arXiv:2605.07149v1 [cs.CV] 8 May 2026

Real-IAD MVN: A Multi-View Normal Vector Dataset and Benchmark for High-Fidelity Industrial Anomaly Detection

Wenbing Zhu[1,4†], Jianing Liang[3†], Linjie Cheng[1], Yurui Pan[1],
Zhuhao Chen[1], Qingwang Yan[1], Yudong Cheng[1], Jianghui Zhang[2], Mingmin Chi[1], Bo Peng[2*]
[1]Fudan University [2]Shanghai Ocean University [3]Donghua University [4]Rongcheer Co., Ltd.
{wbzhu23, ljcheng24, yrpan24}@m.fudan.edu.cn, 231310103@mail.dhu.edu.cn,
{zhuhaochen24, qwyan25, 25113050248, mmchi}@m.fudan.edu.cn, m240751960@st.shou.edu.cn,
bpeng@shou.edu.cn

Abstract

Industrial Anomaly Detection (IAD) is critical for quality control, but existing methods struggle with subtle, geometric defects. Standard 2D (RGB) images are sensitive to texture and lighting but often miss fine geometric anomalies. While 3D point clouds capture macro-shape, they are typically too sparse to detect micro-defects like scratches or pits. We address this fundamental data limitation by introducing Real-IAD-MVN (Multi-View Normal), a large-scale industrial dataset. By upgrading our acquisition system, Real-IAD-MVN captures high-fidelity surface normal maps from five distinct viewpoints, replacing sparse 3D data entirely. This provides a comprehensive geometric representation at a micro-detail level, making previously invisible side-wall and occluded defects explicitly detectable. Our experiments, conducted on this new dataset, first provide evidence that incorporating dense, multi-view pseudo-3D (surface normals) yields significantly better detection performance than using sparse 3D point cloud data. To further validate the dataset and provide a strong benchmark, we introduce a baseline method based on reconstruction, which learns to extract cross-modal unified prototypes from the image and normal map streams. We demonstrate that this unified prototype approach surpasses existing state-of-the-art multimodal fusion methods, highlighting the rich potential of our new dataset for advancing geometric anomaly detection.

1. Introduction

Unsupervised anomaly detection (AD) is a cornerstone of modern industrial manufacturing, ensuring product quality and safety by identifying defects without requiring vast datasets of failure examples [2, 23]. The primary goal is to model the distribution of normal products and then identify any deviation as an anomaly, both at the image level (classification) and pixel level (segmentation).

Initial AD methods relied on 2D (RGB) images [2, 23]. However, these approaches face a critical limitation: they are adept at finding *textural* or *color* anomalies but are fundamentally weak at identifying purely *geometric* defects, such as fine scratches, dents, or deformations. An anomaly's 2D appearance can be masked by challenging illumination, low-contrast, or reflective surfaces.

This weakness spurred the development of multimodal datasets, most notably those combining RGB with 3D data. Datasets like MVTec 3D-AD [3] and Real3D-AD [11] utilize 3D point clouds to capture the macro-structural shape of objects. While this helps detect large-scale geometric anomalies (e.g., missing parts, severe warping), 3D point cloud data is often sparse, noisy, and at a resolution far too coarse to register the micro-geometric defects that are critical in high-precision manufacturing.

A more promising direction for capturing fine geometry is photometric stereo (PS), which reconstructs high-fidelity surface normal maps [10]. Surface normals provide a dense, per-pixel representation of local geometry, making them extremely sensitive to subtle surface changes. Our prior work, Real-IAD D3 [22], pioneered this by introducing a top-down pseudo-3D (normal map) modality. This addition proved highly effective for detecting defects on surfaces facing the camera. However, its single, top-down viewpoint remained a critical bottleneck. As illustrated in Figure 1, a defect on the side of an object is geometrically invisible to a top-down sensor, rendering it undetectable.

In this paper, we directly address this fundamental limitation. We present **Real-IAD-MVN**, a significant evolution of our previous work, detailed in Section 3. We have engineered a new acquisition system that captures high-fidelity

† Equal contribution. ∗ Corresponding author.

photometric stereo data from **five calibrated viewpoints** (e.g., top, front, back, left, right). This multi-view normal (MVN) data provides a complete, 360-degree geometric profile of each sample, effectively eliminating the blind spots of single-view systems. We have extended the 20 categories from Real-IAD D3 [22], creating a large-scale, challenging benchmark where view-dependent geometric anomalies are now the focus.

A new dataset demands validation via a strong baseline. We propose a reconstruction-based model designed to leverage this rich, multi-modal, multi-view data. Inspired by the challenge of cross-modal inconsistencies [18], our baseline learns to extract **cross-modal unified prototypes** from the fused image and normal map streams. These prototypes holistically represent the joint "appearance-geometry" distribution of a normal sample. By tasking the model with reconstructing both modalities from this shared prototype representation, anomalies that violate this learned consistency (e.g., correct appearance but incorrect geometry) fail reconstruction and are thus clearly identified.

We validate our contributions comprehensively. First, experiments on our new dataset provide clear evidence that incorporating dense, multi-view pseudo-3D (surface normals) yields significantly better detection performance than using sparse 3D point cloud data. Second, we demonstrate that our unified prototype baseline surpasses existing state-of-the-art multimodal fusion methods, highlighting the rich potential of our new dataset.

2. Related Work

2.1. Unsupervised Anomaly Detection

Unsupervised AD methods are typically trained only on normal samples. These methods can be broadly categorized into three families. **Embedding-based** methods learn a descriptive embedding for normal features. PaDiM [4] models patch features with a Gaussian distribution, while PatchCore [14] uses a coreset-based memory bank of normal features, achieving state-of-the-art performance with high efficiency. **Reconstruction-based** methods train models like Autoencoders [1] or GANs [15] to reconstruct normal samples, assuming anomalies will have high reconstruction error. Recent works like UniAD [20] and OneNIP [5] have successfully used Transformers for unified, multi-class AD, with OneNIP pioneering the use of a normal image as a prompt. **Discriminator-based** methods, or "pseudo-anomaly" methods, generate synthetic anomalies and train a discriminative model to distinguish them from real normal samples. DRAEM [21] uses Perlin noise, CutPaste [9] pastes image patches, and SimpleNet [12] adds noise in the feature space.

2.2. Multimodal Anomaly Detection

Recognizing the limitations of 2D data, researchers have moved to multimodal AD. **RGB + 3D Point Cloud:** The MVTec 3D-AD dataset [3] and Real3D-AD [11] provided benchmarks for fusing 2D images with 3D point cloud data. Methods like M3DM [17, 18] proposed hybrid fusion techniques to leverage both surface texture (from 2D) and macro-shape (from 3D). However, as noted, point clouds are often too sparse for micro-defects. **RGB + Pseudo-3D (Normals):** Our previous work, Real-IAD D3 [22], was the first to incorporate high-precision, top-down photometric stereo (PS) data. This demonstrated the power of surface normals for detecting fine geometric defects. LINO UniPS [10] further confirmed the utility of PS for surface reconstruction, although not specifically for multi-view AD. Our new work builds directly on this, addressing the key limitation of single-viewpoint PS.

3. The Real-IAD-MVN Dataset

Our new dataset, Real-IAD-MVN, represents a paradigm shift in data acquisition for IAD, moving from single-view or sparse 3D data to dense, comprehensive multi-view surface geometry. It is built upon the established foundation of Real-IAD D3 [22] but introduces a fundamentally richer geometric modality.

3.1. Multi-View Acquisition System

We designed and constructed an integrated multi-view photometric stereo (MVPS) gantry. This system is crucial for capturing the novel data modality.

- **Cameras:** A single high-resolution industrial camera (3648×5472 pixels, similar to [22]) is mounted on a robotic arm capable of precisely positioning it at five pre-calibrated viewpoints around the object (e.g., top, front, back, left, right).
- **Photometric Stations:** Co-located with each camera viewpoint is a photometric stereo station. Each station employs multiple (e.g., 4 or more) directional light sources positioned at known angles relative to the camera [10, 22].
- **Synchronized Capture:** For each sample, the system captures a sequence of images under different lighting directions from each of the five viewpoints. Additionally, a standard diffusely-lit RGB image (I_v) is captured from each view.

This setup ensures precise alignment between the RGB image and the computed normal map for each view, and the calibration between views allows for potential 3D reconstruction or cross-view warping if needed, though our proposed method operates primarily in 2D feature space.

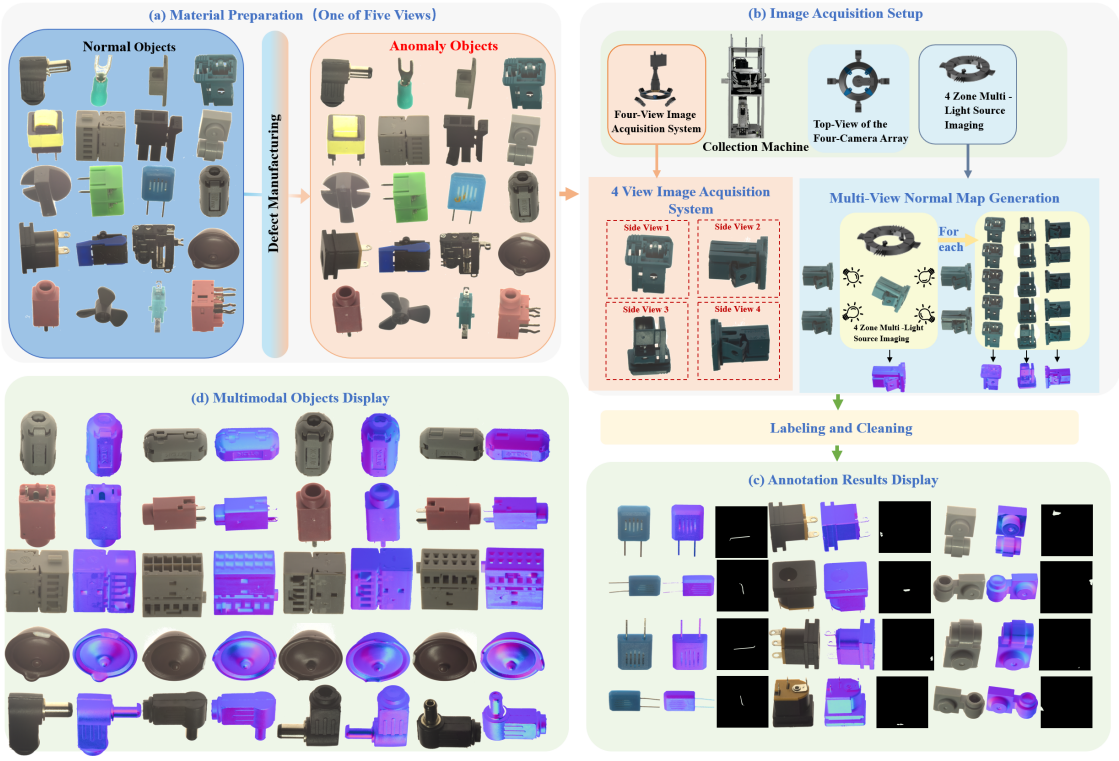

Figure 1. Overview of the Real-IAD-MVN dataset acquisition and annotation pipeline. (a) Material preparation, showing examples of normal and anomalous objects from one of the five views. (b) The integrated multi-view photometric stereo (MVPS) acquisition gantry. (c) Examples of the pixel-perfect annotation masks and cleaning process. (d) A display of the final multimodal data, showing various RGB objects and their corresponding high-fidelity pseudo-3D (normal map) representations.

Table 1. Comparison of Real-IAD-MVN with existing multimodal industrial anomaly detection datasets.

Dataset	Categories	Defect Types	Samples	Modalities	Geometric Resolution	Multi-View	Primary Geometric Data
MVTec 3D-AD [3]	10	33	4147	RGB + 3D Point Cloud	~0.11mm (point precision)	No	Sparse Point Cloud
Real3D-AD [11]	12	40	1254	RGB + 3D Point Cloud	~0.01mm (point precision)	No	Sparse Point Cloud
Real-IAD D3 [22]	20	69	8450	RGB + 3D Point Cloud + Pseudo-3D (Normal)	~0.002mm (normal precision)	No (Top-view normal)	Sparse PC + Single-view Normal
Real-IAD-MVN (Ours)	**20**	**69**	**8450**	**MV-RGB + MV-Normal**	**~0.002mm (normal precision)**	**Yes (5 views)**	**Dense Multi-View Normals**

3.2. Photometric Stereo for Normals

From the sequence of images captured under varying illumination at each viewpoint v, we compute the surface normal map N_v. Assuming Lambertian reflectance for simplicity (though more complex models can be used [10]), the intensity $I(x, y)$ at a pixel (x, y) under a known light source direction $L \in \mathbb{R}^3$ is related to the surface normal $n(x, y) \in \mathbb{R}^3$ and albedo $\rho(x, y)$ by:

$$I(x, y) = \rho(x, y) \max(0, L \cdot n(x, y))$$

With at least three non-collinear light source directions $\{L_k\}_{k=1}^K$ $(K \geq 3)$ and their corresponding intensities $\{I_k(x, y)\}_{k=1}^K$, we can solve for the scaled normal vector $g(x, y) = \rho(x, y)n(x, y)$ using a least-squares approach [19, 22]. Let $\mathbf{I}$ be the $K \times 1$ vector of intensities and $\mathbf{L}$ be the $K \times 3$ matrix of light directions. Then:

$$\mathbf{I} = \mathbf{L}g$$

$$g = (\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T\mathbf{I}$$

The final normal $n(x, y)$ is obtained by normalizing $g(x, y)$: $n = g/||g||_2$. This process is repeated for each viewpoint v, yielding the set $\{N_v\}_{v=1}^5$. The resulting normal maps offer micrometer-level precision in capturing surface geometry, far exceeding the detail available from typical 3D scanners used in prior datasets [22].

3.3. Dataset Statistics and Comparison

Real-IAD-MVN retains the 20 object categories and 69 defect types from Real-IAD D3 [22]. The total number of samples remains 8,450 (5,000 normal, 3,450 abnormal). Each sample now consists of:

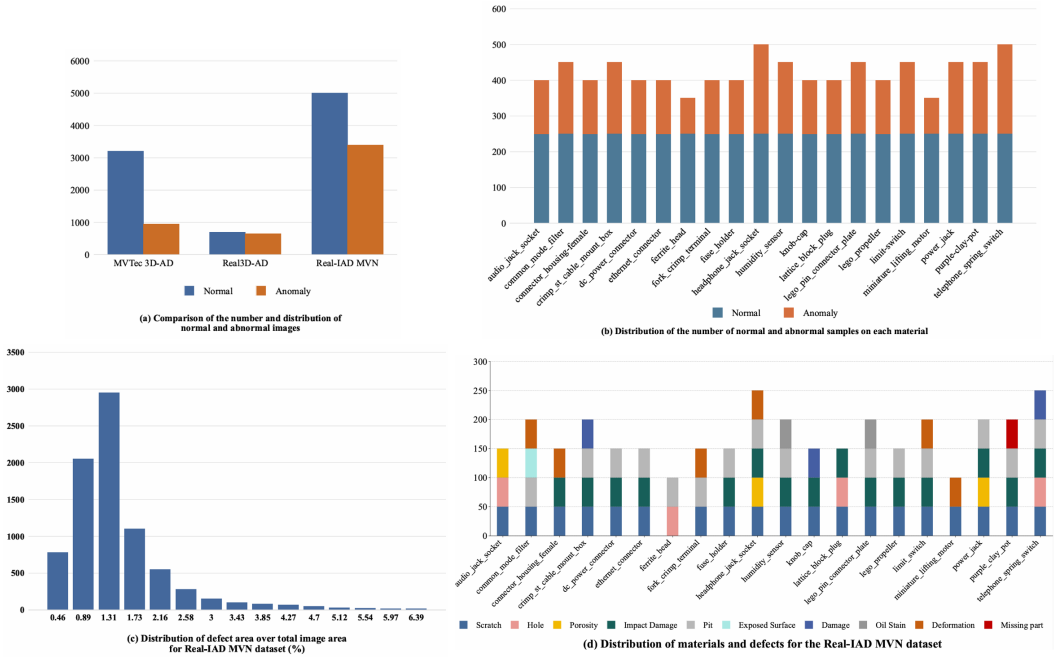

Figure 2. Statistical overview of the Real-IAD MVN dataset in comparison to MVTec 3D-AD and Real3D-AD, illustrating sample counts, normal/anomalous distribution per product, defect area ratios, and defect distribution patterns.

- 5 RGB Images $\{I_v\}_{v=1}^5$
- 5 Surface Normal Maps $\{N_v\}_{v=1}^5$
- Pixel-level anomaly masks $\{M_v\}_{v=1}^5$ (where available for abnormal samples)

This constitutes a dataset size roughly $5\times$ larger than Real-IAD D3 in terms of image/map count. The critical difference is the replacement of the single top-down normal map and sparse 3D point cloud with five high-fidelity normal maps, enabling comprehensive geometric analysis.

Table 1 compares Real-IAD-MVN with prior multi-modal IAD datasets. Our key innovations are the **multi-view capture** and the reliance on **dense surface normals** as the primary geometric modality, offering superior resolution and coverage compared to point clouds or single-view systems.

4. Methodology: A Unified Prototype Baseline

To validate the descriptive power of our new Real-IAD-MVN dataset, we establish a strong baseline model, the **Cross-modal Prototype Reconstruction Network (CPRN)**. This model is designed to leverage the multi-modal (RGB and Normal Vector) data from a single viewpoint, as our dataset's primary comparison is against single-view 2D+3D methods which require strict alignment. The core hypothesis, inspired by recent work on intrinsic prototype learning [8, 13], is that a robust model of normality must capture the *joint probability distribution* $P(\text{appearance}, \text{geometry})$.

Our baseline is a reconstruction-based network, illustrated in Figure 3. It learns a set of Unified Cross-Modal Prototypes (UCMPs) that represent holistic, normal "concepts" (e.g., "this specific shiny, flat surface"). Anomalies are detected as features that cannot be well-reconstructed from this learned dictionary of normality. The network consists of three main stages: (1) Multi-Modal Feature Extraction, (2) Unified Cross-Modal Prototype Extraction (UCP), and (3) Cross-Modal Prototype-Guided Reconstruction.

4.1. Multi-Modal Feature Extraction

Given an input RGB image $I_{rgb} \in \mathbb{R}^{H\times W\times 3}$ and its corresponding surface normal (NV) map $I_{nv} \in \mathbb{R}^{H\times W\times 3}$, we first extract multi-level features. We use two parallel, pre-trained Vision Transformer (ViT) encoders, E_{rgb} and E_{nv}, with frozen weights [4, 14]. The encoders produce multi-level feature sequences:

$$E_{rgb} = \{E_{rgb,l}\}_{l=1}^{L} \quad \text{and} \quad E_{nv} = \{E_{nv,l}\}_{l=1}^{L}$$

where l is the layer index, N is the number of patch tokens, and D is the feature dimension. We denote the final-layer features as $F_{rgb} = E_{rgb,L}$ and $F_{nv} = E_{nv,L}$.

4.2. Unified Cross-Modal Prototype Extraction (UCP)

The UCP module creates a single, compact set of M prototypes representing the joint distribution. We initialize M learnable tokens $Q_{learn} \in \mathbb{R}^{M\times D}$, which act as queries. As

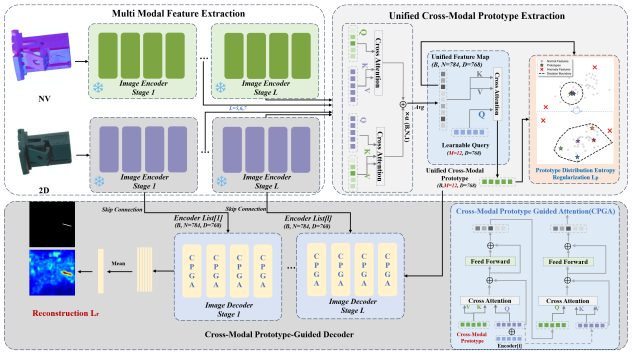

Figure 3. Overview of our CPRN baseline architecture. Multi-modal features (NV and 2D) are extracted by frozen encoders. The UCP module learns a set of unified prototypes P_{ucmp} by attending to both feature streams. The decoder then reconstructs the 2D features using P_{ucmp} as a shared dictionary, guided by skip connections and CPGA blocks. Anomalies are detected via high reconstruction loss (L_r), and the prototype space is regularized by an entropy loss (L_p).

shown in Figure 3 (top-right), these queries attend to both feature streams:

$$P'_{rgb} = \text{Attention}(Q_{learn}, F_{rgb}, F_{rgb}) \tag{1}$$

$$P'_{nv} = \text{Attention}(Q_{learn}, F_{nv}, F_{nv}) \tag{2}$$

The aggregated information is averaged and refined via an FFN to produce the final Unified Cross-Modal Prototypes $P_{ucmp} \in \mathbb{R}^{M \times D}$:

$$P_{ucmp} = \text{FFN}\left(\text{LayerNorm}\left(Q_{learn} + \frac{1}{2}(P'_{rgb} + P'_{nv})\right)\right) \tag{3}$$

This set P_{ucmp} serves as a unified dictionary of all valid appearance-geometry concepts in the normal sample.

4.3. Cross-Modal Prototype-Guided Reconstruction

The final stage uses the learned prototypes P_{ucmp} to reconstruct the original RGB features, forcing the reconstruction to be consistent with the joint prototype concepts. The decoder consists of L stages, mirroring the encoder.

Each decoder stage l employs a Cross-Modal Prototype Guided Attention (CPGA) block. As shown in Figure 3 (bottom-right), this block takes the corresponding skip-connection feature $E_{rgb,l}$ as the Query (Q). The Key (K) and Value (V) are the *same* set of Unified Cross-Modal Prototypes P_{ucmp}.

$$F'_{dec,l} = \text{Attention}(E_{rgb,l}, P_{ucmp}, P_{ucmp}) \tag{4}$$

The output $F'_{dec,l}$ is the "reconstructed" version of $E_{rgb,l}$ using only the normal concepts available in P_{ucmp}. This output is then refined with a residual connection and an FFN. If an input contains an anomaly (e.g., normal color but anomalous geometry), the P_{ucmp} set will be "confused" and will not contain the correct concepts to reconstruct the RGB features, leading to a high reconstruction error.

4.4. Optimization Objective

The network is trained end-to-end with a composite loss function: $L_{total} = L_r + \lambda_p L_p$, where λ_p is a balancing hyperparameter.

Reconstruction Loss (L_r). This primary loss measures the fidelity of the reconstruction. As shown in Figure 3, we compute the mean of all skip connections from the RGB encoder ($F_{rgb,\text{mean}}$) and the mean of all outputs from the decoder ($\tilde{F}_{rgb,\text{mean}}$). The loss is the mean cosine distance between them:

$$L_r = 1 - \text{sim}(F_{rgb,\text{mean}}, \tilde{F}_{rgb,\text{mean}}) \tag{5}$$

where $\text{sim}(\cdot)$ denotes cosine similarity.

Prototype Distribution Entropy Regularization (L_p). This regularization loss, also depicted in the UCP module diagram, prevents "prototype collapse" by encouraging the model to utilize all M prototypes equally. It computes the negative entropy of the prototype assignment distribution q, thereby maximizing diversity:

$$L_p = \sum_{j=1}^{M} q_j \log(q_j + \epsilon) \tag{6}$$

This ensures the learned prototype space is well-structured.

Inference. At test time, an image-level anomaly score is computed from L_r. A pixel-level anomaly map is generated by computing the patch-wise cosine distance between the final layer encoder features F_{rgb} and their corresponding reconstructions from the final decoder layer $\tilde{F}_{rgb,L}$.

5. Experiments

5.1. Experimental Setup

Datasets and Metrics. Our primary experiments are conducted on our newly proposed **Real-IAD-MVN** dataset. We also perform a validation experiment on the public Real-IAD (MV-RGB) benchmark [16] to confirm our baseline's architectural robustness. We evaluate all methods using standard IAD metrics: Image-level AUROC (I-AUROC), Image-level F1-Score (I-F1), Pixel-level AUROC (P-AUROC), and Pixel-level AUPRO (P-AUPRO).

Experimental Settings and Baselines. We validate our contributions with two key experiments on our **Real-IAD-MVN** dataset: (1) A **Single-View (Top-Down)** comparison to fairly test our PS-3D (RGB+Normal) modality against traditional 2D+3D (RGB+Point Cloud) [22]. (2) A **Full**

Table 2. Performance comparison of different multimodal anomaly detection methods on the **Real-IAD-MVN (Ours)** dataset. **Crucially, this table evaluates a single-view (top-down) scenario** to ensure a fair comparison with traditional 2D+3D methods that require strict alignment. The table is grouped by modality: Single Modality (RGB), traditional multimodal (2D+3D), and our proposed modality (RGB+NV). Our baseline, CPRN achieves new state-of-the-art results against all single-view competitors.

Modality	**RGB**						**2D+3D**									**NV+RGB**		
Model	**SimpleNet**			**Cflow**			**AST**			**PatchCore**			**M3DM**			**CPRN (Ours)**		
Metrics	**I-Auroc**	**P-AUROC**	**P-AUPRO**	**I-Auroc**	**P-AUROC**	**P-AUPRO**	**I-Auroc**	**P-AUROC**	**P-AUPRO**	**I-Auroc**	**P-AUROC**	**P-AUPRO**	**I-Auroc**	**P-AUROC**	**P-AUPRO**	**I-Auroc**	**P-AUROC**	**P-AUPRO**
audio_jack_socket	0.973	0.926	0.710	0.943	0.944	0.725	0.860	0.590	0.410	0.926	0.673	0.505	0.981	0.699	0.550	**0.986**	**0.979**	**0.878**
common_mode_filter	0.717	0.822	0.600	0.271	0.847	0.615	0.899	0.802	0.590	0.523	0.922	0.790	0.580	0.934	0.820	**0.945**	**0.942**	**0.741**
connector_housing_female	0.795	0.891	0.690	0.839	0.921	0.705	0.914	0.716	0.510	0.870	0.919	0.785	0.920	0.979	0.900	**0.997**	**0.972**	**0.791**
crimp_st_cable_mount_box	0.372	0.745	0.510	0.180	0.442	0.300	0.485	0.589	0.390	0.713	0.931	0.810	0.749	0.933	0.830	**0.995**	**0.959**	**0.771**
dc_power_connector	0.661	0.725	0.500	0.661	0.726	0.500	0.995	0.770	0.550	0.720	0.921	0.790	0.715	0.950	0.850	**0.994**	**0.901**	**0.702**
ethernet_connector	0.981	0.866	0.650	0.967	0.853	0.640	1.000	0.906	0.710	0.947	0.956	0.840	0.983	0.978	0.910	**0.992**	**0.989**	**0.958**
ferrite_bead	0.408	0.806	0.590	0.529	0.914	0.690	0.894	0.817	0.600	0.913	0.932	0.815	0.965	0.966	0.880	**0.990**	**0.988**	**0.946**
fork_crimp_terminal	0.416	0.945	0.720	0.462	0.657	0.450	0.595	0.773	0.560	0.769	0.952	0.840	0.780	0.964	0.870	**0.950**	**0.981**	**0.825**
fuse_holder	0.564	0.957	0.750	0.853	0.861	0.640	0.597	0.754	0.540	0.736	0.927	0.800	0.770	0.948	0.850	**0.996**	**0.994**	**0.980**
headphone_jack_socket	0.933	0.879	0.670	0.996	0.914	0.700	0.660	0.696	0.490	0.919	0.942	0.830	0.982	0.982	0.920	**0.997**	**0.969**	**0.857**
humidity_sensor	0.737	0.890	0.680	0.781	0.836	0.610	0.565	0.723	0.510	0.689	0.933	0.810	0.717	0.958	0.860	**0.979**	**0.970**	**0.819**
knob_cap	0.672	0.879	0.670	0.637	0.893	0.680	0.919	0.656	0.450	0.903	0.958	0.850	0.925	0.938	0.830	**0.966**	**0.954**	**0.875**
lattice_block_plug	0.790	0.898	0.690	0.833	0.852	0.630	0.842	0.919	0.730	0.911	0.923	0.800	0.917	0.958	0.860	**0.961**	**0.976**	**0.926**
lego_pin_connector_plate	0.857	0.947	0.730	0.828	0.877	0.660	0.847	0.629	0.420	0.662	0.759	0.580	0.681	0.734	0.580	**0.950**	**0.992**	**0.945**
lego_propeller	0.939	0.799	0.580	0.615	0.739	0.510	0.471	0.703	0.500	0.540	0.727	0.560	0.530	0.773	0.610	**0.998**	**0.952**	**0.806**
limit_switch	0.823	0.790	0.570	0.846	0.950	0.740	0.804	0.641	0.430	0.822	0.938	0.820	0.863	0.966	0.890	**0.979**	**0.969**	**0.864**
miniature_lifting_motor	0.402	0.760	0.530	0.402	0.799	0.570	0.766	0.467	0.310	0.948	0.962	0.870	0.975	0.991	0.940	**0.923**	**0.967**	**0.703**
power_jack	0.176	0.489	0.310	0.354	0.664	0.450	0.564	0.645	0.440	0.981	0.923	0.800	0.996	0.902	0.780	**0.992**	**0.937**	**0.662**
purple_clay_pot	0.343	0.938	0.720	0.343	0.571	0.390	0.635	0.445	0.290	0.921	0.961	0.870	0.944	0.953	0.860	**0.896**	**0.981**	**0.951**
telephone_spring_switch	0.627	0.916	0.700	0.575	0.910	0.690	0.951	0.551	0.380	0.827	0.944	0.830	0.856	0.936	0.830	**0.975**	**0.976**	**0.905**
Avg	0.659	0.843	0.629	0.645	0.808	0.595	0.776	0.699	0.521	0.812	0.905	0.790	0.841	0.922	0.830	**0.973**	**0.968**	**0.841**

Table 3. Performance comparison of RGB+NV multimodal methods on the **full Real-IAD-MVN dataset (all 5 views)**. This experiment highlights the performance of different fusion strategies when leveraging the complete multi-view RGB and multi-view NV data. Our baseline, CPRN, sets a new state-of-the-art on this full dataset.

Model	**PatchCore**			**M3DM**			**D3M**			**Dinomaly**			**InpFormer**			**CPRN (Ours)**		
Category	**I-AUROC**	**P-AUROC**	**P-AUPRO**	**I-AUROC**	**P-AUROC**	**P-AUPRO**	**I-AUROC**	**P-AUROC**	**P-AUPRO**	**I-AUROC**	**P-AUROC**	**P-AUPRO**	**I-AUROC**	**P-AUROC**	**P-AUPRO**	**I-AUROC**	**P-AUROC**	**P-AUPRO**
dc_power_connector	0.932	0.991	0.488	0.920	0.990	0.481	0.929	0.991	0.487	0.935	0.992	0.491	0.938	0.993	0.495	**0.9965**	**0.9044**	**0.7117**
fork_crimp_terminal	0.843	0.963	0.769	0.835	0.961	0.758	0.832	0.962	0.765	0.838	0.965	0.771	0.841	0.966	0.775	**0.9536**	**0.9845**	**0.8361**
ethernet_connector	0.969	0.994	0.582	0.962	0.992	0.575	0.967	0.994	0.580	0.972	0.995	0.588	0.974	0.996	0.591	**0.9954**	**0.9922**	**0.9684**
fuse_holder	0.846	0.981	0.574	0.900	0.980	0.565	0.919	0.982	0.551	0.925	0.983	0.559	0.928	0.984	0.563	**0.9983**	**0.9972**	**0.9901**
common_mode_filter	0.915	0.988	0.728	0.902	0.986	0.715	0.949	0.993	0.743	0.940	0.990	0.735	**0.9500**	**0.994**	0.748	0.9479	0.9455	**0.7511**
crimp_st_cable_mount_box	0.985	0.998	0.570	0.980	0.997	0.562	0.985	0.998	0.594	0.988	0.999	0.601	0.990	0.999	0.605	**0.9980**	**0.9629**	**0.7804**
connector_housing_female	0.995	0.999	0.610	0.993	0.998	0.595	0.999	0.999	0.602	0.999	0.999	0.605	**1.000**	**1.000**	0.608	**1.0000**	0.9752	**0.8005**
lego_pin_connector_plate	0.689	0.948	0.809	0.680	0.945	0.795	0.717	0.952	0.812	0.725	0.955	0.818	0.730	0.956	0.821	**0.9526**	**0.9947**	**0.9564**
knob_cap	0.907	0.977	0.609	0.898	0.975	0.601	0.921	0.981	0.617	0.925	0.982	0.622	0.928	0.983	0.625	**0.9686**	**0.9572**	**0.8861**
lattice_block_plug	0.695	0.956	0.578	0.681	0.952	0.565	0.759	0.964	0.587	0.768	0.966	0.593	0.772	0.968	0.598	**0.9640**	**0.9797**	**0.9364**
humidity_sensor	0.844	0.973	0.617	0.835	0.971	0.609	0.855	0.975	0.619	0.860	0.976	0.624	0.863	0.977	0.627	**0.9819**	**0.9732**	**0.8298**
ferrite_bead	0.907	0.974	0.580	0.899	0.972	0.571	0.930	0.981	0.577	0.935	0.982	0.583	0.938	0.983	0.587	**0.9922**	**0.9911**	**0.9560**
power_jack	0.947	0.993	0.526	0.941	0.991	0.518	0.960	0.994	0.545	0.963	0.995	0.551	0.965	0.995	0.554	**0.9951**	**0.9406**	**0.6729**
telephone_spring_switch	0.781	0.969	0.672	0.772	0.965	0.665	0.834	0.977	0.692	0.840	0.979	0.698	0.843	0.980	0.702	**0.9783**	**0.9794**	**0.9162**
miniature_lifting_motor	0.856	0.937	0.604	0.847	0.933	0.595	0.850	0.927	0.638	**0.9300**	0.931	0.645	0.928	**0.972**	0.710	0.9262	0.9706	**0.7141**
purple_clay_pot	0.833	0.966	0.686	0.825	0.964	0.678	0.838	0.969	0.707	0.843	0.971	0.713	0.845	0.972	0.716	**0.8993**	**0.9843**	**0.9617**
headphone_jack_socket	0.970	0.997	0.555	0.963	0.995	0.546	0.987	0.997	0.531	0.990	0.998	0.535	0.992	0.998	0.538	**0.9991**	**0.9722**	**0.8671**
lego_propeller	0.993	0.999	0.562	0.990	0.998	0.553	**1.000**	**1.000**	0.572	0.998	**1.000**	0.575	0.999	**1.000**	0.578	0.9999	0.9557	**0.8171**
limit_switch	0.878	0.980	0.635	0.869	0.978	0.626	0.914	0.986	0.619	0.918	0.987	0.623	0.920	0.988	0.626	**0.9824**	**0.9728**	**0.8748**
audio_jack_socket	0.922	0.992	0.570	0.913	0.990	0.562	0.931	0.993	0.571	0.935	0.994	0.576	0.938	0.995	0.580	**0.9889**	**0.9815**	**0.8897**
Mean	0.885	0.978	0.616	0.878	0.975	0.608	0.904	0.981	0.620	0.913	0.983	0.630	0.916	0.984	0.633	**0.9759**	**0.9707**	**0.8558**

Dataset (5-View) comparison using our complete RGB + Normal data to test SOTA 2D+PS fusion methods. We compare against a comprehensive suite of baselines, including single-modality (SimpleNet [12], Cflow[6]), 2D+3D (AST, PatchCore [14], M3DM [18]), and 2D+PS methods (D3M [22], Dinomaly [7], INP-Former [13]). We also validate our baseline on the Real-IAD (MV-RGB) benchmark [16].

5.2. Modality Comparison: RGB-NV vs. RGB-3D Point Cloud

Our first experiment (Table 2) investigates the most effective geometric modality in a single-view (top-down) setting for fair alignment. The results are conclusive: 2D+3D methods (e.g., M3DM at 0.841 I-Auroc, 0.830 P-AUPRO) improve over RGB-only (SimpleNet at 0.659 I-Auroc, 0.629 P-AUPRO). However, our CPRN baseline using RGB+NV data achieves a massive leap in performance (0.973 I-Auroc, 0.841 P-AUPRO). This demonstrates that for industrial inspection, the dense, high-fidelity geometric detail from photometric stereo (PS) normals is a far superior data source than sparse 3D point clouds, especially for localizing fine-grained defects (P-AUPRO).

5.3. Full Multi-View Dataset Evaluation

Our second experiment (Table 3) evaluates the performance of different single-view methods when applied across the full 5-view dataset. All metrics are computed by treating each view from all samples as an independent image. Here, our CPRN baseline again sets a new state-of-the-art (0.9759 I-Auroc, 0.8558 P-AUPRO), significantly outperforming strong recent competitors like INP-Former [13] (0.916 I-Auroc) and Dinomaly [7] (0.913 I-Auroc). While the dataset remains challenging, with other methods showing competitive performance on specific categories (e.g., INP-Former [13] on 'common_mode_filter' and D3M [22] on 'lego_propeller'), our strong average performance validates the robustness of our unified prototype fusion mecha-

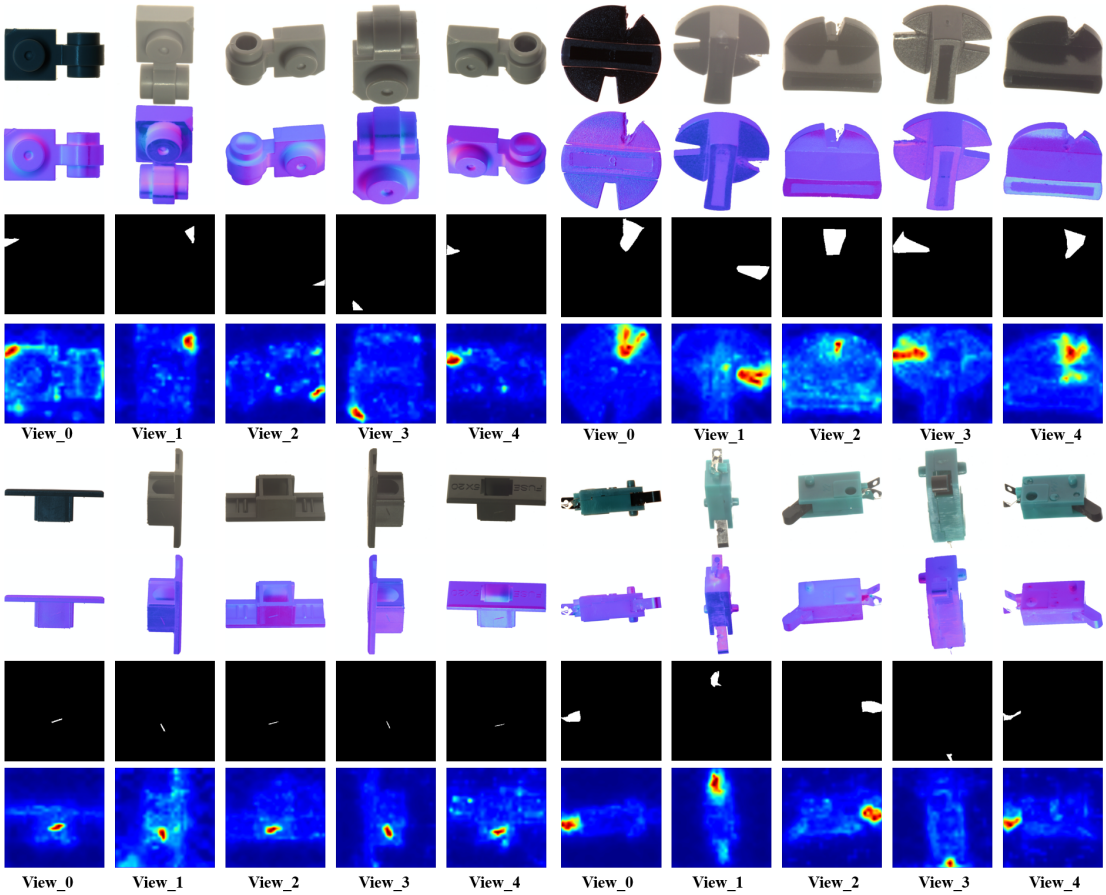

Figure 4. Qualitative results of our single-view CPRN model applied independently to multiple views from the Real-IAD-MVN dataset. For two categories. For two categories (top: 'lego_pin_connector_plate', bottom: 'limit_switch'), we show the RGB image, corresponding PS normal map, ground truth mask, and our CPRN's predicted anomaly map from multiple viewpoints. Note how defects invisible in one view (e.g., top-down) become clearly visible from a side view, demonstrating the necessity of our multi-view normal dataset.

nism for this new, challenging modality.

5.4. Qualitative and Visual Analysis

Figure 4 illustrates the core advantage of our Real-IAD-MVN dataset: defects on object sides (e.g., 'lego_pin_connector_plate' and 'limit_switch') are invisible in the top-down view but are clearly revealed in the side-view normal maps, eliminating critical blind spots. Furthermore, Figure 5 provides a qualitative comparison on the top-down view, showing our CPRN baseline produces cleaner, more precisely localized heatmaps than SOTA methods like INP-Former [13] and M3DM [18], which exhibit more diffuse, false-positive activations. This suggests our unified prototype approach learns a more constrained and accurate representation of normality.

5.5. Ablation Study

We validate our CPRN baseline's design on the full 5-view dataset in Table 4. The study shows a clear progression: the 'NV only' (multi-view normal) reconstruction (0.7952 P-AUPRO) performs notably better than 'RGB only' (0.641 P-AUPRO), confirming the geometric stream's power. A 'Naive Fusion (Concat)' baseline (0.8104 P-AUPRO) improves upon single modalities, proving fusion is beneficial. The major performance leap comes from our 'UCP (Ours)' module (0.8420 P-AUPRO), which replaces naive concatenation, demonstrating the superiority of learning unified prototypes. Finally, adding the Prototype Distribution Entropy Regularization (L_p) provides the last significant boost to our final performance (0.8558 P-AUPRO), confirming its importance in preventing prototype collapse and ensuring a diverse set of normal concepts.

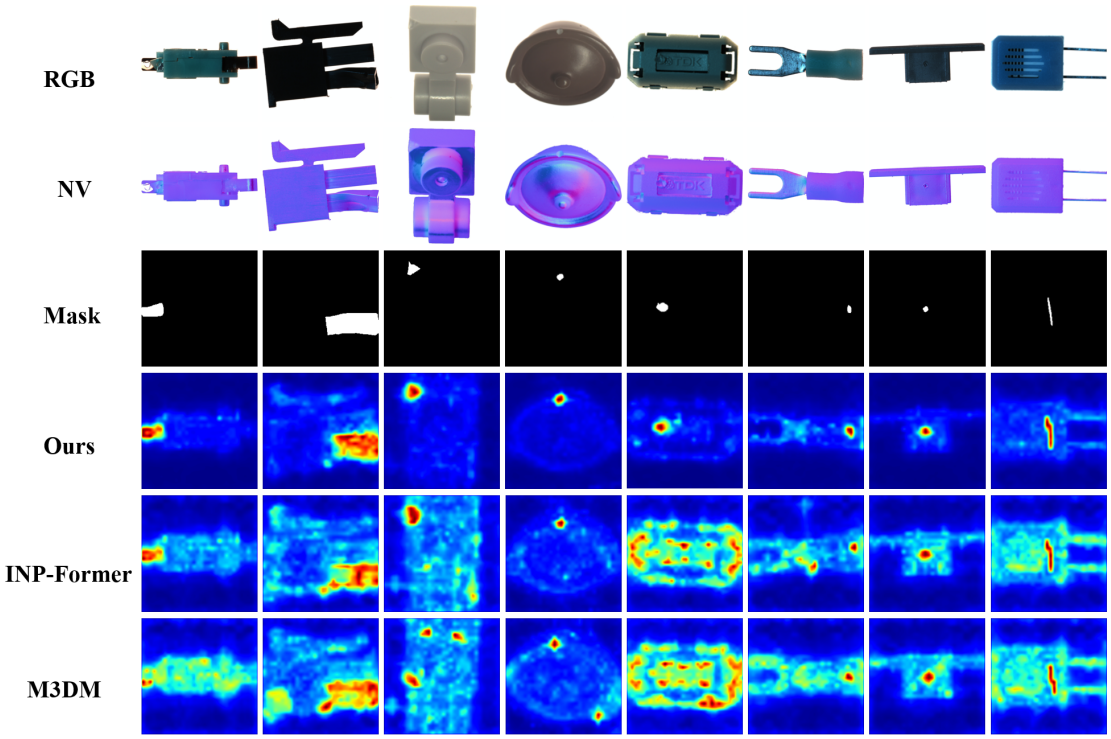

Figure 5. Qualitative comparison with state-of-the-art methods on various categories from the Real-IAD-MVN dataset (top-down view). From top to bottom: Input RGB, input normal map, ground truth mask, our CPRN (Ours) heatmap, INP-Former [13] heatmap, and M3DM [18] heatmap. Our method consistently produces more accurate and tightly localized anomaly maps with fewer false positives.

Table 4. Ablation study of our CPRN baseline. Performance is evaluated on the full Real-IAD-MVN dataset, with our single-view model applied to all 5 views and results aggregated.

Configuration	RGB Stream	NV Stream	Fusion Strategy	I-Auroc	I-F1	P-AUROC	P-AUPRO
Baseline (RGB only)	✓		None (Reconstruction)	0.9152	0.9413	0.9421	0.6410
Baseline (NV only)		✓	None (Reconstruction)	0.9215	0.9458	0.9503	0.6952
Naive Fusion (Concat)	✓	✓	Concatenation	0.9521	0.9614	0.9550	0.8104
Ours (w/o L_p)	✓	✓	UCP (Ours)	0.9712	0.9805	0.9681	0.8420
CPRN (Full Model)	✓	✓	**UCP + L_p (Ours)**	**0.9759**	**0.9833**	**0.9707**	**0.8558**

6. Conclusion

In this paper, we introduced Real-IAD-MVN, a new large-scale industrial anomaly detection dataset. Its primary contribution is the introduction of high-fidelity, dense surface normal maps captured from five distinct viewpoints, replacing traditional sparse 3D point cloud data. We demonstrated that this new multi-view normal (PS-3D) modality is a far superior data source for detecting subtle, geometric defects compared to both RGB-only and RGB + 3D point cloud approaches. To validate this new dataset, we proposed a strong baseline, the Cross-modal Prototype Reconstruction Network (CPRN). This model leverages the concept of unified prototypes to effectively leverages the joint distribution of normality.

Acknowledgements

This research was supported by the Shanghai Agricultural Science and Technology Project (grant number T20252016), and the Shanghai Science and Technology Project (grant number 24YF2716900) and Suzhou Major Project ("Jiebang Guashuai") for Transformation of Scientific and Technological Achievements (Grant No. SZC2024020).

References

[1] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018. 2

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 1

[3] Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. *arXiv preprint arXiv:2112.09045*, 2021. 1, 2, 3

[4] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 2, 4

[5] Bin-Bin Gao. Learning to detect multi-class anomalies with just one normal image prompt. In *European Conference on Computer Vision*, pages 454–470. Springer, 2024. 2

[6] Denis Gudovskiy et al. Cflow-ad: Real-time unsupervised anomaly detection with compact neural flow. *IEEE Transactions on Artificial Intelligence*, page 2022, 2022. 6

[7] Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Hongen Liao, and Huiqi Li. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2405.14325*, 2024. 6

[8] Jaemin Lee, Dong-ju Lee, Jihun Lee, Hyunjin Song, Jongheon Kim, Sung-eui Park, and Chan-hyung Choi. Inpformer: A new paradigm for training-free anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 562–572, 2024. 4

[9] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on CVPR*, pages 9664–9674, 2021. 2

[10] Hong Li, Houyuan Chen, Chongjie Ye, Zhaoxi Chen, Bohan Li, Shaocong Xu, Xianda Guo, Xuhui Liu, Yikai Wang, Baochang Zhang, et al. Light of normals: Unified feature representation for universal photometric stereo. *arXiv preprint arXiv:2506.18882*, 2025. 1, 2, 3

[11] Jiaqi Liu, Guoyang Xie, Ruitao Chen, Xinpeng Li, Jinbao Wang, Yong Liu, Chengjie Wang, and Feng Zheng. Real3d-ad: A dataset of point cloud anomaly detection. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3

[12] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on CVPR*, pages 20402–20411, 2023. 2, 6

[13] Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9974–9983, 2025. 4, 6, 7, 8

[14] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2022. 2, 4, 6

[15] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017. 2

[16] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. 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*, pages 22883–22892, 2024. 5, 6

[17] Chengjie Wang, Haokun Zhu, Jinlong Peng, Yue Wang, Ran Yi, Yunsheng Wu, Lizhuang Ma, and Jiangning Zhang. M3dm-nr: Rgb-3d noisy-resistant industrial anomaly detection via multimodal denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[18] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8032–8041, 2023. 2, 6, 7, 8

[19] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980. 3

[20] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022. 2

[21] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8330–8339, 2021. 2

[22] Wenbing Zhu, Lidong Wang, Ziqing Zhou, Chengjie Wang, Yurui Pan, Ruoyi Zhang, Zhuhao Chen, Linjie Cheng, Bin-Bin Gao, Jiangning Zhang, et al. Real-iad d3: A real-world 2d/pseudo-3d/3d dataset for industrial anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15214–15223, 2025. 1, 2, 3, 5, 6

[23] Zongyi Zou, Qiang Qiu, and Weiming Shen. Spot-the-difference: A novel benchmark for image anomaly detection in industrial inspection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2608–2616, 2022. 1

A. Generative Potential of Real-IAD-MVN

To showcase the richness of the learned representation in our new dataset and to explore future research directions, we present an **Awareness-Guided Generative Module**. This module is designed to validate the deep, intrinsic understanding of 3D structure that can be learned from our high-fidelity 2D (RGB) and Pseudo-3D (Normal Vector) data.

Its purpose is to generate a high-quality geometric representation (e.g., a single-channel depth image) using only the RGB and normal vector (NV) inputs provided by our Real-IAD-MVN dataset. The success of this complex cross-modal generation task serves as compelling evidence that our new data modality (MV-RGB + MV-NV) captures a profound understanding of object structure, potentially enabling hardware-agnostic applications where expensive 3D scanners can be replaced.

Concretely, the module follows a two-stage pipeline conditioned by a hierarchical fusion mechanism. First, RGB and NV features are encoded and fused by a **Hierarchical Dual Conditioner (HDC)**. The HDC performs multi-level cross-modal interaction and injects the resulting conditional features into a **Denoising Unet**, which predicts a latent geometric representation through a conditional diffusion process. Second, the denoised latent code is forwarded to an **Adversarial Dense Generation** decoder that reconstructs a high-fidelity depth image. The overall design therefore combines cross-modal prototype-aware conditioning, diffusion-based latent generation, and adversarial dense decoding in a unified generation framework.

A.1. Two-Stage Generation for High-Fidelity Depth

We first generate a latent geometric representation using diffusion, then decode it into a dense depth image using an adversarial network.

1. Diffusion-based Latent Generation. We adopt a conditional denoising diffusion probabilistic model (DDPM) to generate a latent representation of the geometry. A forward process gradually adds Gaussian noise to a ground truth latent vector, and a **Denoising Unet** is trained to reverse this process. The reverse process is conditioned on the fused features from the RGB image (I) and the pseudo-3D normal vectors (P):

$$p_\theta(x_{t-1}|x_t, I, P) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, I, P), \sigma_t^2 \mathbf{I}) \quad \text{(A1)}$$

The Unet, ϵ_θ, is trained to predict the noise added at each timestep t:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, I, P)\|_2^2\right] \quad \text{(A2)}$$

2. Adversarial Dense Generation. The denoised latent representation from the Unet is passed to a final **Decoder** network, which generates the high-resolution, single-channel depth image $\hat{I}_{\text{depth}}$. This stage is trained with a combination of a reconstruction loss and an adversarial loss to ensure high fidelity and realism:

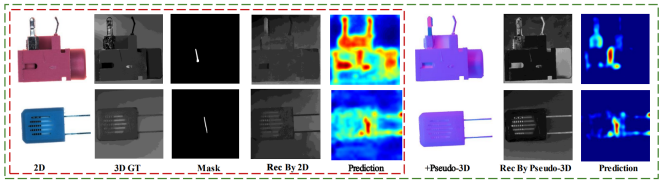

Figure A1. Qualitative comparison of anomaly segmentation from the generative module's output on the Real-IAD D³ dataset. The right panel (+Pseudo-3D) shows that generating geometry from normal vectors (Rec By Pseudo-3D) leads to a sharper, more accurate anomaly prediction compared to the 2D-only baseline (left panel).

$$\mathcal{L}_{\text{reconstruct}} = \mathbb{E}\left[\|\hat{I}_{\text{depth}} - I_{\text{depth_GT}}\|_2^2\right] \quad \text{(A3)}$$

$$\mathcal{L}_{\text{adv}} = \mathbb{E}\left[\log(1 - D(\hat{I}_{\text{depth}}))\right] + \mathbb{E}\left[\log D(I_{\text{depth_GT}})\right] \quad \text{(A4)}$$

where D is a discriminator network distinguishing generated images from ground truth (GT) depth images.

A.2. Hierarchical Conditioning for Input Fusion

The conditioning of the Denoising Unet is orchestrated by a **Hierarchical Dual Conditioner (HDC)**, which effectively fuses the image and normal vector modalities. It uses resolution-aware information routing, extracting features $F_{\mathcal{M}}^l$ at L hierarchical levels. The core is a Bi-Modal Cross-Attention (BCA) module that uses features from the Unified Cross-Modal Prototype Extraction (UCP) module (detailed in our main paper's baseline) as a rich feature source.

$$\text{CA}_{\mathcal{M}}(Q, \tilde{F}_{\mathcal{M}}^l) = \text{Softmax}\left(\frac{QW_Q^{\mathcal{M}}(W_K^{\mathcal{M}}\tilde{F}_{\mathcal{M}}^l)^T}{\sqrt{d_k}}\right) W_V^{\mathcal{M}}\tilde{F}_{\mathcal{M}}^l \quad \text{(A5)}$$

To dynamically balance the modalities, a weighting coefficient α_l is computed for each level l, based on the estimated mutual information $\mathcal{I}$ that each modality's features provide with respect to the Unet's query features Q:

$$\alpha_l = \sigma\left(\beta_l \cdot \frac{\mathcal{I}(Q; \tilde{F}_P^l)}{\mathcal{I}(Q; \tilde{F}_P^l) + \mathcal{I}(Q; \tilde{F}_I^l)}\right) \quad \text{(A6)}$$

The final conditional features injected into the Unet are a weighted sum of the attention outputs, ensuring that the most relevant information from each modality (RGB and NV) is used at the most appropriate stage of generation.

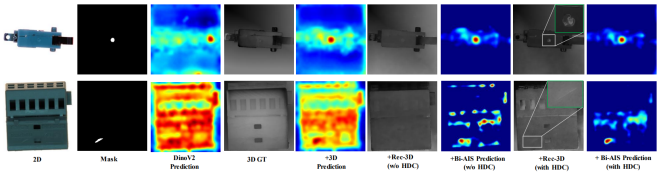

Figure A2. Qualitative results of the generative module's ablation study on Real-IAD D³. The baseline 2D prediction (DinoV2) is diffuse. Adding 3D reconstruction (+Rec-3D) improves focus. The full model (+Bi-AIS Prediction (with HDC)) demonstrates the sharpest and most accurate localization by effectively fusing the RGB and normal vector data.

A.3. Qualitative Analysis of Generative Module

While the quantitative results of this generative model are reserved for future work, we provide qualitative results here to demonstrate its effectiveness, which further validates the quality of our Real-IAD-MVN data.

The qualitative results in Figure A1 (conducted on the similar Real-IAD D³ dataset) visually demonstrate the principle of defect enhancement. The left panel shows the prediction from a 2D-only baseline, which produces a diffuse and inaccurate anomaly map. In contrast, the right panel shows that after generating a geometric representation guided by pseudo-3D information (normal vectors), the model's reconstruction is sharper, and the final prediction is precisely focused on the true anomalous regions (e.g., the bent pins), providing a much clearer signal for detection.

The progressive improvement in localization accuracy is also visualized in the ablation study in Figure A2. The baseline model's (DinoV2) heatmap is noisy and unfocused. Adding the reconstruction module with pseudo-3D data (+Rec-3D) begins to consolidate the activation. Finally, the full model with the Hierarchical Dual Conditioner (+Bi-AIS Prediction (with HDC)) produces a remarkably clean and accurate anomaly map, demonstrating the critical role of sophisticated fusion of the RGB and normal vector data.

B. Real-IAD-MVN Dataset Visualizations

To provide a comprehensive overview of the data included in Real-IAD-MVN, Figure A3 visualizes all 20 object categories. For each category, we display the anomaly-free top-down view, followed by all five distinct, calibrated viewpoints (Side View 0 through Side View 4). Each viewpoint is represented by its (1) RGB image, (2) high-fidelity PS Normal Vector (NV) map, and (3) the corresponding ground truth anomaly mask (shown in white, if an anomaly is present in that view).

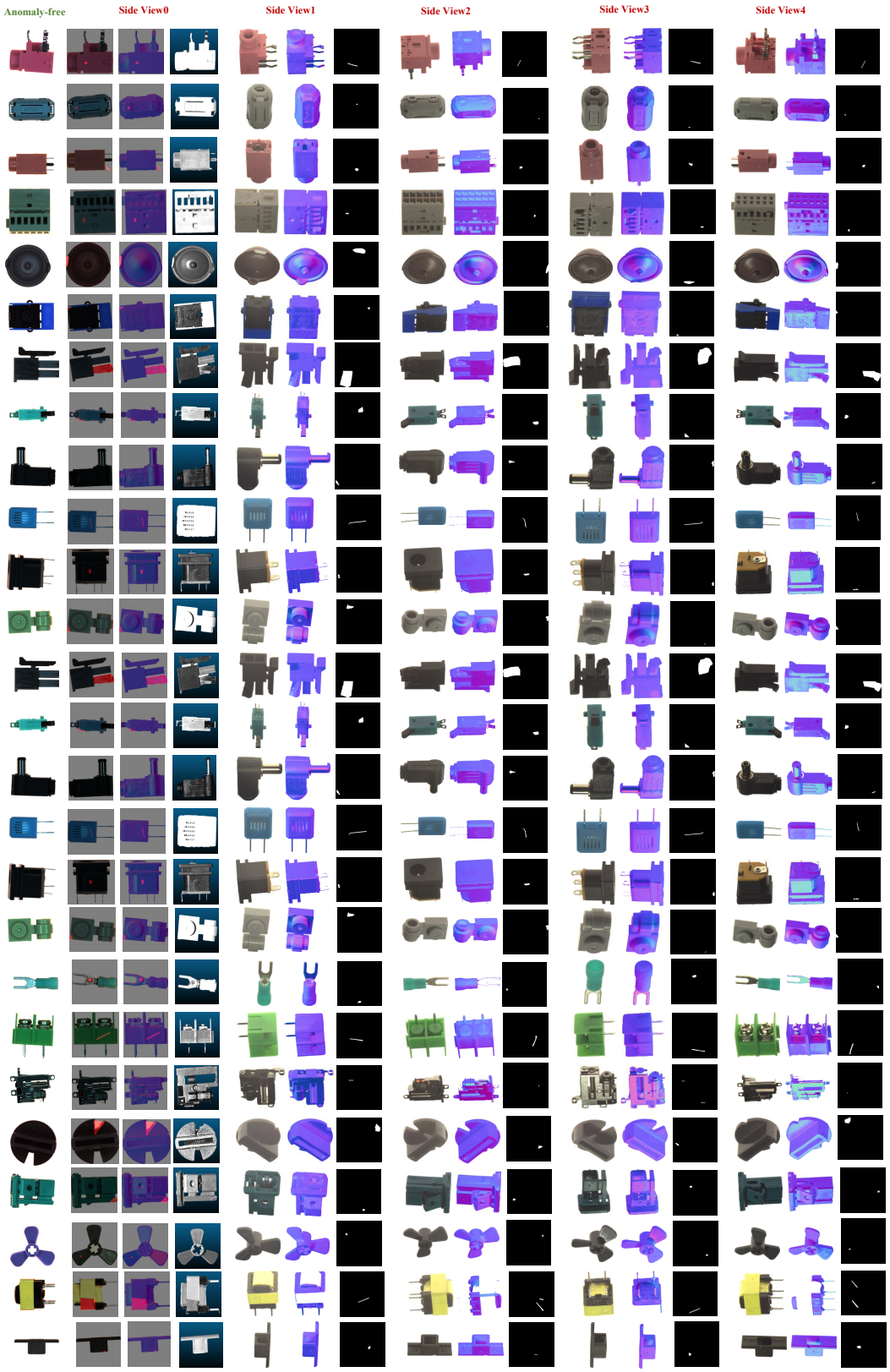

Figure A3. Comprehensive visualization of all 20 categories in the Real-IAD-MVN dataset. Each row corresponds to one object category. The first column shows the anomaly-free top-down RGB image. The subsequent columns (Side View 0-4) display the five captured viewpoints. For each viewpoint, we show the RGB image, its corresponding high-fidelity PS Normal Map (NV), and the pixel-level ground truth anomaly mask.

C. Camera Calibration Parameters

In Section 3.1 of our main paper, we describe our integrated multi-view photometric stereo (MVPS) gantry, which operates with five pre-calibrated viewpoints. To enable and encourage future research in multi-view 3D reconstruction, geometric registration, and novel view synthesis (e.g., NeRF-based tasks), we provide the full calibration parameters for four of our primary acquisition cameras below. The parameters are provided in the '.yml' format, compatible with standard calibration tools.

C.1. Camera 1 Parameters


```
calibration_file_ver: 3.1
calibration_ver: 3.1.359264
calibration_time: 2025-1-24 15:6:57
calibration_type: MVP_CALIB_TYPE_NORMAL
frame_num: 1
image_width: 4096
image_height: 3000
board_width: 0
board_height: 0
board_dx: 2.0000000000000000e+000
board_dy: 2.0000000000000000e+000
points_num: 0
H_matrix:
  rows: 3
  cols: 3
  data:
    - 7.5142908289968972e+001
    - 4.9162084131020807e+000
    - 8.2109891582446537e+001
    - -1.7980084333606616e+000
    - 5.7724585197937522e+001
    - 1.0859333933073324e+003
    - 8.2071064611873192e-005
    - 2.3720123199760068e-003
    - 1.0000000000000000e+000
distortion_model: 3
distortion_level: 2
distortion_matrix:
  rows : 12
  cols : 1
  data:
    - -7.0994615634284253e-001
    - 1.5565761184853105e+001
    - -6.7707702880165407e-002
    - -1.2152307380140845e-002
    - -1.1598420108419958e+002
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
intrinsic_matrix:
  rows : 3
  cols : 3
  data:
    - 1.9426147574516781e+004
    - 0.0000000000000000e+000
    - 1.9428998682433419e+003
    - 0.0000000000000000e+000
    - 2.0040639975377839e+004
    - -1.4103926763176123e+003
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 1.0000000000000000e+000
extrinsic_matrix:
  rows : 3
  cols : 4
  data:
    - 9.9953780819267646e-001
    - 4.1006888450079194e-003
    - -3.0122323024239783e-002
    - -2.4804538528430975e+001
    - -2.1737068720812217e-002
    - 7.8910936074326898e-001
    - -6.1386799609588383e-001
    - 3.2255977192006476e+001
    - 2.1252525421862352e-002
    - 6.1423904234291860e-001
    - 7.8883377781687780e-001
    - 2.5895272008921592e+002
```


C.2. Camera 2 Parameters


```
calibration_file_ver: 3.1
calibration_ver: 3.1.359264
calibration_time: 2025-1-24 15:4:51
calibration_type: MVP_CALIB_TYPE_NORMAL
frame_num: 1
image_width: 4096
image_height: 3000
board_width: 0
board_height: 0
board_dx: 2.0000000000000000e+000
board_dy: 2.0000000000000000e+000
points_num: 0
H_matrix:
  rows: 3
  cols: 3
  data:
    - 7.4755050461534353e+001
    - 6.2458574231518309e+000
    - 1.0981382636031776e+003
    - -1.6285738267351551e+000
```

```
    - 5.7629797788179566e+001
    - 1.7663434952193629e+002
    - -4.8201648078641062e-004
    - 2.8343758580036344e-003
    - 1.0000000000000000e+000
distortion_model: 3
distortion_level: 2
distortion_matrix:
  rows : 12
  cols : 1
  data:
    - -2.4410037940517855e+000
    - 4.3015853395902596e+001
    - 2.1299813348685793e-002
    - 8.8716663242255092e-002
    - -3.5257553640216827e+002
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
intrinsic_matrix:
  rows : 3
  cols : 3
  data:
    - 1.9798511319027260e+004
    - 0.0000000000000000e+000
    - -4.3049246222583417e+002
    - 0.0000000000000000e+000
    - 2.0271741030735127e+004
    - 2.4744169291864846e+003
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 1.0000000000000000e+000
extrinsic_matrix:
  rows : 3
  cols : 4
  data:
    - 9.9188951598251840e-001
    - 9.9339040733826808e-002
    - 7.9290245743398075e-002
    - 2.0339137531916062e+001
    - -5.6639905656809125e-003
    - 6.5775238666472491e-001
    - -7.5321292942150830e-001
    - -2.9859351241362621e+001
    - -1.2697679825400085e-001
    - 7.4665490879183360e-001
    - 6.5298035183474057e-001
    - 2.6342833350187118e+002
```


C.3. Camera 3 Parameters


```
calibration_file_ver: 3.1
calibration_ver: 3.1.359264
calibration_time: 2025-1-24 15:54:51
calibration_type: MVP_CALIB_TYPE_NORMAL
frame_num: 1
image_width: 4096
image_height: 3000
board_width: 0
board_height: 0
board_dx: 2.0000000000000000e+000
board_dy: 2.0000000000000000e+000
points_num: 0
H_matrix:
  rows: 3
  cols: 3
  data:
    - 7.5368561341388229e+001
    - 6.6572215304853710e+000
    - 2.2010105906328986e+003
    - 1.2956815846651235e+000
    - 5.6687607049282498e+001
    - 4.9215373641409366e+002
    - 1.3453776666924063e-004
    - 3.0541386844316651e-003
    - 1.0000000000000000e+000
distortion_model: 3
distortion_level: 2
distortion_matrix:
  rows : 12
  cols : 1
  data:
    - 1.8004248377403349e-002
    - -3.0044854730835597e+000
    - -2.5340313073236019e-003
    - 1.5663430597834375e-002
    - 3.9016750900379250e+001
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
intrinsic_matrix:
  rows : 3
  cols : 3
  data:
    - 1.7313015180332859e+004
    - 0.0000000000000000e+000
    - 2.9813759579949819e+003
    - 0.0000000000000000e+000
    - 1.7258484049521408e+004
```

```
    - 1.2130817833681663e+003
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 1.0000000000000000e+000
extrinsic_matrix:
  rows : 3
  cols : 4
  data:
    - 9.9940303465816926e-001
    - -3.2638864426226488e-002
    - -1.1326025119533373e-002
    - -1.0403174684703622e+001
    - 1.5144923326588894e-002
    - 7.0855235884226586e-001
    - -7.0549570237988890e-001
    - -9.6411733094746737e+000
    - 3.1051660398014748e-002
    - 7.0490301411472722e-001
    - 7.0862369074036691e-001
    - 2.3080255579353312e+002
```


C.4. Camera 4 Parameters


```
calibration_file_ver: 3.1
calibration_ver: 3.1.359264
calibration_time: 2025-1-24 15:56:5
calibration_type: MVP_CALIB_TYPE_NORMAL
frame_num: 1
image_width: 4096
image_height: 3000
board_width: 0
board_height: 0
board_dx: 2.0000000000000000e+000
board_dy: 2.0000000000000000e+000
points_num: 0
H_matrix:
  rows: 3
  cols: 3
  data:
    - 7.2093441984252976e+001
    - 6.3939899920335943e+000
    - 1.2086169768470568e+003
    - 9.8103425121876583e-003
    - 5.3897464619363276e+001
    - 1.2205660967156343e+003
    - 3.1503871509957305e-004
    - 2.7843508343726778e-003
    - 1.0000000000000000e+000
distortion_model: 3
distortion_level: 2
distortion_matrix:
  rows : 12
  cols : 1
  data:
    - 1.8131330524292905e-001
    - -1.2469765989270881e+001
    - 6.3753197148693971e-003
    - 9.0223021157360868e-003
    - 1.5616130476247724e+002
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
intrinsic_matrix:
  rows : 3
  cols : 3
  data:
    - 1.7832381835705353e+004
    - 0.0000000000000000e+000
    - 3.6234241894269994e+003
    - 0.0000000000000000e+000
    - 1.7843555944438864e+004
    - 1.0788558390330131e+003
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 1.0000000000000000e+000
extrinsic_matrix:
  rows : 3
  cols : 4
  data:
    - 9.9686931222851527e-001
    - -5.1912991967135176e-002
    - -5.9637367497796648e-002
    - -3.3927864116395583e+001
    - -4.6345694251529318e-003
    - 7.1460287931792632e-001
    - -6.9951500744213679e-001
    - 1.9897718916501284e+000
    - 7.8930951491100951e-002
    - 6.9760143788236939e-001
    - 7.1212508645683870e-001
    - 2.5054365609050163e+002
```