DepthGate: Confidence-Gated Depth Verification for Multimodal Industrial Anomaly Detection

Soyeon Kim, Tae-Yong Kim, and Jongpil Jeong(✉)

Department of Smart Factory Convergence, Sungkyunkwan University, Suwon-si, Gyeonggi-do 16419, Republic of Korea
{kimsoyeon744,jpjeong}@skku.edu, skywin94@naver.com

Abstract. Multimodal industrial anomaly detection often combines RGB and depth, but most existing methods fuse the two modalities at every inference step regardless of whether depth actually helps. We show this is harmful: on MVTec 3D-AD, naive fusion drops mean Image-Area Under the Receiver Operating Characteristic curve (I-AUROC) by 5.2 points compared to an RGB-only baseline. We propose DepthGate, which pairs a training-free depth memory bank with EfficientAD and gates depth at two levels. A category-level gate uses a mini-AUROC check to decide whether depth is discriminative for a given object class; an unreliable class is permanently bypassed. A sample-level gate then applies a small residual depth correction only when the depth score exceeds the RGB score. Across all 10 MVTec 3D-AD categories, DepthGate improves mean I-AUROC by 3.1 points while leaving Pixel-AUROC (P-AUROC) unchanged. For geometric-defect categories such as cable_gland and cookie, gains reach 16.9 and 16.4 points. Because the depth branch needs no gradient-based training, deploying to a new product class only requires populating the memory bank with normal samples.

Keywords: Anomaly Detection · Multimodal Fusion · Confidence-based Gating · Industrial Inspection · RGB-D Fusion

1 Introduction

Industrial anomaly detection aims to identify defective products in manufacturing environments through automated visual inspection [1]. As modern production lines demand ever-higher throughput and quality standards, the ability to detect subtle defects—ranging from surface scratches and discolorations to geometric deformations and structural cracks—has become a critical requirement. With the increasing availability of 3D sensing technologies such as structured-light scanners and time-of-flight cameras [2], recent approaches have begun to exploit both RGB imagery and geometric depth information for more comprehensive defect detection [3–5]. The combination of appearance and geometry is particularly appealing because certain defect types manifest primarily in one

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16768, pp. 107–125, 2027.
https://doi.org/10.1007/978-3-032-30488-9_7

modality: scratches, stains, and color variations are visible in RGB, while dents, warping, and surface deformations are better captured through depth measurements [4].

Current state-of-the-art multimodal methods, including M3DM [3], AST [6], and Shape-Guided [7], adopt a full-fusion strategy in which both RGB and 3D features are always processed jointly. While this approach can in principle capture complementary information from both modalities, it suffers from three fundamental limitations that restrict its practical applicability. First, these methods require training separate feature extractors for the 3D modality, typically involving pretrained point cloud encoders such as Point-MAE [8] or PointBERT [9] that must be fine-tuned on the target domain. This increases both computational cost and deployment complexity, preventing rapid adaptation to new product categories. Second, unconditional fusion means that even when depth information provides no discriminative value for a given category, it is still computed and integrated at every inference step, wasting computational resources. Third, and most critically, when the depth branch produces unreliable or near-random anomaly scores, full fusion can actively degrade detection performance below what the RGB-only baseline achieves. As we demonstrate in our experiments, unconditional fusion reduces the mean Image-AUROC (I-AUROC) by 5.2% points compared to the RGB baseline across the MVTec 3D-AD benchmark.

Recent work has begun to recognize this issue: for instance, Daci et al. [10] introduce reliability-gated mapping and confidence-weighted reconstruction to mitigate unreliable cross-modal cues, suggesting that adaptive use of depth—rather than unconditional fusion—is an emerging design principle for multimodal industrial anomaly detection. The MVTec 3D-AD benchmark [4] makes this issue easy to see. Among its 10 object categories, depth-based anomaly detection achieves strong performance for categories with predominantly geometric defects—for example, cable_gland achieves a depth-only I-AUROC of 0.99 and cookie achieves 0.80. However, for categories where defects are primarily chromatic or textural, depth performs at or near random chance: peach achieves only 0.19, tire 0.38, and foam 0.44. An ideal multimodal system should leverage depth information when it provides reliable geometric cues and gracefully ignore it when it does not, rather than blindly fusing both modalities at all times.

In this paper, we address these limitations with DepthGate. The framework has three components:

- **Training-free depth branch.** We replace learned depth encoders with a PatchCore-style [11] depth memory bank that requires no gradient-based training. Normal depth patch features are simply stored during the setup phase, and anomaly scores are computed via k-Nearest Neighbor (KNN) distance at test time. This enables instant deployment to new product categories by simply collecting normal samples and populating the memory bank, with no optimization loop required.
- **Two-level confidence gate.** We introduce a hierarchical gating mechanism operating at two complementary granularities. At the category level, a mini-AUROC validation procedure—a small-scale AUROC computed on

a held-out subset of normal and anomalous samples—automatically assesses whether the depth branch produces reliable anomaly signals for each object category. Categories where depth is unreliable (mini-AUROC below a threshold $\tau = 0.65$) have their gate permanently closed, completely bypassing all depth computation. At the sample level, within gate-open categories, depth correction is applied only when the normalized depth anomaly score exceeds the normalized RGB score.

– **Selective score fusion with map preservation.** Our fusion strategy applies depth correction exclusively at the image level through a residual score addition formula, while the pixel-level anomaly map relies solely on the RGB branch output. This asymmetric design reflects our empirical finding that the depth memory bank operates at 64×64 resolution, producing coarse anomaly maps that degrade localization accuracy by 2.5–4.7% points in Pixel-AUROC (P-AUROC) when fused with high-resolution RGB maps.

We evaluate our framework on the complete MVTec 3D-AD benchmark across all 10 object categories. The results demonstrate a mean I-AUROC improvement of +3.1% points over the EfficientAD RGB-only baseline while maintaining zero P-AUROC degradation across every category. For categories with geometric defects, improvements are substantial: cable_gland improves by +16.9 and cookie by +16.4% points. For the seven categories where depth is unreliable, the confidence gate automatically closes and preserves the RGB baseline with negligible deviation (maximum ±0.15 points). Additionally, when the gate is closed, inference throughput improves substantially compared to unconditional depth processing, as the computationally expensive k-Nearest Neighbor (KNN) memory bank queries are entirely bypassed.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed method in detail. Section 4 presents comprehensive experimental results. Section 5 concludes with a discussion of limitations and future directions.

2 Related Work

2.1 RGB-Based Anomaly Detection

Unsupervised anomaly detection methods for RGB images have evolved through several paradigms [12]. Reconstruction-based methods train autoencoders [13] or generative models [14,15] on normal data, detecting anomalies through reconstruction error. However, deep autoencoders can generalize well enough to reconstruct some anomalous patterns, limiting their discriminative power [16].

Knowledge distillation methods [17] represent a more effective paradigm. These approaches train a student network to mimic a pretrained teacher on normal samples, exploiting the observation that the student will fail to replicate the teacher's representations for anomalous inputs [18,19]. EfficientAD [20] advances this paradigm with a dual-branch architecture combining teacher-student distillation with autoencoder reconstruction. The distillation error captures semantic

anomalies where the student diverges from the teacher, while the reconstruction error captures structural anomalies. Both error maps are normalized and summed to produce the final anomaly map. Critically, EfficientAD achieves inference speeds under 2 ms per image through its lightweight Patch Description Network (PDN) architecture, making it suitable for real-time industrial deployment. We adopt EfficientAD as the RGB backbone of our framework.

Memory-bank approaches, exemplified by PatchCore [11], store representative normal patch features extracted from a pretrained backbone [21] and compute anomaly scores via nearest-neighbor distance. SPADE [22] similarly leverages multi-resolution feature correspondences for anomaly localization. These methods require no gradient-based training—only forward-pass feature extraction and storage. This training-free property is central to our depth branch design, where we adapt the memory-bank paradigm to operate directly on raw depth patch statistics rather than deep network features. Normalizing flow-based methods such as CFLOW-AD [23] and FastFlow [24] model the distribution of normal features through invertible transformations, achieving competitive performance with efficient density estimation.

2.2 Multimodal 3D Anomaly Detection

The release of MVTec 3D-AD [4] catalyzed research into multimodal anomaly detection leveraging both RGB and 3D information. M3DM [3] pioneered this direction by combining RGB features from a pretrained vision transformer [25] with 3D point cloud features from Point-MAE [8] through a decision-layer fusion mechanism. While effective, both encoders require separate pretraining and the fusion operates unconditionally regardless of modality reliability.

AST [6] introduced asymmetric student-teacher networks where separate student-teacher pairs process RGB and 3D modalities with architecturally distinct networks. CPMF [26] proposed cross-modal pseudo-feature mapping to bridge the representation gap between RGB and depth. Shape-Guided [7] leveraged geometric shape priors derived from depth to guide RGB-based detection. BTF [5] demonstrated that classical 3D features such as FPFH [27] remain competitive with learned representations, suggesting that complex 3D encoders may not always be necessary.

Despite their architectural differences, many of these methods keep both modalities active during inference and rely on a dedicated trained 3D feature extractor. This unconditional fusion implicitly assumes that depth is universally beneficial—an assumption our experiments directly challenge. Our work differs from prior multimodal industrial anomaly detection methods by introducing a two-level confidence gate—operating at both category and sample granularity—for selective depth utilization with a training-free depth branch.

2.3 Confidence-Based and Adaptive Fusion

Selective modality weighting based on reliability has been explored in autonomous driving [28], where sensor fusion frameworks dynamically weight

Light Detection And Ranging (LiDAR), camera, and radar inputs based on conditions. Recent surveys also highlight the importance of adaptive multimodal fusion strategies in anomaly detection. In medical imaging, multi-modal segmentation networks employ attention mechanisms for modality-specific contribution weights [29]. In multi-task learning, Kendall et al. [30] proposed uncertainty-based weighting that adjusts loss contributions based on homoscedastic uncertainty estimates.

Our approach differs in two respects. First, our gate operates at two distinct granularities: a category-level binary gate that completely eliminates unreliable modalities before they can introduce any noise, and a sample-level binary gate that provides finer-grained control within reliable categories. This is more conservative but also more robust than soft-weighting approaches, which may still propagate noisy signals with merely reduced weight. A small but nonzero weight on a fundamentally unreliable modality can accumulate across many samples to produce measurable degradation—a problem our binary gate avoids entirely.

Second, our gate requires no learned parameters: the category-level gate uses a straightforward AUROC threshold and the sample-level gate uses direct score comparison. This parameter-free design aligns with our philosophy of a training-free depth pathway. The entire depth pipeline—from feature extraction to memory bank construction to fusion gating—is deployable without any gradient-based optimization.

3 DepthGate Framework

The proposed DepthGate framework comprises three stages operating in sequence. First, an EfficientAD-based RGB branch outputs an image-level score and a pixel-level anomaly map. Second, a depth memory bank—built without any training—gives an image-level geometric score. The fusion module then combines the two scores under a confidence gate, while the pixel map comes only from the RGB branch. Figure 1 presents the overview of DepthGate framework.

3.1 Stage 1: RGB Branch (EfficientAD)

The RGB branch is based on the EfficientAD [20] architecture, serving as both the primary detector and the baseline that the depth branch aims to improve.

In the teacher-student distillation branch, a pretrained teacher network based on the PDN-Small extracts feature maps from 256×256 RGB images.. A lightweight student network with identical architecture is trained to replicate the teacher's representations on normal samples. At test time, the element-wise squared difference yields the distillation error map St_map, highlighting regions where the student fails to mimic the teacher. The teacher's weights are frozen while the student's weights are optimized during training.

In the autoencoder reconstruction branch, a separate encoder-decoder network reconstructs the teacher's features. The reconstruction error map Ae_map

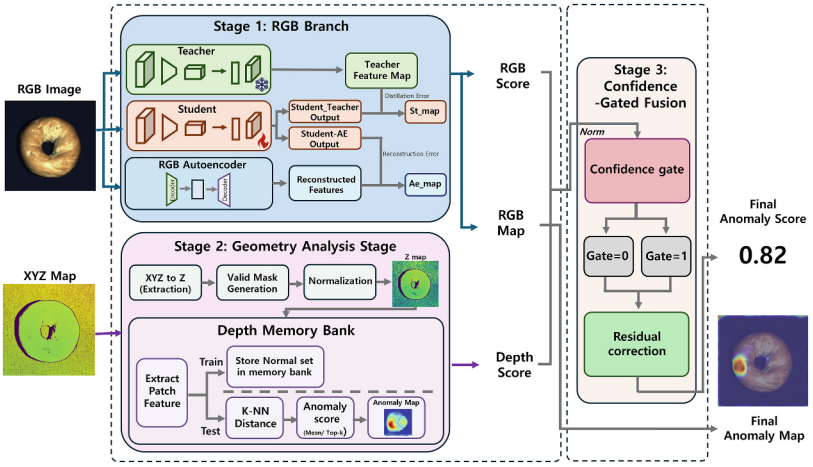


Fig. 1. Overview of DepthGate framework. The RGB branch (top, blue) uses EfficientAD with teacher-student distillation and autoencoder reconstruction to produce RGB anomaly scores and maps. The depth branch (bottom, gray/purple) uses a training-free memory bank with k-NN distance to produce depth scores. The fusion module (right, pink) applies score normalization, a confidence gate (Gate = 0 for RGB-sufficient cases, Gate = 1 for depth-assisted cases), and residual correction. The pixel-level anomaly map uses only the RGB branch output. (Color figure online)

captures structural anomalies. Both maps are normalized and combined:

$$\mathbf{M}_{\text{rgb}} = \frac{\text{St_map}}{q_{\text{st}}} + \frac{\text{Ae_map}}{q_{\text{ae}}} \tag{1}$$

where q_{st} and q_{ae} are the 99.5th percentile normalization factors from training.

The image-level RGB score is computed via top-k aggregation:

$$S_{\text{rgb}} = \text{mean}\left(\text{Top}_k(\text{M}_{\text{rgb}})\right), \quad k = 0.01 \times |\mathbf{M}_{\text{rgb}}| \tag{2}$$

Both S_{rgb} and Mrgb are forwarded to the fusion module.

3.2 Stage 2: Geometry Analysis Stage

Unlike existing multimodal approaches that train separate neural network encoders for depth [3,6], our depth branch requires no gradient-based optimization, following a PatchCore-inspired [11] design adapted for raw depth maps.

Given an XYZ point cloud map, we extract the Z (depth) channel. Invalid pixels (sensor noise, occlusion) are masked via validity check ($z \neq 0$), NaN values are zeroed, and the map is downscaled to 64×64 via bilinear interpolation. The validity mask is downscaled using nearest-neighbor interpolation. This pipeline is identical for training and testing.

During the training phase, each 64×64 depth map is padded and divided into overlapping $p \times p$ patches ($p = 5$) centered at each valid pixel. Per patch, we extract a feature vector comprising:

- Raw patch values: The $p^2 = 25$ pixel values flattened, capturing local surface geometry.
- Patch standard deviation: A scalar measuring local surface roughness.
- Center deviation: $|z_{\text{center}} - \bar{z}_{\text{patch}}|$, capturing local point anomalies.

This yields a $(p^2 + 2)$-dimensional feature vector. Features from normal samples are accumulated into memory bank $\mathcal{M}$ (up to 500 per image, 100 images, ≈50,000 features per category). A k-NN index is built over $\mathcal{M}$.

At test time, test features are queried against $\mathcal{M}$ using k-NN ($k = 3$) to compute an image-level geometric anomaly score S_{geom}, which quantifies how much the test sample's depth patches deviate from the normal depth distribution stored in the memory bank. Specifically, S_{geom} aggregates the top 1% highest patch distances:

$$S_{\text{geom}} = \text{mean}\left(\text{Top}_k\left(d(f_i, \mathcal{M})\right)\right), \quad i = 1, \ldots, N \tag{3}$$

where $d(f_i, \mathcal{M})$ denotes the mean k-NN distance for patch feature f_i and N is the number of valid patches. A higher S_{geom} indicates stronger geometric deviation from normal samples, i.e., a higher likelihood of an anomaly based on depth information alone.

The entire depth branch—preprocessing, memory bank construction, k-NN scoring—requires no gradient-based optimization. Deploying to a new category requires only collecting normal samples and populating the memory bank, avoiding any category-specific gradient-based optimization.

3.3 Stage 3: Confidence-Gated Fusion

The fusion module is the core contribution of our framework. It combines RGB and depth scores through a two-level gating mechanism that leverages depth when informative while protecting the RGB baseline when depth is unreliable.

Score Normalization. Since the RGB and depth scores operate on incomparable scales, both are mapped to a common normalized score $S_{\text{norm}} \in [0, 1]$ using percentile-based normalization:

$$S_{\text{norm}} = \text{clip}\left(\frac{S - P_5}{P_{95} - P_5 + \epsilon},\ 0,\ 1\right) \tag{4}$$

where P_5, P_{95} are the 5th/95th percentiles across test samples and $\epsilon = 10^{-8}$.

Category-Level Reliability Gate. Before full evaluation, we assess depth discriminability per category using a *mini-AUROC*, a lightweight validation metric defined as the AUROC of depth-only anomaly scores (S_{geom}) computed on a small held-out subset, distinct from the test set used for final evaluation. Specifically, we use 20 held-out normal samples and up to 10 samples per test subcategory. The category-level gate $G_{\text{cat}} \in \{0, 1\}$ indicates whether depth is used (1) or bypassed (0) for a given category, and is defined as:

$$G_{\text{cat}} = \mathbb{1}\left[\text{AUROC}_{\text{mini}}(S_{\text{geom}}) \geq \tau\right] \tag{5}$$

where $\tau = 0.65$. When $G_{\text{cat}} = 0$, the framework defaults to RGB-only inference, providing both performance protection and substantial computational savings (Sect. 4.3). In our experiments, the mini-AUROC values show a clear separation between reliable and unreliable categories: reliable categories cluster between 0.74–1.00 while unreliable categories cluster between 0.15–0.60, making the threshold choice relatively insensitive within the 0.60–0.70 range.

Sample-Level Residual Correction. For categories with $G_{\text{cat}} = 1$, a sample-level gate $g_i \in \{0, 1\}$ determines whether depth correction is applied to sample i, activating only when the depth score exceeds the RGB score:

$$g_i = \mathbb{1}\left[S_{\text{geom}}^{\text{norm}}(i) > S_{\text{rgb}}^{\text{norm}}(i)\right] \tag{6}$$

The final score combines both branches via residual correction:

$$S_{\text{final}}(i) = S_{\text{rgb}}^{\text{norm}}(i) + \alpha \cdot G_{\text{cat}} \cdot g_i \cdot S_{\text{geom}}^{\text{norm}}(i) \tag{7}$$

where $\alpha = 0.5$ controls depth contribution. When either gate closes, the product vanishes and $S_{\text{final}} = S_{\text{rgb}}^{\text{norm}}$, reducing to the normalized RGB-only score. In our experiments, the sample-level gate activates for approximately 64–66% of samples within gate-open categories.

Map Preservation Strategy. For pixel-level localization, we use only the RGB branch:

$$\text{M}_{\text{final}} = \text{M}_{\text{rgb}} \tag{8}$$

The depth memory bank operates at 64×64 resolution; upscaled depth maps introduce spatial blurring that degrades localization by 2.5–4.7 points in P-AUROC. This asymmetric strategy—score fusion with map preservation—achieves image-level improvements without sacrificing pixel-level precision.

Algorithm Summary. Algorithm 1 summarizes the inference procedure. Lines 12–14 implement early termination: when the category gate is closed, depth computation is entirely bypassed, yielding substantial runtime savings.

Algorithm 1. DepthGate Inference Procedure

Require: RGB image $\mathbf{x}_{\text{rgb}}$, XYZ map $\mathbf{x}_{\text{xyz}}$, category c
Parameters:
  $G_{\text{cat}}[c] \in \{0, 1\}$: pre-computed category gate (Eq. 5)
  $\mathcal{M}_c$: depth memory bank for category c
  $\alpha = 0.5$: depth contribution weight
Intermediate variables:
  $S_{\text{rgb}}, \text{M}_{\text{rgb}}$: RGB anomaly score and map (from EfficientAD)
  Z, F: preprocessed depth map and patch features
  S_{geom}: depth anomaly score (Eq. 3)
  $S_{\text{rgb}}^{\text{norm}}, S_{\text{geom}}^{\text{norm}}$: normalized scores (Eq. 4)
  $g \in \{0, 1\}$: sample-level gate (Eq. 6)
Ensure: Final score S_{final}, Final map M_{final}
$S_{\text{rgb}}, \text{M}_{\text{rgb}} \leftarrow \text{EfficientAD}(\mathbf{x}_{\text{rgb}})$ ▷ Stage 1
if $G_{\text{cat}}[c] = 0$ **then**
  return $\text{Normalize}(S_{\text{rgb}}), \text{M}_{\text{rgb}}$ ▷ Skip depth (gate closed)
end if
$\text{Z} \leftarrow \text{Preprocess}(\mathbf{x}_{\text{xyz}})$ ▷ Stage 2
$\text{F} \leftarrow \text{ExtractPatchFeatures}(\text{Z})$
$S_{\text{geom}} \leftarrow \text{TopKMeanDistance}(\text{F}, \mathcal{M}_c)$
$S_{\text{rgb}}^{\text{norm}} \leftarrow \text{Normalize}(S_{\text{rgb}})$ ▷ Stage 3
$S_{\text{geom}}^{\text{norm}} \leftarrow \text{Normalize}(S_{\text{geom}})$
$g \leftarrow \mathbb{1}[S_{\text{geom}}^{\text{norm}} > S_{\text{rgb}}^{\text{norm}}]$ ▷ Sample gate
$S_{\text{final}} \leftarrow S_{\text{rgb}}^{\text{norm}} + \alpha \cdot g \cdot S_{\text{geom}}^{\text{norm}}$ ▷ Residual correction
$\text{M}_{\text{final}} \leftarrow \text{M}_{\text{rgb}}$ ▷ Map preserved
return $S_{\text{final}}, \text{M}_{\text{final}}$

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate on the MVTec 3D-AD benchmark [4], which contains 10 object categories (bagel, cable_gland, carrot, cookie, dowel, foam, peach, potato, rope, tire) with both RGB images and structured-light 3D scans. Each category includes normal training samples and test samples with various defect types. The dataset provides pixel-level ground truth annotations for anomaly localization.

Implementation Details. The RGB branch uses EfficientAD with a small PDN architecture (out_channels = 384) and pretrained teacher weights. The depth memory bank uses patch size $p = 5$, $k = 3$ nearest neighbors, and a downscaled resolution of 64×64. Normal features are subsampled to 500 per image with a maximum of 100 images, yielding approximately 50,000 features per category. The fusion parameters are set to $\alpha = 0.5$ and $\tau = 0.65$ unless otherwise specified. The k-NN index uses scikit-learn's NearestNeighbors with Euclidean distance and parallel computation. All experiments are conducted on a single NVIDIA GPU.

Evaluation Metrics. We report I-AUROC for anomaly detection, P-AUROC for anomaly localization, and Area Under the Per-Region Overlap curve (AUPRO) for region-level evaluation. We additionally report the category-level gate status (OPEN/CLOSE), the mini-AUROC from depth reliability validation, and the sample-level gate activation ratio.

4.2 Results

Table 1 presents the complete evaluation results across all 10 categories of MVTec 3D-AD.

Table 1. Main results on MVTec 3D-AD. I-AUROC and P-AUROC are reported as percentages. Gate indicates category-level depth reliability status. Δ denotes improvement over the RGB-only baseline. mini-AUROC is the depth reliability score used for gating.

Category	Gate	I-RGB	I-Fused	I-Δ	P-RGB	P-Fused	P-Δ	mini-AUROC
bagel	CL	82.13	81.97	−0.15	96.69	96.69	0.00	0.545
cable_gland	OP	74.55	91.41	+16.86	98.15	98.15	0.00	1.000
carrot	CL	87.18	87.40	+0.22	99.28	99.28	0.00	0.600
cookie	OP	58.46	74.83	+16.37	93.46	93.46	0.00	0.765
dowel	OP	96.15	94.01	−2.14	99.50	99.50	0.00	0.745
foam	CL	76.19	76.19	0.00	82.49	82.49	0.00	0.473
peach	CL	71.73	71.66	−0.07	95.57	95.57	0.00	0.188
potato	CL	77.42	77.40	−0.02	98.86	98.86	0.00	0.400
rope	CL	97.64	97.64	0.00	99.51	99.51	0.00	0.433
tire	CL	64.97	65.03	+0.07	98.42	98.42	0.00	0.380
Mean	—	78.64	**81.75**	**+3.11**	96.19	**96.19**	**0.00**	—

Three key findings emerge from Table 1. First, for categories where the depth branch achieves high standalone discriminative power (cable_gland: mini-AUROC = 1.000, cookie: mini-AUROC = 0.765), the confidence gate correctly opens and substantial I-AUROC improvements are achieved (+16.86 and +16.37% points respectively). These two categories are characterized by geometric defects—surface deformations and structural irregularities—that are well captured by depth measurements but may be subtle or invisible in RGB. Second, for categories where depth is unreliable (mini-AUROC < 0.65), the gate remains closed and the RGB baseline is preserved with negligible deviation (maximum ± 0.15 points). The gate handles this protection automatically—we never had to tune anything per category. Third, P-AUROC is preserved at $\Delta = 0.00$ across all categories, confirming that the map preservation strategy (Eq. 8) successfully decouples image-level score improvements from pixel-level localization (Table 2).

Table 2. Comparison on MVTec 3D-AD. Mean image-level AUROC (%). "Depth Train" indicates whether the depth branch requires training. "Selective" indicates conditional modality usage. "P-AUROC Δ" indicates pixel-level degradation from fusion. For EfficientAD, we report our own reproduction on MVTec 3D-AD, since the original paper does not report this benchmark.

Method	Depth Train	Selective	I-AUROC	P-AUROC Δ
BTF [5]	Yes	No	87.3	—
AST [6]	Yes	No	93.7	—
M3DM [3]	Yes	No	94.5	—
Shape-Guided [7]	Yes	No	94.7	—
EfficientAD (our reproduction)	—	—	78.6	—
DepthGate (Ours)	**No**	**Yes**	81.8	**0.00**

DepthGate achieves 81.8% mean I-AUROC, a +3.1 point improvement over its EfficientAD baseline. While fully-trained multimodal methods such as M3DM (94.5%) and AST (93.7%) achieve higher absolute performance on MVTec 3D-AD, they require dedicated 3D encoders and process both modalities for all categories. In contrast, DepthGate requires no depth-branch training, automatically bypasses depth for unreliable categories, and maintains zero P-AUROC degradation in our setting. To the best of our knowledge, existing methods do not combine all of these properties in a single framework. The gating mechanism is orthogonal to the fusion backbone and could potentially be applied on top of methods such as M3DM or AST to improve robustness against unreliable depth categories.

Figure 2 presents anomaly localization results of the RGB branch across representative categories from MVTec 3D-AD.

The RGB branch achieves high-quality pixel-level localization across diverse defect types, successfully identifying anomalous regions that align closely with the ground truth masks. This strong localization performance—reflected in the mean P-AUROC of 96.19% across all categories—motivates our map preservation strategy. Since the RGB branch already provides precise spatial localization, fusing depth maps (which operate at 64×64 resolution) would only introduce blurring artifacts without improving localization accuracy. Instead, depth contributes exclusively through image-level score correction, where its geometric signal is most valuable.

4.3 Ablation Study

Gate Confidence. To validate the contribution of the confidence gate, we compare three configurations: (a) RGB-only baseline (EfficientAD), (b) unconditional fusion without gating ($\tau = 0$, all categories open), and (c) the full method with confidence gating ($\tau = 0.65$). Table 3 presents results for representative categories.

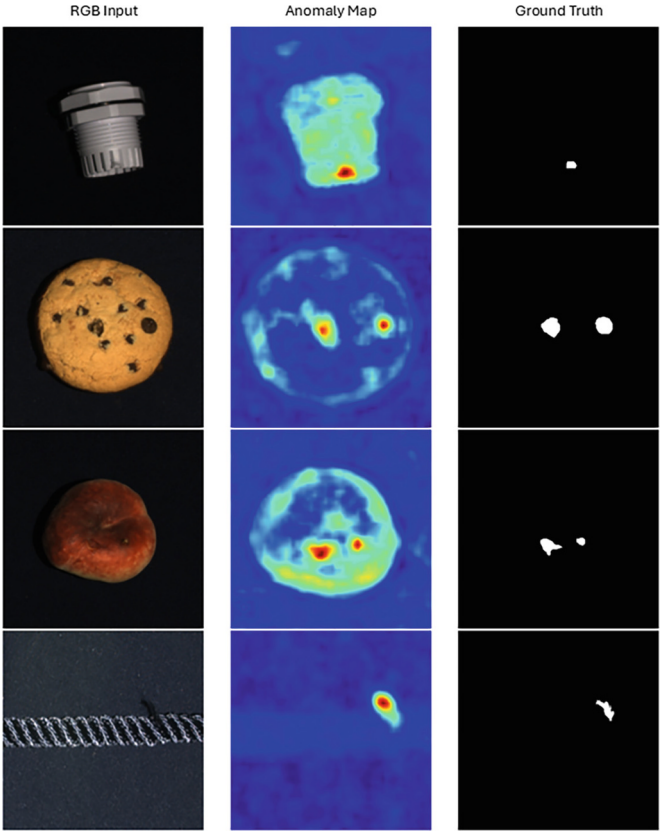


Fig. 2. Anomaly localization results of the RGB branch (EfficientAD) across representative categories. Left: RGB input. Center: predicted anomaly heatmap. Right: ground truth mask. The RGB branch provides precise pixel-level localization across diverse defect types including geometric deformations (cable_gland), textural anomalies (cookie, rope), and chromatic defects (peach). This strong localization capability motivates our map preservation strategy ($M_{final} = M_{rgb}$), where depth contributes only to image-level scoring.

Table 3. Ablation study (I-AUROC %). Configuration (b) without gating shows that unconditional fusion catastrophically degrades depth-unreliable categories, while configuration (c) with gating preserves the baseline. Mean Δ is computed across all 10 categories.

Config.	cable_gl.	cookie	carrot	foam	peach	rope	dowel	Mean Δ
(a) RGB only	74.55	58.46	87.18	76.19	71.73	97.64	96.15	—
(b) No gate	90.20	76.07	71.72	63.00	57.15	89.27	93.86	−5.23
(c) Full (ours)	91.41	74.83	87.40	76.19	71.66	97.64	94.01	+3.11

The ablation reveals the critical role of the confidence gate. Without gating (configuration b), depth fusion improves cable_gland (+15.65) and cookie (+17.61) but catastrophically degrades carrot (−15.46), peach (−14.59), foam (−13.19), and rope (−8.38), resulting in a mean Δ of −5.23 across all 10 categories. The confidence gate (configuration c) resolves this by automatically closing depth for unreliable categories, converting the −5.23 mean degradation into a +3.11 improvement—a swing of over 8% points. This demonstrates that the gate is not merely a minor refinement but the essential mechanism that makes multimodal fusion viable.

Map Fusion. We investigated whether depth anomaly maps could improve pixel-level localization by applying the same gating mechanism to maps: $\mathbf{M}_{\text{final}} = \mathbf{M}_{\text{rgb}}^{\text{norm}} + \alpha \cdot g \cdot \mathbf{M}_{\text{geom}}^{\text{norm}}$. Table 4 presents the results for gate-open categories.

Table 4. Impact of depth map fusion on P-AUROC (%) for gate-open categories. Fusing depth maps degrades localization due to resolution limitations.

Category	Map preserved	Map fused	Δ
cable_gland	98.15	95.65	−2.50
cookie	93.46	88.76	−4.70

Depth map fusion causes P-AUROC drops of 2.50 and 4.70% points for cable_gland and cookie respectively. The degradation stems from the memory bank's 64×64 operating resolution: the resulting depth anomaly maps, when bilinearly upscaled to 256×256, introduce spatial blurring that corrupts the precise localization boundaries provided by the RGB branch.

Hyperparameter Sensitivity. We analyze sensitivity to the two key hyperparameters: depth contribution weight α and reliability threshold τ.

Effect of α (Depth Weight). Table 5 shows the mean I-AUROC Δ across all 10 categories for varying α with fixed $\tau = 0.65$.

A clear trade-off emerges: increasing α amplifies the depth contribution, yielding larger improvements for geometric-defect categories (cable_gland, cookie) but also amplifying depth-induced degradation for dowel. The overall mean peaks at $\alpha = 0.7$ (+3.54), but the dowel degradation at this setting (−6.21) is substantially worse than at $\alpha = 0.5$ (−2.63). We retain $\alpha = 0.5$ as the default because it achieves near-optimal mean improvement (+3.09) while limiting the worst-case per-category degradation. This reflects a practical design choice: in industrial deployment, avoiding catastrophic failure in any single category is often more important than maximizing average performance.

Table 5. Sensitivity to α (depth contribution weight) with $\tau = 0.65$. I-AUROC Δ reported for gate-open categories and overall mean. Results may differ slightly from Table 1 due to differences in sampling and evaluation splits used for the sensitivity analysis.

α	0.1	0.3	0.5	0.7	1.0
cable_gland	+5.09	+11.82	+16.42	+19.70	+21.35
cookie	+4.82	+12.03	+17.06	+21.84	+23.61
dowel	+1.00	−0.15	−2.63	−6.21	−11.43
Mean (all 10)	+1.10	+2.38	+3.09	**+3.54**	+3.36

Table 6. Sensitivity to τ (reliability threshold) with $\alpha = 0.5$. "Open" indicates the number of gate-open categories. Results may differ slightly from Table 1 due to differences in sampling and evaluation splits used for the sensitivity analysis.

τ	0.50	0.60	0.65	0.70	0.75	0.80
Open cats	5	3	3	3	2	1
Mean I-Δ	+0.85	+3.20	+3.06	+3.10	**+3.40**	+1.69
cable_gland	+16.91	+17.41	+16.69	+16.48	+16.09	+16.80
cookie	+17.23	+17.30	+16.75	+17.20	+17.82	0.00
dowel	−2.22	−2.74	−2.88	−2.70	+0.07	+0.07

Effect of τ (Reliability Threshold). Table 6 shows the effect of varying τ with fixed $\alpha = 0.5$. Lower thresholds open more category gates, risking depth-induced degradation; higher thresholds are more conservative.

The results reveal an interesting pattern. At $\tau = 0.50$, the gate opens for 5 categories including bagel and carrot, whose unreliable depth causes severe degradation that offsets the gains, yielding only +0.85 mean improvement. Increasing τ to 0.60 closes these harmful categories, jumping to +3.20. Performance remains stable across $\tau \in [0.60, 0.70]$, confirming the insensitivity of the threshold within this range due to the bimodal distribution of mini-AUROC values.

At $\tau = 0.75$, the dowel gate closes (mini-AUROC$\approx$0.71), eliminating its degradation and achieving the highest mean improvement of +3.40. However, at $\tau = 0.80$, the cookie gate also closes (mini-AUROC$\approx$0.77), losing its +17 point improvement and dropping the mean to +1.69. We retain $\tau = 0.65$ as the default for its balance between improvement and robustness, though $\tau = 0.75$ represents an attractive alternative that eliminates the dowel failure case.

Per-Category Depth Reliability Analysis. To provide deeper insight into the relationship between depth reliability and defect characteristics, Table 7 presents a comprehensive per-category analysis.

Table 7. Per-category depth reliability analysis. Depth-only I-AUROC indicates standalone depth discriminability. Primary defect type explains why depth is or is not effective. Gate status: OP (open) = depth used, CL (close) = depth bypassed.

Category	Depth AUROC	Gate	Defect Type	Why Depth Helps/Fails
cable_gland	0.99	OP	Geometric	Surface deformations visible in Z
cookie	0.80	OP	Geometric	Dough deformations alter depth
dowel	0.68	OP	Mixed	Some geometric; RGB already strong
carrot	0.52	CL	Textural	Color/texture defects on flat surface
bagel	0.54	CL	Textural	Surface blemishes; consistent shape
foam	0.44	CL	Textural	Porosity variations; noisy depth
rope	0.43	CL	Textural	Fiber damage; no depth signature
potato	0.40	CL	Textural	Skin defects; organic shape variance
tire	0.38	CL	Textural	Surface marks; consistent geometry
peach	0.19	CL	Chromatic	Color defects only; no depth info

A clear pattern emerges: categories with geometric defects (cable_gland, cookie) achieve high depth-only AUROC and benefit from fusion, while categories with primarily textural or chromatic defects (peach, foam, tire) show near-random depth performance. This aligns with the physical characteristics of structured-light depth sensors, which capture surface geometry but provide no signal for color-based or texture-based anomalies. The confidence gate automatically discovers this pattern without requiring manual annotation of defect types.

Inference Speed. A key practical advantage of the confidence gate is computational savings when depth is bypassed. Table 8 compares inference throughput for gate-closed and gate-open categories.

Table 8. Inference speed comparison. Gate-closed categories skip depth computation entirely, yielding substantial speedup.

Configuration	Throughput (samples/s)	Relative
Gate CLOSE (RGB only)	∼25–57	1.0×
Gate OPEN (RGB + Depth)	∼3.0–5.2	0.06–0.21×

When the category-level gate is closed, inference throughput ranges from approximately 25 to 57 samples per second depending on the category, compared to 3.0–5.2 samples per second when depth memory bank queries are executed. Since 7 out of 10 categories have their gate closed in our experiments, the framework achieves substantial average throughput gains across the full benchmark.

Failure Case. The dowel category presents an instructive failure case. Despite a mini-AUROC of 0.745 (above $\tau = 0.65$), fusion decreases I-AUROC by 2.14% points. This occurs because the RGB baseline is already very high (96.15%), leaving minimal headroom for improvement while retaining capacity for depth-induced false alarms at the sample level.

Specifically, within dowel's gate-open inference, approximately 64.6% of samples have their sample-level gate activated. Among these, a subset of normal samples receive elevated final scores due to depth noise, pushing them above the decision boundary and increasing the false positive rate. The fundamental issue is that the category-level gate assesses depth *discriminability* (whether depth can separate normal from anomalous), but does not account for the *marginal value* of depth given an already-strong RGB baseline.

Two potential mitigations exist. First, conditioning the category gate on both depth reliability and RGB baseline performance: if the RGB baseline already exceeds a threshold (e.g., 95% I-AUROC), the gate closes regardless of depth reliability. Second, increasing τ to 0.75 would close the dowel gate (mini-AUROC $= 0.745$) while keeping cable_gland (1.000) and cookie (0.765) open. We leave the integration of these refinements into a unified gating criterion for future work.

5 Conclusion

A few practical lessons emerged from our experiments. Naive fusion of RGB and depth is not safe: depth gains in geometric-defect classes were wiped out by losses in classes where depth carries no useful signal, costing 5.23 points on average. A simple binary gate flips this picture, turning a 5.23-point loss into a 3.11-point gain—and the gate-closed classes incur no depth computation at all. We also found that image-level and pixel-level fusion behave very differently: depth helps the score but hurts the localization map at 64×64 resolution, which is why we keep the RGB map untouched. Finally, because the depth pathway needs no gradient-based training, adding a new product category is just a matter of populating the memory bank.

In summary, we presented DepthGate, a confidence-gated depth verification framework for multimodal industrial anomaly detection. By combining a training-free depth memory bank with a two-level gating mechanism, our approach selectively leverages depth information only when it provides reliable discriminative value. Evaluated on MVTec 3D-AD across all 10 categories, the framework achieves a mean I-AUROC improvement of +3.1% points over the EfficientAD baseline while maintaining zero P-AUROC degradation. For geometric-defect categories, improvements reach +16.9 and +16.4% points for cable_gland and cookie respectively.

Several directions remain for future work. First, the gating criterion could be extended to incorporate the RGB baseline performance, addressing failure cases where the baseline is already very strong (e.g., dowel). Second, the depth feature representation could be enriched beyond raw patch statistics, potentially using lightweight pretrained geometric encoders within the gated framework.

Third, the confidence-gated fusion paradigm could be applied to other multimodal fusion pipelines such as M3DM [3] or AST [6], potentially improving their robustness without modifying their core architectures. Finally, evaluation on additional industrial datasets with diverse sensor modalities would test the generalizability of the gating approach beyond RGB-depth fusion.

Acknowledgments. This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-ICT Creative Consilience Program grant funded by the Korea government (MSIT) (IITP-2026-RS-2020-II201821), This work was supported by the Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government under the Industrial Innovation Talent Growth Support Project (KIAT-2026-RS-2026-25532917).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Li, Z., Yan, Y., Wang, X., Ge, Y., Meng, L.: A survey of deep learning for industrial visual anomaly detection. Artif. Intell. Rev. **58**(9), 279 (2025). https://doi.org/10.1007/s10462-025-11287-7
2. Lin, Y., et al.: A survey on RGB, 3D, and multimodal approaches for unsupervised industrial image anomaly detection. Inf. Fusion **121**, 103139 (2025). https://doi.org/10.1016/j.inffus.2025.103139
3. Wang, Y., Peng, J., Zhang, J., Yi, R., Wang, Y., Wang, C.: Multimodal industrial anomaly detection via hybrid fusion. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8032–8041 (2023). https://doi.org/10.1109/CVPR52729.2023.00776
4. Bergmann, P., Jin, X., Sattlegger, D., Steger, C.: The MVTec 3D-AD dataset for unsupervised 3D anomaly detection and localization. In: VISAPP, pp. 202–213 (2022). https://doi.org/10.5220/0010865000003124
5. Horwitz, E., Hoshen, Y.: Back to the feature: classical 3D features are (almost) all you need for 3D anomaly detection. In: CVPR Workshops, pp. 2968–2977 (2023). https://doi.org/10.1109/CVPRW59228.2023.00298
6. Rudolph, M., Wehrbein, T., Rosenhahn, B., Wandt, B.: Asymmetric student-teacher networks for industrial anomaly detection. In: WACV, pp. 2591–2601 (2023). https://doi.org/10.1109/WACV56688.2023.00262
7. Chu, Y., Liu, C., Hsieh, T., Chen, H., Liu, T.: Shape-guided dual-memory learning for 3D anomaly detection. In: ICML. Proceedings of Machine Learning Research, vol. 202, pp. 6185–6194. PMLR (2023). https://proceedings.mlr.press/v202/chu23b.html
8. Pang, Y., Wang, W., Tay, F.E.H., Liu, W., Tian, Y., Yuan, L.: Masked autoencoders for point cloud self-supervised learning. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. ECCV 2022. LNCS, vol. 13662, pp. 604–621. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20086-1_35
9. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J.: Point-BERT: pre-training 3D point cloud transformers with masked point modeling. In: CVPR, pp. 19291–19300 (2022). https://doi.org/10.1109/CVPR52688.2022.01871

10. Daci, R., et al.: Cross-modal mapping and dual-branch reconstruction for 2D-3D multimodal industrial anomaly detection. arXiv preprint arXiv:2603.03939 (2026). https://arxiv.org/abs/2603.03939
11. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: CVPR, pp. 14298–14308 (2022). https://doi.org/10.1109/CVPR52688.2022.01392
12. Shukla, V., Shukla, A., Surya Prakash, S.K., Shukla, S.: A systematic survey: role of deep learning-based image anomaly detection in industrial inspection contexts. Front. Robot. AI **12**, 1554196 (2025). https://doi.org/10.3389/frobt.2025.1554196
13. Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D., Steger, C.: Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In: VISIGRAPP, pp. 372–380 (2019). https://doi.org/10.5220/0007364503720380
14. Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: GANomaly: semi-supervised anomaly detection via adversarial training. In: Jawahar, C., Li, H., Mori, G., Schindler, K. (eds.) Computer Vision – ACCV 2018. ACCV 2018. LNCS, vol. 11363, pp. 622–637. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-20893-6_39
15. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-AnoGAN: fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. **54**, 30–44 (2019). https://doi.org/10.1016/j.media.2019.01.010
16. Zavrtanik, V., Kristan, M., Skocaj, D.: DRAEM – a discriminatively trained reconstruction embedding for surface anomaly detection. In: ICCV, pp. 8310–8319 (2021). https://doi.org/10.1109/ICCV48922.2021.00822
17. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015). https://arxiv.org/abs/1503.02531
18. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Uninformed students: student-teacher anomaly detection with discriminative latent embeddings. In: CVPR, pp. 4182–4191 (2020). https://doi.org/10.1109/CVPR42600.2020.00424
19. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: CVPR, pp. 9727–9736 (2022). https://doi.org/10.1109/CVPR52688.2022.00951
20. Batzner, K., Heckler, L., König, R.: EfficientAD: accurate visual anomaly detection at millisecond-level latencies. In: WACV, pp. 127–137 (2024). https://doi.org/10.1109/WACV57701.2024.00020
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
22. Cohen, N., Hoshen, Y.: Sub-image anomaly detection with deep pyramid correspondences. arXiv preprint arXiv:2005.02357 (2020). https://arxiv.org/abs/2005.02357
23. Gudovskiy, D., Ishizaka, S., Kozuka, K.: CFLOW-AD: real-time unsupervised anomaly detection with localization via conditional normalizing flows. In: WACV, pp. 1819–1828 (2022). https://doi.org/10.1109/WACV51458.2022.00188
24. Yu, J., et al.: FastFlow: unsupervised anomaly detection and localization via 2D normalizing flows. arXiv preprint arXiv:2111.07677 (2021). https://arxiv.org/abs/2111.07677
25. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. In: ICLR (2021). https://openreview.net/forum?id=YicbFdNTTy

26. Cao, Y., Xu, X., Shen, W.: Complementary pseudo multimodal feature for point cloud anomaly detection. Pattern Recognition **156**, 110761 (2024). https://doi.org/10.1016/j.patcog.2024.110761
27. Rusu, R.B., Blodow, N., Beetz, M.: Fast point feature histograms (FPFH) for 3D registration. In: ICRA, pp. 3212–3217 (2009). https://doi.org/10.1109/ROBOT.2009.5152473
28. Qian, H., Wang, M., Zhu, M., Wang, H.: A review of multi-sensor fusion in autonomous driving. Sensors **25**(19), 6033 (2025). https://doi.org/10.3390/s25196033
29. Zhou, T., Ruan, S., Canu, S.: A review: deep learning for medical image segmentation using multi-modality fusion. Array **3–4**, 100004 (2019). https://doi.org/10.1016/j.array.2019.100004
30. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: CVPR, pp. 7482–7491 (2018). https://doi.org/10.1109/CVPR.2018.00781