PTAD: Pose and Texture Agnostic Anomaly Detection

Wei Zhuo[1,2*] Jianen Xiang[1,2] Miaomiao Liu[3] Huajun Lu[1,2]
School of Artificial Intelligence, Shenzhen University, Shenzhen 518060, China[1]
National Engineering Laboratory of Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, China[2]
Australian National University[3]
weizhuo@szu.edu.cn {2410673035, 2400671004}@mails.szu.edu.cn miaomiao.liu@anu.edu.au

Abstract

Anomaly detection based on images is a core task in industrial inspection and defect recognition. Recent methods perform well in controlled environments with pose alignment and texture consistency. However, when faced with real-world conditions where poses and surface textures vary, existing methods degrade significantly. This paper introduces the pose and texture-agnostic anomaly detection problem, which generalizes the pose-unknown setting by using a texture-free 3D reference as a geometric prior to detect and localize structural anomalies from arbitrary viewpoints. Unlike training feature embedding based on large-scale multi-view images [4], we propose a novel render-for-detect framework that targets a challenging setting of inspecting geometric anomalies depending on texture-less references. Specifically, we first reconstruct a 3DGS from texture-less references in anomaly detection and render the target pose image from mask-guided pose estimation with differential rendering. Unlike prior single-modal detection, an efficient multi-modal anomaly detection network is introduced, that is proven boundary-sensitive. Benefiting from the general embedding, we can gain competitive results with boundary feature improvements and only using 20% training data, making the framework more feasible in real application scenarios.

1. Introduction

Anomaly detection [7, 21, 24] aims to identify samples that deviate from normal patterns. Due to the scarcity and unpredictability of abnormal samples, most research has focused on unsupervised anomaly detection, which learns the distribution of normal samples in the image or feature space and regards regions deviating from this distribution as anomalies. Current unsupervised methods [3, 40] typically assume that the testing and training images are well aligned and that the same object instance exhibits only minor variations in texture or shape. However, such assumptions severely constrain the capability of detecting anomalies under arbitrary poses, and limit the generalization ability when facing real-world scenarios with complex geometry and appearance. To overcome these limitations, Conditional Anomaly Detection (CAD) [4, 7, 41] is proposed. Specifically, pose-agnostic anomaly detection [18, 25, 41] provides greater flexibility across viewpoints, while texture-agnostic

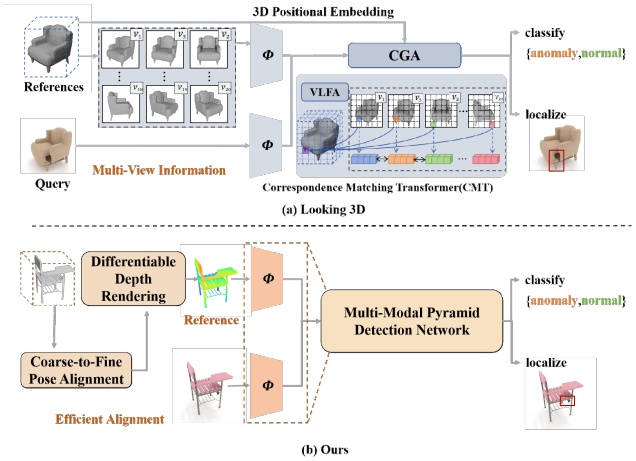

Figure 1. **Comparison between Our Framework and Correspondence Matching Transformer (CMT) [4].** (a) Looking 3D relies on multi-view reference images and a cross-attention network with 3D position encoding to enforce view-consistent embeddings for geometric correspondence, which is computationally intensive and requires large-scale training. (b) Our method simplifies conditional anomaly detection by replacing multi-view alignment with efficient differentiable pose retrieval and depth rendering from a single texture-less 3D model. This makes our cross-modal detection easy to train and data-efficient, even in the presence of significant texture and pose variations.

*Corresponding author

anomaly detection [7] enhances generalization under varying surface appearances. In practical scenarios, anomalies may occur from any viewpoint. For products such as furniture, functionality-oriented anomalies (e.g., missing backrest) are of greater concern than differences in texture or material. This makes the CAD significantly valuable in realistic applications.

Despite the extensive studies of unimodal anomaly detection methods [9, 10, 19, 30, 38], they are often ill-suited for the conditional anomaly detection scenarios, due to their appearance-based nature of modeling textural features for reconstruction or statistical estimation of normal objects. Similar to traditional unimodal paradigms, multimodal anomaly detection methods can be primarily categorized into embedding-based [1, 14, 34] and reconstruction-based [20, 39] approaches. The former fuses image and point cloud features via cross-modal attention to measure distances from normal prototypes, while the latter learns joint geometry–appearance distributions and defines anomalies by reconstruction errors. These methods strongly rely on the assumption of data alignment, and by fusing appearance and the object geometry into a unified representation, they struggle to distinguish benign texture variations from genuine geometric anomalies.

PAD [41] first reconstructs a NeRF [27] model from multi-view texture-rich images and renders the target pose image for anomaly detection. Upon it, both the reference and query images are imported to a pretrained network to extract texture features for similarity calculation and anomaly detection. However, this method is still based on texture correspondence. Beyond it, Looking 3D [4] is introduced to solely address geometric anomalies. This setting stems from the manufacturing practices that models are often designed without textures to allow for the subsequent application of arbitrary surface finishes. Specifically, this method [4] introduces the BrokenChairs-180K dataset, containing about 180k images with five anomaly types. For anomaly detection between the texture-less references and texture-rich query images, a 3D position-encoded cross-attention network is trained on large-scale images to ensure that embeddings from different perspective views are consistent if they correspond to the same geometric position.

For the pose- and texture-agnostic anomaly detection, current conditional anomaly detection methods still face the following challenges: 1) Matching a query with references in multi-view feature alignment methods like [4] is computationally intensive. This stems primarily from the difficulty in establishing correspondence for an absolute position across distinct views, while enforcing embedding consistency among them despite appearance differences; 2) An effective strategy to circumvent the need for multi-view feature alignment training is to directly align the pose between the query and reference. However, significant textural differences between the query and reference make such alignment and subsequent detection difficult to achieve; 3) Existing methods [18, 41] are typically designed for scenarios where training and testing data share similar object types (e.g., the same LEGO instance). In contrast, our setting involves testing on fine-grained object categories with markedly different geometric structures. This domain shift poses a significant generalization challenge for approaches.

To this end, we propose a render-for-detect method for conditional anomaly detection. This task requires only texture-less references and aims to detect geometric anomalies on texture-rich queries captured at arbitrary pose. To reduce the computational cost, we propose to simplify the full multi-view feature consistency training [4] to single-view similarity computation. By transitioning from using multi-view images as references to employing 3D Gaussian Splatting (3DGS) representation, our method realizes differentiable query pose retrieval and reference depth rendering on the retrieved pose. As a result, we can directly apply the cross-modal matching between the query image and the single rendered reference view. Notably, the proposed pose retrieval and rendering process is highly parameter-efficient and fully differentiable. In addition, by solving the pose alignment, the subsequent detection network is relieved from learning complex viewpoint variations, enabling training with far fewer images.

Specifically, realizing this pipeline faces two major challenges: significant textural differences and the challenge of reconstructing texture-less objects. To address this, we treat each Gaussian primitive as a disk floating on the surface, enabling a direct transformation from the mesh. Upon it, a coarse-to-fine pose retrieval strategy is proposed, where the pose is first initialized using the chamfer distance and then refined through differential optimization under the guidance of a mask. The referring geometry of the depth image on the query pose is rendered for subsequent matching and anomaly detection. Due to the modality gap between the depth maps and the query images, we adopt a shared feature encoding strategy between the two modalities. The detection network can be pushed to learn general embeddings, such as boundaries, to better understand images from both modalities. An interesting observation is that our method outperforms a uinmodal network that adopts depth generation tools [36], proving the effectiveness and generality of the proposed cross-modal detection network.

In summary, our main contributions are as follows:

- We build a novel and efficient framework for conditional anomaly detection via rendering for detection.
- We are the first to introduce the idea of converting texture-less models into 3DGS representations with differentiable rendering for anomaly detection, motivated by the challenge of cross-model view alignment.
- We propose a coarse-to-fine pose alignment method

with chamfer-based matching and mask-guided refinement, to effectively address the challenges of pose estimation under texture-less conditions.
- We propose a lightweight and efficient anomaly detection network. On the BrokenChairs-180K dataset, with only 10% of training samples, our method achieves accuracy comparable to state-of-the-art methods, demonstrating strong generalization and practical potential.

2. Related Work

2.1. Industrial Anomaly Detection

In industrial manufacturing and quality inspection, anomaly detection aims to automatically identify and localize defects under the condition of limited or even no defective samples [2, 9, 24, 31]. Traditional unimodal anomaly detection typically relies on large quantities of normal samples for training and identifies anomalies via reconstruction errors [10, 38] or feature deviations [8, 9]. However, these methods struggle to capture complex defects such as structural deformations or assembly inconsistencies. Recent studies have explored various directions for improvement, including memory-based feature contrast [30], self-supervised reconstruction for anomaly localization [19, 38], and synthetic anomaly generation for data enhancement [5, 17, 29]. Yet, they still suffer from heavy memory consumption, excessive sensitivity to texture variations, and limited generalization caused by unrealistic defect patterns. Similarly, existing multimodal anomaly detection approaches can be broadly categorized into embedding-based and reconstruction-based paradigms. They align information across modalities through feature fusion and compute anomaly scores based on feature-space distances or reconstruction discrepancies. However, their performance degrades sharply when confronted with texture-sparse regions or unknown poses. The PAD task [18, 41] first introduced differentiable pose optimization to explicitly align query images, yet it remains heavily dependent on texture cues. Looking 3D [4] further extended this idea to texture-less scenarios, but it demands large-scale datasets to establish view-consistent embeddings within a unified space—resulting in prohibitive computational overhead. To overcome these limitations, we propose a unified framework that integrates 3D representations with a multimodal detection network, effectively reducing dependency on texture information and computational burden of detection.

2.2. Pose Estimation in 3DGS

The PAD task first introduced pose estimation into the field of anomaly detection, relying on implicit neural representations[27] and the iNeRF[37] framework to achieve novel view pose estimation. However, NeRF requires substantial computational resources and suffers from low rendering efficiency, making training and inference impractical for real-time anomaly detection. With the advent of 3DGS, which replaces implicit volume representations with explicit learnable Gaussian primitives for efficient rendering, this paradigm has inspired a series of 3DGS-based pose-agnostic anomaly detection methods [18, 25]. While these methods can optimize camera pose given a query image, they differ fundamentally from our implementation: they heavily rely on texture information and photometric loss, and require dense multi-view sampling. In contrast, our approach employs mask-guided pose optimization and additionally overcomes the challenge of reconstructing texture-less objects. Finally, there are 3DGS techniques that optimize camera poses in SLAM[15, 26, 35] setting. However, these methods are typically applied to video sequences, where the pose of a new frame is initialized from the previous one. Moreover, they are not designed to handle scenarios with significant texture variations.

3. Method

In this section, we introduce an overview of the render-for-detect framework, as illustrated in Fig. 2. Sec. 3.1 defines the task of conditional anomaly detection in detail. Sec. 3.2 describes 3DGS-based pose alignment and depth rendering. Finally, in Sec. 3.3, we introduce a multi-modal detection network built upon a shared-weight ResNet18 backbone.

3.1. Problem Definition and Overview

The definition of conditional anomaly detection follows the setting of Looking 3D [4]. This task aims to identify and localize shape anomalies in a query image that deviate from a reference 3D model, based on a pose-agnostic query image and a texture-less reference 3D model. Specifically, we denote $q \in \mathbb{R}^{H\times W\times 3}$ as a query image captured from an arbitrary pose, which may contain variations in texture and material. The reference set $V = \{(v_n, P_n)\}$ denotes a collection of rendered images v_n and their corresponding camera poses P_n, obtained from the reference (defect-free) 3D model under hemispherical or uniformly sampled poses. The reference model is a texture-less normal instance, which is converted into a 3DGS representation using *mesh2splat*[32] tool, enabling the rendering of corresponding geometric views including masks and depth maps.

Unlike the reference set, the pose of the query image is unknown and may fall outside the coverage of the reference pose set. This leads to a key issue: although about 20 reference poses are available, the query image may be captured from arbitrary poses. Directly using all rendered images from the 20 poses for matching would introduce significant parallax interference and texture inconsistency noise. Moreover, the differences in pose as well as possible variations in texture and illumination make it challenging to directly find a reference image that is "sufficiently similar" to the query image within the reference set, which in

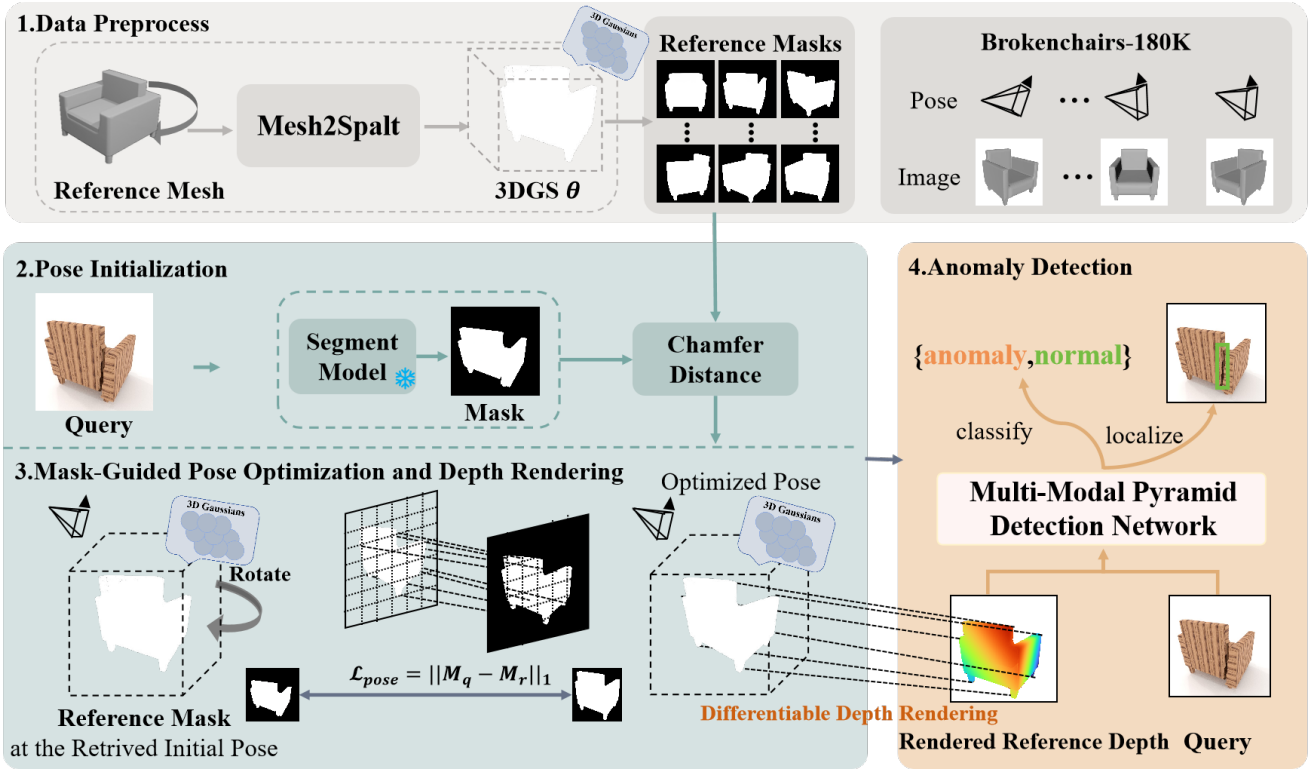

Figure 2. **Overview of Our Framework.** The framework consists of four stages: (1) We first convert the texture-less 3D model into a learnable 3DGS representation Θ using the *mesh2splat* tool, and render reference masks M_v from the reference pose. (2) According to the chamfer distance between the reference mask M_v and the query mask M_q, the best-matching candidate pose is selected as the initial pose $T^{(0)}$. (3) The initial pose $T^{(0)}$ is iteratively optimized using the L_1 loss $\mathcal{L}_{pose}$ between the query mask M_q and the rendered mask M_r, to obtain the target pose $T^{(k)}$. The estimated pose is then used to generate the reference depth map I_r. (4) A retrained ResNet18 network is employed to compare the multi-scale similarities between the query and reference image, which are then fused in a bottom-up manner for anomaly detection.

turn increases the difficulty of subsequent anomaly detection and localization. Therefore, the first step of the task is to gain the reference image rendered from the pose most similar to the query image in terms of shape or geometry, enabling more accurate pose refinement and anomaly comparison under similar viewing conditions.

3.2. Coarse-to-Fine Pose Alignment

Traditional 3D reconstruction methods, such as NeRF[27], model color and density in continuous space through implicit volumetric rendering. Although they can achieve high-fidelity novel view synthesis, the computational complexity of volume rendering limits their real-time performance and scalability in industrial inspection tasks. To overcome the rendering bottleneck of NeRF, 3DGS[16] addresses this by replacing implicit density fields with a learnable set of 3D Gaussians:

$$G_i = (\mu_i, \Sigma_i, c_i, \alpha_i), \tag{1}$$

where μ_i denotes the 3D spatial position, Σ_i denotes the covariance matrix, c_i denotes the color, and α_i denotes the opacity. These Gaussians are projected onto screen space and α-blended for fast, high-quality rendering:

$$C_{pixel} = \sum_{i \in N} c_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j). \tag{2}$$

Here, C_{pixel} is the accumulated pixel color along the ray, and N represents the set of all Gaussians intersected by the ray. c_i is the color of the i^{th} Gaussian. The term $\prod_{j=1}^{i-1}(1 - \alpha'_j)$ is the accumulated transmittance along the ray.

Data Preprocess. Although 3DGS is an emerging technique for 3D scene representation and novel view synthesis, reference models in industrial scenarios are typically texture-less meshes. In regions with sparse or nearly absent textures, 3DGS produces inaccurate geometric and surface reconstructions [33]. Specifically, this limitation arises from insufficient color cues, which fail to effectively guide Gaussian distributions to align with the true surface geometry. To address this issue, we adopt *mesh2splat*, which directly converts triangular meshes into 3DGS rep-

resentations. Specifically, each triangle is projected onto an axis-aligned plane via triplanar orthogonal projection. The Jacobian from UV space to 3D space is used to derive anisotropic Gaussian scales along the surface tangential directions, while the normal-direction scale is fixed to a small constant to enforce a thin surface approximation. During rasterization, Gaussian parameters are interpolated across the triangle, effectively generating one 3D Gaussian per fragment. Each Gaussian is defined by its position, anisotropic scale, and orientation, enabling efficient mesh-to-3DGS conversion without iterative optimization.

Pose Initialization. Before pose optimization, it is necessary to provide an appropriate initial camera pose for the query image. Since the reference model lacks texture, we adopt the geometric consistency principle and measure the set difference between the reference mask and the query mask using the chamfer distance. Specifically, the reference mask is rendered from the 3DGS representation under an initial pose, while the query mask is obtained through threshold-based segmentation. The chamfer distance is defined as:

$$\begin{aligned}\mathcal{L}_{cd}(M_q, M_v) = &\frac{1}{|M_q|}\sum_{x\in M_q}\min_{y\in M_v}\|x-y\|_2 \\ &+\frac{1}{|M_v|}\sum_{y\in M_v}\min_{x\in M_q}\|x-y\|_2,\end{aligned} \tag{3}$$

where M_q and M_v denote the pixel coordinate sets in image space of the query and reference masks, respectively. The chamfer distance between two masks, M_q and M_v, is calculated by summing the average shortest distance from all pixels in M_q to M_v and the average from M_v to M_q. Among all candidate poses[1], the one with the smallest distance is selected as the initialization, providing a stable starting point for subsequent optimization (see Fig.3).

Mask-Guided Pose Optimization and Depth Rendering. To address the challenge of accurate pose optimization caused by texture-less 3DGS, we introduce a fully differentiable pipeline that integrates mask-guided pose estimation with 3DGS-based depth rendering, providing geometric references that are optimally aligned with the pose of the query image for subsequent anomaly detection.

To achieve precise alignment between the query image and the reference model, we keep the camera fixed at the initial pose, which is achieved via the prior process of *pose initialization*, and then transform the reference Gaussian. Formally, a transformation is defined as $T \in SE(3)$, which consists of a rotation $R \in SO(3)$ and a translation vector $t \in \mathbb{R}^3$. To ensure continuity and differentiability, we parameterize the transformation in the Lie algebra space using a screw axis $S = (\omega, \boldsymbol{v}) \in \mathbb{R}^6$ and a rotation angle θ. Based

[1]We use the same 20 multi-view poses as in [4].

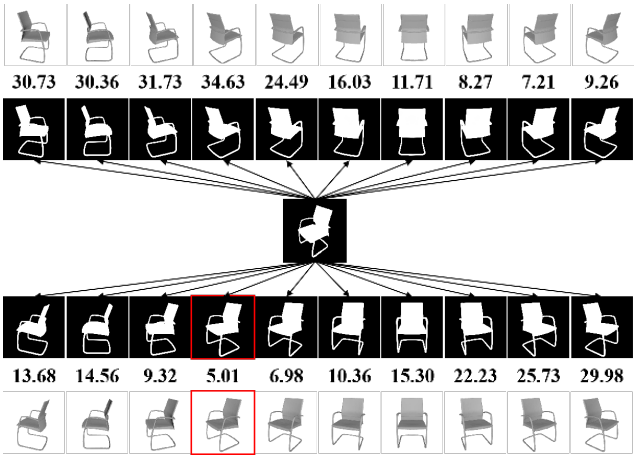

Figure 3. **Visualization of Chamfer Distances.** This illustration shows the chamfer distances between the query mask and 20 reference masks with their corresponding reference images. The bold numbers indicate the computed distances, and the reference pair with the smallest distance is highlighted with a red bounding box.

on Rodrigues' rotation formula, the rotation matrix can be expressed as:

$$R = Rot(\omega, \theta) = I + \sin\theta[\omega] + (1-\cos\theta)[\omega]^2, \tag{4}$$

where $[\omega]$ denotes the skew-symmetric matrix of ω. Furthermore, the full transformation T can be parametrized:

$$T = \begin{bmatrix} Rot(\omega,\theta) & t \\ 0 & 1 \end{bmatrix}, \tag{5}$$

where,

$$t = F(\theta)\boldsymbol{v} = (I\theta + (1-\cos\theta)[\omega] + (\theta - \sin\theta)[\omega]^2)\boldsymbol{v}. \tag{6}$$

This Lie algebra-based parameterization can be directly applied to 3D Gaussians, enabling precise alignment between the query image and the reference 3D model.

Unlike SplatPose[18], which leverages photometric consistency loss between the rendered image and the query image during pose estimation, our method employs a binarized object mask as the optimization guidance signal. This design provides a more robust geometric constraint, especially under varying textures and lighting conditions. Specifically, we apply the transformation T to the reference Gaussian model and render a predicted mask M_r. The pose optimization objective is defined by minimizing the L_1 distance between the rendered mask M_r and the query image mask M_q:

$$\mathcal{L}_{\text{pose}} = ||M_r - M_q||_1. \tag{7}$$

This mask-based loss focuses purely on geometric alignment, enabling the optimization to be invariant to appearance differences such as texture or illumination. The reference mask is rendered from the transformed 3DGS using

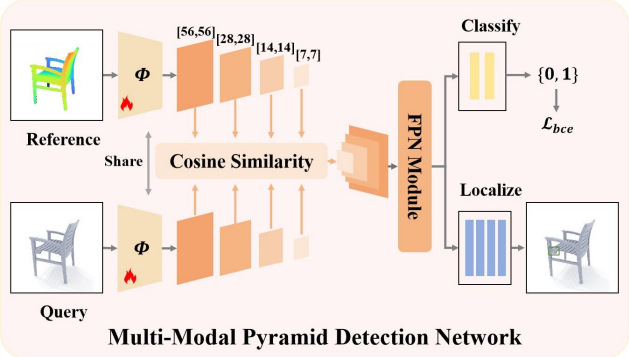

Figure 4. **Overview of the Proposed Multi-Modal Pyramid Detection Network (MMPDN).** The reference and query images are encoded by a shared ResNet18 backbone. Multi-scale cosine similarity maps are upsampled and fused by the FPN module for anomaly classification and localization.

Eq.2. Therefore, the loss allows gradients to flow from the opacity values in Eq.1 to the Jacobian matrix, and subsequently back-propagate to the transformation parameters T. After the pose optimization converges, we obtain the optimal alignment between the reference model and the query image in 3D space.

Leveraging the *differentiable* rendering property of 3DGS, we then render the corresponding depth map from this optimized pose:

$$D_{pixel} = \sum_{i \in N} d_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j). \tag{8}$$

Eq. (8) is derived from the principle of volumetric rendering, where the final depth value is computed by accumulating the weighted contributions of all visible 3D Gaussians along each camera ray. Specifically, d_i denotes the z-coordinate of the center of the i^{th} Gaussian on the current pixel ray in the camera coordinate system. In this manner, the model maintains differentiability while achieving hierarchical depth fusion, resulting in a continuous and physically consistent depth map.

As a result, our method introduces a mask-guided pose optimization strategy, effectively mitigating optimization noise caused by photometric inconsistencies. Combined with the proposed *mesh2splat* conversion module, this design forms a closed-loop system that enables stable and accurate geometric alignment in texture-less, illumination-varying, and pose-unknown industrial scenarios.

3.3. Multi-Modal Pyramid Detection Network

Although the depth maps rendered by 3DGS can effectively provide geometric references, establishing a reliable correspondence across different modalities remains a key challenge in anomaly detection. We propose a geometry-aware multi-scale similarity fusion network (see Fig.4) that

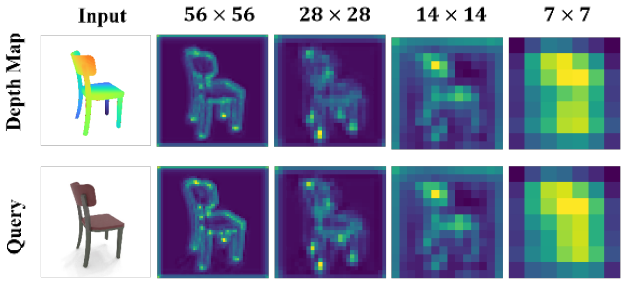

Figure 5. **Visualization of Multi-Scale Feature Maps.** In this figure, we show the strong boundary extraction capability of our retrained ResNet18.

learns structural consistency between the query image and the rendered depth within a shared feature space, enabling the identification of fine-grained geometric anomalies and structural deviations. This module not only achieves robust anomaly detection but also significantly improves classification accuracy. Conventional detection methods typically rely on single-scale features or pixel-wise error metrics, making them highly sensitive to geometric perturbations and false anomalies caused by illumination or texture variations. In contrast, our approach reformulates the detection task as cross-modal geometric consistency learning: within a shared feature space, cosine similarity is calculated to measure the structural consistency between the reference depth rendering and the query image, thereby capturing geometric anomalies while suppressing the influence of texture. Furthermore, to address anomaly patterns at different spatial scales, we incorporate a Feature Pyramid Network[23] (FPN) to aggregate multi-level similarity maps, enabling the model to maintain robustness across both local anomalies (e.g., missing parts) and global deviations (e.g., structural misalignment).

Feature Extraction. We propose a geometry-aware multi-scale similarity fusion network, which takes as input the query image I_q and the rendered reference depth map I_r, where the depth of each pixel is achieved by Eq.8. This network outputs an image-level anomaly score y and bounding boxes for candidate abnormal regions. We adopt a retrained ResNet18 [13] as the feature encoder, which produces four hierarchical feature maps, $\{F^1, F^2, F^3, F^4\}$ from $\Phi(I)$, where Φ denotes the encoder and $F^i \in \mathbb{R}^{C_i \times H_i \times W_i}$ represents the feature map at the i-th block. Like a siamese network [6], our encoder shares weights between the query and reference branches to ensure geometric consistency in the learned feature space. This allows the cosine similarity to directly reflect structural discrepancies. Accordingly, the feature maps of the reference image and query image, i.e., F_r^i and F_q^i, are obtained by $\Phi(I_r)$ and $\Phi(I_q)$, respectively.

Multi-Scale Similarity Map Fusion. At each feature level in Fig.5, we compute the normalized cosine distance

between the query and reference features:

$$D^i = 1 - \frac{F_q^i \cdot F_r^i}{||F_q^i||_2||F_r^i||_2}, \tag{9}$$

which measures the geometric and semantic consistency between the two images. When the corresponding regions share similar shapes or boundary structures, D_i approaches zero; in contrast, geometric anomalies such as missing parts, rotations, or misalignments lead to a significant increase in D_i. Single-layer features are insufficient to represent anomalies across different spatial scales. Therefore, we adopt FPN to progressively integrate low-level local details with high-level structural semantics. This hierarchical fusion enables full-scale perception from geometry to semantics, allowing the model to maintain stable responses even under complex shape variations. Upon it, we design a classification head, implemented as a two-layer MLP, to perform anomaly discrimination by aggregating global features to predict the final image-level anomaly score.

4. Experiments

This section details the experimental setup and implementation, followed by ablation studies on the key components of the proposed render-for-detect framework, and concludes with a comparison against state-of-the-art anomaly detection methods to validate the effectiveness of our approach.

4.1. Experimental Setup

Dataset. We use the BrokenChairs-180K dataset [4], which is constructed from the chair category of PartNet[28]. It contains 8,143 3D chair models, each rendered into several pose-agnostic query images, totaling approximately 180K images. In addition, each 3D model is rendered from 20 multi-view reference images with associated camera parameters to support geometric alignment learning. The dataset defines five types of structural anomalies: Positional, Rotational, Broken, Swapped, and Missing.

Implementation Details. For each chair instance, we preset N = 20 candidate poses and resize input images to 256×256. Reference models are converted from mesh to 3DGS using *mesh2splat* with a Gaussian scale of 0.65 and sampling density of 0.5. Pose estimation applies $k = 150$ iterations of differentiable optimization to obtain the optimal pose. The optimized 3DGS is rendered into depth maps as geometric priors for the detection module. The detection network is based on ResNet18, extracting multi-layer features and computing cross-layer cosine similarity, which are fused into an anomaly map. During training, the model is trained for 20 epochs on single NVIDIA RTX 5880 Ada Generation with a batch size of 16, using the Adam optimizer with a learning rate of 2×10^{-5}.

Table 1. Quantitative comparison of the proposed framework on anomaly classification with baselines in terms of area under the ROC curve (AUC) and accuracy score. For both metrics, higher values indicate better performance. The results highlighted in bold represent the best performance.

Methods	**AUC (%)** (↑)	**Accuracy (%)** (↑)
ResNet18-FPN [23]	74.6	64.7
ResNet18-FPN w/ SA blocks	75.2	65.1
Vision Transformer (ViT) [11]	75.4	65.2
LFD [12]	-	64.9
Lim et al. [22]	-	67.8
Looking 3D [4]	84.7	75.4
SplatPose [18]	86.7	79.4
Ours	**91.0**	**83.0**

Table 2. Ablation study of the CFPA and MMPDN Modules. Results are reported in AUC and Accuracy.

	CFPA	**MMPDN**	**AUC (%)** (↑)	**Accuracy (%)** (↑)
A	×	×	76.3	67.1
B	×	✓	75.5	66.6
C	✓	×	79.1	71.0
D	✓	✓	**91.0**	**83.0**

4.2. Experimental Results

We report the quantitative results using two evaluation metrics: the area under the ROC curve (AUC) and accuracy score in Tab. 1, and provide qualitative results in Fig. 6.

Comparison with Related Work. As shown in Tab.1, our method achieves the best overall performance among all baselines. Traditional 2D-based networks (ResNet18-FPN [23], ViT [11]) rely solely on the query image for anomaly detection and perform poorly due to the lack of explicit 3D model references. In contrast, our render-for-detect framework improves the baseline by more than 16%. Existing methods [12, 22] retrieve the best-matching 3D model based on feature-distance matching, while we perform anomaly classification by computing the distance on shape embeddings. As shown in Tab.1, These methods struggle to detect subtle geometric anomalies, mainly because fine-grained local alignment is difficult to establish. Looking 3D [4] leverages unsupervised cross-view feature alignment, but its global contrastive objectives limit its ability to detect fine-grained geometric changes, resulting in lower accuracy. We adapt SplatPose [18] to this task to achieve accurate pose alignment. However, when texture is absent, the method produces large feature discrepancies, which introduce unnecessary noise and interference. In contrast, our pose alignment and differentiable depth rendering provide precise geometric references, while the lightweight cross-modal network learns general embeddings that bridge the modality gap.

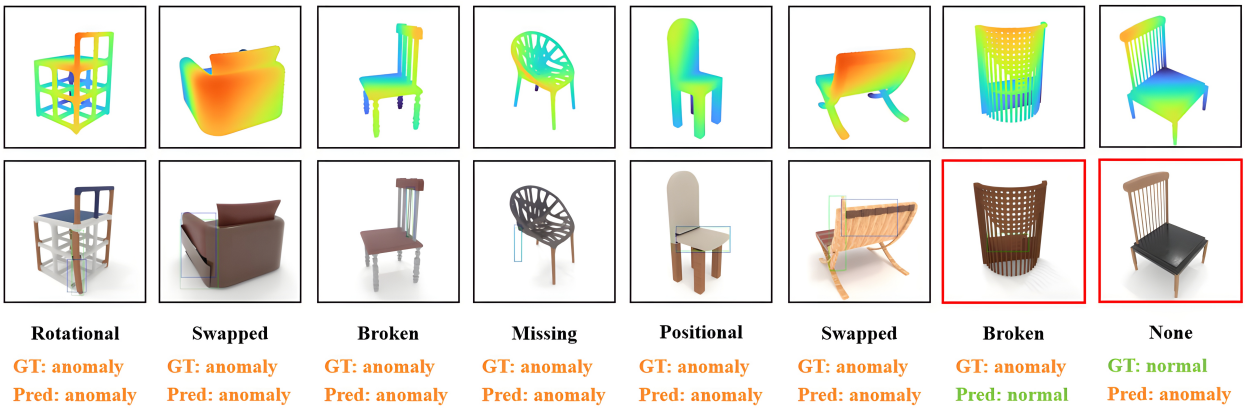

Figure 6. **Qualitative Visualization.** Here we show the anomaly detection results on the test set achieved by our framework. From top to bottom is the reference depth map and the query image, respectively. Red indicates misclassified predictions.

Table 3. Ablation study on pose alignment.

Pose Initialization	Pose Optimization	AUC (%) (↑)	Accuracy (%) (↑)
✓	×	87.0	78.4
✓	✓	91.0	83.0

Ablation Studies. We conduct ablation studies to evaluate the contribution of each component: Coarse-to-Fine Pose Alignment (CFPA) and Multi-Modal Pyramid Detection Network (MMPDN). Our first *baseline A* includes none of our two components but the Correspondence-Guided Attention (CGA) component with cross-attention from Looking 3D, which matches references and the query based on all local features. That is also a baseline provided by [4]. *Baseline B* replaces the transformer-based detection network in Baseline *A* with our proposed MMPDN. However, replacing the transformer with MMPDN alone does not improve performance, as the multi-view similarity maps introduce noise that weakens anomaly discrimination. *Baseline C* builds upon A by integrating CFPA, which significantly improves accuracy by reducing redundant reference views and enforcing pose-consistent feature alignment. *Baseline D* combines CFPA with MMPDN, yielding the largest performance gain. This confirms the strong complementarity between CFPA and MMPDN, that CFPA ensures pose-consistent geometric correspondence, while MMPDN performs multi-scale anomaly discrimination. When removing the pose optimization from the *Baseline D*, the performance severely degrades. This underscores the necessity of accurate pose calibration for reliable reference–query matching.

Anomaly Localization. Building on the anomaly detection network, we introduce a bounding-box regression branch for anomaly localization. This branch adopts a 4-layer MLP regression head and is jointly optimized with the rest of the network using L_1 loss and generalized IoU loss. The method achieves a mean IoU of 47.8%, indicating that the detection module beyond global anomaly discrimination, can effectively capture spatial anomaly patterns.

Table 4. Ablation study on the effect of training sampling ratio.

Sampling Ratio	AUC (%)	Accuracy (%)
5%	82.1	73.9
10%	83.9	75.6
20%	84.9	76.5
50%	88.8	80.2
80%	90.2	81.9
100%	91.0	83.0

Sensitivity to Training Sampling Ratio. To evaluate the sensitivity of our model to the scale of training data, we conducted experiments under different sampling ratios, and the results are presented in Tab.4. When using smaller training data, model performance exhibits a gradual and stable decline. When trained with only 50% of the training samples, the accuracy decreases by merely 2.8%, demonstrating that the proposed method maintains robust generalization under limited data conditions. At a 20% ratio, the degradation becomes more noticeable, yet the accuracy remains comparable to Looking 3D. When the sampling ratio drops below 10%, the model performance shows a notable decline compared with the full-data setting, indicating that its discriminative capability is somewhat affected under extremely limited training samples. Overall, these results demonstrate that our method maintains strong robustness and stability even in low-data regimes.

5. Conclusion

In this paper, we have introduced a novel render-for-detect framework for conditional anomaly detection. Despite these advantages, our method has several limitations. In rare cases where the missing regions are excessively large, the pose optimization becomes difficult to converge, resulting in inaccurate anomaly localization.

Acknowledgments.

This project is supported by National Key R&D Program of China under Grant 2024YFF0618403, National Natural Science Foundation of China under Grant 62306183, Guangdong Natural Science Foundation under Grant 2024A1515010194, Guangdong-Macao Science and Technology Innovation Joint Funding Program under Grant 2024A0505090003, Shenzhen Natural Science Foundation under Grant JCYJ20240813141807010, the Special Fund for the Cultivation of Independent Innovation Achievements of Postgraduate Students at Shenzhen University (315-000066010846), and the Intelligent Computing Center of Shenzhen University.

References

[1] Paul Bergmann and David Sattlegger. Anomaly detection in 3d point clouds using deep geometric descriptors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2613–2623, 2023. 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. 3

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020. 1

[4] Ankan Bhunia, Changjian Li, and Hakan Bilen. Looking 3d: Anomaly detection with 2d-3d alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17263–17272, 2024. 1, 2, 3, 5, 7, 8

[5] David Breuss, Karel Rusỳ, Maximilian Götzinger, and Axel Jantsch. Generation of synthetic image anomalies for analysis. In *International Conference on Intelligent Systems and Pattern Recognition*, pages 13–27. Springer, 2024. 3

[6] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993. 6

[7] Shang-Fu Chen, Yu-Min Liu, Chia-Ching Liu, Trista Pei-Chun Chen, and Yu-Chiang Frank Wang. Domain-generalized textured surface anomaly detection. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 01–06. IEEE, 2022. 1, 2

[8] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020. 3

[9] 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, 3

[10] David Dehaene, Oriel Frigo, Sébastien Combrexelle, and Pierre Eline. Iterative energy-based projection on a normal data manifold for anomaly localization. In *International Conference on Learning Representations*, 2020. 2, 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*, 2021. 7

[12] Alexander Grabner, Peter M Roth, and Vincent Lepetit. Location field descriptors: Single image 3d model retrieval in the wild. In *2019 international conference on 3d vision (3DV)*, pages 583–593. IEEE, 2019. 7

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[14] Eliahu Horwitz and Yedid Hoshen. Back to the feature: classical 3d features are (almost) all you need for 3d anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2968–2977, 2023. 2

[15] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21357–21366, 2024. 3

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 4

[17] Hyuntae Kim and Changhee Lee. Enhancing anomaly detection via generating diversified and hard-to-distinguish synthetic anomalies. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1089–1098, 2024. 3

[18] Mathis Kruse, Marco Rudolph, Dominik Woiwode, and Bodo Rosenhahn. Splatpose & detect: Pose-agnostic 3d anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3950–3960. IEEE, 2024. 1, 2, 3, 5, 7

[19] 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 computer vision and pattern recognition*, pages 9664–9674, 2021. 2, 3

[20] Wenqiao Li, Xiaohao Xu, Yao Gu, Bozhong Zheng, Shenghua Gao, and Yingna Wu. Towards scalable 3d anomaly detection and localization: A benchmark via 3d anomaly synthesis and a self-supervised learning network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22207–22216, 2024. 2

[21] Zhuo Li, Yuhao Yan, Xiangheng Wang, Yifei Ge, and Lin Meng. A survey of deep learning for industrial visual

anomaly detection. *Artificial Intelligence Review*, 58(9):279, 2025. 1

[22] Ming-Xian Lin, Jie Yang, He Wang, Yu-Kun Lai, Rongfei Jia, Binqiang Zhao, and Lin Gao. Single image 3d shape retrieval via cross-modal instance and category contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11405–11415, 2021. 7

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 6, 7

[24] Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1):104–135, 2024. 1, 3

[25] Yizhe Liu, Yan Song Hu, Yuhao Chen, and John Zelek. Splatpose+: Real-time image-based pose-agnostic 3d anomaly detection. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 378–391. Springer, 2024. 1, 3

[26] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18039–18048, 2024. 3

[27] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3, 4

[28] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019. 7

[29] Adin Ramirez Rivera, Adil Khan, Imad Eddine Ibrahim Bekkouch, and Taimoor Shakeel Sheikh. Anomaly detection based on zero-shot outlier synthesis and hierarchical feature distillation. *IEEE Transactions on Neural Networks and Learning Systems*, 33(1):281–291, 2020. 3

[30] 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, 3

[31] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021. 3

[32] Stefano Scolari. Mesh2splat: Fast mesh to 3d gaussian splat conversion. https://github.com/electronicarts/mesh2splat, 2025. Extended and updated version of the author's Master's thesis at KTH. 3

[33] Jiepeng Wang, Yuan Liu, Peng Wang, Cheng Lin, Junhui Hou, Xin Li, Taku Komura, and Wenping Wang. Gaussurf: Geometry-guided 3d gaussian splatting for surface reconstruction. *arXiv preprint arXiv:2411.19454*, 2024. 4

[34] 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

[35] Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19595–19604, 2024. 3

[36] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 2

[37] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330. IEEE, 2021. 3

[38] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draema discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8330–8339, 2021. 2, 3

[39] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Cheating depth: Enhancing 3d surface anomaly detection via depth simulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2164–2172, 2024. 2

[40] Ye Zheng, Xiang Wang, Rui Deng, Tianpeng Bao, Rui Zhao, and Liwei Wu. Focus your distribution: Coarse-to-fine non-contrastive learning for anomaly detection and localization. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 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. *Advances in Neural Information Processing Systems*, 36:44558–44571, 2023. 1, 2, 3