreted accordingly.

Computational Cost. In SGRD, the frozen DINOv2 and FASB are also executed at inference, so we compare its inference cost with that of RD in Table 14. The measurements use the same single NVIDIA RTX 4090 GPU with a batch size of 1 and the 256 × 256 input in FP32 precision (PyTorch 2.0.0 default settings). Latency is the mean forward-pass time of the networks over 600 runs after warm-up, measured with CUDA synchronization, and memory is the peak allocated GPU memory. SGRD has 87.56 M more parameters and 67.22 GFLOPs more than RD, its latency increases from 4.62 ms to 10.37 ms (216.4 to 96.4 FPS), and its peak memory increases from 769 MB to 1114 MB, so SGRD is more expensive than RD at inference. This increase comes almost entirely from the forward pass of the frozen DINOv2-ViT-B/14 that provides the features used by FASB: it contributes 86.58 M parameters, 66.65 GFLOPs, and 5.53 ms, whereas the FASB module itself adds 0.98 M parameters, 0.57 GFLOPs, and 0.22 ms. FSRA adds no inference cost because its projection heads are discarded, so RD+FSRA has the same inference cost as RD. Reducing the cost of obtaining the foundation-model features while retaining the semantic grounding is a direction for future work.

Table 14. Computational cost at inference.

Method	Params (M)	FLOPs (G)	Latency (ms)	FPS	Memory (MB)
RD	150.64	66.68	4.62 ± 0.24	216.4	769
SGRD	238.20	133.90	10.37 ± 0.27	96.4	1114
Increase over RD	+87.56	+67.22	+5.75	–	+345
DINOv2 (frozen)	+86.58	+66.65	+5.53	–	–
FASB	+0.98	+0.57	+0.22	–	–

Note: FLOPs count one multiply–accumulate as two operations. Latency is the mean ± standard deviation over 600 forward passes, and FPS is 1000 divided by the latency in ms. The latencies of DINOv2 and FASB are measured separately. Memory is the peak allocated GPU memory. All counts are measured on the implementation used in our experiments. FSRA is used only in training and adds no inference cost.

5. Conclusions

This paper proposes SGRD, a semantically grounded reverse distillation framework for unsupervised anomaly detection. The proposed FASB redesigns the RD bottleneck as a semantically conditioned encoding pathway by leveraging FM patch-token representations to disentangle structurally meaningful CNN activations and produce a semantically coherent one-class code. Meanwhile, FSRA addresses the geometric underdetermination of pairwise distillation by jointly anchoring teacher-side and student-side representations to the FM semantic subspace, constraining the co-learned manifold toward a semantically structured coordinate system. Extensive experiments on the MVTec AD and VisA datasets demonstrate

the effectiveness of the proposed method. On MVTec AD, SGRD achieves 98.5% image-level AUROC, 98.3% pixel-level AUROC, and 94.6% AUPRO, while on VisA, it achieves 99.8% image-level AUROC and 94.2% AUPRO. These results demonstrate effective anomaly detection and localization performance, together with strong robustness and consistently competitive performance across structurally diverse industrial inspection scenarios.

Author Contributions: Conceptualization, C.L., X.Y. and J.W.; methodology, C.L., X.Y. and J.W.; software, C.L., X.Y.; validation, C.L., X.Y., J.W., R.Y. and H.Z.; formal analysis, C.L. and X.Y.; investigation, X.Y.; resources, X.Y.; data curation, X.Y.; writing—original draft preparation, X.Y.; writing—review and editing, C.L., X.Y. and J.W.; visualization, X.Y.; supervision, C.L., J.W., R.Y. and H.Z.; project administration, X.Y.; and funding acquisition, C.L. and J.W. and R.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (62472463), Henan Key Research and Development Projects (241111220700), Henan Province Science and Technology Research and Development Program Joint Fund (252103810264), Natural Science Foundation of Henan (262300422595), Natural Science Foundation of Zhongyuan University of Technology (K2025YB004), Zhongyuan University of Technology 2026 Major Project Achievement Cultivation Program, and China National Textile and Apparel Council Technology Guiding Projects (2025055, 2025010).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The MVTec AD dataset used in this study is publicly available from the official website: https://www.mvtec.com/research-teaching/datasets/mvtec-ad (accessed on 15 August 2026). The VisA dataset is publicly available from the official GitHub repository: https://github.com/amazon-science/spot-diff (accessed on 15 August 2026) and the Registry of Open Data on AWS: https://registry.opendata.aws/visa (accessed on 15 August 2026). The MPDD dataset is publicly available from the official GitHub repository: https://github.com/stepanje/MPDD (accessed on 13 October 2026). The BTAD dataset is publicly available from the official GitHub repository: https://github.com/pankajmishra000/VT-ADL (accessed on 13 October 2026). The source code of SGRD will be made publicly available at https://github.com/Yuanyuan677/SGRD (accessed on 13 October 2026) upon acceptance of the manuscript.

Acknowledgments: The authors sincerely appreciate the helpful comments and constructive suggestions provided by the academic editors and reviewers. All individuals mentioned have consented to this acknowledgement.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Batzner, K.; Heckler, L.; König, R. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 128–138.
2. Hendria, W.F.; Kim, H.; Seo, D. Multi-model anomaly detection for industrial inspection with dynamic loss weighting and soft-hard features loss. *Neural Comput. Appl.* **2025**, *37*, 17031–17054. [CrossRef]
3. Wang, J.; Xu, G.; Li, C.; Wang, Z.; Yan, F. Surface defects detection using non-convex total variation regularized RPCA with kernelization. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5007013. [CrossRef]
4. Huang, Z.; Wang, Z.; Li, C. FESA-ViT: Frequency-Enhanced Staged Attention Vision Transformer With Decoupled Training for UAV Insulator Inspection. *IEEE Trans. Ind. Inform.* **2026**, *22*, 7043–7054. [CrossRef]
5. Cai, Y.; Zhang, W.; Chen, H.; Cheng, K.T. MedIAnomaly: A comparative study of anomaly detection in medical images. *Med. Image Anal.* **2025**, *102*, 103500. [CrossRef] [PubMed]
6. Shiri, M.; Beyan, C.; Murino, V. MadCLIP: Few-shot medical anomaly detection with CLIP. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 416–426.

7. Qian, Y.; Ye, S.; Wang, C.; Cai, X.; Qian, J.; Wu, J. Ucf-crime-dvs: A novel event-based dataset for video anomaly detection with spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25–4 March 2025*; AAAI Press: Palo Alto, CA, USA, 2025; Volume 39, pp. 6577–6585.
8. Ye, M.; Liu, W.; He, P. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 8679–8688.
9. He, H.; Zhang, J.; Chen, H.; Chen, X.; Li, Z.; Chen, X.; Wang, Y.; Wang, C.; Xie, L. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024*; AAAI Press: Palo Alto, CA, USA, 2024; Volume 38, pp. 8472–8480.
10. You, Z.; Cui, L.; Shen, Y.; Yang, K.; Lu, X.; Zheng, Y.; Le, X. A unified model for multi-class anomaly detection. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 4571–4584. [CrossRef]
11. Zavrtanik, V.; Kristan, M.; Skočaj, D. Dsr–a dual subspace re-projection network for surface anomaly detection. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 539–554.
12. Guo, J.; Lu, S.; Zhang, W.; Chen, F.; Li, H.; Liao, H. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 20405–20415.
13. Liu, W.; Chang, H.; Ma, B.; Shan, S.; Chen, X. Diversity-measurable anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 12147–12156.
14. Liu, Z.; Zhou, Y.; Xu, Y.; Wang, Z. Simplenet: A simple network for image anomaly detection and localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 20402–20411.
15. Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; Gehler, P. Towards total recall in industrial anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14318–14328.
16. Wu, S.; Wang, Y.; Liu, X.; Yang, Y.; Wang, R.; Guo, G.; Doermann, D.; Zhang, B. DFM: Differentiable feature matching for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 10–17 June 2025; pp. 15224–15233.
17. Zhang, X.; Li, S.; Li, X.; Huang, P.; Shan, J.; Chen, T. DeSTSeg: Segmentation guided denoising student-teacher for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 3914–3923.
18. Zhou, Q.; He, S.; Liu, H.; Chen, T.; Chen, J. Pull & push: Leveraging differential knowledge distillation for efficient unsupervised anomaly detection and localization. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 2176–2189. [CrossRef]
19. Deng, H.; Li, X. Anomaly detection via reverse distillation from one-class embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9737–9746.
20. Liu, X.; Wang, J.; Leng, B.; Zhang, S. Unlocking the potential of reverse distillation for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2025; Volume 39, pp. 5640–5648.
21. Tien, T.D.; Nguyen, A.T.; Tran, N.H.; Huy, T.D.; Duong, S.; Nguyen, C.D.T.; Truong, S.Q. Revisiting reverse distillation for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 24511–24520.
22. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision. *Trans. Mach. Learn. Res.* **2024**.
23. Zhang, Q.; Chen, Q. UCRT: A two-stage noisy label learning framework with uniform consistency selection and robust training. *Appl. Intell.* **2026**, *56*, 21. [CrossRef]
24. Rolih, B.; Fučka, M.; Skočaj, D. Supersimplenet: Unifying unsupervised and supervised learning for fast and reliable surface defect detection. In *Proceedings of the International Conference on Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 47–65.
25. Zhou, Y.; Xu, X.; Song, J.; Shen, F.; Shen, H.T. Msflow: Multiscale flow-based framework for unsupervised anomaly detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 2437–2450. [CrossRef] [PubMed]
26. Wan, Q.; Gao, L.; Li, X.; Wen, L. Unsupervised image anomaly detection and segmentation based on pretrained feature mapping. *IEEE Trans. Ind. Inform.* **2022**, *19*, 2330–2339. [CrossRef]
27. Rudolph, M.; Wehrbein, T.; Rosenhahn, B.; Wandt, B. Asymmetric student-teacher networks for industrial anomaly detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 2592–2602.
28. Li, N.; Liu, A.; Jiang, C.; Tang, S.; Li, Y.; Liang, Y. Dual-teacher guided denoising distillation for anomaly detection. *IEEE Access* **2025**, *13*, 174910–174921. [CrossRef]

29. Zhou, Y.; Huang, Z.; Zeng, D.; Qu, Y.; Wu, Z. Dual-branch knowledge distillation via residual features aggregation module for anomaly segmentation. *IEEE Trans. Instrum. Meas.* **2024**, *74*, 3503411. [CrossRef]
30. Chen, Z.; Lyu, C.; Zhang, L.; Li, S.; Xia, B. RDMS: Reverse distillation with multiple students of different scales for anomaly detection. *IET Image Process.* **2024**, *18*, 3815–3826. [CrossRef]
31. Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; Dabeer, O. Winclip: Zero-/few-shot anomaly classification and segmentation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 19606–19616.
32. Zhou, Q.; Pang, G.; Tian, Y.; He, S.; Chen, J. AnomalyCLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection. In Proceedings of the The Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2023.
33. Chen, X.; Han, Y.; Zhang, J. A Zero-/Few-Shot Anomaly Classification and Segmentation Method for CVPR 2023 VAND Workshop Challenge Tracks 1&2: 1st Place on Zero-shot AD and 4th Place on Few-shot AD. *arXiv* **2023**, arXiv:2305.17382.
34. Damm, S.; Laszkiewicz, M.; Lederer, J.; Fischer, A. Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. In Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 26 February–6 March 2025; pp. 1319–1329.
35. Zhang, H.; Chen, H.; Cheng, Y.; Wu, S.; Sun, L.; Han, L.; Shi, Z.; Qi, L. SuperAD: A training-free anomaly classification and segmentation method for CVPR 2025 VAND 3.0 workshop challenge track 1: Adapt & detect. *arXiv* **2025**, arXiv:2505.19750.
36. Dalmonte, F.; Bayar, E.; Akbas, E.; Georgescu, M.I. Q-former autoencoder: A modern framework for medical anomaly detection. In Proceedings of the 2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 6–10 March 2026; pp. 7985–7995.
37. Bergmann, P.; Fauser, M.; Sattlegger, D.; Steger, C. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9592–9600.
38. Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; Dabeer, O. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 392–408.
39. Jezek, S.; Jonak, M.; Burget, R.; Dvorak, P.; Skotak, M. Deep learning-based defect detection of metal parts: Evaluating current methods in complex conditions. In Proceedings of the 2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Brno, Czech Republic, 25–27 October 2021; pp. 66–71.
40. Mishra, P.; Verk, R.; Fornasier, D.; Piciarelli, C.; Foresti, G.L. VT-ADL: A vision transformer network for image anomaly detection and localization. In Proceedings of the 2021 IEEE 30th International Symposium on Industrial Electronics (ISIE), Kyoto, Japan, 20–23 June 2021; pp. 1–6.
41. Bergmann, P.; Fauser, M.; Sattlegger, D.; Steger, C. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4182–4191.
42. Gudovskiy, D.; Ishizaka, S.; Kozuka, K. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 1819–1828.
43. Zavrtanik, V.; Kristan, M.; Skočaj, D. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual Event, 11–17 October 2021; pp. 8330–8339.
44. Yu, J.; Zheng, Y.; Wang, X.; Li, W.; Wu, Y.; Zhao, R.; Wu, L. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv* **2021**, arXiv:2111.07677.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.