This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Wavelet-Driven 3D Anomaly Detection under Pose-Agnostic and Sparse-View

Mingwen Shao[1,2] Qiao Zhang[2*] Xinyuan Chen[2] Xiang Lv[2] Lingzhuang Meng[2]
Chang Liu[2] Qinglin Zhan[2] Ling Jian[3]
[1] Artificial Intelligence Research Institute, Shenzhen University of Advanced Technology
[2] Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software, Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China)
[3] School of Economics and Management, China University of Petroleum (East China)
smw278@126.com, bz24070008@s.upc.edu.cn, b24070011@s.upc.edu.cn, lvxiang1997@126.com, lzhmeng1688@163.com, upc_liuchang@163.com, s24070001@s.upc.edu.cn, bebetter@upc.edu.cn

Abstract

Pose-agnostic anomaly detection (PAD) achieves strong performance in localizing anomalies from arbitrary viewpoints when trained on densely sampled normal data. However, under sparse-view conditions, existing methods face two key challenges: (1) sparse observations lead to overfitting and geometric detail loss in 3D reconstruction; (2) limited visual cues lead to inaccurate pose estimation, compromising the reliability of subsequent anomaly localization. To address these challenges, we propose ***Wave-Pose3D****, a wavelet-driven 3D anomaly detection framework tailored for PAD under sparse-view conditions. First, we design a structure-aware and wavelet-optimized Gaussian modeling strategy that dynamically filters unreliable regions via structural priors to mitigate overfitting and leverages high-frequency supervision to restore fine-grained geometric details. Second, to improve pose estimation under sparse views, we develop a wavelet-based pose estimator that integrates low-frequency structural cues and high-frequency details to enhance both initialization and refinement accuracy. Finally, we introduce a wavelet difference-aware anomaly detector that computes frequency-domain anomaly scores, improving localization robustness against pose and geometric variations. By integrating these strategies, Wave-Pose3D achieves robust and accurate anomaly localization under sparse views. Extensive experiments validate that the proposed approach achieves state-of-the-art performance under 10% and 20% sparse-view configurations.*

1. Introduction

Visual anomaly detection automatically identifies image or video regions that deviate from expected patterns, with significant research value and broad applications in defect inspection [16, 20, 33, 34, 38], medical analysis [1, 21], video surveillance [2, 7, 25], and autonomous driving [24, 36].

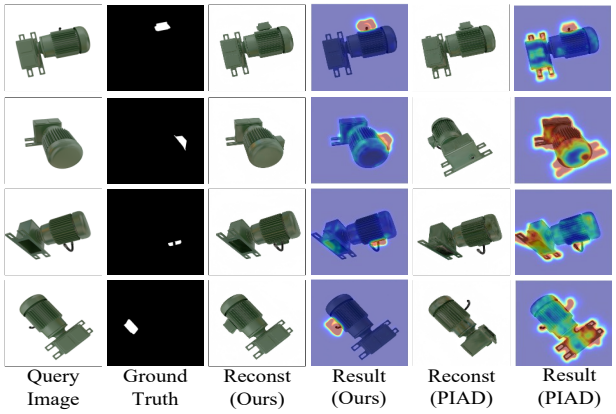

Figure 1. Compared with the spatial domain method PIAD [29], our frequency domain modelling approach (Wave-Pose3D) generates more complete reconstructions, more accurate pose estimates, and more precise anomaly localization under sparse views.

Most existing 2D anomaly detection methods [4, 5, 14, 15, 27, 30, 32, 35, 41] assume consistent poses between training and testing images, an assumption often violated in real-world industrial scenarios. Consequently, their performance degrades significantly under unseen pose variations. To address this limitation, Zhou et al. [40] introduced pose-agnostic anomaly detection (PAD) and proposed OmniPoseAD, which leverages Neural Radiance Fields (NeRF) [17] to learn 3D representations of normal objects from multi-view observations, enabling detection and localization under unknown poses. However, its reliance on NeRF entails high computational cost and slow convergence, limiting real-time industrial applicability.

Recently, 3D Gaussian Splatting (3DGS) [9] has emerged as an efficient and high-quality 3D representa-

*Corresponding author

tion method, offering advantages in differentiable rendering and advancing PAD. For instance, Kruse et al. [10] proposed SplatPose, a 3DGS-based PAD framework that leverages multi-view inputs for high-fidelity reconstruction and robust anomaly localization. Jiang et al. [8] introduced an inversion strategy for 3DGS to address pose estimation in query images, while Yang et al. [29] extended PAD to illumination-agnostic settings by jointly modeling reflectance and color. Despite their promise, these methods assume dense multi-view inputs, which are costly and often impractical in real-world industrial scenarios. Motivated by this limitation, we focus on the more realistic and challenging task of PAD under sparse views.

To investigate this challenge, we first review the standard PAD paradigm and its limitations under sparse views. Existing PAD methods [10, 29, 40] typically follow three spatial-domain steps: (1) training a 3D representation from multi-view normal images, (2) estimating the query pose via spatial matching to reference views, and (3) computing anomaly scores based on spatial differences. While effective with dense views, this paradigm becomes unreliable under sparse views. As shown in Figure 1, PIAD [29] suffers from severe pose estimation errors under sparse conditions, resulting in rendering artifacts and inaccurate anomaly detection. The degradation stems from sparse training observations, which lead to overfitting and structurally incomplete 3D models, thereby limiting generalization and fine-grained reconstruction. Moreover, sparse views hinder reliable keypoint correspondence, compounding pose errors and further degrading anomaly localization.

To address the above challenges, we propose Wave-Pose3D, a wavelet-driven PAD framework under sparse-view settings. Specifically, to mitigate overfitting and enhance geometric fidelity under sparse views, we propose a Structure-aware Wavelet-optimized Gaussian Modeling (SWGM) strategy. SWGM identifies reliable regions via structural complexity and multi-view depth consistency, selectively regularizing the Gaussian representation to reduce noise, while supervising high-frequency wavelet coefficients to recover fine geometric details. To reduce pose estimation errors under sparse views, we propose a wavelet-based pose estimator (WPE) with two stages: wavelet matching verification (WMV) for initial estimation and wavelet pose optimization (WPO) for refinement. By leveraging multi-scale wavelet decomposition, WPE uses low-frequency components for robust global alignment and high-frequency cues for precise local matching, thereby improving pose stability under sparse viewpoints. Furthermore, we employ a wavelet difference-aware anomaly detector (WDAD) that computes anomaly scores in the frequency domain. Operating on wavelet coefficients, WDAD inherently suppresses artifacts caused by spatial misalignment and disentangles structural inconsistencies from texture noise, enhancing robustness to pose variations and structural ambiguity.

Leveraging these components, our method achieves more complete reconstruction, accurate pose estimation, and improved anomaly localization under sparse-view conditions, as depicted in Figure 1. Our main contributions can be summarized as:

- We propose Wave-Pose3D, a pose-agnostic 3D anomaly detection framework that exploits wavelet transforms to jointly mitigate overfitting, stabilize pose estimation, and enhance detection robustness under sparse views.
- We develop a structure-aware and wavelet-optimized Gaussian modeling strategy tailored for sparse-view scenarios to enhance reconstruction robustness and preserve fine geometric details.
- A wavelet-based pose estimator is designed to leverage multi-scale information, mitigating pose misalignment and structural ambiguity caused by sparse viewpoints and enabling more refined alignment.
- A wavelet difference-aware anomaly detector is devised to highlight subtle and large-scale defects by comparing frequency components, enabling accurate localization.

2. Related Work

2.1. Pose-agnostic Anomaly Detection

Pose-agnostic anomaly detection addresses the challenge of identifying 3D defects without known poses, a crucial step toward practical deployment in unconstrained environments. To spur research in this area, Zhou et al. [40] introduced the MAD dataset and proposed OmniposeAD, a NeRF-based approach tailored for this task. However, this method incurs substantial training time and suffers from low efficiency during pose optimization. To address these limitations, SplatPose [10] and IGSPAD [8] adopt 3D Gaussian Splatting as their underlying representation, substantially reducing the computational cost of PAD. Recently, PIAD [29] further extends PAD to handle illumination variations between training and query images. Despite these advances, all rely on densely sampled multi-view inputs, which are costly and impractical in real-world settings. In contrast, our work focuses on PAD under sparse views, tackling the more practical scenario of limited viewpoints.

2.2. Sparse-view Reconstruction in 3DGS

Sparse-view reconstruction with 3DGS [13] has garnered increasing interest for its promise in reducing capture costs and improving efficiency. Nonetheless, sparse views often exacerbate overfitting and hinder the faithful reconstruction of fine details, curtailing the practical utility of existing techniques. Prior work in this area falls into two main categories: (1) approaches that inject external priors, such as depth maps [11, 37, 39], semantic features [3, 6, 12], or op-

tical flow [18, 22], to compensate for missing viewpoints; and (2) methods that directly optimize the Gaussian splat representation for enhanced robustness under sparse conditions [19, 28, 31]. Similar to methods in the second category, our framework performs direct optimization of the Gaussian representation without external priors, while additionally enforcing structure- and wavelet-guided constraints to achieve robust reconstruction and enhanced preservation of fine details under sparse views.

2.3. Pose Estimation in PAD

Pose estimation is a critical component of PAD, as accurate camera alignment directly affects 3D reconstruction quality and defect localization. SplatPose [10] represents the scene via 3D Gaussian Splatting and optimizes the query pose, while IGSPAD [8] uses Inverse Gaussian Splatting for direct pose estimation of anomalous views. PIAD [29] performs an initial pose estimate via reflectance-based filtering and refines it through MAE-driven optimization over reflectance and RGB inputs. However, all operate exclusively in the spatial domain, making them sensitive to textureless regions and local ambiguities, which degrade both global alignment and local consistency under sparse views. To address this, we propose a wavelet-based pose estimation strategy that leverages low-frequency components for robust global alignment and high-frequency components for precise local refinement, significantly enhancing pose-estimation robustness under sparse-view conditions.

3. Method

3.1. Overview

The pipeline of Wave-Pose3D is illustrated in Figure 2, comprising three core components: Structure-aware and Wavelet-optimized Gaussian Modeling (SWGM), Wavelet-based Pose Estimator (WPE), and Wavelet Difference-aware Anomaly Detector (WDAD). Given multi-view normal images $\{R_c^i\}_{i=1}^N$, we first construct a structure-aware 3D representation using SWGM. Then, WPE estimates the object pose of a query image R_q based on wavelet-domain verification and optimization. Finally, WDAD compares the rendering image R_r and query image R_q in the wavelet domain to produce an anomaly map. This end-to-end framework enables accurate and robust anomaly localization under sparse-view conditions.

3.2. Preliminaries

3D Gaussian Splatting. 3DGS [9] replaces implicit neural representations with discrete 3D Gaussian primitives, offering an explicit and efficient 3D reconstruction paradigm with high rendering fidelity and real-time view synthesis. Each primitive is defined by a multivariate Gaussian distribution $G(\mathbf{x})$ in 3D space:

$$G(\mathbf{x}) = exp(-\frac{1}{2}(\mathbf{x}-\mu)^\top \sum^{-1}(\mathbf{x}-\mu)), \tag{1}$$

where $\mu \in \mathbb{R}^3$ denotes the mean position and $\sum \in \mathbb{R}^{3\times3}$ controls the spatial covariance. For differentiable optimization, $\sum$ is decomposed into a rotation matrix $\mathbf{R} \in \mathrm{SO}(3)$ and a diagonal scaling matrix $\mathbf{S}$:

$$\sum = \mathbf{RSS}^\top \mathbf{R}^\top. \tag{2}$$

During rendering, the 3D covariance $\sum$ is projected to 2D camera coordinates via viewing transformation $\mathbf{W}$ and projective Jacobian $\mathbf{J}$:

$$\sum{}' = \mathbf{JW} \sum \mathbf{W}^\top \mathbf{J}^\top. \tag{3}$$

Following established projection mechanics. Each Gaussian is parameterized by: Position $\mathbf{x} \in \mathbb{R}^3$, Opacity $\alpha \in \mathbb{R}$, Scale $\mathbf{s} \in \mathbb{R}^3$, Rotation $\mathbf{r} \in \mathbb{R}^4$, and Color $\mathbf{c} \in \mathbb{R}^k$. The final pixel color C is computed by alpha-blending N overlapping Gaussians in depth order:

$$C = \sum_{i=1}^{N} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1} (1-\alpha_j), \tag{4}$$

where $\mathbf{c}_i$ and α_i denote the color and opacity of the i-th Gaussian. Through iterative optimization of these parameters, 3DGS achieves photorealistic rendering at interactive rates. We employ 3DGS to form an efficient, compact 3D representation that underpins sparse-view pose optimization and wavelet-based anomaly detection.

Wavelet Transform. The 2D discrete wavelet transform decomposes an image into one low-frequency component (LL) and three high-frequency components (LH, HL, HH), capturing vertical, horizontal, and diagonal details with joint spatial–frequency localization. Formally, this can be expressed as:

$$\begin{aligned} f(x,y) = &\sum_{j,k} c_{j,k}\, \phi_{j,k}(x,y) \\ &+ \sum_{j,k} \sum_{d\in\{H,V,D\}} d_{j,k}^d\, \varphi_{j,k}^d(x,y). \end{aligned} \tag{5}$$

where ϕ denotes the scaling function (low-frequency), φ^d denotes wavelet functions in different directions, and $c_{j,k}$, $d_{j,k}^d$ are the corresponding coefficients.

Wavelet transform plays three key roles in our framework: (1) providing high-frequency supervision during 3D Gaussian training to preserve details; (2) enabling multi-scale pose estimation, using low-frequency components for global alignment and high-frequency components for robust local matching; and (3) driving anomaly detection by leveraging high-frequency cues for subtle texture defects and low-frequency cues for structural anomalies. In this work, we employ the Haar wavelet transform.

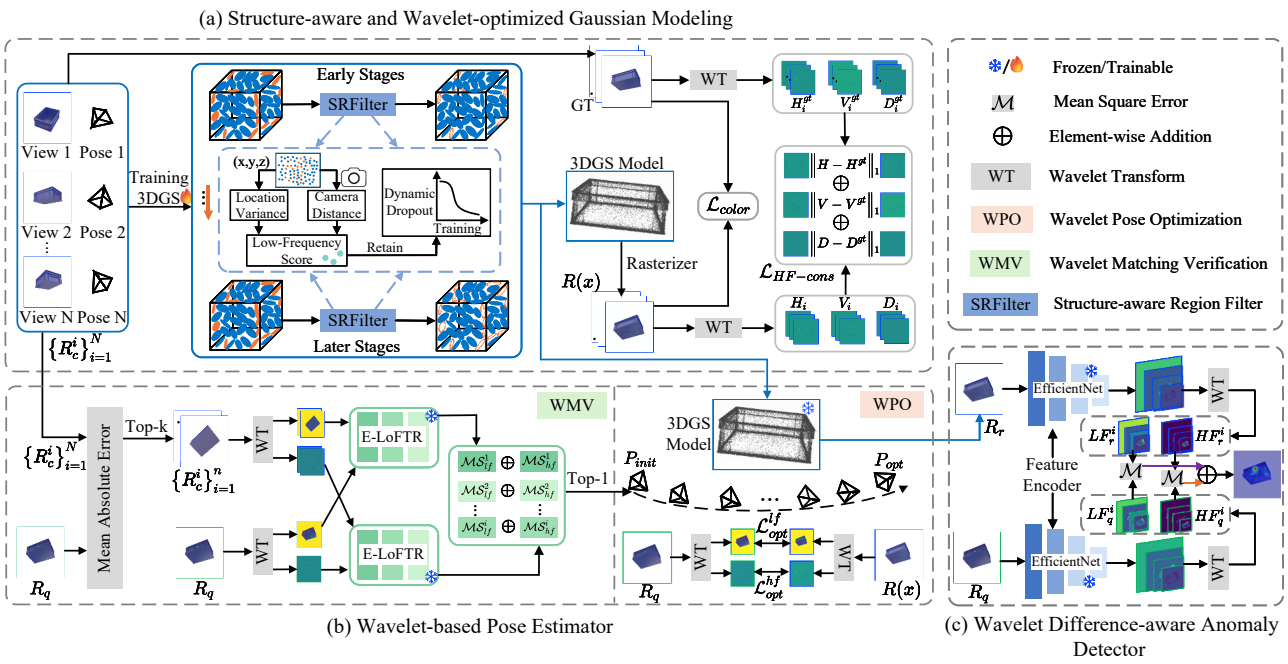

Figure 2. Overall architecture of Wave-Pose3D. Wave-Pose3D comprises three modules: (a) structure-aware wavelet-optimized Gaussian modeling (SWGM), (b) wavelet-based pose estimator (WPE), and (c) wavelet difference-aware anomaly detector (WDAD).

3.3. Structure-aware and Wavelet-optimized Gaussian Modeling

3D reconstruction is essential for PAD, yet sparse views often cause overfitting and loss of structural details. To address this, we propose Structure-aware and Wavelet-optimized Gaussian Modeling (SWGM), which integrates structure-aware modeling with wavelet-guided frequency optimization. SWGM incorporates structure-aware region filtering (SRFilter) to extract reliable structural cues from sparse images while suppressing ambiguous regions, mitigating overfitting, and improving generalization. Additionally, SWGM employs high-frequency supervision during training to recover fine-grained geometric details. The detailed process of SWGM is shown in Figure 2 (a).

Structure-aware Region Filtering. To mitigate overfitting in sparse-view 3D reconstruction, we introduce structure-aware region filtering that selectively suppresses noisy Gaussian points during training. Filtering scores are computed based on two cues: *structural complexity* and *depth consistency*. Specifically, we adopt the variance of point coordinates as a proxy for structural complexity, under the assumption that higher variance corresponds to greater local geometric fluctuations. Depth consistency is estimated using the distances between corresponding points across views, where smaller distances are indicative of more reliable multi-view alignment. These measures provide a principled basis for weighting points during training, helping the model focus on structurally informative and geometrically consistent regions.

Given the 3D positions of all Gaussians as $\mathbf{X} = \{x_i \in \mathbb{R}^3\}_{i=1}^{m}$, we compute the variance of its global positions as a measure of structural complexity. The definition is as follows:

$$\begin{aligned} \mathbf{v} &= \mathrm{Var}(\mathbf{X}) \in \mathbb{R}^3, \\ s_{\text{var}} &= \frac{\|\mathbf{v}\|_2}{\max\left(\|\mathbf{v}\|_2 + \epsilon\right)}. \end{aligned} \tag{6}$$

where $s_{\text{var}} \in [0, 1]$ denotes the the normalized structural complexity, and ϵ is a constant to ensure numerical stability. Furthermore, given the camera center $\mathbf{\Delta}_c \in \mathbb{R}^3$, we compute the Euclidean distance between each Gaussian point and $\mathbf{\Delta}_c$ to quantify depth consistency:

$$\begin{aligned} d_i &= \|x_i - \mathbf{\Delta}_c\|_2, \\ s_{\text{depth},i} &= \frac{d_i - \min(d)}{\max(d) - \min(d) + \epsilon}, \end{aligned} \tag{7}$$

where $s_{\text{depth},i} \in [0, 1]$ represents the normalized depth distance, serving as an indicator of depth consistency. To jointly promote points with low structural complexity and high depth consistency while balancing the retention of distant structures and suppression of unreliable regions, we define a low-frequency prior and compute a weighted retention score as follows:

$$\begin{aligned} s_{\text{low},i} &= (1 - s_{\text{depth},i}) \cdot (1 - s_{\text{var},i}), \\ r_i &= \omega \cdot s_{\text{depth},i} + (1 - \omega) \cdot s_{\text{low},i}, \end{aligned} \tag{8}$$

where r_i denotes the retention score of the i-th Gaussian point, with higher values indicating greater geometric reliability and structural consistency. ω represents the balance coefficient.

To adaptively suppress unreliable points during training, we employ a cosine annealing strategy to dynamically adjust the drop rate δ_t over training iterations:

$$\delta_t = \delta_{\min} + \frac{1}{2}(\delta_{\max} - \delta_{\min}) \cdot \left(1 + \cos\left(\frac{t}{T_{\max}}\pi\right)\right). \tag{9}$$

The final keep probability is computed as $p_i = \delta_t \cdot r_i$, and used for stochastic region filtering. Finally, the p_i are applied to the opacity of each Gaussian point, enabling structure-aware filtering to suppress overfitting-induced degradation of generalization during training.

In addition, to alleviate detail loss in 3D representations under sparse viewpoints, we introduce a high-frequency detail supervision mechanism. Specifically, we apply wavelet transforms to both the rendered images $R(x)$ and ground-truth (GT) to extract directional high-frequency components (horizontal, vertical, and diagonal), thereby capturing critical edges and textures. To ensure frequency-domain consistency, we define the loss as:

$$\mathcal{L}_{\text{HF-cons}} = \sum_{d\in\{H,V,D\}} \lambda^d \left\| c_d^{\text{render}} - c_d^{\text{gt}} \right\|_1, \tag{10}$$

where c_d^{render} and c_d^{gt} denote the directional high-frequency components of the rendered images and GT, respectively, and λ^d denotes the weight assigned to each direction to balance their contributions. In our implementation, λ^H, λ^V, and λ^D are set to 0.4, 0.4, and 0.2, respectively. By enforcing high-frequency supervision in the wavelet domain, the model is guided to retain fine-grained structural cues, thereby improving the visual fidelity of reconstructed images under sparse-view conditions.

Ultimately, the total loss of optimizing SWGM is expressed as follows.

$$\mathcal{L}_{total} = \mathcal{L}_{color} + \mathcal{L}_{\text{HF-cons}}, \tag{11}$$

where $\mathcal{L}_{color}$ is the color loss of 3DGS.

3.4. Wavelet-based Pose Estimator

Existing PAD methods [29, 40] typically perform pose initialization and optimization in the spatial domain. However, under sparse views, the reduced number of reliable feature correspondences leads to significant pose estimation errors and misaligned structural reconstruction. To overcome these challenges, we propose a Wavelet-based Pose Estimator (WPE), which leverages low-frequency components for robust global alignment and high-frequency components for fine-grained refinement. This frequency-aware design improves both the robustness and accuracy of pose estimation under sparse-view conditions. WPE consists of two key modules: wavelet matching verification (WMV) and wavelet pose optimization (WPO).

Wavelet Matching Verification. To enable robust and discriminative pose initialization, we employ wavelet transforms to decompose images into global structures and local details. First, we compute the mean absolute error (MAE) between the query image R_q and the set of anomaly-free reference images $\{R_c^i\}_{i=1}^N$ to filter out irrelevant views, reducing the subsequent matching overhead. This yields a refined subset $\{R_c^i\}_{i=1}^n$ with $n < N$. Next, both R_q and the selected references are decomposed via wavelet transform into high- and low-frequency components. These components are independently matched using EfficientLoFTR [26], producing frequency-specific matching scores. The process is formulated as follows:

$$\begin{aligned} F_q^{\text{LF}}, F_q^{\text{HF}} &= \mathbf{WT}(R_q); F_{c,i}^{\text{LF}}, F_{c,i}^{\text{HF}} = \mathbf{WT}(R_c^i), \\ \mathcal{M}_{\text{HF}}^i &= \text{EfficientLoFTR}(F_q^{\text{HF}}, F_{c,i}^{\text{HF}}), \\ \mathcal{M}_{\text{LF}}^i &= \text{EfficientLoFTR}(F_q^{\text{LF}}, F_{c,i}^{\text{LF}}), \end{aligned} \tag{12}$$

where $\mathbf{WT}(\cdot)$ denotes the wavelet transform. F_q^{HF} and $F_{c,i}^{\text{HF}}$ represent the weighted combination of the horizontal, vertical, and diagonal high-frequency components. The resulting high-frequency ($\mathcal{M}_{\text{HF}}^i$) and low-frequency ($\mathcal{M}_{\text{LF}}^i$) matching scores are jointly evaluated to identify the most confident match. The initial pose is then assigned as that of the anomaly-free image with the highest combined score. The matching strategy is detailed as follows:

$$\begin{aligned} \mathcal{M}_{\text{fusion}}^i &= \eta \cdot \mathcal{M}_{\text{HF}}^i + (1-\eta) \cdot \mathcal{M}_{\text{LF}}^i, \\ P_{\text{init}} &= P_{i^*}, \quad i^* = \arg\max_i(\mathcal{M}_{\text{fusion}}^i), \end{aligned} \tag{13}$$

where η is a weighting coefficient that balances structural consistency and detail alignment. We set $\eta = 0.8$ to emphasize high-frequency details during matching. P_{init} denotes the selected initial pose.

Wavelet Pose Optimization. To further improve pose accuracy, we refine the initial estimate P_{init} in the wavelet domain by minimizing the discrepancy between the wavelet decompositions of the rendered image $R(x)$ and the query image R_q. This optimization leverages low-frequency components for robust global alignment and high-frequency details for fine-grained refinement. The camera pose $\mathbf{x}$ is parameterized on the SE(3) manifold using exponential coordinates, where $\mathbf{x} = [\mathbf{S}, \theta]$ with screw axis $\mathbf{S} = [\boldsymbol{\omega}, \mathbf{v}]^\top \in \mathbb{R}^6$ and rotation angle $\theta \in \mathbb{R}$. The corresponding rigid body transformation is defined as:

$$\mathbf{T}(\mathbf{x}) = e^{[\mathbf{S}]\theta}. \tag{14}$$

Given the current pose $\mathbf{x}$, the rendered image $R(\mathbf{x})$ is obtained from the trained 3DGS model via a differentiable

rasterization process:

$$R(\mathbf{x}) = \mathbf{Raster}\big(V(\mathbf{x}); \mathbf{\Theta}\big), \\ V(\mathbf{x}) = \mathbf{T}(\mathbf{x})^{-1} T^{(0)}, \tag{15}$$

where $\mathbf{Raster}(\cdot)$ denotes the differentiable rendering operation, and $T^{(0)}$ is the canonical camera transformation.

To exploit multi-scale alignment cues, we transform both $R(\mathbf{x})$ and R_q into wavelet space and compute low-frequency and high-frequency reconstruction losses separately:

$$\mathcal{L}_{opt}^{lf} = \left\| R^{lf}(\mathbf{x}) - R_q^{lf} \right\|_2^2, \\ \mathcal{L}_{opt}^{hf} = \sum_{i=1}^{2} \left\| R^{hf}(\mathbf{x}) - R_q^{hf} \right\|_1, \\ \mathcal{L}_{opt} = \lambda_{opt} \cdot \mathcal{L}_{opt}^{lf} + (1 - \lambda_{opt}) \cdot \mathcal{L}_{opt}^{hf}, \tag{16}$$

where λ_{opt} controls the trade-off between global structure alignment and local detail refinement. In our experiments, we set $\lambda_{opt} = 0.4$. Here, i denotes the decomposition level of the wavelet transform. After the WPE stage, a refined camera pose is obtained, enabling the 3D Gaussian model to render more accurate and view-consistent reference images for subsequent anomaly detection.

3.5. Wavelet Difference-ware Anomaly Detector

Existing PAD methods [10, 29, 40] typically compute anomaly scores in the spatial domain, making them sensitive to pose misalignment and geometric distortions under sparse views. To improve robustness and better capture subtle structural anomalies, we propose a wavelet difference-aware anomaly detector, which computes anomaly scores in the wavelet domain by analyzing both low- and high-frequency components. Specifically, we adopt EfficientNet-B4 [23] to extract multi-scale features from the query image R_q, and the rendered image R_r, followed by wavelet transforms to obtain corresponding frequency components. The process is defined as:

$$F_r^l = \mathbf{E}\left(R_r\right), F_q^l = \mathbf{E}\left(R_q\right), \\ LF_r^l, HF_r^l = \mathbf{WT}\left(F_r^l\right), \\ LF_q^l, HF_q^l = \mathbf{WT}\left(F_q^l\right), \\ S_L^l = \left\| LF_r^l - LF_q^l \right\|_2^2, \\ S_H^l = \left\| HF_r^l - HF_q^l \right\|_2^2, \\ S = \alpha \cdot S_L^l + \beta \cdot S_H^l, \tag{17}$$

where $\mathbf{E}(\cdot)$ refers to EfficientNet-B4. l denotes the number of feature levels. Here, we set α=1, β=3 to enhance the sensitivity to subtle local anomalies, thereby improving the accuracy of anomaly detection.

4. Experiments and Analysis

4.1. Datasets

We evaluate our method on two datasets: MAD_sim [40] and PIAD_synt [29]. MAD_sim comprises 20 toy objects with diverse shapes and textures, simulating fine-grained structural and appearance variations. PIAD_synt contains industrial objects with various surface and structural defects, offering a more challenging benchmark. Since existing PAD datasets are constructed under dense multi-view settings, we simulate sparse-view scenarios by uniformly sampling 10% and 20% of the available views for each object. For example, for the 20% sparse-view setting, we retain 0.2*N training views per category (N = total training views), uniformly sampled over the full view range, while keeping the test views unchanged.

4.2. Evaluation Metrics

Following prior work [10, 29, 40], we adopt both image-level and pixel-level AUROC as evaluation metrics. Image-level AUROC assesses the ability to distinguish normal and anomalous images based on aggregated anomaly scores, while pixel-level AUROC evaluates localization accuracy by comparing predicted anomaly maps with ground truth. Higher AUROC indicates better performance.

4.3. Implementation details

For each category in PIAD_synt, the 3DGS model is trained for 10k iterations at a resolution of 400×400, while MAD_sim categories are trained for 15k iterations. Pose optimization in WPO is performed over 300 iterations, with an initial learning rate of 0.02 and an exponential decay schedule. All experiments use an NVIDIA RTX A6000 GPU.

4.4. Comparison with SoTAs

We compare our method with state-of-the-art PAD approaches, including OmniPoseAD [40], SplatPose [10], PIAD [29], and DropGaussian [19]. For fairness, we reproduce sparse-view results using their official code with a uniform sampling strategy. It is noteworthy that DroupGuass was originally a sparse-view reconstruction method. To enable anomaly detection, we integrate PIAD's pose estimation and anomaly computation modules.

Quantitative Comparison. Table 1 presents a quantitative comparison between our method and existing PAD approaches under 10% and 20% sparse-view settings on PIAD_synt and MAD_sim. Under all sparse views, our approach consistently outperforms existing PAD methods. Furthermore, when compared to DropGaussian, a sparse-view reconstruction method that incorporates pose estimation and anomaly detection components, our method still maintains a leading position. This validates that modeling 3D representations in the wavelet domain effectively miti-

Table 1. Quantitative Comparison on PIAD_synt and MAD_sim under 10% and 20% Sparse-View Settings. Results are averaged over all categories. The best performance is highlighted in **bold**, while the second-best is underlined.

Method	MAD_sim(10%)		MAD_sim(20%)		PIAD_synt(10%)		PIAD_synt(20%)	
	P_AUROC	I_AUROC	P_AUROC	I_AUROC	P_AUROC	I_AUROC	P_AUROC	I_AUROC
OmniAD [40]	89.1	57.0	90.2	59.7	92.6	61.9	93.8	69.7
SplatPose [10]	90.3	58.4	91.1	60.0	94.4	65.3	95.3	71.0
PIAD [29]	92.4	60.2	93.0	64.5	95.2	67.4	96.4	75.4
DropGaussian [19]	93.1	62.3	93.6	67.5	95.6	70.1	97.3	77.8
Ours	**94.0**	**64.6**	**95.1**	**69.1**	**96.3**	**73.2**	**97.8**	**82.3**

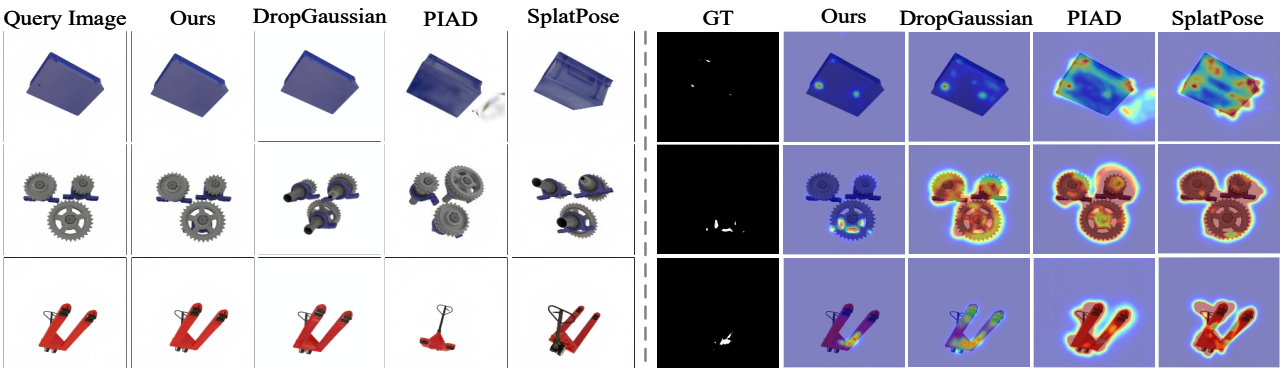

Figure 3. Qualitative visualization of pixel-level anomaly segmentation. The left side compares the rendered images produced by our method and baselines, while the right side visualizes the corresponding anomaly segmentation results.

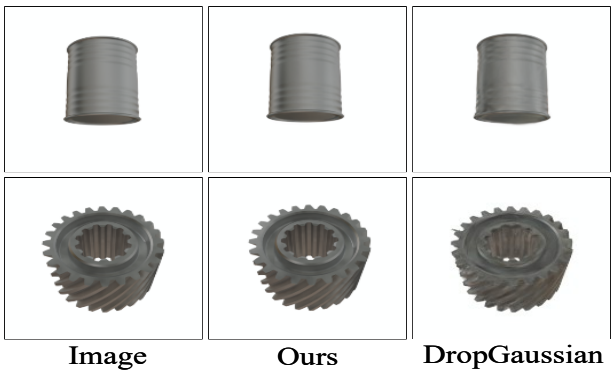

Figure 4. Comparison with the reconstruction results of DropGaussian. Our method can better reconstruct object surfaces.

Table 2. Quantitative Comparison on PIAD_real [29] under 20% Sparse-View Settings. Best results are shown in **bold**.

Method	P_ROCAUC	I_ROCAUC
PIAD [29]	98.6	81.8
DropGaussian [19]	98.4	83.8
Our	**98.8**	**87.6**

gates pose ambiguity and preserves geometric fidelity under sparse observations.

Validation on real-world dataset. Due to nonnegligible camera pose estimation errors (typically from COLMAP) in real-world PAD datasets, reliable evaluation is limited. Nevertheless, we find that PIAD_real [29] serves as a reliable benchmark for real-world PAD evaluation. As shown in Tab 2, our method outperforms PIAD and DropGaussian on PIAD_real (20%), achieving the best performance.

Table 3. Ablation studies of the proposed components. The best results are denoted using **bold**.

	Setting	P_AUROC	I_AUROC
(a)	Baseline	95.89	71.89
(b)	w/o SWGM	97.69	80.05
(c)	w/o WPE	97.51	78.65
(d)	w/o WDAD	97.65	80.20
(e)	All	**97.75**	**82.34**

Qualitative Comparison. Figure 3 presents a qualitative comparison between our method and existing PAD approaches on PIAD_synt (20%). Our method reconstructs more complete object geometry, estimates more accurate poses, and achieves more precise anomaly localization. In contrast, PIAD and SplatPose suffer from surface artifacts or inaccurate pose estimation, leading to suboptimal localization results. Although some pose estimations produced by DropGaussian are accurate, the reconstructed 3D object surfaces contain artifacts, which subsequently lead to inaccurate detection results. Moreover, compared with DropGaussian reconstruction results, our method reduces surface artefacts (Figure 4), highlighting the benefits of wavelet-domain optimization for 3D representations.

Table 4. Ablation studies of SWGM. "→" denotes substitution. The best results are denoted using **bold**.

	Setting	P_AUROC	I_AUROC
(a)	w/o SRFilter	97.69	80.12
(b)	w/o $\mathcal{L}_{\text{HF-cons}}$	97.71	80.08
(c)	Dropout→SRFilter	97.31	78.76
(d)	All	**97.75**	**82.34**

Table 5. Ablation experiment of balance coefficient ω in SWGM. The best results are denoted using **bold**.

	Setting	Pixels AUROC	Images AUROC
(a)	w=0.5	97.60	81.89
(b)	w=0.6	**97.75**	**82.34**
(f)	w=0.7	97.71	82.21

Table 6. Ablation studies of WPE. "MV" and "PO" refer to spatial-domain matching verification and pose optimization, respectively. The best results are denoted using **bold**.

	Setting	P_AUROC	I_AUROC
(a)	MV → WMV	97.70	80.80
(b)	PO → WPO	97.61	80.14
(c)	All	**97.75**	**82.34**

Table 7. Ablation experiment of η in WPE. The best results are denoted using **bold**.

	Setting	Pixels AUROC	Images AUROC
(a)	η =0.7	97.53	82.01
(b)	η =0.8	**97.75**	**82.34**
(c)	η =0.9	97.48	81.89

Table 8. Ablation experiments for settings (α, β) in WDAD. The best results are denoted using **bold**.

	Setting	Pixels AUROC	Images AUROC
(a)	(1, 1)	97.70	80.36
(b)	(1, 0.5)	97.69	79.53
(c)	(0.5, 1)	97.69	80.31
(d)	(2, 1)	97.70	81.95
(e)	(1, 2)	97.71	81.33
(f)	(1, 3)	**97.75**	**82.34**
(g)	(1, 4)	97.74	82.28

4.5. Ablation Studies

We perform ablation studies on the PIAD_synt (20%) dataset to validate the effectiveness of each component. The baseline model does not include SWGM and replaces WPE with spatial-domain pose estimation and WDAD with spatial-domain anomaly computation.

Effectiveness of SWGM. As shown in Table 3 (b) and (e), removing SWGM leads to a 2.29% and 0.06% drop in image and pixel AUROC, respectively, demonstrating its effectiveness in enhancing 3D representation under sparse views. In addition, we further perform ablation in Table 4 to study the contribution of each component in SWGM. Removing either the SRFilter or the $\mathcal{L}_{HF-cons}$ component results in performance degradation, indicating that incorporating structural priors and preserving fine-grained details during training are crucial under sparse-view conditions. Moreover, replacing the SRFilter with Dropout [28] results in a significant performance drop, further demonstrating that SRFilter offers better generalization under sparse-view conditions. Finally, in Table 4, we conduct an ablation study on the balance coefficient ω. The results demonstrate that the best performance is achieved when ω = 0.6.

Effectiveness of WPE. As shown in Table 3(c) and (e), replacing the proposed WPE with a spatial-domain matching approach results in a 3.69% and 0.24% drop in image and pixel AUROC, respectively, indicating the superior robustness of wavelet-based pose estimation under sparse views. Moreover, we further validate the effectiveness of WMV and WPO in WPE (Table 6). Removing either component leads to a performance drop, indicating that wavelet-domain processing better supports sparse-view conditions by leveraging low-frequency structural cues and high-frequency details. Furthermore, we evaluate the effectiveness of the parameter η in WPE. As shown in Table 7, the model achieves the best performance when η=0.8, further indicating the increased importance of high-frequency components in determining the initial pose.

Effectiveness of WDAD. To assess the effectiveness of WDAD, we replace it with a spatial-domain anomaly scoring method. As shown in Table 3 (c) and (e), this substitution causes a 2.14% and 0.1% drop in image- and pixel-level AUROC, demonstrating the advantage of frequency-domain anomaly detection under sparse views. Table 8 further studies the impact of (α, β), where $(1, 3)$ yields the best results, indicating that assigning a higher weight to high-frequency components at the final detection stage improves anomaly localization.

5. Conclusion

In this paper, we propose a wavelet-driven 3D anomaly detection framework, dubbed Wave-Pose3D, tailored for pose-agnostic anomaly detection under sparse-view conditions. Unlike spatial-domain PAD methods that rely solely on appearance cues, Wave-Pose3D explicitly leverages multi-scale frequency information to achieve robust pose-agnostic anomaly detection. The framework combines a structure-aware and wavelet-optimized Gaussian modeling strategy to enhance generalization and preserve fine-grained geometry, with a wavelet-based pose estimator that facilitates accurate initialization and refinement via multi-scale frequency-domain alignment. A wavelet difference-aware anomaly detector is incorporated to capture structural deviations and subtle details, enabling robust anomaly localization. Comprehensive experiments show the effectiveness of each component, with Wave-Pose3D achieving SOTA performance under both 10% and 20% sparse views.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (2021YFA1000102), in part by the grants from the National Natural Science Foundation of China (Nos. 62376285, 61673396), and the Innovation Fund Project for Graduate Students at China University of Petroleum (East China), and the Fundamental Research Funds for the Central Universities (No. 26CX04038A).

References

[1] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020. 1

[2] Yanliang Ge, Qiao Zhang, Tian-Zhu Xiang, Cong Zhang, and Hongbo Bi. Tcnet: Co-salient object detection via parallel interaction of transformers and cnns. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(6):2600–2615, 2022. 1

[3] Jun Guo, Xiaojian Ma, Yue Fan, Huaping Liu, and Qing Li. Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting. *arXiv preprint arXiv:2403.15624*, 2024. 2

[4] 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. *Advances in Neural Information Processing Systems*, 37:71162–71187, 2024. 1

[5] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8472–8480, 2024. 1

[6] Zongqi He, Zhe Xiao, Kin-Chung Chan, Yushen Zuo, Jun Xiao, and Kin-Man Lam. See in detail: Enhancing sparse-view 3d gaussian splatting with local depth and semantic regularization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2

[7] Varun Jain, Zongwei Wu, Quan Zou, Louis Florentin, Henrik Turbell, Sandeep Siddhartha, Radu Timofte, Qifan Gao, Linyan Jiang, Qing Luo, et al. Ntire 2025 challenge on video quality enhancement for video conferencing: Datasets, methods and results. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1184–1194, 2025. 1

[8] Bolin Jiang, Yuqiu Xie, Jiawei Li, Naiqi Li, Bin Chen, and Shu-Tao Xia. Igspad: Inverting 3d gaussian splatting for pose-agnostic anomaly detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10229–10237, 2024. 2, 3

[9] 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. 1, 3

[10] Mathis Kruse, Marco Rudolph, Dominik Woiwode, and Bodo Rosenhahn. Splatpose & detect: Pose-agnostic 3d anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3950–3960, 2024. 2, 3, 6, 7

[11] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20775–20785, 2024. 2

[12] Guibiao Liao, Qing Li, Zhenyu Bao, Guoping Qiu, and Kanglin Liu. Spc-gs: Gaussian splatting with semantic-prompt consistency for indoor open-world free-view synthesis from sparse inputs. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11264–11274, 2025. 2

[13] Haitian Liu, Binglin Liu, Qianchao Hu, Peilun Du, Jing Li, Yang Bao, and Feng Wang. A review on 3d gaussian splatting for sparse view reconstruction. *Artificial Intelligence Review*, 58(7):215, 2025. 2

[14] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20402–20411, 2023. 1

[15] Wei Luo, Haiming Yao, Yunkang Cao, Qiyu Chen, Ang Gao, Weiming Shen, and Wenyong Yu. Inp-former++: Advancing universal anomaly detection via intrinsic normal prototypes and residual learning. *arXiv preprint arXiv:2506.03660*, 2025. 1

[16] Yuxin Ma, Jiaxing Yin, Feng Huang, and Qipeng Li. Surface defect inspection of industrial products with object detection deep networks: A systematic review. *Artificial Intelligence Review*, 57(12):1–48, 2024. 1

[17] 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. 1

[18] Avinash Paliwal, Wei Ye, Jinhui Xiong, Dmytro Kotovenko, Rakesh Ranjan, Vikas Chandra, and Nima Khademi Kalantari. Coherentgs: Sparse novel view synthesis with coherent 3d gaussians. In *European Conference on Computer Vision*, pages 19–37. Springer, 2024. 3

[19] Hyunwoo Park, Gun Ryu, and Wonjun Kim. Dropgaussian: Structural regularization for sparse-view gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21600–21609, 2025. 3, 6, 7

[20] Je-Kang Park, Bae-Keun Kwon, Jun-Hyub Park, and Dong-Joong Kang. Machine learning-based imaging system for surface defect inspection. *International Journal of Precision Engineering and Manufacturing-Green Technology*, 3 (3):303–310, 2016. 1

[21] Olamilekan Shobayo and Reza Saatchi. Developments in deep learning artificial neural network techniques for med-

ical image analysis and interpretation. *Diagnostics*, 15(9): 1072, 2025. 1

[22] Su Sun, Cheng Zhao, Zhuoyang Sun, Yingjie Victor Chen, and Mei Chen. Splatflow: Self-supervised dynamic gaussian splatting in neural motion flow field for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27487–27496, 2025. 3

[23] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6

[24] Sifan Tu, Xin Zhou, Dingkang Liang, Xingyu Jiang, Yumeng Zhang, Xiaofan Li, and Xiang Bai. The role of world models in shaping autonomous driving: A comprehensive survey. *arXiv preprint arXiv:2502.10498*, 2025. 1

[25] Xiaogang Wang. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3–19, 2013. 1

[26] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient loftr: Semi-dense local feature matching with sparse-like speed. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21666–21675, 2024. 5

[27] Yuexin Wang, Xiaolei Wang, Yizheng Gong, and Jimin Xiao. Normal-abnormal guided generalist anomaly detection. *arXiv preprint arXiv:2510.00495*, 2025. 1

[28] Yexing Xu, Longguang Wang, Minglin Chen, Sheng Ao, Li Li, and Yulan Guo. Dropoutgs: Dropping out gaussians for better sparse-view rendering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 701–710, 2025. 3, 8

[29] Kaichen Yang, Junjie Cao, Zeyu Bai, Zhixun Su, and Andrea Tagliasacchi. Piad: Pose and illumination agnostic anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4734–4743, 2025. 1, 2, 3, 5, 6, 7

[30] Xincheng Yao, Zixin Chen, Chao Gao, Guangtao Zhai, and Chongyang Zhang. Resad: A simple framework for class generalizable anomaly detection. *Advances in Neural Information Processing Systems*, 37:125287–125311, 2024. 1

[31] Jiawei Zhang, Jiahe Li, Xiaohan Yu, Lei Huang, Lin Gu, Jin Zheng, and Xiao Bai. Cor-gs: sparse-view 3d gaussian splatting via co-regularization. In *European Conference on Computer Vision*, pages 335–352. Springer, 2024. 3

[32] Jiangning Zhang, Haoyang He, Zhenye Gan, Qingdong He, Yuxuan Cai, Zhucun Xue, Yabiao Wang, Chengjie Wang, Lei Xie, and Yong Liu. A comprehensive library for benchmarking multi-class visual anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3524–3533, 2025. 1

[33] Qiao Zhang, Yanliang Ge, Cong Zhang, and Hongbo Bi. Tprnet: camouflaged object detection via transformer-induced progressive refinement network. *The Visual Computer*, 39 (10):4593–4607, 2023. 1

[34] Qiao Zhang, Xiaoxiao Sun, Yurui Chen, Yanliang Ge, and Hongbo Bi. Attention-induced semantic and boundary interaction network for camouflaged object detection. *Computer Vision and Image Understanding*, 233:103719, 2023. 1

[35] Qiao Zhang, Mingwen Shao, Xinyuan Chen, Xiang Lv, and Kai Xu. Wave-mambaad: Wavelet-driven state space model for multi-class unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20868–20877, 2025. 1

[36] Jingyuan Zhao, Wenyi Zhao, Bo Deng, Zhenghong Wang, Feng Zhang, Wenxiang Zheng, Wanke Cao, Jinrui Nan, Yubo Lian, and Andrew F Burke. Autonomous driving system: A comprehensive survey. *Expert Systems with Applications*, 242:122836, 2024. 1

[37] Liang Zhao, Zehan Bao, Yi Xie, Hong Chen, Yaohui Chen, and Weifu Li. Tsgaussian: Semantic and depth-guided target-specific gaussian splatting from sparse views. *Image and Vision Computing*, page 105706, 2025. 2

[38] Xiaoqing Zheng, Song Zheng, Yaguang Kong, and Jie Chen. Recent advances in surface defect inspection of industrial products using deep learning techniques. *The International Journal of Advanced Manufacturing Technology*, 113(1):35–58, 2021. 1

[39] Yulong Zheng, Zicheng Jiang, Shengfeng He, Yandu Sun, Junyu Dong, Huaidong Zhang, and Yong Du. Nexusgs: Sparse view synthesis with epipolar depth priors in 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26800–26809, 2025. 2

[40] 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, 5, 6, 7

[41] Ziqing Zhou, Yurui Pan, Lidong Wang, Wenbing Zhu, Mingmin Chi, Dong Wu, and Bo Peng. Pro-ad: Learning comprehensive prototypes with prototype-based constraint for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2506.13097*, 2025. 1