C-MTSS: Compact Multi-teacher Single-Student Framework for Real-Time Multimodal Industrial Anomaly Detection

Dongwoo Kim, Taeheon Jin, Seokhyun Gong, Jieun Lee, and Jongpil Jeong(✉)

Department of Smart Factory Convergence, Sungkyunkwan University, Suwon-si, Gyeonggi-do 16419, Republic of Korea
{kdw5641,kongsh95,jpjeong}@skku.edu, {honey7844,lu3873}@g.skku.edu

Abstract. Recently, multimodal anomaly detection techniques that combine 2D texture and 3D shape information have been widely studied to automate quality inspections in smart factories. However, existing 3D-2D models suffer from feature scale discrepancy between modalities, which causes one modality to dominate the anomaly score, and are highly vulnerable to the shape variance of objects. Furthermore, their massive parameters and high computational costs make real-time inference on resource-constrained environments infeasible. To address these industrial constraints, this paper proposes the Compact Multi-Teacher Single-Student framework (C-MTSS), a lightweight 3D-2D multimodal anomaly detection framework capable of inference in less than 1 *ms*. The proposed framework ensures robustness against shape variance by converting raw depth maps into surface normal maps. In addition, to enable a single-student network with fewer than 1M parameters to efficiently distill knowledge from two teacher models, we introduce a dual-stem structure that prevents initial interference between modalities, along with a Masked Mean Square Error (MSE) loss function to suppress background noise. During the inference stage, a Z-Score Calibration module is applied to normalize and weight-fuse the error distributions of the two heterogeneous modalities, which improves detection accuracy. Experiments on the MVTec 3D-AD benchmark demonstrate that C-MTSS achieves 91.3% image-level AUROC and 96.7% pixel-level AUROC with only 2.15M parameters and 38MB memory footprint, attaining over 1,000 FPS on a single GPU. These results indicate that C-MTSS is designed as a lightweight framework for real-time industrial inspection, where low latency and small memory footprint are critical.

Keywords: Anomaly Detection · 3D Surface Normal · Knowledge Distillation · Smart Manufacturing · Real-time Inference

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16768, pp. 333–348, 2027.
https://doi.org/10.1007/978-3-032-30488-9_21

1 Introduction

Recently, unsupervised anomaly detection technology has become essential for automating quality inspections in smart factories. Existing models relying solely on 2D RGB images have inherent limitations in detecting structural defects such as missing parts or deep surface scratches. To overcome these limitations, multimodal anomaly detection that combines 2D texture information with 3D shape information has attracted growing research interest [1–4].

However, the latest 3D-2D multimodal models face two critical limitations when deployed on resource-constrained environments in real-world industrial settings. First, simply concatenating 2D and 3D features without accounting for their scale discrepancy allows one modality's noise to dominate anomaly scoring, thereby degrading overall performance. In particular, on the MVTec 3D-AD dataset, for objects with high morphological variance—such as Carrot or Bagel—3D models that rely on absolute Z-axis depth tend to misclassify normal data as abnormal [5]. Second, models require inference speeds in the millisecond range, similar to those proposed in 2D vision [6], and recent high-resolution 3D anomaly detection studies have also considered real-time processing for subtle industrial defects [7]. However, existing multimodal approaches that utilize massive multi-memory banks to prevent modality scale conflicts [4] have excessively high computational and memory requirements, making real-time processing infeasible.

Therefore, this paper proposes the C-MTSS framework to address these issues. The main contributions are as follows:

- We propose a Dual-Stem Late Fusion structure that effectively distills 2D and 3D knowledge into a single-student network of less than 1M parameters without gradient interference.
- We improve robustness against shape variance by utilizing a 3D Surface Normal map and propose a Masked MSE Loss to optimize the model's capacity by suppressing background noise.
- We introduce a Z-score-based Calibration module that normalizes the scale discrepancy between modality-specific error distributions, enabling practical detection performance within sub-millisecond inference latency ($< 1\,ms$).

The remainder of this paper is organized as follows. Section 2 reviews related work on unsupervised anomaly detection in manufacturing, knowledge distillation-based anomaly detection, and multimodal 3D-2D anomaly detection. Section 3 describes the proposed C-MTSS framework, including surface normal extraction, dual-stem single-student knowledge distillation, masked MSE loss, and Z-score-based anomaly scoring. Section 4 presents the experimental results on the MVTec 3D-AD dataset, including quantitative comparisons, computational efficiency analysis, CPU-only benchmarking, sensitivity analysis, qualitative visualization, and ablation study. Finally, Sect. 5 concludes the paper and discusses future research directions (Fig. 1).

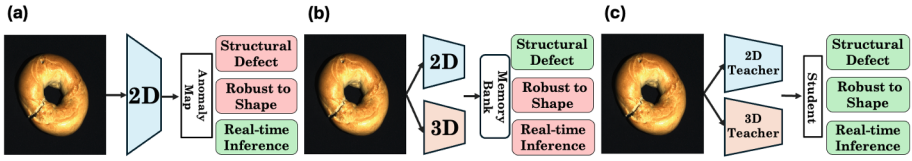

Fig. 1. Comparison summary of the proposed C-MTSS framework and existing anomaly detection paradigms. Unlike single 2D models or existing 3D-2D multimodal models with massive computations, the proposed C-MTSS framework simultaneously satisfies three core industrial requirements: structural defect detection, robustness to shape variance, and real-time inference speed (< 1 ms) for resource-constrained environments.

2 Related Work

2.1 Unsupervised Visual Anomaly Detection in Manufacturing

Unsupervised anomaly detection models for quality inspection in smart factory environments have evolved primarily by learning the distribution of normal data and treating patterns that deviate from it as anomalous [8]. Early approaches relied on Generative Adversarial Network (GAN)-based adversarial reconstruction [9,10], and more recently, hyperbolic space representations have been introduced to strengthen the separability of normal features [11]. In a different line of work, Normalizing Flow methods that directly estimate the density of high-dimensional features have led to notable gains in detection performance [12], and diffusion-based anomaly localization has also shown promising results [13,14].

Meanwhile, anomaly detection research is expanding beyond single-domain and closed-set settings toward broader applicability. Strategies such as distorted-feature attention, continuous memory representations, and self-supervised outlier synthesis have been explored for cross-domain generalization and open-world detection [15–17], along with few-shot prompting strategies [18].

2.2 Knowledge Distillation for Anomaly Detection

Knowledge Distillation (KD)-based approaches, in which a student model is trained to mimic the outputs of a pre-trained teacher with strong feature extraction capabilities, have emerged as a prominent direction in anomaly detection [19,20]. Since the student is trained exclusively on normal data, it fails to accurately reproduce the teacher's output for anomalous inputs at inference time; the resulting feature discrepancy between teacher and student serves as the anomaly score. Early work adopted a forward distillation scheme that transfers multi-resolution teacher features to the student, enabling the capture of anomalies at various scales [20]. Subsequently, a reverse distillation structure was introduced in which the student decoder reconstructs the teacher encoder's output, further amplifying representational discrepancy in anomalous regions [19]. As real-time processing demands in manufacturing have grown, lightweight KD

models with millisecond-level inference have been actively investigated. EfficientAD [6] achieves substantial speedup through a Patch Description Network (PDN), and heterogeneous distillation models [21] exploit structural asymmetry between teacher and student for improved efficiency.

However, existing KD-based anomaly detection models are predominantly designed for a single 2D RGB modality. Extending this paradigm to a multimodal setting introduces new challenges: scale discrepancy between heterogeneous modalities allows one modality to dominate learning, and gradient conflicts arise when distinct modalities are processed within a single student network. These issues have not been addressed in conventional single-modality KD frameworks, calling for dedicated architectural designs for multimodal KD.

2.3 Multimodal 3D-2D Anomaly Detection

In industrial environments, structural defects such as missing parts, dents, and assembly failures occur frequently, in addition to surface-level scratches or discoloration. Since such defects are difficult to detect using a single 2D modality, multimodal anomaly detection research combining 3D geometric features has become essential [1–3,22]. Recent 3D-2D multimodal methodologies have attempted to exploit complementary information through cross-modal feature mapping, attention mechanisms, or score-level fusion [2–4,23]. Memory bank-based approaches such as BTF [24], M3DM [4], and hypergraph-enhanced memory methods [25] achieve high localization accuracy by storing and retrieving normal feature prototypes, while KD-based methods such as AST [26] and MMRD [27] offer faster inference by avoiding large-scale storage. More recently, CFM [3] and LSFA [5] have proposed lightweight cross-modal mapping and self-supervised feature adaptation strategies, respectively, advancing performance on standard benchmarks.

However, existing methods still require complex network structures or pre-trained backbone models with substantial parameters, making them difficult to deploy in real-time resource-constrained environments. Moreover, models relying on raw, absolute Z-axis depth are highly vulnerable to the natural shape variances that occur within the same object class [5].

3 C-MTSS Framework

In this paper, we propose the C-MTSS, a lightweight 3D-2D multimodal anomaly detection framework capable of operating in real-time on resource-constrained environments.

3.1 Overall Structure

Figure 2 illustrates the overall structure of the proposed C-MTSS framework. This framework consists of four core pipelines. First, a preprocessing step converts the raw depth map acquired from the 3D sensor into a surface normal map

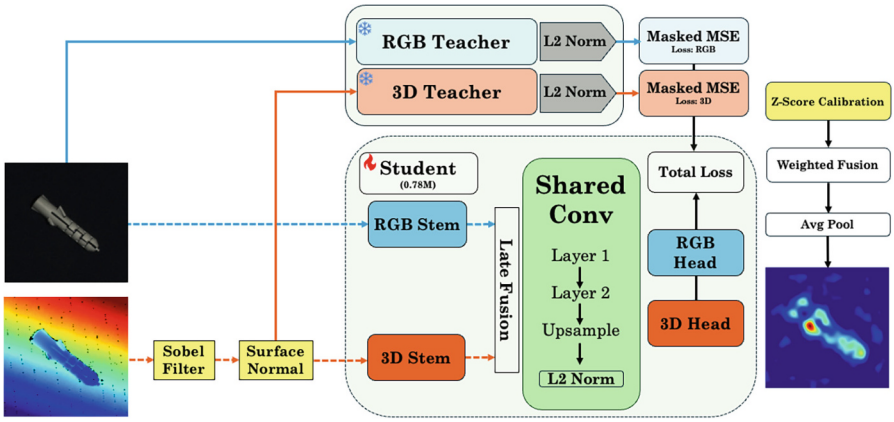

Fig. 2. Overall structure of the proposed C-MTSS framework. The input raw depth map is converted into a surface normal map via a Sobel filter to overcome shape variance. The single-student model, which mimics the knowledge of the frozen multi-teacher models, employs a dual-stem input layer to prevent initial gradient interference between modalities. L2 normalization and a Masked MSE loss function are applied during training to suppress background noise and control scale, while a Z-Score-based fusion module normalizes the errors of both modalities to the same scale during inference.

to mitigate shape variance. Second, multi-teacher models with frozen parameters extract the target features of the RGB and 3D Surface Normal domains. Third, a lightweight single-student model with fewer than 1M parameters (approximately 0.78M) simultaneously distills knowledge from both teacher models through a dual-stem structure that prevents interference between modalities. Finally, in the inference stage, a Z-Score Calibration fusion module normalizes and weight-combines the scale differences between the two modalities.

3.2 Surface Normal Extraction for Shape Variance

In industrial environments, shape variance frequently occurs when parts are placed at a slant or randomly rotated on a conveyor belt. In such cases, the overall height distribution of the raw 3D depth map $D \in \mathbb{R}^{H \times W}$ obtained from the sensor is deformed, which is a major cause of deep learning models misclassifying normal objects as defective (false positives).

To ensure robustness against such shape variance, this paper converts the depth map into a surface normal vector in 3D space. The gradients along the x-axis and y-axis at each pixel (x, y) of the depth map are calculated using a Sobel filter, and the result is normalized to extract the normal map $N(x, y)$:

$$N(x,y) = \frac{(-\nabla_x D,\ -\nabla_y D,\ 1)}{\|(-\nabla_x D,\ -\nabla_y D,\ 1)\|_2}. \tag{1}$$

Since the extracted normal vector has a range of $[-1, 1]$, a linear transformation $N'(x, y) = (N(x, y) + 1)/2$ is applied to scale it to the $[0, 1]$ range so that the network can learn in the same feature space as the RGB image. Through this process, the model can ignore the overall tilt of the object and focus only on essential surface defects, such as dents or scratches. Furthermore, a binary validity mask M is defined as $M(x, y) = 1$ if $D(x, y) > 0$, and 0 otherwise, to exclude sensor-invalid background regions from both training loss and inference scoring.

3.3 Dual-Stem Single-Student Knowledge Distillation

To maximize computational efficiency, this framework adopts a single-student network. However, if information from two heterogeneous modalities is simply merged in the initial layers, representation learning can collapse due to gradient interference.

Dual-Stem Structure. To prevent initial interference, the input part of the student model is separated into a dual-stem structure consisting of two shallow convolutional layers that independently process RGB and 3D inputs. The two branches are fused in the subsequent shared layers, which mitigates the gradient conflict inherent in a naïve single-input design.

L2 Normalization. The output feature maps of the teacher and student models exhibit different scales for each modality. To prevent the error of a particular modality from dominating the entire loss function, L2 normalization is applied to the intermediate feature maps f_1 and f_2 prior to concatenation ($f_{\text{norm}} = f_1 \,/\, \|f\|_2$), for both teacher and student networks.

Masked MSE Loss. The background region of 3D data contains meaningless noise, which wastes the representational capacity of the model. To mitigate this, we propose a Masked MSE Loss using a binary mask M in which only the valid region containing the object from the depth sensor is assigned a value of 1. For modality $m \in \{\text{RGB, 3D}\}$, let T_m denote the teacher features and S_m denote the student features. The loss function is defined as follows:

$$\mathcal{L}_m = \frac{\displaystyle\sum_{i,j} \left(M_{i,j} \cdot \frac{1}{C} \sum_{c=1}^{C} \left(S_m^c(i,j) - T_m^c(i,j) \right)^2 \right)}{\displaystyle\sum_{i,j} M_{i,j} + \epsilon}. \tag{2}$$

The network is optimized via $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RGB}} + \mathcal{L}_{\text{3D}}$, enabling the student to distill multimodal surface knowledge while suppressing background interference. The same mask is applied to the RGB modality to ensure that the student focuses its limited capacity on the object region, avoiding unnecessary learning of background patterns such as table surfaces or lighting artifacts.

3.4 Z-Score Calibration and Anomaly Scoring

During the inference stage, for each modality m, the mean squared error (MSE) between the teacher and student features is computed to generate the error map E_m. However, since the RGB error and the 3D error differ in their value ranges and distributions, they cannot be directly summed.

In this paper, we apply Z-Score Calibration, which normalizes the two error maps into standard normal distributions with mean 0 and variance 1, based on the error mean (μ_m) and standard deviation (σ_m) of each modality extracted from the normal training set. Considering the relative importance of RGB texture defects and 3D surface defects, we assign weights of 0.6 to RGB and 0.4 to 3D to compute the final fused anomaly map A_{fused}:

$$A_{\text{fused}} = 0.6 \cdot \frac{E_{\text{RGB}} - \mu_{\text{RGB}}}{\sigma_{\text{RGB}}} + 0.4 \cdot \frac{E_{\text{3D}} - \mu_{\text{3D}}}{\sigma_{\text{3D}}}. \tag{3}$$

Finally, average pooling with an 11×11 kernel and replicate padding is applied to A_{fused} to suppress local noise artifacts. The image-level anomaly score is then defined as the mean of the top 1% pixel values within the valid object region (where $M = 1$).

4 Experiments

This section evaluates the anomaly detection performance and computational efficiency of the proposed C-MTSS framework. We conduct quantitative comparisons against representative 3D-2D baselines and state-of-the-art multimodal models spanning diverse paradigms—including memory bank-based, knowledge distillation-based, and cross-modal mapping approaches—and analyze both detection accuracy and real-time inference capability.

4.1 Dataset and Experimental Setup

Dataset. Performance evaluation was conducted on the MVTec 3D-AD dataset [1], a standard benchmark for 3D industrial anomaly detection. This dataset comprises 10 diverse object categories, each providing paired high-resolution RGB images and corresponding 3D XYZ depth maps. In our experiments, the depth maps were converted into surface normal maps through the preprocessing pipeline described in Sect. 3.2.

Evaluation Metrics. For objective evaluation, we adopted widely used standard metrics [8]. Image AUROC (I-AUROC) was used to measure image-level classification performance. Pixel AUROC (P-AUROC) was employed to assess pixel-level localization accuracy.

Implementation Details. The proposed framework was implemented in PyTorch. To ensure fair benchmarking of inference latency against competing models, all training and evaluation were performed on an NVIDIA A100 Tensor

Table 1. Quantitative comparison on MVTec 3D-AD (mean across 10 categories). Results for BTF, AST, M3DM, and CFM are from [3]; MMRD from [27]; LSFA from [5]. The best result in each column is in **bold**.

Method	Venue	Paradigm	I-AUROC (%) ↑	P-AUROC (%) ↑
BTF [24]	CVPRW'23	MB	86.5	99.2
AST [26]	WACV'23	KD	93.7	97.6
M3DM [4]	CVPR'23	MB	94.5	99.2
MMRD [27]	AAAI'24	KD	94.0	98.7
CFM [3]	CVPR'24	FM	95.4	99.3
LSFA [5]	ECCV'24	FA	**97.1**	**99.2**
C-MTSS (Ours)	—	KD	91.3	96.7

KD = Knowledge Distillation, MB = Memory Bank, FM = Feature Mapping, FA = Feature Adaptation.

Core GPU (80 GB VRAM) with 16 vCPUs and 192 GB RAM. The Adam optimizer was used with a learning rate of 1×10^{-3}, scheduled via StepLR with a step size of 5,000 and decay factor of 0.1. The student network was trained for 10,000 iterations per category with a batch size of 32 and an input resolution of 256×256 to allow stable calibration of inter-modality scale differences. Both teacher networks use a ResNet-18 backbone pre-trained on ImageNet, truncated at Layer 2.

4.2 Quantitative Comparison with State-of-the-Art

To validate the proposed framework, we compare C-MTSS against six representative multimodal anomaly detection methods spanning three distinct paradigms: (1) *Memory bank-based* methods—BTF [24] and M3DM [4]—which store and retrieve normal feature prototypes for anomaly scoring; (2) *Knowledge distillation-based* methods—AST [26] and MMRD [27]—which train student networks to mimic frozen teacher representations; and (3) *Cross-modal mapping/adaptation* methods—CFM [3] and LSFA [5]—which learn inter-modality feature correspondences.

As shown in Table 1 and Table 2, the proposed C-MTSS framework achieves an I-AUROC of 91.3%, surpassing the memory bank-based baseline BTF (86.5%) by +4.8 pp and narrowing the gap to within 2.7 pp of the AAAI'24 knowledge distillation model MMRD (94.0%) and 2.4 pp of AST (93.7%). This is particularly noteworthy considering that the entire C-MTSS inference pipeline comprises only 2.15M parameters—two orders of magnitude fewer than competing approaches.

The per-category analysis in Table 2 further reveals that C-MTSS achieves strong I-AUROC on categories where the Surface Normal representation effectively captures geometric defects, such as Rope (98.1%) and Dowel (96.4%), with

Table 2. Per-category I-AUROC (%) on MVTec 3D-AD. Values for BTF, AST, M3DM, and CFM are from [3]. MMRD and LSFA are omitted as per-category breakdowns are not publicly available. The best result per category is in **bold**.

Method	Bagel	Cable G.	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
BTF [24]	91.8	74.8	96.7	88.3	93.2	58.2	89.6	91.2	92.1	88.6	86.5
AST [26]	98.3	87.3	97.6	97.1	93.2	88.5	97.4	98.1	**100.0**	79.7	93.7
M3DM [4]	**99.4**	**90.9**	97.2	97.6	96.0	**94.2**	97.3	89.9	97.2	85.0	94.5
CFM [3]	**99.4**	88.8	**98.4**	**99.3**	**98.0**	88.8	94.1	**94.3**	98.0	**95.3**	**95.4**
C-MTSS (Ours)	96.0	**91.5**	93.7	89.7	96.4	87.9	87.7	88.1	98.1	83.9	91.3

several categories approaching or matching the performance of models with substantially larger capacity. The accuracy gap is more pronounced on categories with extreme texture complexity (e.g., Cookie 89.7%, Tire 83.9%), which we attribute to the limited capacity of the compact student network for encoding fine-grained texture patterns.

On the localization side, C-MTSS achieves 96.7% P-AUROC—a pixel-level localization score that remains within 2.6 pp of the best-performing models—indicating that the proposed Z-Score Calibration and Masked MSE Loss effectively preserve fine-grained defect localization capability even within an ultra-lightweight structure. As will be demonstrated in Sect. 4.3, this accuracy is accompanied by a throughput advantage exceeding 200× over memory bank-based methods.

4.3 Computational Efficiency and Real-Time Inference Analysis

In real-world manufacturing lines, not only detection accuracy but also real-time processing speed and low memory footprint on resource-constrained environments are critical requirements. This subsection benchmarks the total inference parameters, trainable parameters, and inference speed (FPS) of C-MTSS against the baseline models.

As shown in Table 3, memory bank-based methods (BTF, M3DM) deliver the highest localization accuracy but require over 100–300M parameters and incur severe computational bottlenecks due to feature matching during inference, yielding only 0.5–3.2 FPS. Even knowledge distillation and feature mapping approaches (AST, MMRD, CFM) require 24–50M parameters and achieve moderate throughput of 5–30 FPS, which remains insufficient for millisecond-level response requirements.

In contrast, the entire inference footprint of C-MTSS—including two frozen teacher models and the trainable student—amounts to only **2.15M** parameters, representing an approximately 140× reduction compared to M3DM and roughly 11× reduction compared to AST. Among these, only **0.78M** parameters are trainable. The high throughput of C-MTSS (**1031.8 FPS**) can be largely attributed to the teacher network design: rather than executing the full ResNet-18 backbone, each teacher performs a *truncated forward pass* through only the shallow layers (up to Layer 2), discarding the computationally expensive deeper

Table 3. Computational efficiency and inference speed comparison. C-MTSS was measured on an NVIDIA A100 GPU. Total Params includes all parameters loaded during inference (including frozen teachers). Values marked with ∼ are cited from respective publications or estimated under comparable settings. Memory values for BTF, AST, M3DM, and CFM are from [3].

Method	Paradigm	Total Params (M) ↓	Trainable Params (M) ↓	Memory (MB) ↓	Speed (FPS) ↑	P-AUROC (%) ↑
BTF [24]	MB	> 100.0	–	381.1	∼3.2	99.2
M3DM [4]	MB	> 300.0	> 100.0	6526.1	∼0.5	99.2
AST [26]	KD	≈ 24.5	≈ 24.5	463.9	∼5.0	97.6
MMRD [27]	KD	≈ 50.0	≈ 25.0	–	∼30	98.7
CFM [3]	FM	≈ 40.0	≈ 5.0	437.9	∼21.8	99.3
C-MTSS (Ours)	KD	**2.15**	**0.78**	**38.0**	**1031.8**	96.7

layers (Layers 3 and 4). Combined with the absence of any memory bank lookups, this design results in a per-image latency below 1 ms.

To further examine deployment scenarios without dedicated GPU acceleration, we additionally report a CPU-only reference benchmark. This benchmark is divided into two settings. The end-to-end setting includes file loading, RGB/depth preprocessing, surface normal extraction, model inference, Z-Score fusion, and anomaly scoring. The model-scoring setting excludes file loading and preprocessing, and measures only the inference and anomaly scoring stages using preprocessed inputs.

Table 4. CPU-only reference benchmark of C-MTSS. End-to-end includes preprocessing and anomaly scoring, while model scoring excludes file loading and preprocessing. Measurements are conducted with batch size 1 and an input resolution of 256×256 on the CPU of the same workstation.

Setting	Latency (ms/image) ↓	FPS ↑
CPU-only End-to-End	75.87	13.18
CPU-only Model Scoring	34.47	29.01

As shown in Table 4, the CPU-only end-to-end setting achieves 75.87 *ms* per image, corresponding to 13.18 FPS. In contrast, the model-scoring setting achieves 34.47 *ms* per image and 29.01 FPS, approaching the common video-rate criterion of 30 FPS. These results indicate that the full CPU-only pipeline is still affected by file I/O and surface normal preprocessing, while the computational burden of the C-MTSS inference and scoring stages remains relatively low even without GPU acceleration.

In summary, high-accuracy models such as M3DM or LSFA may be more suitable for offline analysis or accuracy-oriented inspection scenarios. In contrast, C-MTSS targets real-time manufacturing inspection environments where latency, memory footprint, and throughput are critical constraints. Therefore,

the proposed framework should be viewed not as a model that solely maximizes benchmark accuracy, but as a practical lightweight structure for real-time inspection. It achieves 96.7% P-AUROC while exceeding 1,000 FPS on an A100 GPU, and the additional CPU-only benchmark further shows that its model-scoring stage approaches video-rate throughput.

4.4 Sensitivity Analysis of Z-Score Fusion Weights

To examine whether the Z-Score fusion weights are overly dependent on a single heuristic setting, we conduct a sensitivity analysis by varying the RGB and 3D weights while keeping their sum equal to 1.0. This experiment is performed on all 10 categories of the MVTec 3D-AD dataset, with the RGB-to-3D ratio varied from 0.4:0.6 to 0.8:0.2.

Table 5. Sensitivity analysis of Z-Score fusion weights on MVTec 3D-AD. I-AUROC values are averaged across all 10 categories. The default setting used in C-MTSS is highlighted in bold.

RGB	3D	I-AUROC (%) ↑
0.4	0.6	87.32
0.5	0.5	89.75
0.6	**0.4**	**91.30**
0.7	0.3	89.81
0.8	0.2	89.60

As shown in Table 5, the default RGB-to-3D weight ratio of 0.6:0.4 achieves the highest I-AUROC of 91.30%. When the 3D branch is assigned a larger weight than the RGB branch, the performance drops to 87.32%. This suggests that excessive reliance on 3D residuals can amplify shape-related variations or sensor noise in the final anomaly score. Conversely, increasing the RGB weight beyond 0.6 also reduces the detection accuracy, indicating that surface-normal cues still provide complementary information for structural defects.

These results show that the 0.6:0.4 fusion weight used in C-MTSS is not an arbitrary choice, but provides the best balance between RGB texture information and 3D surface-normal information in our experiments. The analysis further suggests that Z-Score Calibration is not merely a scale-normalization step, but also affects how modality-specific error distributions contribute to the final anomaly score.

4.5 Qualitative Visualization

To verify the anomaly localization capability across all categories of the proposed C-MTSS framework, we present qualitative visualization results for all 10

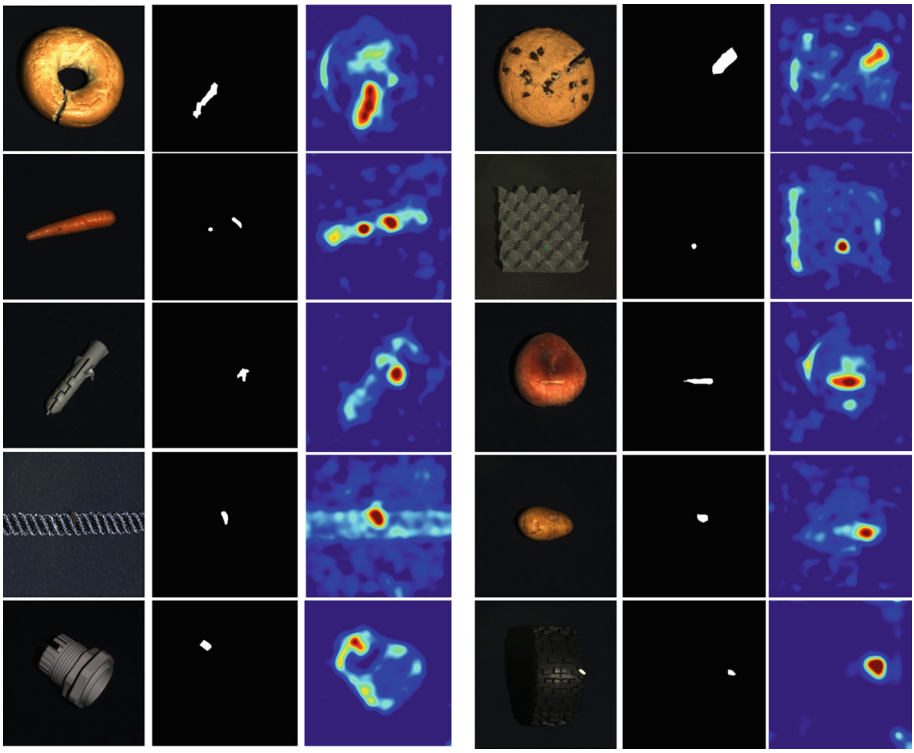

Fig. 3. Qualitative anomaly localization on all 10 categories of MVTec 3D-AD. Each group of three images (from left to right: input RGB, ground truth mask, anomaly heatmap) corresponds to one category, arranged in a 5×2 grid.

categories of the MVTec 3D-AD dataset in Fig. 3. Each group of three images (from left to right: input RGB, ground truth, anomaly heatmap) corresponds to one category, arranged in a 5×2 grid.

The overall results can be analyzed according to three physical property groups. For natural object categories with significant intra-class shape variance—such as **Bagel**, **Carrot**, **Cookie**, **Peach**, and **Potato**—C-MTSS successfully focuses the heatmap on localized surface defects (e.g., cracks and dents) through the Surface Normal conversion to disregard global variations in tilt and height distribution. For rigid body categories with fixed geometries—such as **Dowel**, **Cable Gland**, and **Tire**—the predicted anomaly regions exhibit high spatial correspondence with the ground truth masks, demonstrating the precise localization capability of the lightweight student network even for fine-grained structural defects. Finally, for deformable object categories whose spatial layout varies across samples—such as **Rope** and **Foam**—texture-based anomalous regions are clearly activated in the heatmaps, which suggests that the RGB and 3D Surface Normal modalities provide complementary cues even under spatial deformation.

Table 6. Ablation study on MVTec 3D-AD. Each row progressively adds one component.

Variant	Dual-Stem	Masked MSE	Z-Score	I-AUROC (%)	P-AUROC (%)
A				87.68	95.80
B	✓			90.34	96.10
C	✓	✓		90.50	96.14
D	✓	✓	✓	**91.30**	**96.70**

4.6 Ablation Study

To quantify the contribution of each proposed component, we conduct an ablation study with four progressive variants on MVTec 3D-AD, as summarized in Table 6.

Variant *A* serves as the baseline, where a single-stem student network processes both modalities through a shared input layer, trained with a standard MSE loss, and the two error maps are directly summed without normalization. Replacing the single-stem input with the proposed dual-stem structure ($A \rightarrow B$) yields the largest individual gain of +2.66 in I-AUROC and +0.30 in P-AUROC, confirming that separating the initial feature pathways reduces gradient interference between the heterogeneous RGB and 3D modalities. Adding the Masked MSE loss ($B \rightarrow C$) produces a modest increase of +0.16 in I-AUROC and +0.04 in P-AUROC. While its standalone effect is limited, the primary purpose of this module is not to directly boost classification accuracy but to suppress background noise in the 3D modality, which refines the per-modality error distributions. This benefit becomes apparent when combined with Z-Score Calibration: applying Z-Score fusion on top of Masked MSE ($C \rightarrow D$) yields +0.80 in I-AUROC and +0.56 in P-AUROC. Without background suppression, the calibration statistics are contaminated by sensor-invalid regions, which diminishes the effectiveness of the normalization-based fusion. Overall, the full C-MTSS configuration (Variant *D*) outperforms the baseline by +3.62 in I-AUROC and +0.90 in P-AUROC, validating the complementary contributions of all three proposed components.

5 Conclusion

This paper presented C-MTSS, a lightweight framework for 3D-2D multimodal industrial anomaly detection. The framework uses three design choices: dual-stem late fusion, Masked MSE Loss, and Z-Score Calibration. These components jointly address gradient interference, background noise, and scale discrepancy when processing heterogeneous modalities within an ultra-lightweight structure.

On the MVTec 3D-AD benchmark, C-MTSS achieves 91.3% I-AUROC and 96.7% P-AUROC. It also requires only 38 MB of GPU memory and delivers over 1,000 FPS. These results suggest that C-MTSS provides a practical balance between detection performance, inference speed, and memory efficiency for real-time industrial inspection.

However, a performance gap remains on texture-heavy categories due to the limited representational capacity of the compact student network. Future work aims to narrow this gap by adopting foundation models as teachers to improve distillation quality and by leveraging cross-modal attention mechanisms to exploit complementary information between modalities, while preserving the real-time inference capability.

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. Bergmann, P., et al.: The MVTec 3D-ad dataset for unsupervised 3D anomaly detection and localization. In: VISAPP, pp. 202–213 (2022). https://doi.org/10.5220/0010935500003123
2. Bhunia, A., et al.: Looking 3D: anomaly detection with 2D-3D alignment. In: CVPR, pp. 17263–17272 (2024). https://doi.org/10.1109/CVPR52733.2024.01634
3. Costanzino, A., et al.: Multimodal industrial anomaly detection by crossmodal feature mapping. In: CVPR, pp. 17234–17243 (2024). https://doi.org/10.1109/CVPR52733.2024.01631
4. Wang, Y., et al.: Multimodal industrial anomaly detection via hybrid fusion. In: CVPR, pp. 8032–8041 (2023). https://doi.org/10.48550/arXiv.2303.00601
5. Tu, Y., et al.: Self-supervised feature adaptation for 3d industrial anomaly detection. In: ECCV, pp. 75–91. Springer (2024). https://doi.org/10.1007/978-3-031-72627-9_5
6. Batzner, K., et al.: EfficientAD: accurate visual anomaly detection at millisecond-level latencies. In: WACV, pp. 128–138 (2024). https://doi.org/10.1109/WACV57701.2024.00020
7. Cheng, Y., et al.: Towards high-resolution 3D anomaly detection: a scalable dataset and real-time framework for subtle industrial defects. In: AAAI, vol. 40, pp. 3327–3334 (2026). https://doi.org/10.1609/aaai.v40i5.37328
8. Bergmann, P., et al.: MVTec AD–a comprehensive real-world dataset for unsupervised anomaly detection. In: CVPR, pp. 9584–9592 (2019). https://doi.org/10.1109/CVPR.2019.00982

9. Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: GANomaly: semi-supervised anomaly detection via adversarial training. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11363, pp. 622–637. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20893-6_39
10. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.-T., Shen, D. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 146–157. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_12
11. Li, H., et al.: Hyperbolic anomaly detection. In: CVPR, pp. 17511–17520 (2024). https://doi.org/10.1109/CVPR52733.2024.01658
12. Gudovskiy, D., et al.: CFlow-AD: real-time unsupervised anomaly detection with localization via conditional normalizing flows. In: WACV, pp. 98–107 (2022). https://doi.org/10.48550/arXiv.2107.12571
13. Fučka, M., et al.: Transfusion – a transparency-based diffusion model for anomaly detection. In: ECCV, pp. 91–108. Springer (2024). https://doi.org/10.1007/978-3-031-72761-0_6
14. Liu, C., et al.: Learning diffusion models for multi-view anomaly detection. In: ECCV, pp. 328–345. Springer (2024). https://doi.org/10.1007/978-3-031-73414-4_19
15. Isaac-Medina, B.K.S., et al.: Towards open-world object-based anomaly detection via self-supervised outlier synthesis. In: ECCV, pp. 196–214. Springer (2024). https://doi.org/10.1007/978-3-031-73209-6_12
16. Lee, J.C., et al.: Continuous memory representation for anomaly detection. In: ECCV, pp. 438–454. Springer (2024). https://doi.org/10.1007/978-3-031-72983-6_25
17. Sträter, L.P.J., et al.: GeneralAD: anomaly detection across domains by attending to distorted features. In: ECCV, pp. 448–465. Springer (2024). https://doi.org/10.1007/978-3-031-72913-3_25
18. Zhu, J., Pang, G.: Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In: CVPR, pp. 17826–17836 (2024). https://doi.org/10.48550/arXiv.2403.06495
19. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: CVPR, pp. 9737–9746 (2022). https://doi.org/10.1109/CVPR52688.2022.00950
20. Salehi, M., et al.: Multiresolution knowledge distillation for anomaly detection. In: CVPR, pp. 14902–14912 (2021). https://doi.org/10.1109/CVPR46437.2021.01465
21. Wu, L., Zhou, J.: Heterogeneous knowledge distillation for anomaly detection. IEEE Access **12**, 161490–161499 (2024). https://doi.org/10.1109/ACCESS.2024.3415503
22. Zhu, W., et al.: Real-IAD D^3: a real-world 2D/pseudo-3D/3D dataset for industrial anomaly detection. In: CVPR, pp. 15214–15223 (2025). https://doi.org/10.1109/CVPR52734.2025.01417
23. Tao, C., et al.: G2SF: geometry-guided score fusion for multimodal industrial anomaly detection. In: ICCV, pp. 20551–20560 (2025). https://doi.org/10.48550/arXiv.2503.10091
24. Horwitz, E., et al.: 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

25. Lin, Y., et al.: Commonality in few: few-shot multimodal anomaly detection via hypergraph-enhanced memory. In: AAAI, vol. 40, pp. 7015–7023 (2026). https://doi.org/10.1609/aaai.v40i9.37636
26. Rudolph, M., et al.: Asymmetric student-teacher networks for industrial anomaly detection. In: WACV, pp. 2592–2602 (2023). https://doi.org/10.1109/WACV56688.2023.00261
27. Gu, Z., et al.: Rethinking reverse distillation for multi-modal anomaly detection. In: AAAI, vol. 38, pp. 8218–8226 (2024). https://doi.org/10.1609/aaai.v38i8.28687