

Article

Cross-Modal Disagreement-Guided Reliability-Aware Scoring for RGB-3D Industrial Anomaly Detection

Jing Xu ^{1,2,3} , Pengfei Xiu ^{1,2,3}, Kun Shi ^{4,5}, Lei Xu ^{1,2,3} and Hongliang Wang ^{1,2,3,*}

¹ Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China; xujing21a@mails.ucas.ac.cn (J.X.); xiupengfei@sict.ac.cn (P.X.); xulei212@mails.ucas.ac.cn (L.X.)

² University of Chinese Academy of Sciences, Beijing 100049, China

³ Liaoning Province Human-Computer Interaction System Engineering Research Center Based on Digital Twin, Shenyang 110168, China

⁴ State Key Laboratory of Advanced Casting Technologies, Shenyang 110022, China; shikun77@163.com

⁵ China Academy of Machinery Shenyang Research Institute of Foundry Co., Ltd., Shenyang 110022, China

* Correspondence: wanghl@sict.ac.cn

Abstract

RGB-3D industrial anomaly detection seeks to jointly exploit texture and geometric cues for robust defect inspection. However, existing multimodal fusion methods still face two practical limitations: modality-specific anomaly evidence is often weakened after direct fusion, and image-level decisions remain unstable on difficult categories. To address these issues, this study develops a reliability-aware scoring enhancement on top of the released Hybrid Fusion/M3DM memory-bank pipeline. The method constructs a disagreement cue from RGB and point-cloud anomaly responses to enhance suspicious local regions and introduces a dual-branch image-level score calibration that combines a sensitive fusion branch with a robust statistical branch. Evaluated on MVTec 3D-AD under the official released-code full setting, the proposed method achieves 0.800 image-level ROCAUC, 0.980 pixel-level ROCAUC, and 0.926 AU-PRO, compared with 0.779, 0.975, and 0.915 for the corresponding released-code baseline in our environment. Additional evaluation on Eyecandies improves pixel-level ROCAUC and AU-PRO, while showing that image-level calibration remains dataset-sensitive. On a supplementary three-category Real-IAD D3 subset, the mean image-level ROCAUC, pixel-level ROCAUC, and AU-PRO improve from 0.963, 0.979, and 0.921 to 0.980, 0.988, and 0.941, respectively. These results indicate that explicit cross-modal disagreement modeling improves localization consistency, while image-level score calibration provides dataset-dependent gains rather than a uniform cross-dataset guarantee.

Keywords: industrial anomaly detection; RGB-3D fusion; point cloud; cross-modal disagreement; score calibration; multimodal inspection



Academic Editors: Douglas O'Shaughnessy and Yutaka Ishibashi

Received: 20 April 2026

Revised: 25 May 2026

Accepted: 27 May 2026

Published: 1 June 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Industrial anomaly detection is an essential technique for automated quality inspection in intelligent manufacturing [1–3]. In practical production lines, defects are often expressed through both texture variation and geometric deformation. A reliable inspection system should therefore make joint use of RGB appearance cues and 3D structural cues, while providing accurate localization together with stable image-level decisions. In industrial inspection, such decisions directly affect downstream sorting, rework, and production efficiency.

Feature-memory-based anomaly detection methods have become a strong paradigm for industrial vision because they model the distribution of normal samples and identify defects through feature deviation at inference time [4,5]. However, most early methods are designed for single-modal inputs and consequently face limitations when defect evidence is incomplete in only one modality. RGB images provide rich appearance cues but may miss subtle geometric deformation, while 3D point clouds provide structural information but are often less sensitive to fine texture irregularities. As a result, RGB–3D multimodal anomaly detection has become an important research direction for industrial inspection [2,6].

Although multimodal methods exploit complementary information from RGB and 3D data, existing fusion strategies are still not fully satisfactory in practical deployment [2,7]. The core difficulty is not simply how to concatenate two modalities, but how to preserve informative texture–geometry discrepancy while maintaining stable image-level decisions. In practice, some defects are mainly reflected in the RGB branch, whereas others are more evident in the point-cloud branch. When the two modalities are fused by direct concatenation or simple weighted aggregation, discriminative modality-specific evidence may be weakened. At the same time, many multimodal methods produce reasonable pixel-level maps but still rely on image-level statistics that are overly sensitive to local peaks, branch fluctuations, and category-specific score shift. This instability is especially problematic for difficult categories and weak anomalies.

To address these issues, we develop a cross-modal disagreement-guided reliability-aware enhancement for RGB–3D industrial anomaly detection. Instead of redesigning the full multimodal backbone, the proposed method strengthens a strong fusion baseline from two aspects. First, it explicitly models the discrepancy between RGB and 3D anomaly responses as an auxiliary cue for local anomaly enhancement. Second, it replaces a fragile single-branch image score with a stability-aware dual-branch scoring calibration that combines a sensitive fusion branch with a robust statistical branch. The final decision is obtained through adaptive blending according to reliability-related cues.

This study is developed on top of a multimodal RGB–3D fusion baseline [7] and focuses on empirical reliability enhancement rather than a complete paradigm replacement or a theoretical robustness guarantee. In this paper, robustness refers to improved empirical stability of anomaly evidence and decision aggregation under the released-code full-setting evaluation. This positioning is appropriate for industrial anomaly detection, where a controlled improvement over an existing inspection backbone is often more practical than a more complex but brittle redesign [2]. The main contributions are summarized as follows:

1. A cross-modal disagreement modeling strategy is introduced to explicitly characterize the discrepancy between RGB and point-cloud anomaly responses, providing an additional anomaly cue beyond direct fusion.
2. A dual-branch image-level score calibration strategy is proposed to adaptively combine a base fusion branch and a robust statistical branch, improving image-level performance under the released-code MVTec 3D-AD setting while revealing dataset sensitivity on additional benchmarks.
3. A practical enhancement pipeline is constructed to improve local anomaly emphasis and global anomaly discrimination without replacing the original Hybrid Fusion backbone.
4. Experimental results on MVTec 3D-AD demonstrate that the proposed enhancement improves the released-code Hybrid Fusion/M3DM baseline under the same official full-setting command in image-level ROCAUC, pixel-level ROCAUC, and AU-PRO, while additional Eyecandies and Real-IAD D3 subset evaluations examine cross-dataset transfer behavior.

2. Related Work

2.1. Memory-Based Industrial Anomaly Detection

Memory-based anomaly detection methods describe normality through representative feature banks and score test samples by their deviation from stored normal patterns. PaDiM models patch-wise Gaussian statistics of normal features, whereas PatchCore improves retrieval quality through a memory bank combined with coreset compression [4,5]. Similar memory-bank and registration ideas have also been extended to high-resolution 3D point-cloud anomaly detection, for example in Real3D-AD/Reg3D-AD and PointCore [8,9]. These methods are attractive because they are training-efficient and often provide strong localization quality. However, their standard formulation is mainly designed for single-modal image or point-cloud inspection and does not explicitly address cross-modal inconsistency between appearance and geometry.

2.2. Teacher–Student, Reconstruction, and Self-Supervised Strategies

A second line of research improves anomaly detection by amplifying the discrepancy between normal and abnormal representations through auxiliary learning objectives. Representative examples include Uninformed Students, which models normality through student–teacher disagreement, DRAEM, which combines reconstruction with discriminative embedding, and reverse distillation, which transfers one-class representations through a reverse student architecture [10–12]. Self-supervised and lightweight discriminative methods such as CutPaste and SimpleNet further show that anomaly-sensitive representations can be learned effectively without anomaly labels [13,14]. More recent variants, including AEKD, continue this line by explicitly strengthening asymmetric teacher–student discrepancy for industrial inspection [15]. Although these approaches are effective for 2D anomaly detection, they do not directly address multimodal RGB–3D fusion or the practical instability of multimodal image-level scoring.

2.3. Multimodal RGB–3D Anomaly Detection

Compared with 2D industrial anomaly detection, multimodal RGB–3D anomaly detection has been studied much less extensively. The MVTec 3D-AD benchmark established a standardized setting for jointly evaluating appearance and geometry cues [6], and subsequent analysis on this benchmark showed that effective use of 3D information is non-trivial even when strong handcrafted or classical feature representations are available [16]. Recent datasets such as Real-IAD and Real-IAD D3 further emphasize the need to evaluate industrial anomaly detection under larger-scale, real-world, and multimodal sensing conditions [17,18]. The Hybrid Fusion baseline then demonstrated that multimodal fusion can outperform unimodal reasoning by combining RGB and point-cloud branches within a memory-bank framework [7]. More recent work has extended this direction through density-aware point-cloud enhancement, multimodal feature reconstruction, latent-bridged cross-modal prediction, noise-resistant RGB–3D modeling, prototype-driven normality communication, cross-modal feature mapping, and geometry-guided score fusion [19–25]. In particular, PIRN introduces balanced prototype assignment, adaptive prototype refinement, and multimodal normality communication for few-shot multimodal anomaly detection, and evaluates the method on MVTec 3D-AD, Eyecandies, and Real-IAD [23]. Crossmodal Feature Mapping models 2D–3D feature consistency directly and reports strong standard and few-shot results on MVTec 3D-AD, while G²SF-MIAD studies geometry-guided score fusion on MVTec 3D-AD and Eyecandies [24,25]. These studies indicate that recent multimodal anomaly detection has moved beyond simple feature concatenation toward explicit cross-modal interaction, reconstruction, score fusion, and normality modeling. At the same time, recent surveys and reviews have highlighted the growing

relevance of 2D, 3D, and multimodal anomaly detection for industrial inspection and manufacturing [2,3,26]. Nevertheless, direct fusion alone does not guarantee that complementary modality-specific evidence is preserved, and strong pixel-level maps do not necessarily translate into reliable image-level decisions. The proposed method is therefore positioned as a reliability-oriented enhancement of multimodal fusion, with explicit focus on disagreement-aware local modulation and reliability-aware image-level scoring.

3. Materials and Methods

3.1. Overall Framework

The proposed enhancement contains three feature branches: an RGB branch, a point-cloud branch, and a fusion branch. Given an RGB image and its corresponding 3D point cloud, the RGB branch extracts appearance features, the point-cloud branch extracts geometric features, and the fusion branch produces joint multimodal patch embeddings. A memory bank is constructed from normal training samples, and anomaly maps are generated according to the deviation of test features from normal memory features. The final output includes a pixel-level anomaly map and an image-level anomaly score.

Compared with the baseline multimodal fusion framework, the proposed method introduces two additional components: cross-modal disagreement modeling and dual-branch image-level score calibration. The first component enhances local anomaly responses, while the second improves global decision stability. Figure 1 illustrates the overall architecture of the proposed enhancement.

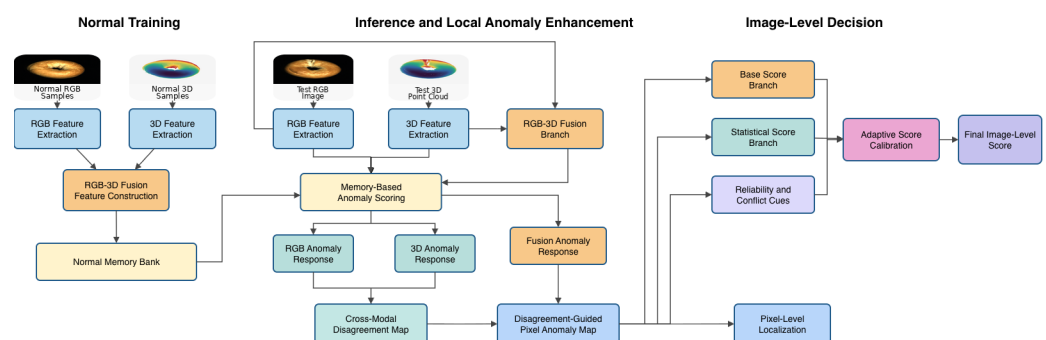


Figure 1. Overall pipeline of the proposed method. The method preserves the multimodal fusion memory-bank pipeline and augments it with two additional modules: cross-modal disagreement modeling for local anomaly enhancement and dual-branch image-level score calibration for robust global decision making. Arrows indicate the information flow among input modalities, feature branches, memory-bank scoring, disagreement modulation, and final decision output; different colors distinguish the RGB, point-cloud, fusion, and scoring components.

3.2. RGB and Point-Cloud Feature Extraction

Let I denote the RGB image and P denote the corresponding point cloud. The RGB backbone extracts feature maps from the visual modality, while the point-cloud backbone produces geometric descriptors from the 3D modality. In this work, DINO ViT-B/8 is adopted as the RGB backbone and Point-MAE is used as the point-cloud backbone [27,28]. The extracted patch features are then aligned into a common spatial layout for subsequent multimodal fusion and anomaly scoring.

The point-cloud input is first converted from organized representation to valid unorganized points. Invalid points are removed through finite-value checking and zero-point filtering, and the number of valid points is constrained to satisfy the grouping requirements of the point-cloud encoder. This preprocessing improves stability during point interpolation and feature extraction.

3.3. Fusion Branch and Normal Memory Construction

For each normal training sample, RGB patch features and point-cloud patch features are fused into multimodal patch embeddings. The normal training embeddings are accumulated into a memory bank. During inference, the anomaly strength of each test patch is estimated by its distance to the normal memory bank. This fusion memory serves as the core anomaly branch and is retained as the stable basis of the final system.

To avoid excessive memory redundancy, the method uses memory compression through a coreset-style selection strategy [4]. The compressed normal memory bank maintains representative normal patterns while keeping the inference cost acceptable.

3.4. Cross-Modal Disagreement Modeling

Although fusion features are effective, they may hide the discrepancy between modality-specific responses. To preserve such complementary information, we compute separate anomaly response maps from the RGB and point-cloud branches and normalize them to a comparable range. For a map M , min–max normalization is defined as

$$N(M) = \frac{M - \min(M)}{\max(M) - \min(M) + \epsilon}, \quad (1)$$

where ϵ is a small constant for numerical stability. The disagreement map is then defined as the absolute difference between the normalized modality-specific maps:

$$M_{\text{dis}} = |N(M_{\text{rgb}}) - N(M_{\text{xyz}})|. \quad (2)$$

This disagreement map highlights local regions where the two modalities respond differently. Such regions are often informative in industrial scenarios because certain anomalies may be more visible in one modality than in the other. Rather than replacing the fusion anomaly map, the disagreement cue is introduced as a modulation term on top of the fusion branch:

$$M_{\text{pix}} = M_{\text{fusion}} \odot (1 + \lambda M_{\text{dis}}), \quad (3)$$

where λ controls the modulation strength and $\odot$ denotes element-wise multiplication. The value $\lambda = 0.5$ was selected through preliminary module-progression experiments under the released-code full-setting evaluation and then fixed for all reported runs. In this way, the fusion branch remains the primary anomaly carrier, while cross-modal disagreement selectively amplifies locally inconsistent regions that are more likely to be anomalous.

3.5. Dual-Branch Image-Level Score Calibration

The image-level decision should not rely solely on a single maximum activation or a single Top-K statistic. For this reason, a dual-branch image-level score calibration strategy is introduced.

The first branch is the base branch. It is derived from the modulated fusion anomaly map and preserves sensitivity to dominant abnormal responses. The second branch is the robust statistical branch. This branch integrates additional calibrated cues, including percentile-based responses, tail-sensitive signals, map statistics, and branch-distribution evidence. These signals are more stable than a raw peak score and help suppress accidental fluctuations. In the implementation used for the paper, the base branch is defined as the average of the Top-K responses on the modulated fusion map:

$$S_{\text{base}} = \frac{1}{K} \sum_{i \in \mathcal{T}_K(M_{\text{pix}})} M_{\text{pix}}^{(i)}, \quad (4)$$

where $\mathcal{T}_K(\cdot)$ denotes the indices of the K largest responses. The value $K = 64$ was selected in the same preliminary released-code setting experiments and then kept unchanged. The robust branch is computed from percentile and distribution statistics:

$$S_{\text{robust}} = \alpha Q_{95}(M_{\text{pix}}) + \beta \mu(M_{\text{pix}}) + \gamma \mu(M_{\text{dis}}), \quad (5)$$

where $Q_{95}(\cdot)$ denotes the 95th percentile, $\mu(\cdot)$ denotes the spatial mean, and α, β, γ are calibration weights. The weights $(\alpha, \beta, \gamma) = (0.60, 0.25, 0.15)$ were determined through the preliminary released-code setting experiments and then fixed for the final evaluation.

Let S_{base} denote the score from the base branch and S_{robust} denote the score from the robust statistical branch. The final image score is obtained by adaptively blending the two branches:

$$S_{\text{final}} = (1 - w)S_{\text{base}} + wS_{\text{robust}} + S_{\text{rec}}, \quad (6)$$

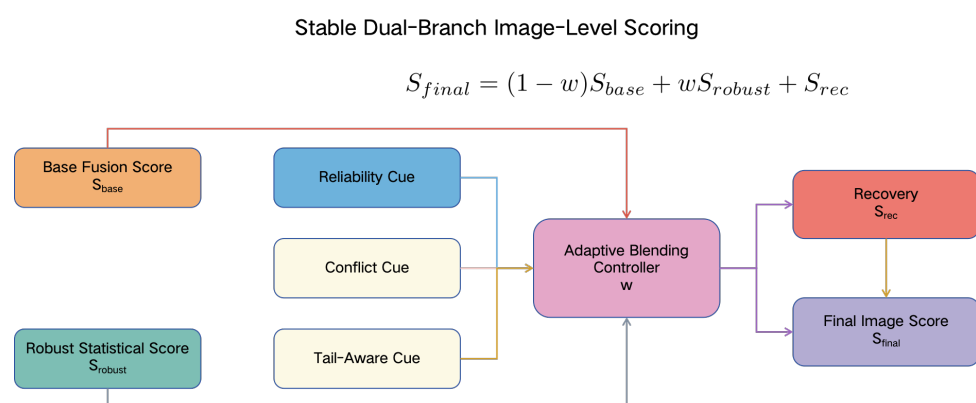
where w is the adaptive blending coefficient and S_{rec} is a recovery term used to maintain anomaly sensitivity under stable but difficult conditions. In the paper implementation, w is computed from a reliability cue R and a conflict cue C :

$$w = \sigma(\eta_1 C - \eta_2 R), \quad (7)$$

where $\sigma(\cdot)$ is the sigmoid function and η_1, η_2 are scaling parameters. The scaling parameters were set to $\eta_1 = 1.0$ and $\eta_2 = 1.0$ according to the preliminary released-code setting experiments. The recovery term is activated only when the sample is stable but the base branch is likely to underestimate the anomaly:

$$S_{\text{rec}} = \delta \max(0, Q_{95}(M_{\text{pix}}) - S_{\text{base}}). \quad (8)$$

where δ controls the recovery strength. The recovery strength was selected as $\delta = 0.20$ in the preliminary released-code setting experiments and then kept fixed. In this way, the method automatically increases the influence of the robust branch when the base branch becomes unstable while preserving sensitivity on hard but reliable cases. Figure 2 presents the proposed score calibration strategy.



The controller balances sensitivity and stability, while the recovery term preserves anomaly evidence on difficult but reliable samples.

Figure 2. Dual-branch image-level score calibration strategy. The final image score is obtained by adaptively blending a sensitive base fusion branch and a robust statistical branch using reliability and conflict cues, together with a recovery term to preserve sensitivity in difficult cases. Arrows indicate the score flow, and different colors denote the base branch, robust branch, reliability/conflict cues, and recovery term.

3.6. Stability Cues for Adaptive Blending

The adaptive blending mechanism is driven by several stability-related cues. First, branch reliability is estimated from the agreement between dominant branch scores and distribution statistics. Second, branch conflict is measured to detect cases where different anomaly evidence sources strongly disagree. Third, tail-sensitive statistics are used to characterize whether a sample lies in an extreme response region. These signals are jointly used to determine the final blending ratio and score correction intensity. In implementation terms, the reliability cue is defined as

$$R = 1 - \left| \tilde{S}_{\text{base}} - \tilde{Q}_{95}(M_{\text{pix}}) \right|, \quad (9)$$

where tildes denote min–max normalized scalar statistics under the released-code evaluation setting. The conflict cue is defined as

$$C = \left| \tilde{S}_{\text{base}} - \tilde{S}_{\text{robust}} \right|. \quad (10)$$

This keeps the calibration logic explicit and reproducible while preserving the intended interpretation of reliability as branch consistency and conflict as branch disagreement.

When branch evidence is reliable and consistent, the model relies more on the base branch; when branch responses are unstable or conflicting, the robust branch contributes more strongly. This design improves the stability of image-level anomaly prediction without sacrificing the localization ability of the underlying fusion map. In this sense, the proposed strategy can be interpreted as a reliability-aware decision calibration module built on top of multimodal anomaly evidence rather than a replacement of the original fusion representation.

3.7. Training and Inference Procedure

The method uses only normal training samples. During training, RGB patch features, point-cloud patch features, and fusion features are extracted to construct the memory bank and fit the statistical calibration components. During inference, anomaly maps are computed for RGB, point-cloud, and fusion branches. The disagreement map is generated from the RGB and point-cloud responses and used to modulate the fusion anomaly map. Image-level scoring is then performed through the proposed dual-branch score calibration strategy.

3.8. Implementation Details

The implementation follows the released Hybrid Fusion/M3DM configuration to isolate the contribution of the proposed scoring and modulation strategy. The RGB branch uses DINO ViT-B/8 features, the 3D branch uses Point-MAE features, and the multimodal fusion branch uses the released UFF module with multiple memory banks. The baseline command follows the public M3DM repository full-setting command using `-method_name DINO+Point_MAE+Fusion`, `-use_uff`, `-memory_bank multiple`, `-rgb_backbone_name vit_base_patch8_224_dino`, `-xyz_backbone_name Point_MAE`, and `-fusion_module_path checkpoints/uff_pretrain.pth`. The RGB input resolution is 224×224 , and the dataloader uses batch size 1 with four workers. For the point-cloud branch, the adopted setting uses 1024 groups with group size 128 and enforces at least 1024 valid points after preprocessing. Patch retrieval relies on Euclidean distance computed with the `cdist` operator in PyTorch (version 1.10.0+cu113), while local 3D grouping is built with kNN neighborhoods. The memory bank is compressed with sparse-random-projection coreset selection using $f_{\text{coreset}} = 0.2$ and $\epsilon_{\text{coreset}} = 0.9$. The final journal configuration uses $\lambda = 0.5$, $K = 64$, $(\alpha, \beta, \gamma) = (0.60, 0.25, 0.15)$, $(\eta_1, \eta_2) = (1.0, 1.0)$, and $\delta = 0.20$. These

scoring hyperparameters were selected from a small fixed candidate set in preliminary module-progression experiments and were then frozen for all final MVTec 3D-AD, Eyecandies, and Real-IAD D3 subset evaluations; no category-specific or dataset-specific retuning was used. Table 1 summarizes the parameter roles and the representative candidate ranges considered during preliminary selection. All paper-facing runs were executed on a single CUDA device under the same released-code evaluation setting. The comparison in this paper is therefore interpreted as an add-on scoring enhancement comparison against the released-code Hybrid Fusion/M3DM baseline, rather than as a claim of improving the literature-reported numbers in the original CVPR paper.

Table 1. Scoring hyperparameter roles and preliminary selection ranges. The final values were fixed before all reported MVTec 3D-AD, Eyecandies, and Real-IAD D3 subset evaluations.

Parameter	Role	Representative Candidate Range	Final Value
λ	Strength of disagreement-guided local modulation	{0.25, 0.50, 0.75, 1.00}	0.50
K	Number of top responses used by the base image-score branch	{32, 64, 128}	64
(α, β, γ)	Weights of percentile, map-mean, and disagreement-mean statistics in the robust branch	Convex combinations emphasizing Q_{95} while retaining mean and disagreement terms	(0.60, 0.25, 0.15)
(η_1, η_2)	Scaling of conflict and reliability cues in adaptive blending	Unit-scale and nearby balanced settings	(1.0, 1.0)
δ	Recovery strength for underestimated high-percentile responses	{0.00, 0.10, 0.20, 0.30}	0.20

4. Results

4.1. Experimental Setup

Experiments are conducted mainly on the MVTec 3D-AD dataset [6], which contains ten industrial categories and provides aligned RGB images and 3D point clouds for anomaly detection. To provide additional evidence beyond a single benchmark, a supplementary evaluation is also conducted on Eyecandies [29], a synthetic multimodal anomaly detection dataset containing ten candy-like object categories with RGB images, depth information, and pixel-level anomaly masks. In the Eyecandies experiment, depth maps are converted to point-cloud inputs following the same released-code RGB–3D pre-processing interface. In response to the need for additional real-world validation, a limited Real-IAD D3 subset experiment is also included using three downloaded categories, namely Fork_Crimp_Terminal, Knob_Cap, and Limit_Switch [18]. This subset experiment is reported as supplementary evidence only and is not intended to replace a full Real-IAD D3 benchmark evaluation. Following standard practice, three metrics are reported: image-level ROCAUC, pixel-level ROCAUC, and AU-PRO.

The proposed method is implemented with DINO ViT-B/8 as the RGB backbone and Point-MAE as the point-cloud backbone. The fusion module follows the released Hybrid Fusion/M3DM full-setting command from the public repository, <https://github.com/nomewang/M3DM> (accessed on 26 May 2026). To maintain consistency, the released-code baseline, the proposed method, and the ablation variants are evaluated with the same backbone, UFF checkpoint, multiple-memory-bank setting, input preprocessing, and evaluation scripts. For context, the main comparison includes the original Hybrid Fusion/M3DM result reported in the baseline paper [7]. The benchmark contains ten categories, namely Bagel, Cable_Gland, Carrot, Cookie, Dowel, Foam, Peach, Potato, Rope, and Tire, covering diverse combinations of texture-sensitive and geometry-sensitive

defects [6]. This category diversity makes the dataset suitable for evaluating whether a multimodal method can remain reliable when the dominant anomaly cue changes across object types. During evaluation, the model is trained only on normal samples and is tested on both normal and anomalous samples following the official protocol.

4.2. Official Released-Code Setting and Parameter Sensitivity

The main baseline is evaluated with the public Hybrid Fusion/M3DM released-code full-setting command. Specifically, the released UFF checkpoint is enabled by `-use_uff`, the decision layer uses `-memory_bank multiple`, and the DINO ViT-B/8 and Point-MAE backbones follow the released configuration. The proposed method is added after the same feature extraction, UFF fusion, and memory-bank construction stages. Therefore, the comparison tests whether the proposed disagreement-guided scoring improves the released-code baseline output under the same implementation and evaluation scripts.

Our local released-code run does not exactly match the numerical values reported in the original CVPR paper. Several factors may contribute to this discrepancy, including differences in software dependency versions, CUDA/PyTorch behavior, checkpoint loading details, point-cloud preprocessing and valid-point filtering, random-projection coreset sampling, local grouping implementation, random seeds, and metric interpolation or mask-processing details. Because these implementation-level factors can affect memory-bank anomaly detection, we report the original literature value only as context and make all claims relative to the released-code baseline actually executed in the same environment as the proposed method. To make the experimental setting transparent, Table 2 summarizes the released-code setting used for the baseline and the proposed add-on.

Table 2. Released-code Hybrid Fusion/M3DM setting used for the baseline and the proposed add-on comparison.

Item	Released-Code Hybrid Fusion/M3DM Baseline	Proposed Add-On Setting
Code source	Public M3DM repository full-setting command	Same released code path with the proposed scoring module added
Backbones	DINO ViT-B/8 and Point-MAE	Same
Fusion module	Released UFF checkpoint, <code>-use_uff</code>	Same
Memory mode	<code>-memory_bank multiple</code>	Same
RGB input	224×224 RGB input	Same
Point grouping	1024 groups with group size 128	Same
Memory compression	Sparse-random-projection coreset with $f_{\text{coreset}} = 0.2$ and $\epsilon_{\text{coreset}} = 0.9$	Same
Image score	Released-code fusion decision score	Disagreement-guided reliability-aware calibrated score
Interpretation	Local execution of the released full-setting pipeline	Add-on comparison under the same released-code setting

To answer whether the results are sensitive to released-code setting parameters, Table 3 reports representative released-code baseline variants that change the image-score rule, point grouping, or multiscale option. The proposed add-on is compared against the default released-code full setting. The baseline variants show that the released-code performance varies with these parameters, especially at the image level, while the proposed setting remains above the tested released-code baseline variants in all three mean metrics under the reported MVTec 3D-AD evaluation.

Table 3. Parameter sensitivity of representative released-code Hybrid Fusion/M3DM variants on MVTEC 3D-AD.

Setting	Image ROCAUC	Pixel ROCAUC	AU-PRO
Released-code baseline, fusion score, group 128, num 1024	0.779	0.975	0.915
Released-code baseline, pixel-max score, group 128, num 1024	0.754	0.975	0.915
Released-code baseline, pixel-TopK score, group 128, num 1024	0.731	0.975	0.915
Released-code baseline, fusion score, group 64, num 784, multiscale	0.781	0.976	0.917
Proposed add-on, same default released-code setting	0.800	0.980	0.926

4.3. Comparison with the Baseline

Table 4 reports the overall quantitative comparison among the original Hybrid Fusion/M3DM result reported in the baseline paper [7], our local released-code Hybrid Fusion/M3DM baseline, and the proposed add-on under the same released-code setting.

Table 4. Overall quantitative comparison among the original Hybrid Fusion/M3DM report, the released-code Hybrid Fusion/M3DM baseline, and the proposed add-on on MVTEC 3D-AD.

Method	Image ROCAUC	Pixel ROCAUC	AU-PRO
Hybrid Fusion/M3DM (reported in original paper)	0.945	0.992	0.964
Hybrid Fusion/M3DM (released-code baseline, local run)	0.779	0.975	0.915
Proposed Add-On (same released-code setting)	0.800	0.980	0.926

Table 4 shows that the original Hybrid Fusion/M3DM report remains higher than the local released-code baseline and the proposed add-on in pixel-level ROCAUC and AU-PRO. Therefore, this paper does not claim to outperform the original literature-reported Hybrid Fusion/M3DM result. Instead, the main comparison focuses on whether the proposed disagreement-guided modulation and score calibration improve the released-code baseline when both are evaluated under the same local environment, command configuration, and evaluation scripts.

Compared with the released-code baseline, the proposed add-on improves image-level ROCAUC on MVTEC 3D-AD under the same released-code full setting. Pixel-level ROCAUC and AU-PRO are also improved relative to the released-code baseline, showing that the local disagreement modulation strengthens localization while the image-level calibration improves decision aggregation in this dataset-specific setting. This result should be interpreted as controlled evidence under the released-code local run rather than as a general claim of uniform image-level improvement across datasets.

4.4. Additional Evaluation on Eyecandies

To further examine whether the proposed disagreement-guided enhancement transfers beyond MVTEC 3D-AD, an additional experiment is conducted on Eyecandies using the same released-code evaluation interface. For a fair controlled comparison on Eyecandies, the baseline and the proposed add-on use the same depth-to-point-cloud conversion, preprocessing, memory construction, and metric scripts. Table 5 reports the comparison

between the released-code Hybrid Fusion/M3DM baseline and the proposed method on the Eyecandies public test split.

Table 5. Additional quantitative comparison on Eyecandies under the released-code RGB–3D evaluation setting.

Method	Image ROCAUC	Pixel ROCAUC	AU-PRO
Hybrid Fusion/M3DM (released-code baseline)	0.810	0.963	0.849
Proposed Add-On (same released-code setting)	0.794	0.971	0.870

The Eyecandies results show a mixed but informative transfer behavior. The proposed method improves the two localization-oriented metrics, indicating that cross-modal disagreement modulation remains useful for local anomaly concentration on an additional RGB–D benchmark. However, the image-level metric decreases slightly. This suggests that the localization component transfers more robustly than the image-level calibration component, and that the score calibration strategy remains sensitive to dataset-specific score distributions. A likely reason is that Eyecandies differs from MVTec 3D-AD in object appearance, synthetic rendering characteristics, depth-to-point-cloud conversion, and the distribution of anomaly sizes and score tails. The robust branch uses percentile and mean statistics, which can shift when the global score distribution differs across datasets even if the local anomaly map remains well localized. Therefore, the Eyecandies experiment is interpreted as supplementary cross-dataset evidence for localization robustness, rather than as a claim of uniform improvement across all metrics.

4.5. Supplementary Evaluation on a Real-IAD D3 Subset

To further examine transfer behavior on real industrial data beyond MVTec 3D-AD and Eyecandies, a supplementary experiment is conducted on a limited Real-IAD D3 subset [18]. This experiment uses three categories that were available in the local evaluation environment: Fork_Crimp_Terminal, Knob_Cap, and Limit_Switch. The normal samples are used for memory construction, and the test split contains both normal and anomalous samples following the same released-code RGB–3D evaluation interface. Because this subset covers only three categories, it is reported as additional evidence rather than as a full Real-IAD D3 benchmark comparison. Table 6 reports the corresponding quantitative comparison.

Table 6. Supplementary quantitative comparison on a three-category Real-IAD D3 subset under the released-code RGB–3D evaluation setting.

Method	Image ROCAUC	Pixel ROCAUC	AU-PRO
Hybrid Fusion/M3DM (released-code baseline)	0.963	0.979	0.921
Proposed Add-On (same released-code setting)	0.980	0.988	0.941

The Real-IAD D3 subset results provide limited but useful supplementary evidence. On the selected three categories, the proposed method improves all three mean metrics over the released-code baseline under the same preprocessing and evaluation setting. At the same time, the experiment is intentionally interpreted conservatively because the subset is much smaller than the full Real-IAD D3 benchmark. Therefore, the result supports the potential transferability of the proposed disagreement-guided enhancement, but it does not replace a full-scale Real-IAD D3 evaluation.

4.6. Category-Wise Analysis

Category-wise image-level results on MVTec 3D-AD under the released-code setting show that the proposed method provides clear gains on several challenging categories compared with the released-code baseline. Specifically, improvements are observed on Bagel, Cable_Gland, Carrot, Cookie, Dowel, Foam, Potato, and Rope. Among them, the gain on Potato is particularly large, while Rope and Cable_Gland also benefit substantially. These improvements suggest that the proposed design is effective when multimodal evidence is complementary but unevenly distributed across appearance and geometry.

At the same time, the category-wise results also show that the improvement is not uniform across all categories. In particular, Peach and Tire remain difficult in the released-code setting, and their image-level performance does not consistently exceed the released-code baseline. These results indicate that the proposed disagreement cue is not equally informative for all product types. A plausible reason is that the proposed local modulation is most beneficial when RGB and 3D responses are complementary or asymmetric. For categories in which the two modality responses are weak, highly correlated, or affected by broad normal structural variation, the disagreement map may provide less separable evidence for image-level scoring. In addition, image-level calibration can be affected by category-specific score distributions when abnormal regions are small or when normal geometric variability produces high tail responses. This observation motivates future category-adaptive scoring mechanisms. Figure 3 reports the category-wise image-level comparison on MVTec 3D-AD.

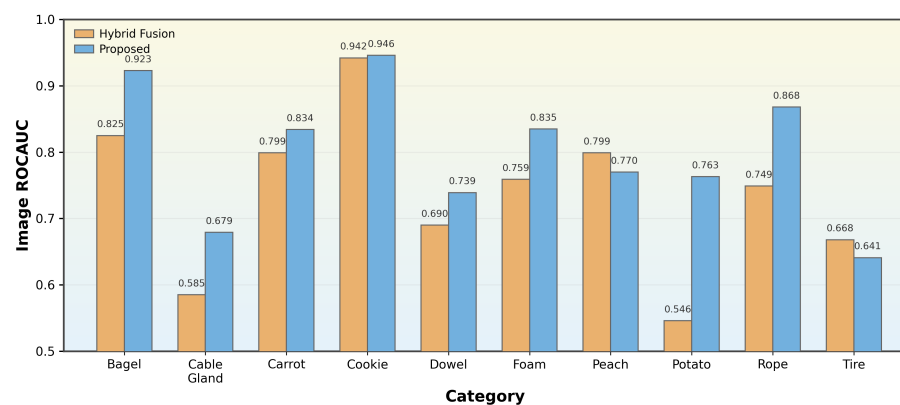


Figure 3. Category-wise image-level ROCAUC comparison between the released-code Hybrid Fusion/M3DM baseline and the proposed method on MVTec 3D-AD under the same released-code setting. Higher values indicate better image-level discrimination; the comparison highlights both improved and remaining difficult categories.

4.7. Pixel-Level Localization Performance

On MVTec 3D-AD, compared with the released-code baseline, the proposed method improves both pixel-level ROCAUC and AU-PRO. This result indicates that disagreement-guided local modulation strengthens the anomaly map without destabilizing it under the released-code setting. The gains on Cable_Gland, Cookie, Dowel, Peach, Rope, and Tire further suggest that cross-modal discrepancy is useful for enhancing local anomaly concentration.

Representative qualitative results on Rope, Dowel, Cable_Gland, and Potato are shown in Figure 4.

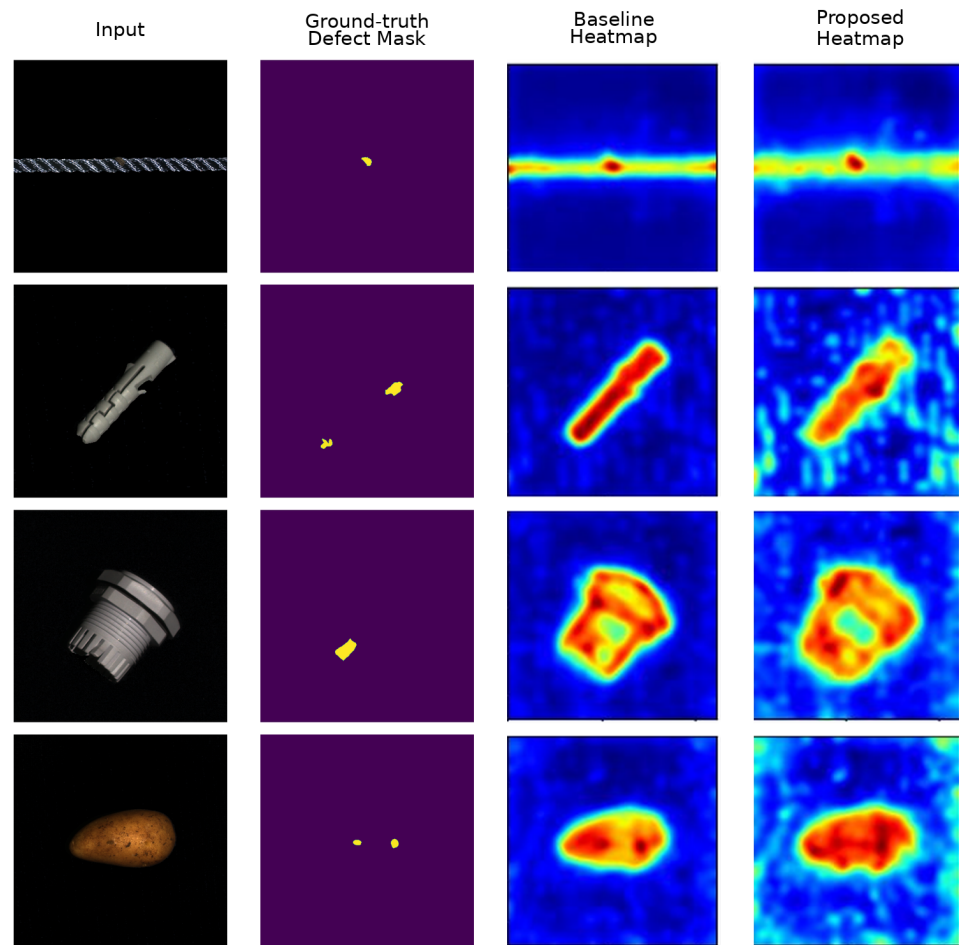


Figure 4. Representative qualitative comparison between the released-code baseline method and the proposed method on four challenging MVtec 3D-AD categories. The columns show the input image, the ground-truth defect mask, the baseline anomaly heatmap, and the proposed anomaly heatmap. Warmer colors in the heatmaps indicate higher anomaly responses.

4.8. Module Progression and Ablation Analysis

In addition to the main released-code comparison, a module progression comparison was conducted under the same released-code evaluation setting using the released-code baseline and three representative module configurations. This comparison is intended to show how disagreement modulation and score calibration affect the released pipeline, rather than to compare independent release versions.

To further validate the role of each component, a formal ablation block was evaluated under the same released-code evaluation setting with four representative settings: the released-code baseline, a disagreement-only variant, a disagreement-plus-score-calibration variant, a partial statistical calibration variant, and the full statistical calibration variant. The disagreement-only variant improves the localization-oriented metrics but slightly decreases image-level ROCAUC, while the later variants maintain most of the segmentation advantage and shift the emphasis toward more reliable image-level aggregation.

The ablation results are non-monotonic but informative. Disagreement modeling mainly strengthens localization, whereas reliability-aware statistical calibration is needed to recover and improve image-level discrimination. The final proposed setting provides the best overall balance and is the only reported variant that improves all three mean metrics over the released-code baseline. The module progression results are reported in Table 7, and the formal ablation settings are reported in Table 8. The “Full Statistical Calibration” row denotes the most complete calibration variant within the ablation naming system

only; it is not identical to the final paper setting used for the main comparison. Figure 5 summarizes the overall ablation trend across the three reported metrics.

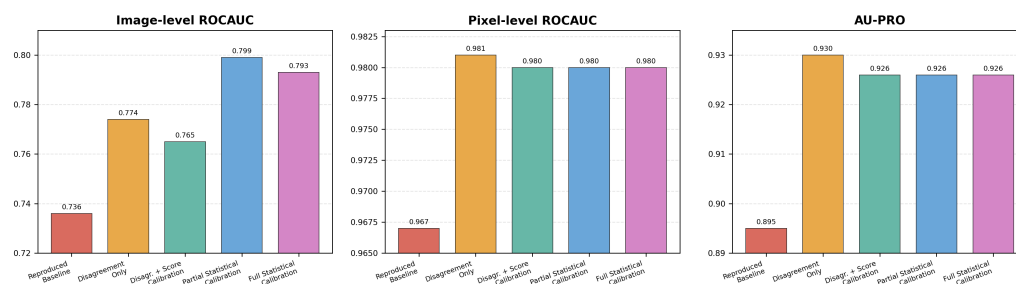


Figure 5. Overall trend of the formal ablation results. Disagreement modeling mainly improves localization, while later calibration variants place greater emphasis on reliable image-level aggregation.

Table 7. Module progression comparison on MVTec 3D-AD under the released-code evaluation setting.

Method	Image ROCAUC	Pixel ROCAUC	AU-PRO
Released-Code Hybrid Fusion/M3DM Baseline	0.779	0.975	0.915
+ Disagreement Modulation	0.774	0.981	0.930
+ Reliability-Aware Score Calibration	0.799	0.980	0.926
Final Proposed Setting	0.800	0.980	0.926

Table 8. Formal ablation study on MVTec 3D-AD under the released-code evaluation setting. The ablation labels denote component settings rather than internal release names; “Full Statistical Calibration” is an ablation variant and is not identical to the final proposed setting in Table 7.

Variant	Disagr.	Score Calib.	Stat. Calib.	Img AUC	Pix AUC	AU-PRO
Released-Code Baseline	No	No	No	0.779	0.975	0.915
Disagreement Only	Yes	No	No	0.774	0.981	0.930
Disagreement + Score Calibration	Yes	Yes	No	0.765	0.980	0.926
Partial Statistical Calibration	Yes	Yes	Partial	0.799	0.980	0.926
Full Statistical Calibration	Yes	Yes	Full	0.793	0.980	0.926

5. Discussion

The results indicate that simple multimodal fusion can be insufficient when RGB and 3D responses are asymmetric. The proposed method addresses this problem from two complementary perspectives. First, it preserves modality discrepancy rather than suppressing it. Second, it does not rely on a single image-level statistic but instead builds a stability-aware score aggregation mechanism. In other words, the method is not merely a stronger feature extractor, but a more cautious decision layer for multimodal anomaly evidence.

A main strength of the proposed method is its reliability-oriented design. In industrial settings, image-level false alarms or missed detections may be more costly than slight differences in pixel-level map smoothness. Therefore, the gain in image-level ROCAUC on MVTec 3D-AD is meaningful within the released-code local evaluation setting. The method also remains relatively lightweight because it does not require anomaly samples during training and is implemented as an enhancement over a memory-bank baseline. This makes it suitable for scenarios in which reusing an existing inspection backbone is more practical than replacing the full pipeline.

The empirical evidence in this paper should be interpreted with a clear boundary. On MVTEC 3D-AD, disagreement-aware local modulation and reliability-aware score calibration improve the released-code Hybrid Fusion/M3DM baseline under the same released-code setting. On Eyecandies, localization-oriented metrics improve but image-level ROCAUC decreases slightly, showing that the image-level calibration component is sensitive to dataset-specific score distributions. This behavior is consistent with the design of the proposed method: disagreement-guided modulation operates locally on the anomaly map, whereas image-level calibration depends on dataset-level score statistics such as percentiles, means, reliability cues, and conflict cues. Consequently, localization can transfer more reliably than global score aggregation when the test dataset has different rendering properties, depth quality, anomaly area distribution, or normal-score tails. On the three-category Real-IAD D3 subset, the proposed method improves all three mean metrics, but this experiment remains a limited subset evaluation rather than a full benchmark. Taken together, the results support the usefulness of disagreement-guided local enhancement more consistently than they support a universal image-level improvement claim.

To avoid overinterpreting the results, Table 9 summarizes the main claims supported by the experiments and their explicit boundaries.

Table 9. Supported claims and explicit boundaries of the present study.

Claim	Supporting Evidence	Boundary
The proposed add-on improves the released-code Hybrid Fusion/M3DM baseline on MVTEC 3D-AD	Tables 4, 7 and 8 show improvements over the local released-code baseline	This is not a claim of outperforming the original CVPR-reported Hybrid Fusion/M3DM result
Cross-modal disagreement is useful for local anomaly enhancement	Pixel-level ROCAUC and AU-PRO improve on MVTEC 3D-AD, Eyecandies, and the Real-IAD D3 subset	Localization gains are more consistent than image-level gains
Reliability-aware score calibration can improve image-level aggregation	MVTEC 3D-AD image-level ROCAUC improves from 0.779 to 0.800 under the released-code setting	Eyecandies shows that image-level calibration is dataset-sensitive
The method is a lightweight enhancement of an existing pipeline	The same backbones, UFF fusion module, and memory-bank structure are retained	The method is not presented as a new theoretical anomaly detection paradigm
The Real-IAD D3 result provides additional real-world evidence	The three-category subset improves from 0.963/0.979/0.921 to 0.980/0.988/0.941	The subset result does not replace a full Real-IAD D3 benchmark

5.1. Practical Implications for Industrial Inspection

From an engineering viewpoint, the proposed method is suitable for deployment scenarios in which product categories share the same inspection hardware but exhibit heterogeneous defect characteristics. The method does not require anomaly supervision, preserves the baseline memory-bank pipeline, and improves decision aggregation on MVTEC 3D-AD while maintaining strong localization quality across the additional evaluations. These properties are relevant for industrial quality control because they reduce the cost of data annotation, preserve compatibility with existing multimodal inspection systems, and provide a configurable pass/fail decision layer for downstream automation.

The qualitative results further support this practical value. In categories such as Rope, Dowel, Cable_Gland, and Potato, the proposed method produces anomaly responses that are visually more concentrated and less noisy than the baseline method. This behavior is also important for human-in-the-loop inspection, because concentrated anomaly maps are easier to interpret during manual verification and root-cause analysis.

5.2. Limitations and Future Work

This study still has limitations. First, the main method selection and primary quantitative validation are based on MVTec 3D-AD. Although Eyecandies and a Real-IAD D3 subset are additionally evaluated, the Real-IAD D3 experiment covers only three categories and therefore should not be interpreted as a full Real-IAD D3 benchmark result. Second, although the baseline is now executed from the public M3DM released-code full-setting command, our local baseline remains lower than the original Hybrid Fusion/M3DM value reported in the CVPR paper. The present results should therefore be interpreted as an add-on comparison against the released-code baseline in the same local environment, not as a claim of outperforming the original literature-reported result. Third, the parameter sensitivity analysis covers representative image-scoring and point-grouping choices, but it is not an exhaustive search over all possible implementation details, random seeds, dependency versions, and hardware environments. Fourth, some categories such as Peach and Tire remain difficult at the image level in the released-code MVTec 3D-AD setting, and the Eyecandies experiment also shows a slight decrease in image-level ROCAUC. These observations indicate that disagreement cues and score calibration are not equally beneficial across all product types and datasets. Fifth, the current paper focuses on an engineering reliability enhancement rather than a new theoretical anomaly detection paradigm, and the selected hyperparameters are obtained from preliminary module-progression experiments rather than from an exhaustive multi-seed search. Sixth, few-shot RGB–3D anomaly detection is not evaluated in the present study because this work focuses on the released-code full normal-training memory-bank setting. Extending the proposed disagreement-guided calibration strategy to few-shot multimodal anomaly detection will be investigated in future work.

Nevertheless, this positioning is still meaningful for industrial deployment, because many production settings value reproducibility, interpretability, and stable pass/fail decisions more than aggressive architectural novelty. Future work can build on this study by investigating category-adaptive cue weighting, stronger multimodal calibration strategies, full-scale evaluation on additional RGB–3D industrial datasets or real production data, and repeated multi-seed validation to further quantify stability. In particular, complete evaluation on Real-IAD D3 and comparison with more recent multimodal methods would be necessary before making stronger cross-dataset claims [18]. More broadly, the growing literature on out-of-distribution detection in 3D applications also indicates that future industrial anomaly detection systems should be evaluated not only for within-benchmark accuracy but also for robustness under distribution shift [30].

6. Conclusions

This study proposed a cross-modal disagreement-guided reliability-aware scoring enhancement for RGB–3D industrial anomaly detection. By explicitly modeling the discrepancy between RGB and point-cloud anomaly responses and by introducing a dual-branch image-level score calibration strategy, the proposed method enhances a released Hybrid Fusion/M3DM pipeline without replacing its core memory-bank structure. Experiments on MVTec 3D-AD show that the final proposed setting reaches 0.800 image-level ROCAUC, 0.980 pixel-level ROCAUC, and 0.926 AU-PRO under the released-code setting, improving the released-code local baseline on all three reported metrics. Additional evaluation on Eyecandies improves pixel-level ROCAUC and AU-PRO, although image-level ROCAUC decreases slightly. A supplementary three-category Real-IAD D3 subset experiment further shows improvements over the released-code baseline, but it is not a substitute for a full benchmark evaluation. Overall, the results support the value of disagreement-aware local enhancement and indicate that image-level calibration can be useful under a controlled

released-code setting, while also showing that such calibration is dataset-sensitive. Future work will consider full-scale cross-dataset validation, comparison with more recent multimodal methods, repeated multi-seed evaluation, and more adaptive multimodal calibration mechanisms.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The MVTec 3D-AD, Eyecandies, and Real-IAD D3 datasets are available from their official sources. The implementation code, configuration files, run commands, and processed experimental records supporting the findings of this study are available from the corresponding author upon reasonable request for academic research purposes and will be released in a public repository after publication.

Conflicts of Interest: Author Kun Shi was employed by the company China Academy of Machinery Shenyang Research Institute of Foundry Co. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

RGB	Red, green, and blue image modality
3D	Three-dimensional
ROCAUC	Area under the receiver operating characteristic curve
AU-PRO	Area under the per-region overlap curve
DRAEM	Discriminatively trained reconstruction embedding for anomaly detection
AEKD	Auto-encoder knowledge distillation
PIRN	Prototypical-based intra-modal reconstruction with normality communication
MAD	Multimodal anomaly detection

References

1. Bergmann, P.; Batzner, K.; Fauser, M.; Sattlegger, D.; Steger, C. The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. *Int. J. Comput. Vis.* **2021**, *129*, 1038–1059. [\[CrossRef\]](#)
2. Lin, Y.; Chang, Y.; Tong, X.; Yu, J.; Liotta, A.; Huang, G.; Song, W.; Zeng, D.; Wu, Z.; Wang, Y.; et al. A Survey on RGB, 3D, and Multimodal Approaches for Unsupervised Industrial Image Anomaly Detection. *Inf. Fusion* **2025**, *121*, 103139. [\[CrossRef\]](#)
3. Li, G.; Jiang, C.; Li, M.; Li, J.; Han, D.; Zhou, M. Industrial-Application-Oriented 2D Image and 3D Object Anomaly Detection Technology: A Comprehensive Review. *Appl. Intell.* **2025**, *55*, 938. [\[CrossRef\]](#)
4. 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 (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 14318–14328. [\[CrossRef\]](#)
5. Defard, T.; Setkov, A.; Loesch, A.; Audigier, R. PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection and Localization. In *Pattern Recognition. ICPR International Workshops and Challenges*; Springer: Cham, Switzerland, 2021; pp. 475–489. [\[CrossRef\]](#)
6. Bergmann, P.; Jin, X.; Sattlegger, D.; Steger, C. The MVTec 3D-AD Dataset for Unsupervised 3D Anomaly Detection and Localization. In Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), Virtual, 5–7 February 2022; pp. 202–213. [\[CrossRef\]](#)

7. Wang, Y.; Peng, J.; Zhang, J.; Yi, R.; Wang, Y.; Wang, C. Multimodal Industrial Anomaly Detection via Hybrid Fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 8032–8041. [\[CrossRef\]](#)
8. Liu, J.; Xie, G.; Chen, R.; Li, X.; Wang, J.; Liu, Y.; Wang, C.; Zheng, F. Real3D-AD: A Dataset of Point Cloud Anomaly Detection. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2023; pp. 30402–30415. Available online: https://proceedings.neurips.cc/paper_files/paper/2023/hash/611b896d447df43c898062358df4c114-Abstract-Datasets_and_Benchmarks.html (accessed on 26 May 2026).
9. Zhao, B.; Xiong, Q.; Zhang, X.; Guo, J.; Liu, Q.; Xing, X.; Xu, X. PointCore: An Efficient Framework for Unsupervised Point Cloud Anomaly Detection Using Joint Local–Global Features. *Neural Netw.* **2026**, *197*, 108446. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Bergmann, P.; Fauser, M.; Sattlegger, D.; Steger, C. Uninformed Students: Student–Teacher Anomaly Detection with Discriminative Latent Embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 4183–4192. [\[CrossRef\]](#)
11. 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 (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 8330–8339. [\[CrossRef\]](#)
12. 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 (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 9737–9746. [\[CrossRef\]](#)
13. Li, C.-L.; Sohn, K.; Yoon, J.; Pfister, T. CutPaste: Self-Supervised Learning for Anomaly Detection and Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 9664–9674. [\[CrossRef\]](#)
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 (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 20402–20411. [\[CrossRef\]](#)
15. Wu, Q.; Li, H.; Tian, C.; Wen, L.; Li, X. AEKD: Unsupervised Auto-Encoder Knowledge Distillation for Industrial Anomaly Detection. *J. Manuf. Syst.* **2024**, *73*, 182–194. [\[CrossRef\]](#)
16. Horwitz, E.; Hoshen, Y. Back to the Feature: Classical 3D Features Are (Almost) All You Need for 3D Anomaly Detection. *arXiv* **2022**, arXiv:2203.05550. [\[CrossRef\]](#)
17. Wang, C.; Zhu, W.; Gao, B.-B.; Gan, Z.; Zhang, J.; Gu, Z.; Qian, S.; Chen, M.; Ma, L. Real-IAD: A Real-World Multi-View Dataset for Benchmarking Versatile Industrial Anomaly Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 22883–22892. Available online: https://openaccess.thecvf.com/content/CVPR2024/html/Wang_Real-IAD_A_Real-World_Multi-View_Dataset_for_Benchmarking_Versatile_Industrial_Anomaly_CVPR_2024_paper.html (accessed on 26 May 2026).
18. Zhu, W.; Wang, L.; Zhou, Z.; Wang, C.; Pan, Y.; Zhang, R.; Chen, Z.; Cheng, L.; Gao, B.-B.; Zhang, J.; et al. Real-IAD D3: A Real-World 2D/Pseudo-3D/3D Dataset for Industrial Anomaly Detection. *arXiv* **2025**, arXiv:2504.14221. [\[CrossRef\]](#)
19. Li, H.; Niu, Y.; Yin, H.; Mo, Y.; Liu, Y.; Huang, B.; Wu, R.; Liu, J. DAUP: Enhancing Point Cloud Homogeneity for 3D Industrial Anomaly Detection via Density-Aware Point Cloud Upsampling. *Adv. Eng. Inform.* **2024**, *62*, 102823. [\[CrossRef\]](#)
20. Wang, J.; Niu, Y.; Huang, B. Fusion-Restoration Model for Industrial Multimodal Anomaly Detection. *Neurocomputing* **2025**, *637*, 130073. [\[CrossRef\]](#)
21. Shangguan, W.; Wu, H.; Niu, Y.; Yin, H.; Yu, J.; Chen, B.; Huang, B. CPIR: Multimodal Industrial Anomaly Detection via Latent Bridged Cross-Modal Prediction and Intra-Modal Reconstruction. *Adv. Eng. Inform.* **2025**, *65*, 103240. [\[CrossRef\]](#)
22. Wang, C.; Zhu, H.; Peng, J.; Wang, Y.; Yi, R.; Wu, Y.; Ma, L.; Zhang, J. M3DM-NR: RGB–3D Noisy-Resistant Industrial Anomaly Detection via Multimodal Denoising. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 9981–9993. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Li, Y.; Yang, X.; Zhang, J.; Tian, S.; Liao, J.; Liu, F. PIRN: Prototypical-Based Intra-Modal Reconstruction with Normality Communication for Multimodal Anomaly Detection. In Proceedings of the International Conference on Learning Representations (ICLR), Rio de Janeiro, Brazil, 24–28 April 2026. Available online: <https://openreview.net/forum?id=7L7kmHHfgf> (accessed on 26 May 2026).
24. Costanzino, A.; Zama Ramirez, P.; Lisanti, G.; Di Stefano, L. Multimodal Industrial Anomaly Detection by Crossmodal Feature Mapping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 17234–17243. [\[CrossRef\]](#)
25. Tao, C.; Cao, X.; Du, J. G²SF-MIAD: Geometry-Guided Score Fusion for Multimodal Industrial Anomaly Detection. *arXiv* **2025**, arXiv:2503.10091. [\[CrossRef\]](#)
26. Du, J.; Tao, C.; Cao, X.; Tsung, F. 3D Vision-Based Anomaly Detection in Manufacturing: A Survey. *Front. Eng. Manag.* **2025**, *12*, 343–360. [\[CrossRef\]](#)

27. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 9650–9660. [[CrossRef](#)]
28. Pang, Y.; Wang, W.; Tay, F.E.H.; Liu, W.; Tian, Y.; Yuan, L. Masked Autoencoders for Point Cloud Self-Supervised Learning. In *Computer Vision—ECCV 2022*; Springer: Cham, Switzerland, 2022; pp. 604–621. [[CrossRef](#)]
29. Bonfiglioli, L.; Toschi, M.; Silvestri, D.; Fioraio, N.; De Gregorio, D. The Eyecandies Dataset for Unsupervised Multimodal Anomaly Detection and Localization. In *Computer Vision—ACCV 2022*; Springer: Cham, Switzerland, 2023; pp. 459–475. [[CrossRef](#)]
30. Li, Z.; Kang, X.; West, J.; Khoshelham, K. Out-of-Distribution Detection in 3D Applications: A Review. *arXiv* **2025**, arXiv:2507.00570. [[CrossRef](#)]

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.