Modulate-and-Map: Crossmodal Feature Mapping with Cross-View Modulation for 3D Anomaly Detection

Alex Costanzino[1] Pierluigi Zama Ramirez[2] Giuseppe Lisanti[1] Luigi Di Stefano[1]
[1]CVLab, University of Bologna [2]Ca' Foscari University of Venice
https://alex-costanzino.github.io/modmap/

Abstract

We present MODMAP*, a natively multiview and multimodal framework for 3D anomaly detection and segmentation. Unlike existing methods that process views independently, our method draws inspiration from the crossmodal feature mapping paradigm to learn to map features across both modalities and views, while explicitly modelling view-dependent relationships through feature-wise modulation. We introduce a cross-view training strategy that leverages all possible view combinations, enabling effective anomaly scoring through multiview ensembling and aggregation. To process high-resolution 3D data, we train and publicly release a foundational depth encoder tailored to industrial datasets. Experiments on SiM3D, a recent benchmark that introduces the first multiview and multimodal setup for 3D anomaly detection and segmentation, demonstrate that* MODMAP *attains state-of-the-art performance by surpassing previous methods by wide margins.*

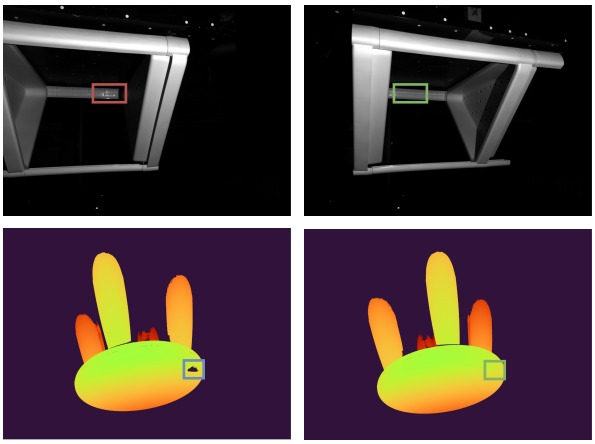

Figure 1. **View-dependent Artefacts.** The first column shows acquisition artefacts observed in an image (top row: specular highlights, **red box**) and a depth map (bottom row: missing depths, **blue box**), dealing with two objects of SiM3D. As shown in the right column, other views of the same object are not affected by artefacts at the positions (**green boxes**) corresponding to those highlighted in the left column.

1. Introduction

The introduction of benchmark datasets such as MVTec AD [3] has significantly accelerated research on *unsupervised* anomaly detection and segmentation (ADS), where the goal is to detect and localise defects by training solely on nominal samples. Most approaches [1, 2, 4, 10, 13, 14, 22, 27, 28, 32, 35, 37, 38, 40] focus on 2D anomaly detection, which consists of processing a single image of the inspected object to produce a 2D anomaly map. While effective for many applications, this paradigm has several limitations. First, geometric defects – such as dents, deformations, or missing parts – may not be clearly visible in RGB images alone, especially under diffuse lighting. Second, 2D anomaly maps do not allow precise localisation of defects in 3D, which is necessary for efficient and possibly automated, additional assessment and rework of high-value products. To address these limitations, more recent benchmarks [6, 7] provide multimodal data, namely RGB images along with pixel-registered 3D information, with methods leveraging both modalities yielding improved performance [11, 19, 36] due to the ability to perceive both colour and geometric anomalies. However, these benchmarks consider a setup where an object is captured from a single viewpoint, which prevents comprehensive inspection when defects may occur over the entire surface. Besides, even partial surface scans may require multiple high-resolution views to detect subtle anomalies. Eventually, in [6, 7] the task is still cast as 2D anomaly detection, and, hence, it does not assess the ability of methods to localise defects in the 3D space precisely.

The recently introduced SiM3D benchmark [12] features the first dataset for multiview multimodal 3D anomaly detection, thereby addressing the limitations highlighted above. Each object is scanned through multiple (12 to 36) multimodal views captured from vantage points designed to ensure comprehensive surface coverage, each view includ-

ing an image along with a pixel-aligned depth map. Unlike previous benchmarks, the task consists of producing a 3D anomaly volume: a voxel grid where each voxel carries an anomaly score. As reported in [12], existing multimodal ADS methods [11, 19, 29, 36] face significant challenges when naively adapted to the multiview scenario due to the lack of mechanisms designed to deploy synergistically multimodal cues gathered from different viewpoints. Instead, one may argue that the co-occurrence of multimodal cues across the views observed in the training data may help establish a more robust, holistic model of the nominal samples compared to learning from individual views in isolation.

In this work, we show how the Crossmodal Feature Mapping (CFM) paradigm, introduced in [11] to tackle single-view anomaly detection, can be extended to obtain an inherently multiview approach and face the challenges set forth by the 3D Anomaly Detection task proposed by SiM3D. CFM relies on the intuition that, given a multimodal view of an object, such as, e.g., an image and depth map, an anomaly manifests itself as an unlikely co-occurrence of the cues – namely the features – observed in the two modalities at corresponding positions. Hence, CFM learns the co-occurrence of features in nominal data by training two neural networks to predict image features from depth features and vice versa. Then, at inference time, it computes per-pixel anomaly scores based on the discrepancy between predicted and observed features. In this paper, we reckon that learning the crossmodal mappings not only within a view but also across views may help robustly handle view-dependent acquisition artefacts, such as specular highlights in images or missing measurements in depth maps (Fig. 1, left column). More in detail, we extend CFM's crossmodal mapping model so that the two neural networks are trained to predict not only a given feature (either an image feature or a depth feature) based on that observed in the other modality in the same view, but also based on those observed in the other modality in all other views. In this way, if at inference time a feature in a view is corrupted by an acquisition artefact unseen at training time, the correct corresponding feature in the other modality may be predicted from another view not affected by the artefact (Fig. 1, right column). Based on this observation, at inference time we compare the observed feature in a view to those predicted from the other modality from all views, and compute the anomaly score based on the closest prediction. Therefore, a feature in a modality can be deemed as nominal if at least one crossmodal mapping from all views can predict it as such. This cross-view and crossmodal feature mapping mechanism is designed to synergistically leverage multimodal cues obtained from all viewpoints, with the goal of maximising robustness to acquisition artefacts. As such, it is a mechanism that inherently trades recall for precision: to counterbalance the potential loss of sensitivity, when computing the final 3D anomaly segmentation and detection outputs, we adopt an aggregation strategy that sifts out the strongest anomaly score, increasing the final recall of our approach.

Thus, the main contribution of this paper is the introduction of the first *natively multiview* multimodal algorithm for 3D anomaly detection, specifically designed to synergistically exploit crossmodal relationships across multiple viewpoints rather than processing views in isolation. Several key technical novelties are instrumental in realising this contribution. We propose a cross-view training strategy that enables effective multiview learning. During training, we consider all pairs of source and target views and, for both modalities, learn to predict target features from source ones at corresponding positions. Crucially, to prevent one-to-many mapping ambiguities, we introduce conditioning on view pairs: an additional network *modulates* the source features, conditioned on both the source and target views, before passing them to the main *mapping* network. Accordingly, we dub our method Modulate-and-Map (MODMAP). Previous multimodal methods [11, 19, 29, 36] rely on point cloud encoders designed for low-resolution inputs (8K points). Conversely, to enable processing of high-resolution depth maps, we deploy multiple industrial datasets and train a *depth foundational model* using self-supervised learning. Our depth encoder can handle high-resolution 3D data (5-7M points) without aggressive downsampling, and provides depth features that are pixel-aligned to image features. Extensive experiments demonstrate that MODMAP achieves state-of-the-art performance on SiM3D, the recent benchmark designed to address multiview and multimodal 3D anomaly detection.

2. Related Work

Anomaly Detection Benchmarks. The standardisation of evaluation protocols through MVTec AD [3] catalysed rapid progress in anomaly detection, fostering the development of numerous methods and subsequent specialised benchmarks. Subsequent benchmarks expanded the scope of ADS to address various challenges: MVTec LOCO [5] introduced logical anomalies, VisA [43] provided high-resolution images of complex scenes, and PAD [41] addressed pose-agnostic detection. Real-IAD [33] recently introduced a large-scale multiview dataset with RGB images, though it still focuses on 2D anomaly maps. To address the limitations of image-only approaches, several benchmarks have incorporated 3D information. MVTec 3D-AD [6] provides RGB images with pixel-aligned XYZ coordinates captured by structured light sensors, while Eyecandies [7] offers synthetic data with pixel-aligned depth and normal maps. Real3D-AD [23] introduced the first point cloud anomaly detection benchmark, using front-and-back scans for training and single-view point clouds for testing. However, the above-described benchmarks still evaluate methods on single-view inputs and yield 2D anomaly maps. SiM3D [12] addresses these limitations

by introducing the first benchmark for multiview 3D anomaly detection, where the task is to produce a 3D anomaly volume by integrating information from multiple views. Moreover, it focuses on the challenging single-instance scenario, where only one nominal object – either real or synthetic – is available for training, closely resembling real industrial scenarios.

Multimodal Anomaly Detection. Multimodal single-view benchmarks have fostered the introduction of approaches that effectively integrate different modalities (e.g., RGB images and 3D data) to perform 2D anomaly detection. BTF [19] extended PatchCore [27], a solution based on memory banks, to multimodal input by combining 2D features from pre-trained CNNs with hand-crafted 3D features (FPFH [30]). M3DM [36] improved upon BTF by using Transformer-based backbones (DINO-v1 [8] for RGB and Point-MAE [25] for point clouds) and learning a fusion function to combine modality-specific features. AST [29] adopted a teacher-student paradigm using normalising flows, processing depth information as additional input channels rather than extracting explicit 3D features. EasyNet [9] introduced a lightweight architecture combining multiple feature extraction strategies. More recently, CFM [11] proposed learning crossmodal feature mappings between 2D and 3D features on nominal samples, detecting anomalies by identifying inconsistencies between predicted and observed features. While these methods achieve strong performance on single-view multimodal benchmarks, they face significant challenges when extended to multiview scenarios. Memory bank methods suffer from computational overhead that increases with the number of views, while reconstruction-based approaches process views independently without leveraging geometric consistency. In addition, solutions based on pre-trained point cloud feature extractors, such as [11, 36], are unable to take full advantage of high-resolution 3D data (e.g., SiM3D provides point clouds of 5–7 million points) due to the severe downsampling performed by [25] (i.e., 2048 points), which limits their ability to detect small geometric defects. Our work extends the crossmodal feature mapping paradigm to the multiview settings by introducing view-conditioned modulation and cross-view training, along with a pre-trained depth encoder, enabling efficient and robust multiview anomaly detection.

3. Method

3.1. Task Definition

Our method addresses the multiview multimodal 3D anomaly detection task introduced by the recent SiM3D benchmark [12]. Specifically, the inputs consist of a set of images $\{I^i\}_{i=1}^n$ and corresponding 3D data $\{D^i\}_{i=1}^n$, captured from n viewpoints of an object instance. The viewpoints remain consistent across different instances of the same object category. In our approach, the 3D data are represented as depth maps, enabling efficient high-resolution processing. For each object instance, the detection task involves predicting a global anomaly score. In contrast, the segmentation task aims to estimate an anomaly volume $\Omega \in \mathbb{R}^{X \times Y \times Z}$, where X, Y, and Z denote the grid dimensions, and each voxel encodes an anomaly score.

3.2. Cross-View Crossmodal Feature Mapping

Our approach draws inspiration from CFM [11] by learning mappings that predict features from one modality to another on nominal samples. At inference time, inconsistencies between the predicted and actual features highlight anomalies. We extend such a mechanism by leveraging multiview inputs to generate more accurate anomaly maps and reduce false positives caused by acquisition artefacts. In particular, we propose a Modulate-and-Map framework, MODMAP, capable of mapping features not only across modalities but also across views.

3.2.1. MODMAP Architecture

MODMAP comprises six main components described in the following sections: an image feature extractor, a depth feature extractor, two modulators and two mapping networks (Fig. 2).

Image Feature Extractor. We employ a frozen DINO-v2 [24] encoder, denoted as $\mathcal{E}_I$, as our image feature extractor. Given an input image $I^i \in \mathbb{R}^{H \times W}$ from view i, $\mathcal{E}_I$ produces a feature map:

$$F_I^i = \mathcal{E}_I(I^i) \in \mathbb{R}^{h \times w \times c_I} \tag{1}$$

where $h = H/p$ and $w = W/p$ correspond to the spatial resolution of the feature map given the patch size p, and c_I denotes the number of feature channels. Finally, the feature map F_I^i is unrolled into $h \cdot w$ elements of dimension c_I for subsequent processing.

Depth Feature Extractor. Existing pre-trained point-cloud feature extractors (e.g., Point-MAE [25]) struggle to handle the high-resolution 3D data provided in SiM3D [12, 16], which contain 5-7M points per sample. This limitation mainly arises from the unstructured nature of point clouds, which makes it computationally expensive to apply local operations and efficiently extract fine-grained geometric features. To overcome this issue, we operate on depth maps, which allow efficient high-resolution processing thanks to their structured, image-like representation [12]. Since no pre-trained depth feature extractors do exist, we train a dedicated encoder $\mathcal{E}_D$ based on a Vision Transformer, adapted to single-channel depth inputs. The encoder is trained from scratch following a self-supervised protocol inspired by DINO-v2, with one key difference: only spatial augmentations are applied during training, including random rotations, horizontal and vertical flips, and random crops with scaling. Photometric augmentations (e.g., brightness, contrast, or colour

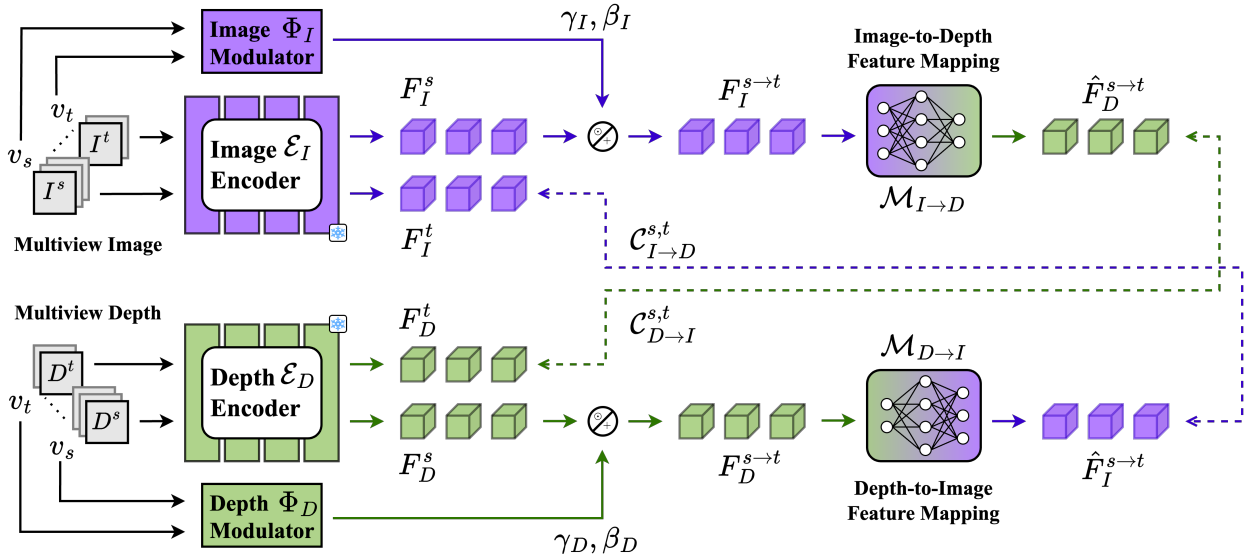

Figure 2. **MODMAP Training.** Starting from the set of images I and depths D from a training sample, we select a source view s and target view t, and forward their images I^s, I^t and depths D^s, D^t to the image, $\mathcal{E}_I$, and depth, $\mathcal{E}_D$, encoders, respectively, so as to compute modality-specific features, F_I^s, F_I^t and F_D^s, F_D^t. Moreover, the one-hot encodings of the view indexes, v_s and v_t, are fed into the feature modulators, Φ_I and Φ_D, to generate modality-specific scale-and-shift parameters, γ_I, β_I and γ_D, β_D. Then, for both modalities, the source features are scale-and-shifted to obtain modulated source features that incorporate view conditioning: $F_I^{s\to t}, F_D^{s\to t}$. The modulated features are passed as inputs to the mapping networks $\mathcal{M}_{I\to D}, \mathcal{M}_{D\to I}$ that predict the corresponding features from the other modality $\hat{F}_D^{s\to t}, \hat{F}_I^{s\to t}$. The predicted features are then compared to the actual target features F_D^t, F_I^t to optimise both the modulators and the mapping networks.

jittering) are deliberately excluded, as they would corrupt the information encoded in depth maps, where pixel intensities correspond to distance values. Beyond efficiency, depth processing provides several advantages over point-cloud processing. First, it produces features that are pixel-aligned with images and architecturally coherent. Indeed, by employing the same Vision Transformer design with shared normalisation layers, positional encodings, and self-attention mechanisms, the depth encoder learns representations that naturally align with image features in both spatial structure and semantic space. Second, it allows leveraging multiple industrial datasets during training, yielding representations specifically tailored for anomaly detection tasks. In particular, $\mathcal{E}_D$ is trained on depth maps from various multimodal industrial anomaly detection datasets, including MVTec 3D-AD [6], Eyecandies [7], and Real-IAD D^3 [34]. We further augment the training set with depth maps obtained from a monocular depth foundation model, DepthAnything-v2 [39], on 2D anomaly detection datasets such as MVTec AD [3], MVTec LOCO [5], MVTec AD 2 [18], and ViSA [43]. After this pre-training, the encoder is frozen and integrated into our pipeline. Since SiM3D – the dataset used in our experimental evaluation – is not included in the training data, our depth encoder is regarded as an off-the-shelf foundation model, similar to DINO-v2, within our framework.

Given a depth map $D^i \in \mathbb{R}^{H\times W}$ from view i, $\mathcal{E}_D$ produces a depth feature map:

$$F_D^i = \mathcal{E}_D(D^i) \in \mathbb{R}^{h\times w\times c_D} \tag{2}$$

where $h = H/p$ and $w = W/p$ correspond to the spatial resolution of the feature map given the patch size p, and c_D denotes the number of feature channels. Finally, the feature map F_D^i is unrolled into $h \cdot w$ elements of dimension c_D for subsequent processing.

Modulate-and-Map. The proposed Modulate-and-Map strategy, depicted in Fig. 2, transforms features, F_m^s, of one modality, $m \in \{I, D\}$, obtained from a source viewpoint, $s \in \{1, .., N\}$, to features, F_n^t, of the other modality, $n \in \{I, D\}$, with $m \neq n$, of a target viewpoint, $t \in \{1, .., N\}$.

First, for each view i, we encode its identity as a one-hot vector $v_i \in \{0, 1\}^N$, where N is the total number of views. Given a source view code, v_s, and a target view code, v_t, we use a modulator Φ to condition F^s based on the desired source to target mapping:

$$F^{s\to t} = \Phi(F^s, v_s, v_t) \tag{3}$$

The modulator, Φ, inspired by FiLM [26], is implemented as a lightweight MLP that takes as input the concatenation of the source and target view codes and produces scale-and-shift parameters:

$$[\gamma, \beta] = \text{MLP}([v_s; v_t]) \tag{4}$$

$$\Phi(F^s, v_s, v_t) = \gamma \odot F^s + \beta \tag{5}$$

where $\gamma, \beta \in \mathbb{R}^d$ are the modulation parameters initialised at $\gamma = \mathbf{1}$ and $\beta = \mathbf{0}$, $\odot$ denotes element-wise multiplication, and $d \in \{c_I, c_D\}$ is the corresponding feature dimension for each modality. We employ feature-wise linear modulations, which enable view-dependent feature adaptation via learnable affine transformations while preserving the geometric structure of pre-trained feature spaces. This preservation is crucial as the mapping networks establish correspondences between image and depth features based on semantic similarity (e.g., edge of hole, flat surface), and these relationships are encoded in the spatial organisation of the feature space. Indeed, disrupting such a structure would prevent learning consistent crossmodal mappings. Moreover, identity initialisation provides a natural starting point, allowing the network to smoothly learn view-specific adaptations without disrupting the rich semantic content encoded by the frozen pre-trained encoders.

Then, we employ two lightweight MLP-based mapping networks that operate feature-wise: $\mathcal{M}_{I\to D} : \mathbb{R}^{c_I} \to \mathbb{R}^{c_D}$ maps image features to depth features, and $\mathcal{M}_{D\to I} : \mathbb{R}^{c_D} \to \mathbb{R}^{c_I}$ performs the opposite mapping. Given modulated features from source view s to target view t, the mapped features are:

$$\hat{F}_D^{s\to t} = \mathcal{M}_{I\to D}\left(\Phi_I(F_I^s, v_s, v_t)\right) \tag{6}$$

$$\hat{F}_I^{s\to t} = \mathcal{M}_{D\to I}\left(\Phi_D(F_D^s, v_s, v_t)\right) \tag{7}$$

3.2.2. MODMAP Training

As introduced in Sec. 1, we realise a natively multiview approach robust to acquisition artefacts by training the mapping networks to predict crossmodal mappings for all possible source-target view pairs. Moreover, this enables us to vastly augment the training samples, which may help prevent overfitting in data scarcity regimes, as is indeed the case of SiM3D, which ships only one training instance, either real or synthetic, per object category. Thus, as shown in Fig. 2, for a training sample with N views, we process all $N \times N$ source-target view pairs. For each pair (s, t), we optimise cosine distances between predicted and actual features:

$$\mathcal{C}_{I\to D}^{s,t} = 1 - \frac{\hat{F}_D^{s\to t} \cdot F_D^t}{\|\hat{F}_D^{s\to t}\| \|F_D^t\|} \tag{8}$$

$$\mathcal{C}_{D\to I}^{s,t} = 1 - \frac{\hat{F}_I^{s\to t} \cdot F_I^t}{\|\hat{F}_I^{s\to t}\| \|F_I^t\|} \tag{9}$$

The overall loss is defined as the sum of the two terms:

$$\mathcal{L} = \mathcal{C}_{I\to D}^{s,t} + \mathcal{C}_{D\to I}^{s,t} \tag{10}$$

3.2.3. MODMAP Inference

As pointed out in Sec. 1, our framework achieves robustness to acquisition artefacts – and thereby maximises precision – by selecting, for each spatial position and modality, the closest prediction across all views. Hence, as depicted in Fig. 3, we predict multiple anomaly maps for each viewpoint and ensemble them so as to maximise precision. In particular, during inference, given a sample with N views, for each of the $N \times N$ combinations of source and target views (s, t), we compute anomaly scores associated with both modalities, $\Psi_I^{s,t}$ and $\Psi_D^{s,t}$, via the cosine distance between predicted and actual features:

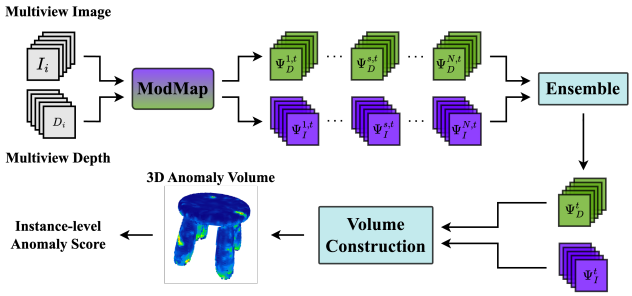

Figure 3. **MODMAP Inference.** We process the set of N images I^i and N depths D^i, obtaining $N \times N$ anomaly maps for each of the two modalities. We ensemble the anomaly scores into N refined 2D anomaly maps for each modality. Finally, we aggregate the 2D anomaly maps to obtain a 3D Anomaly Volume and an Instance-level Anomaly Score.

$$\Psi_I^{s,t} = \mathcal{C}_{D\to I}^{s,t} \quad \Psi_D^{s,t} = \mathcal{C}_{I\to D}^{s,t} \tag{11}$$

with $\Psi_I^{s,t}$ and $\Psi_D^{s,t}$ reshaped as a 2D anomaly map of resolution $h \times w$. Afterwards, for each target view t and for both modalities, we ensemble the anomaly scores computed at each spatial location based on the crossmodal predictions from all source views:

$$\Psi_I^t = \min_{s\in\{1,\dots,N\}} \Psi_I^{s,t} \quad \Psi_D^t = \min_{s\in\{1,\dots,N\}} \Psi_D^{s,t} \tag{12}$$

Following common practice in multimodal anomaly detection [11, 29, 36], we filter out background regions based on depth information.

Rationale. We discuss here in more detail the rationale behind our minimum-based cross-view ensembling strategy, aimed at maximising the robustness of Ψ_D^t and Ψ_I^t with respect to view-dependent image and depth artefacts, respectively. As illustrated in Fig. 4 (first column), if an image artefact corrupts a feature (**red box**) at a certain position in view i, the crossmodal mapping will likely predict an incorrect depth feature (**grey box**), which, therefore, will not match the actual one observed at the considered position in the same view (**green box**, top row). Yet, the image artefact may not corrupt the feature at the same position in view j (**green box**, bottom row), which, therefore, may predict a feature close to the actual one, yielding a low anomaly score in $\Psi_D^{i,j}$. Hence, minimum-based ensembling allows

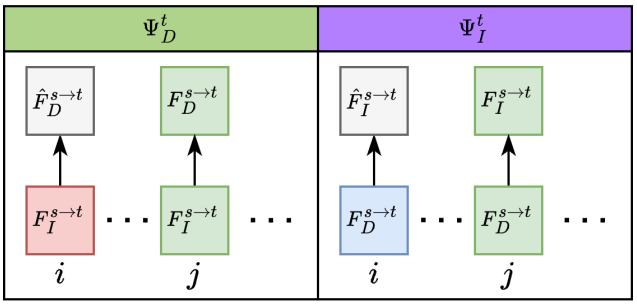

Figure 4. **Rationale for ensembling.** Squares represent features (**green**: uncorrupted, **red**: corrupted by image artefacts, **blue**: corrupted by depth artefacts, **grey**: incorrect prediction), while arrows show mapping predictions.

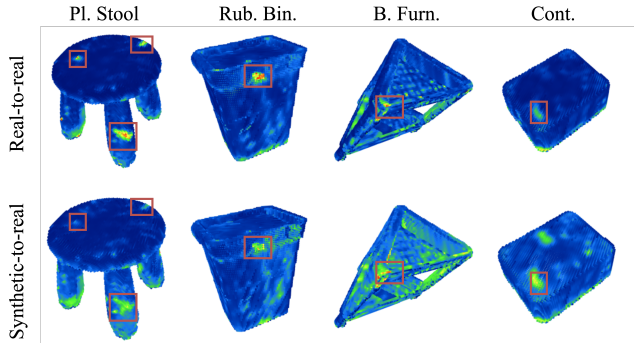

Figure 5. **Qualitative results.** Real-to-real (top) vs. Synthetic-to-real (bottom). Anomalies are highlighted by **red boxes**.

for sifting out the lowest anomaly score at every position, thereby alleviating the impact of the image artefacts in Ψ_D^t. As shown in Fig. 4 (second column), similar considerations may be drawn for a depth artefact and Ψ_I^t.

3.2.4. 3D Anomaly Detection and Segmentation

The anomaly maps Ψ_I^t, Ψ_D^t produced for each viewpoint by our minimum-based ensembling strategy are highly precise but may exhibit insufficient sensitivity. To counterbalance this effect, we adopt a recall-oriented, maximum-based strategy when aggregating the per-view anomaly maps across all viewpoints in order to obtain, for each test sample, the 3D anomaly volume. In particular, for each view t, we project the 2D anomaly maps Ψ_I^t, Ψ_D^t into the 3D space using the known camera intrinsics and extrinsics. Each pixel's anomaly scores are assigned to the corresponding 3D voxel. After processing all views, each voxel accumulates multiple scores from different viewpoints and both modalities. To obtain the final 3D anomaly volume, we take, for each voxel, the *maximum* score. This *maximum*-based aggregation strategy across views ensures that a voxel may be segmented out as anomalous if it contains a defect visible from at least a single viewpoint. At last, the global, instance-level anomaly score is taken as the maximum value across all the scores within the anomaly volume.

4. Experiments

4.1. Multiview ADS with MODMAP

We evaluate our method against state-of-the-art anomaly detection approaches adapted to the multiview and multimodal 3D anomaly detection scenario set forth by SiM3D [12]. The considered competitors include memory bank methods (PatchCore [27], BTF [19], M3DM [36]), teacher-student approaches (EfficientAD [1], AST [29]), and the original, single-view Crossmodal Feature Mapping (CFM [11]). Hence, as proposed in [12], all the competitors are adapted to produce 3D anomaly volumes by processing each view independently and aggregating the resulting 2D anomaly maps into the 3D space using the projection strategy described in the SiM3D paper. Tab. 1 and Tab. 2 report results on the real-to-real and synthetic-to-real setups of SiM3D, respectively, with qualitative results shown in Fig. 5. More experiments are reported in the Supplementary Material.

Real-to-real Setup. In the real-to-real setup (Tab. 1), our method (MODMAP) achieves the best performance in both the detection and segmentation tasks, with a mean I-AUROC of 0.844 and a mean V-AUPRO@1% of 0.804. This represents a substantial improvement – ∼14% in detection and ∼10% in segmentation – over previous multimodal methods, i.e., BTF (WRN-101) and AST, which achieve 0.707 I-AUROC and 0.700 V-AUPRO@1%, respectively. Notably, MODMAP significantly outperforms the original CFM (0.448 I-AUROC, 0.538 V-AUPRO@1%), demonstrating the effectiveness of our *natively multiview* framework. Compared to the runner-up in detection, i.e., PatchCore with WRN-101, MODMAP yields ∼9% improvement (I-AUROC%: 0.844 vs 0.754), with the gap in segmentation being ∼17% (V-AUPRO@1%: 0.804 vs 0.630).

Synthetic-to-real Setup. The synthetic-to-real setup (Tab. 2) presents a more challenging scenario due to the domain shift between synthetic training data and real test samples. Indeed, all methods show a significant performance drop, with MODMAP achieving 0.623 I-AUROC and 0.755 V-AUPRO@1%, outperforming the runner-ups of ∼8% in detection and ∼6% in segmentation. Interestingly, some competitors exhibit erratic behaviour in this setup, i.e., AST achieves 1.000 I-AUROC on Plastic Stool but drops to 0.002 on Rubbish Bin, suggesting instability when faced with domain shift. In contrast, MODMAP demonstrates more consistent performance across categories, which vouches for the robustness of our paradigm.

4.2. Ablation Studies

We conduct comprehensive ablation studies to analyse the contribution of the key components of our framework. All ablations are performed on the real-to-real setup of SiM3D. More ablations are included in the Supplementary Material.

Algorithm	Features		Detection									Segmentation								
	Image	Depth / Point Cloud	Pl. Stool	Rub. Bin	W. Vase	B. Furn.	Cont.	Pl. Vase	W. Stool	Sink Cab.	Mean	Pl. Stool	Rub. Bin	W. Vase	B. Furn.	Cont.	Pl. Vase	W. Stool	Sink Cab.	Mean
PatchCore	WRN-101	✗	0.740	0.987	0.636	0.777	0.774	0.551	1.000	0.566	<u>0.754</u>	0.710	0.461	0.761	0.675	0.694	0.747	0.380	0.609	0.630
	DINO-v2	✗	0.500	0.958	0.636	0.622	0.578	0.563	1.000	0.563	0.678	0.745	0.469	0.775	0.792	0.709	0.753	0.435	0.690	0.671
	✗	WRN-101	–	–	–	–	–	–	–	–	0.582	–	–	–	–	–	–	–	–	0.486
	✗	DINO-v2	–	–	–	–	–	–	–	–	0.600	–	–	–	–	–	–	–	–	0.469
	✗	FPFH	–	–	–	–	–	–	–	–	0.415	–	–	–	–	–	–	–	–	0.464
EfficientAD	PDN	✗	0.280	0.732	0.000	0.878	0.424	0.730	0.928	0.712	0.586	0.682	0.462	0.763	0.534	0.680	0.743	0.407	0.488	0.595
AST	EffNet-B5	✗	0.537	0.566	0.611	0.515	0.573	0.496	0.839	0.537	0.584	0.720	0.502	0.789	0.803	0.716	0.764	0.446	0.758	0.687
	EffNet-B5	EffNet-B5	0.950	0.927	0.785	0.474	0.542	0.470	0.428	0.925	0.688	0.750	0.503	0.792	0.807	0.716	0.764	0.467	0.798	<u>0.700</u>
BTF	DINO-v2	FPFH	0.421	0.217	0.504	0.565	0.545	0.471	0.678	0.424	0.478	0.551	0.402	0.750	0.377	0.614	0.741	0.092	0.030	0.445
	WRN-101	WRN-101	–	–	–	–	–	–	–	–	0.707	–	–	–	–	–	–	–	–	0.609
CFM	DINO-v2	FPFH	0.198	0.301	0.074	0.515	0.483	0.456	0.732	0.825	0.448	0.597	0.415	0.768	0.505	0.640	0.743	0.315	0.321	0.538
	DINO-v2	DINO-v2	–	–	–	–	–	–	–	–	0.548	–	–	–	–	–	–	–	–	0.663
M3DM	DINO-v2	FPFH	0.702	0.988	0.661	0.545	0.556	0.649	0.392	0.475	0.621	0.733	0.452	0.767	0.702	0.702	0.752	0.288	0.126	0.565
	DINO-v2	DINO-v2	–	–	–	–	–	–	–	–	0.659	–	–	–	–	–	–	–	–	0.503
MODMAP	DINO-v2	DINO-Depth	0.909	0.990	0.945	0.647	0.740	0.607	0.916	1.000	**0.844**	0.855	0.707	0.863	0.791	0.831	0.885	0.711	0.785	**0.804**

Table 1. **Results on real-to-real setup of SiM3D.** Best results in **bold**, runner-ups <u>underlined</u>. Missing entries (–) not reported in [12].

Algorithm	Features		Detection									Segmentation								
	Image	Depth / Point Cloud	Pl. Stool	Rub. Bin	W. Vase	B. Furn.	Cont.	Pl. Vase	W. Stool	Sink Cab.	Mean	Pl. Stool	Rub. Bin	W. Vase	B. Furn.	Cont.	Pl. Vase	W. Stool	Sink Cab.	Mean
PatchCore	WRN-101	✗	0.462	0.235	0.537	0.353	0.678	0.576	0.517	0.250	0.451	0.734	0.437	0.751	0.454	0.678	0.741	0.386	0.618	0.600
	DINO-v2	✗	0.958	0.897	0.413	0.464	0.207	0.684	0.607	0.087	0.540	0.701	0.457	0.772	0.563	0.648	0.734	0.321	0.575	0.596
	✗	WRN-101	–	–	–	–	–	–	–	–	0.507	–	–	–	–	–	–	–	–	0.484
	✗	DINO-v2	–	–	–	–	–	–	–	–	0.246	–	–	–	–	–	–	–	–	0.471
	✗	FPFH	–	–	–	–	–	–	–	–	0.313	–	–	–	–	–	–	–	–	0.460
EfficientAD	PDN	✗	0.123	0.673	0.000	0.848	0.008	0.487	0.750	0.662	0.444	0.680	0.449	0.729	0.523	0.669	0.747	0.431	0.361	0.574
AST	EffNet-B5	✗	1.000	0.344	1.000	0.525	0.482	0.388	0.428	0.187	<u>0.544</u>	0.726	0.488	0.791	0.806	0.705	0.764	0.461	0.695	0.680
	EffNet-B5	EffNet-B5	0.636	0.002	0.504	0.474	0.462	0.569	0.428	0.887	0.495	0.751	0.502	0.793	0.805	0.717	0.764	0.450	0.785	<u>0.696</u>
BTF	DINO-v2	FPFH	0.462	0.294	0.520	0.444	0.464	0.424	0.482	0.287	0.422	0.547	0.402	0.756	0.369	0.610	0.741	0.103	0.040	0.446
	WRN-101	WRN-101	–	–	–	–	–	–	–	–	0.448	–	–	–	–	–	–	–	–	0.543
CFM	DINO-v2	FPFH	0.008	0.000	0.190	0.424	0.313	0.459	0.428	0.362	0.273	0.523	0.413	0.753	0.542	0.621	0.741	0.222	0.314	0.516
	DINO-v2	DINO-v2	–	–	–	–	–	–	–	–	0.311	–	–	–	–	–	–	–	–	0.620
M3DM	DINO-v2	FPFH	0.107	0.117	0.735	0.565	0.381	0.586	0.214	0.512	0.402	0.690	0.454	0.770	0.461	0.645	0.734	0.318	0.132	0.526
	DINO-v2	DINO-v2	–	–	–	–	–	–	–	–	0.254	–	–	–	–	–	–	–	–	0.476
MODMAP	DINO-v2	DINO-Depth	0.681	0.707	0.954	0.472	0.805	0.594	0.312	0.458	**0.623**	0.778	0.663	0.861	0.752	0.850	0.864	0.550	0.725	**0.755**

Table 2. **Results on synthetic-to-real setup of SiM3D.** Best results in **bold**, runner-ups <u>underlined</u>. Missing entries (–) not reported in [12].

View Conditioning	Features		Detection	Segmentation
	Image	Point Cloud		
✗	DINO-v2	FPFH	0.448	0.538
✓	DINO-v2	FPFH	**0.548**	**0.663**

Table 3. **Impact of View Conditioning.**

Features		Detection	Segmentation
Image	Depth / Point Cloud		
DINO-v2	FPFH	0.548	0.663
	Point-MAE	0.543	0.545
	DINO-v2	0.575	0.684
	DINO-Depth	**0.804**	**0.735**

Table 4. **Effects of 3D Feature Extractor.**

View Conditioning. Tab. 3 demonstrates the impact of introducing view conditioning into the CFM framework adopted in [12] to more robustly handle one-to-many mappings caused by the multi-view setup. Purposely, we introduce two modality-specific modulators that condition the features passed to the mapping networks based on the one-hot encoded view identity. Without view conditioning, the method achieves 0.448 I-AUROC and 0.538 V-AUPRO@1%. Adding view conditioning via feature modulation does improve performance to 0.548 I-AUROC and 0.663 V-AUPRO@1%, representing gains of +10.0% in detection and +12.5% in segmentation.

3D Feature Extractor. Keeping the improved version of CFM that incorporates view conditioning, in Tab. 4 we compare different choices for the 3D feature extractor. Using FPFH features, as proposed in [12], yields 0.548 I-AUROC and 0.663 V-AUPRO@1%. Employing Point-MAE, as originally proposed in [11], yields lower performance, i.e., 0.543 I-AUROC and 0.545 V-AUPRO@1%. Switching to foundational image features by feeding DINO-v2 with depth maps improves performance to 0.575 I-AUROC and 0.684 V-AUPRO@1%. However, DINO-Depth, our dedicated foundational depth encoder trained on industrial datasets, achieves by far the best results with 0.804 I-AUROC and 0.735 V-AUPRO@1%, with improvements of +22.9% in detection and +5.1% in segmentation compared to DINO-v2. A comparison between DINO-v2 and DINO-Depth features is shown in Fig. 6. The features learned by DINO-v2 are noisier and tend to over-segment depth maps, while those computed by DINO-Depth are significantly smoother and seem more amenable to capturing the semantics of depth maps.

Depth Encoder Training Set. We train DINO-Depth in a self-supervised manner, progressively expanding the composition of the training set, as shown in Tab. 5. We begin with training split A, comprising MVTec 3D-AD [6] and Eyecandies [7], namely the first two datasets for multimodal anomaly detection. Expanding to training split B,

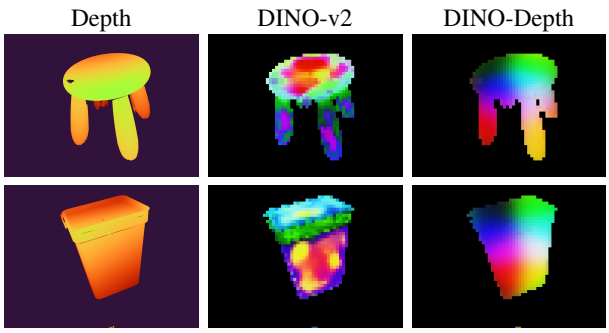

Figure 6. **PCA of Depth Features.**

we incorporate Real-IAD D^3 [34], a more recent large-scale multimodal dataset, obtaining improvements of $+4.1\%$ in detection and $+3.8\%$ in segmentation over split A. Finally, we create training split C by adding depth maps obtained by using DepthAnything-v2 [39] on image-only datasets: MVTec AD [3], MVTec LOCO [5], MVTec AD 2 [18], and ViSA [43]. This substantially increases training scale, delivering $+19.5\%$ in detection and $+3.0\%$ in segmentation relative to split B, demonstrating the value of large-scale pre-training even with pseudo-depth annotations.

Training Split	No. Samples	Detection	Segmentation
A	19647	0.568	0.667
B	23896	0.609	0.705
C	46528	**0.804**	**0.735**

Table 5. **DINO-Depth Training Set.**

Cross-View Conditioning. Tab. 6 compares the CFM approach improved by view conditioning and DINO-Depth to our final proposal, namely MODMAP, that deploys also cross-view conditioning along with the associated *minimum*-based ensembling and *maximum*-based aggregation, to pursue robustness to acquisition artefacts without sacrificing recall. Hence, CFM with view conditioning and DINO-Depth achieves 0.804 I-AUROC and 0.735 V-AUPRO@1%, while MODMAP yields 0.844 I-AUROC and 0.804 V-AUPRO@1%, with a remarkable improvement of $+4.0\%$ in detection and $+6.9\%$ in segmentation.

Cross-View Conditioning	Detection	Segmentation
✗	0.804	0.735
✓	**0.844**	**0.804**

Table 6. **Cross-View Conditioning.**

4.3. Multi-class ADS.

Recently, multi-class ADS, the task of training a single unified model to detect and segment anomalies across multiple object categories, has gained traction in literature, with proposals including both class-agnostic [15, 17] and class-conditioned architectures [20, 42]. This approach offers a particularly MLOps-friendly design as it allows a single model to be trained, deployed and maintained across multiple production lines. MODMAP can be seamlessly extended to multi-class ADS by incorporating one-hot class encodings into the modulators, thereby conditioning the crossmodal mapping networks on both the view pair and object category. We report the comparison between class-specific and multi-class formulations of MODMAP in Tab. 7. We observe that MODMAP is particularly amenable to the SiM3D multi-class scenario, with a comparable performance to the class-specific models.

Model	Real-to-real		Synthetic-to-real	
	Detection	Segmentation	Detection	Segmentation
Class-specific	0.844	**0.804**	**0.623**	0.755
Multi-class	**0.850**	0.801	0.618	**0.758**

Table 7. **Class-specific vs. Multi-class.**

5. Concluding Remarks

We have presented MODMAP, the first *natively* multiview framework for multimodal 3D anomaly detection. Our proposal extends the original crossmodal feature mapping formulation by three key contributions. First, we introduce view-conditioning by feature-wise modulation to handle potential one-to-many crossmodal mappings. Second, we train DINO-Depth, a foundational depth encoder tailored to industrial datasets, which enables processing of high-resolution 3D data. Third, we propose cross-view conditioning with minimum-ensembling and maximum-aggregation, to optimise both robustness to sensing artefacts and sensitivity to defects. Ablation studies validate each contribution's effectiveness, with view-conditioning alone yielding $\sim$10% and $\sim$12% improvements in detection and segmentation, DINO-Depth adding $\sim$26% and $\sim$7% on top of previous improvements, and, finally, cross-view conditioning with min-max further increasing performance by 4% and $\sim$7% in detection and segmentation. Experiments on both setups of SiM3D demonstrate state-of-the-art performance, with large margins versus previous methods. Moreover, MODMAP lends itself to a natural and effective multi-class formulation, which may streamline adoption in real industrial settings. The main limitation of our work pertains to the relatively small variety of the experimental data, due to SiM3D being nowadays the only dataset for multiview and multimodal 3D anomaly detection. Besides, the performance achieved by MODMAP is far from saturating the benchmark, highlighting the need for further research on the challenging ADS setup proposed by SiM3D.

References

[1] Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In *Winter Applications of Computer Vision*, 2024. 1, 6

[2] 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. 1

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 4, 8

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[5] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 2022. 2, 4, 8

[6] Paul Bergmann, Jin Xin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. In *The International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2022. 1, 2, 4, 7

[7] Luca Bonfiglioli, Marco Toschi, Davide Silvestri, Nicola Fioraio, and Daniele De Gregorio. The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In *Asian Conference on Computer Vision*, 2022. 1, 2, 4, 7

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision*, 2021. 3

[9] Ruitao Chen, Guoyang Xie, Jiaqi Liu, Jinbao Wang, Ziqi Luo, Jinfan Wang, and Feng Zheng. Easynet: An easy network for 3d industrial anomaly detection, 2023. 3

[10] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020. 1

[11] Alex Costanzino, Pierluigi Zama Ramirez, Giuseppe Lisanti, and Luigi Di Stefano. Multimodal industrial anomaly detection by crossmodal feature mapping. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2, 3, 5, 6, 7, 11

[12] Alex Costanzino, Pierluigi Zama Ramirez, Luigi Lella, Matteo Ragaglia, Alessandro Oliva, Giuseppe Lisanti, and Luigi Di Stefano. Sim3d: Single-instance multiview multimodal and multisetup 3d anomaly detection benchmark. In *International Conference on Computer Vision*, 2025. 1, 2, 3, 6, 7, 11

[13] 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*, 2021. 1

[14] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Winter Applications of Computer Vision*, 2022. 1

[15] Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025. 8

[16] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[17] Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. In *Advances on Neural Information Processing Systems*, 2024. 8

[18] Lars Heckler-Kram, Jan-Hendrik Neudeck, Ulla Scheler, Rebecca König, and Carsten Steger. The mvtec ad 2 dataset: Advanced scenarios for unsupervised anomaly detection. *arXiv preprint arXiv:2503.21622*, 2025. 4, 8

[19] Eliahu Horwitz and Yedid Hoshen. Back to the feature: classical 3d features are (almost) all you need for 3d anomaly detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3, 6

[20] Xi Jiang, Ying Chen, Qiang Nie, Jianlin Liu, Yong Liu, Chengjie Wang, and Feng Zheng. Toward multi-class anomaly detection: Exploring class-aware unified model against inter-class interference. *arXiv preprint arXiv:2403.14213*, 2024. 8

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 11

[22] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[23] Jiaqi Liu, Guoyang Xie, Xinpeng Li, Jinbao Wang, Yong Liu, Chengjie Wang, Feng Zheng, et al. Real3d-ad: A dataset of point cloud anomaly detection. In *Advances on Neural Information Processing Systems - Datasets & Benchmarks Track*, 2023. 2

[24] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 3, 11

[25] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point

cloud self-supervised learning. In *European Conference on Computer Vision*, 2022. 3

[26] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018. 4

[27] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3, 6

[28] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Winter Applications of Computer Vision*, 2021. 1

[29] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. In *Winter Applications of Computer Vision*, 2023. 2, 3, 5, 6

[30] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *IEEE International Conference on Robotics and Automation*, 2009. 3

[31] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, 2019. 11

[32] Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister. Learning and evaluating representations for deep one-class classification. In *International Conference on Learning Representations*, 2021. 1

[33] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2

[34] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 4, 8

[35] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection. In *British Machine Vision Conference*, 2021. 1

[36] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3, 5, 6

[37] Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1

[38] Jie Yang, Yong Shi, and Zhiquan Qi. Dfr: Deep feature reconstruction for unsupervised anomaly segmentation. *arXiv preprint arXiv:2012.07122*, 2020. 1

[39] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 4, 8

[40] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Asian Conference on Computer Vision*, 2020. 1

[41] Qiang Zhou, Weize Li, Lihan Jiang, Guoliang Wang, Guyue Zhou, Shanghang Zhang, and Hao Zhao. Pad: A dataset and benchmark for pose-agnostic anomaly detection. *arXiv preprint arXiv:2310.07716*, 2023. 2

[42] Yixuan Zhou, Xing Xu, Zhe Sun, Jingkuan Song, Andrzej Cichocki, and Heng Tao Shen. Vq-flow: Taming normalizing flows for multi-class anomaly detection via hierarchical vector quantization. *arXiv preprint arXiv:2409.00942*, 2024. 8

[43] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. *arXiv preprint arXiv:2207.14315*, 2022. 2, 4, 8