# RPE-PAD: Relative Pose Estimation for Pose-agnostic Anomaly Detection

**Zhipeng Zhang[1], Mengzan Qi[1], Rongkang Ma[1], Yingying Fang[2], Guixu Zhang[1], Tieyong Zeng[3], Zhi Li[1]***

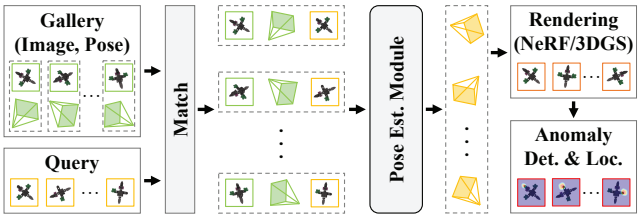
[1]School of Computer Science and Technology, East China Normal University, Shanghai, China
[2]National Heart and Lung Institute, Imperial College London, London, UK
[3]Department of Mathematics, The Chinese University of Hong Kong, Hong Kong, China
51265901066@stu.ecnu.edu.cn, qimengzan@163.com, 51275901010@stu.ecnu.edu.cn,
y.fang@imperial.ac.uk, gxzhang@cs.ecnu.edu.cn, zeng@math.cuhk.edu.hk, zli@cs.ecnu.edu.cn

## Abstract

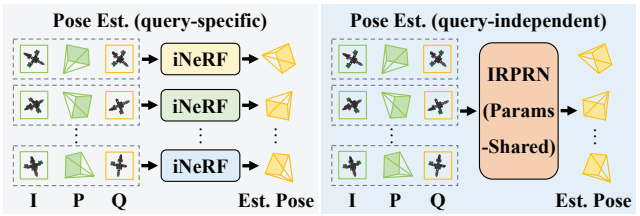
Pose-agnostic Anomaly Detection (PAD) aims to detect anomalies when the poses of query images are unknown and differ from those in the training set. Therefore, accurately estimating the camera poses for the query images in the test set is critical for this task. Existing query-specific framework methods require re-optimizing a new set of parameters for each query image, limiting their generalization and increasing computational burden. To overcome these limitations, we propose a novel method, Relative Pose Estimation for Pose-agnostic Anomaly Detection (RPE-PAD), which enhances both generalization and efficiency with a query-independent framework. Specifically, we propose a Random View Synthesis Scheme (RVSS) that generates new poses by adding Gaussian perturbations to the original poses, then renders the corresponding views to augment the dataset. To estimate the relative camera pose between two input images, we introduce an Iterative Relative Pose Refinement Network (IRPRN), which incorporates a hierarchical coarse-to-fine refinement strategy. Furthermore, we employ a Multi-Pair Training Strategy (MPTS) to train the proposed IRPRN, leveraging multiple image pairs to expand the relative pose transformation space during training. Extensive experiments demonstrate that our method achieves robust anomaly detection performance while significantly improving inference efficiency.

## Introduction

In recent years, unsupervised methods for anomaly detection have attracted significant attention (Dehaene and Eline 2020; Yu et al. 2021; You et al. 2022). These approaches are trained exclusively on normal (defect-free) images due to the limited availability of diverse abnormal samples in real-world scenarios (Bergmann et al. 2019). However, most unsupervised anomaly detection methods assume a high degree of pose alignment between training and test set images, which limits their ability to detect anomalies from arbitrary viewpoints (Zhou et al. 2023). To address this challenge, Pose-agnostic Anomaly Detection (PAD) methods enhance their applicability in real-world scenarios by detecting anomalies in objects from diverse viewpoints (Zhou et al. 2023; Kruse et al. 2024; Jiang et al. 2024).




(a) Inference pipeline for Pose-agnostic Anomaly Detection (PAD)




(b) Existing query-specific PAD framework (c) Our query-independent PAD framework

Figure 1: The inference pipeline and the difference between our method and existing methods. In the previous query-specific framework, a new set of parameters must be re-optimized for pose estimation (pose est.) with each query image, thus restricting its generalization capacity. In contrast, our query-independent framework fixes a set of parameters, enabling efficient pose estimation across multiple query images.

OmniposeAD (Zhou et al. 2023) is a typical method tailored for the PAD task with a query-specific framework. The workflow of this method starts with training a Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) model on normal images from the training set. As shown in Figures 1(a) and 1(b), for each query image in the test set, OmniposeAD matches it with images in the gallery and employs inverted Neural Radiance Fields (iNeRF) (Yen-Chen et al. 2021) to estimate the camera pose. The corresponding normal image is then rendered using the pre-trained NeRF model. At the end of the workflow, OmniposeAD extracts features to compare the query image with the rendered normal image, enabling anomaly detection and localization. However, this approach suffers from two major limitations: **1) Limited Generalization.** OmniposeAD applies iNeRF to camera pose es-

*Corresponding author.

timation for each query image, as illustrated in Figure 1(b). This query-specific framework limits generalization, as a new set of parameters must be re-optimized for pose estimation with each query image. Consequently, this framework significantly reduces the efficiency of the method in large-scale image processing scenarios. **2) Time-consuming Inference.** The inference process for OmniposeAD is time-consuming, as it requires estimating the camera pose using iNeRF for each query image. Furthermore, iNeRF demands substantial time for considerable iterations to estimate the pose. This time-consuming inference restricts the effectiveness of OmniposeAD in real-world applications.

The primary challenge in PAD lies in accurately estimating the camera poses of query images (Zhou et al. 2023; Kruse et al. 2024; Jiang et al. 2024). To address this, we propose an improved PAD approach that integrates relative pose estimation (Melekhov et al. 2017; Von Stumberg et al. 2020; Cho et al. 2023). Our method takes an image pair as input and directly outputs the relative camera pose between the two images, offering two key advantages over OmniposeAD: **1) Improved Generalization Capability.** Our method adopts a query-independent framework, which eliminates the need to re-optimize for each query image and enhances the ability to generalize to unseen data, since it is designed to learn features that are invariant to different query images. As illustrated in Figure 1(c), our method significantly enhances generalization capability, enabling robust performance across diverse images. **2) Accelerated Inference Process.** Our query-independent framework utilizes a fixed set of parameters during inference, enabling efficient pose estimation across multiple query images. Therefore, our approach achieves a significant improvement in inference speed. Experimental results demonstrate that our approach achieves superior generalization capability while significantly accelerating inference.

In this work, we propose a novel framework, Relative Pose Estimation for Pose-agnostic Anomaly Detection (RPE-PAD). We design a Random View Synthesis Scheme (RVSS) that generates new camera poses by adding Gaussian perturbations to training set poses. These generated poses are rendered into corresponding novel views using a pre-trained 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) model. The novel views are then matched with the training set views to form image pairs for model training. To improve camera pose estimation accuracy, we introduce a coarse-to-fine pose iterative refinement strategy into the relative pose estimation process, proposing an Iterative Relative Pose Refinement Network (IRPRN). This hierarchical network iteratively refines the coarse relative pose at each layer, progressively approximating the ground truth relative pose. To further expand the space of relative pose transformations, we adopt a Multi-Pair Training Strategy (MPTS). By leveraging multiple image pairs, the strategy provides a comprehensive understanding of the target object's structure and details in 3D space, allowing accurate relative camera pose estimation, even in challenging real-world scenarios. Our contributions are summarized as follows:

- We propose a novel framework for pose-agnostic anomaly detection based on relative pose estimation. By adopting a query-independent framework, our method allows camera pose estimation for all query images with a fixed set of model parameters. The approach significantly improves generalization and computational efficiency.
- For accurate camera pose estimation, we propose an Iterative Relative Pose Refinement Network (IRPRN). The IRPRN hierarchically refines the coarse relative pose at each layer in a coarse-to-fine manner, progressively approximating the ground truth.
- We introduce a Multi-Pair Training Strategy (MPTS) that enables the model to explore a larger space of relative pose transformations through multiple image pairs. This provides a more comprehensive understanding of the object's 3D structure and details, thereby enhancing robustness in challenging scenarios.
- We propose a Random View Synthesis Scheme (RVSS) that generates new camera poses by applying Gaussian perturbations and renders the corresponding views. This cost-effective strategy augments the training data and improves the generalization of the model to unseen views.
- Extensive experiments demonstrate that our method achieves competitive performance compared to the state-of-the-art method, while delivering an approximately $373\times$ acceleration in inference. Furthermore, our method achieves the best performance in real-world scenarios.

## Related Work

### 2D Image Anomaly Detection

Most current research on 2D image anomaly detection focuses on unsupervised methods (Yan et al. 2021; Leng et al. 2022; Liu et al. 2024b; Dai et al. 2024; Liu et al. 2024a). These methods rely on the training set consisting exclusively of normal images, while the test set contains a mix of normal and abnormal images. Following the introduction of MVTec AD (Bergmann et al. 2019), unsupervised anomaly detection methods for 2D images are generally categorized into feature embedding-based and reconstruction-based methods.

Feature embedding-based methods (Gudovskiy, Ishizaka, and Kozuka 2022; Lee, Lee, and Song 2022) extract image features using traditional machine learning or deep learning techniques. These methods map images into a feature space where normal samples form a compact distribution, while anomalies tend to lie at the periphery of this distribution. Reconstruction-based methods (Dehaene and Eline 2020; You et al. 2022) detect anomalies by reconstructing images through encoder–decoder architectures. Typically, these methods rely on the assumption that the camera poses of training and test images are well-aligned.

### Pose-agnostic Anomaly Detection

To support anomaly detection from different viewpoints and enhance applicability in real-world scenarios, recent methods leverage image camera poses. The Multi-pose Anomaly Detection (MAD) dataset introduces a challenging scenario known as Pose-agnostic Anomaly Detection (PAD) (Zhou

et al. 2023), where the camera poses differ between the training and test sets, and the test set poses are unavailable. To address this issue, OmniposeAD (Zhou et al. 2023) first trains a Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) model on the training set. For each query image, inverted Neural Radiance Fields (iNeRF) (Yen-Chen et al. 2021) estimates the camera pose, which is then used by the NeRF model to render a pose-aligned normal image. Anomalies are detected by comparing the query image with the rendered normal image. Building on OmniposeAD, IGSPAD (Jiang et al. 2024) leverages 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) to render novel views and employs an inverted 3D Gaussian distributions to estimate the pose for each query image individually. Likewise, SplatPose (Kruse et al. 2024) utilizes 3DGS for novel view rendering and transforms 3D point clouds to estimate the pose for each query image. Although these methods enhance certain components, they remain the query-specific framework like OmniposeAD, limiting their generalization capability.

In contrast, our approach adopts a query-independent framework, which enhances both generalization and computational efficiency. To achieve accurate relative camera pose estimation under PAD, we design an Iterative Relative Pose Refinement Network (IRPRN). Additionally, we propose a Multi-Pair Training Strategy (MPTS) that allows the model to explore a larger space of relative pose transformations, thereby improving robustness in challenging scenarios.

### Novel View Synthesis

PAD methods rely on novel view synthesis techniques to synthesize multi-view images for anomaly detection. Recently, implicit neural scene representation methods, exemplified by NeRF (Mildenhall et al. 2021), have attracted significant attention for novel view synthesis from images. Subsequent methods, such as 3DGS (Kerbl et al. 2023), provide more realistic view synthesis and explicit scene representations for complex scenes, while significantly accelerating training and rendering. These innovations (Pumarola et al. 2021; Niemeyer et al. 2022; Gao et al. 2024; Cheng et al. 2024; Yang et al. 2024; Sun et al. 2024; Zhang et al. 2025) significantly advance the field of novel view synthesis, providing a solid foundation for PAD approaches.

### Relative Camera Pose Estimation

To address relative camera pose estimation, Relative PN (Melekhov et al. 2017) demonstrates that feeding image pairs into a neural network enables accurate relative pose estimation. Subsequently, RPNet (En, Lechervy, and Jurie 2018) proposes an end-to-end framework that directly estimates the relative camera poses from image pairs. Based on this framework, RCPNet (Yang, Liu, and Zell 2020) extends it to applications in autonomous navigation for unmanned aerial vehicles. More recently, RelMobNet (Rajendran et al. 2022) introduces a two-stage training strategy that avoids balancing translation and rotation losses, allowing pose estimation without requiring camera parameters. However, these methods fail to leverage multiple image pairs to enhance spatial understanding.

## Method

The overall framework of RPE-PAD is illustrated in Figure 2. In this section, we first introduce the PAD setting and the background of relative camera pose estimation. Subsequently, we describe three core components of RPE-PAD: the Random View Synthesis Scheme (RVSS), the Iterative Relative Pose Refinement Network (IRPRN), and the Multi-Pair Training Strategy (MPTS).

### Preliminary

In the PAD setting, multi-pose anomaly-free images with corresponding camera pose information are utilized for training. During inference, the objective is to detect whether a given query image without camera pose information is anomalous and to localize pixel-level anomalous regions.

Before training, we calculate the relative camera pose between two images. Let $C_1$ and $C_2$ represent the camera coordinate systems for the images $I_1$ and $I_2$, respectively. We define $\mathbf{R}_1$, $\mathbf{R}_2$ as rotation matrices and $\mathbf{t}_1$, $\mathbf{t}_2$ as translation vectors corresponding to $I_1$ and $I_2$. The pairs $(\mathbf{R}_1, \mathbf{t}_1)$ and $(\mathbf{R}_2, \mathbf{t}_2)$ represent the transformations that map 3D points from the world coordinate system to the camera coordinate systems $C_1$ and $C_2$, respectively. The transformation from $C_1$ to $C_2$ is expressed by the relative pose matrix $\mathbf{P}_{1,2}$, where $\mathbf{R}_{1,2}$ is the rotation matrix and $\mathbf{t}_{1,2}$ is the translation vector, as defined below:

$$\mathbf{P}_{1,2} = \begin{bmatrix} \mathbf{R}_{1,2} & \mathbf{t}_{1,2} \\ 0 & 1 \end{bmatrix}; \quad \begin{cases} \mathbf{R}_{1,2} = \mathbf{R}_2 \mathbf{R}_1^T, \\ \mathbf{t}_{1,2} = \mathbf{R}_1(\mathbf{t}_2 - \mathbf{t}_1). \end{cases} \tag{1}$$

We represent the rotation matrices $\mathbf{R}_1$, $\mathbf{R}_2$ and $\mathbf{R}_{1,2}$ using quaternions $\mathbf{q}_1$, $\mathbf{q}_1$ and $\mathbf{q}_{1,2}$, as quaternions can be easily normalized to yield valid rotation from arbitrary 4-D vector.

Given an input image pair, the relative camera pose vector $\mathbf{p}$ predicted by IRPRN is expressed as follows:

$$\mathbf{p} = [\mathbf{q}, \mathbf{t}], \tag{2}$$

where $\mathbf{q}$ is the rotation quaternion and $\mathbf{t}$ denotes the translation vector.

To jointly learn rotation and translation while balancing their weights, a loss function (Yang, Liu, and Zell 2020) with automatic weighting based on homoscedastic uncertainty is formulated as follows:

$$\mathcal{L}_R(\mathbf{p}) = \mathcal{L}_t(\mathbf{p}) \exp(-s_t) + s_t + \mathcal{L}_q(\mathbf{p}) \exp(-s_q) + s_q, \tag{3}$$

where

$$\mathcal{L}_q(\mathbf{p}) = ||\mathbf{q} - \overline{\mathbf{q}}||_2, \quad \mathcal{L}_t(\mathbf{p}) = ||\mathbf{t} - \overline{\mathbf{t}}||_2. \tag{4}$$

Here, $\mathcal{L}_q$ and $\mathcal{L}_t$ represent rotation and translation losses, respectively, and $\mathcal{L}_R$ is the total relative pose loss. The Euclidean norm $||\cdot||_2$ is employed to minimize the discrepancy between the estimation $(\mathbf{q}, \mathbf{t})$ and the ground truth $(\overline{\mathbf{q}}, \overline{\mathbf{t}})$. The weights $s_q$ and $s_t$ serve to balance the contributions of rotation and translation losses, ensuring that the rotation and translation regression errors of the network are unbiased.

During anomaly detection and localization, features are extracted from the query image and corresponding normal image using a fixed feature extractor, followed by pixel-level comparison to generate the anomaly score map.

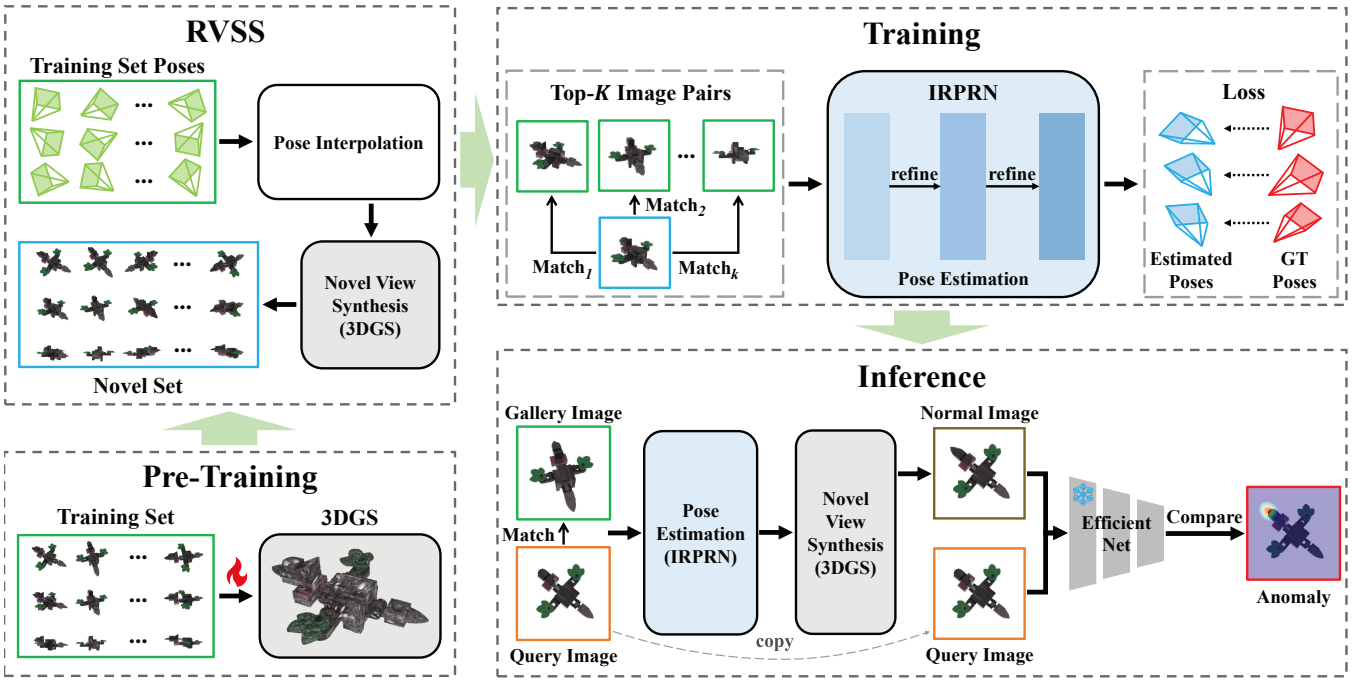



Figure 2: The framework of our RPE-PAD. Our method first pre-trains a 3DGS model using training set poses and images. In the RVSS, we generate new camera poses by adding Gaussian perturbations to the training set poses and employ the pre-trained 3DGS model to render corresponding novel views. These novel views are matched with training set views to form image pairs. Following the MPTS, we utilize multiple image pairs to train the IRPRN. During inference, the query image is matched with training set and novel set images. The pose is then estimated using the IRPRN and the normal image is rendered by the pre-trained 3DGS model. Finally, we apply a fixed feature extractor to compare the query image with the rendered normal image, enabling anomaly detection and localization.

## Random View Synthesis Scheme (RVSS)

We introduce the Random View Synthesis Scheme (RVSS), which generates new camera poses by applying Gaussian perturbations to the training set poses. These generated camera poses are then utilized to render novel views using a pre-trained 3DGS model. The process is shown in Figure 2.

Specifically, given a training set image $I$ with camera pose $\mathbf{P}$, we generate $M$ perturbation poses $\mathbf{P}' = \{\mathbf{P}'_1, \mathbf{P}'_2, \ldots, \mathbf{P}'_M\}$ by adding Gaussian noise to $\mathbf{P}$. The rotation and translation noises are sampled independently from Gaussian distributions with means $\mu_r$, $\mu_t$ and standard deviations $\sigma_r$, $\sigma_t$, respectively. Novel views $I' = \{I'_1, I'_2, \ldots, I'_M\}$ are then rendered using a pre-trained 3DGS model, forming the novel set.

During the training of RPE-PAD, we employ LoFTR (Sun et al. 2021) to match each image in the novel set $I_N = \{I^1_N, I^2_N, \ldots, I^m_N\}$ to the training set $I_T = \{I^1_T, I^2_T, \ldots, I^n_T\}$. The pairs $(I^i_T, I^j_N)$ are utilized as matched image pairs for training IRPRN. During inference, the test set $I_Q = \{I^1_Q, I^2_Q, \ldots, I^m_Q\}$ is matched with the gallery $I_G = \{I^1_G, I^2_G, \ldots, I^n_G\}$, which includes both the training set and the novel set, generating image pairs $(I^i_G, I^j_Q)$.

This scheme offers a cost-effective approach to generating training data. Notably, it effectively enhances the generalization ability of the relative pose estimation network to unseen views and significantly improves performance.

## Iterative Relative Pose Refinement Network (IRPRN)

To accurately estimate the relative camera pose, we incorporate a hierarchical coarse-to-fine pose iterative refinement strategy into the relative pose estimation process, proposing the Iterative Relative Pose Refinement Network (IRPRN). The matched image pairs are first processed through the backbone to extract features and produce coarse relative camera poses. Each subsequent layer iteratively refines the coarse camera poses, ultimately producing the fine relative camera poses as the network outputs. The specific structure of IRPRN is illustrated in Figure 3.

**Feature Extraction.** Given two input images $I_G, I_Q \in \mathbb{R}^{H\times W\times 3}$, where $H$ and $W$ denote height and width, we utilize a ResNet-50 backbone to extract multi-scale features. Among these, the top-level feature $F \in \mathbb{R}^{C\times \frac{H}{32}\times \frac{W}{32}}$ is selected as the image representation. The features are averaged along the channel dimension to produce pose features, which are then fed into the IRPRN. Subsequently, a coarse relative rotation quaternion $\mathbf{q}_{coarse}$ and a coarse relative translation vector $\mathbf{t}_{coarse}$ are generated by the rotation head and translation head, respectively. Both the rotation and translation heads employ multilayer perceptron (MLP) architectures.

**Iterative Relative Pose Refinement.** In the relative pose

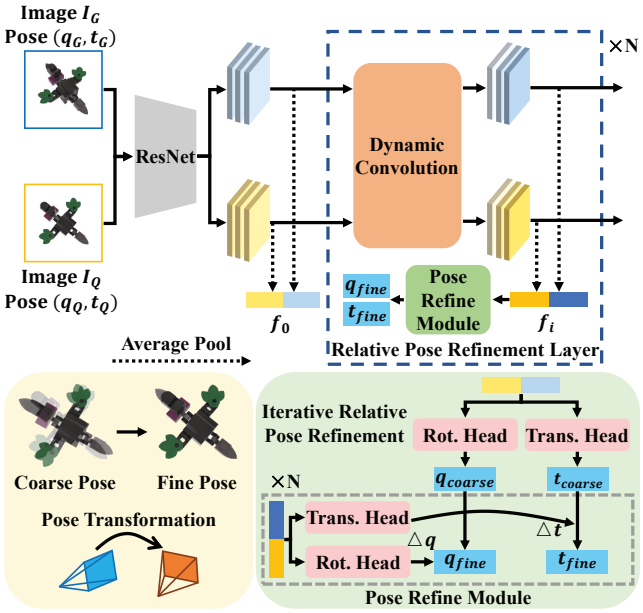



Figure 3: The architecture of our IRPRN. In each layer, the pose refine module refines the relative camera pose from an initial coarse pose or the previous layer. The structure of the pose refine module is displayed at the bottom-right of the figure, and its refinement effect is shown at the bottom-left.

refinement layers, we employ dynamic convolution (Chen et al. 2020), where convolutional kernel weights are adaptively adjusted based on the input image content. This adaptive mechanism improves the ability of the network to represent features, enabling it to focus on regions that are more informative for pose estimation, such as edges and corners.

Dynamic convolution refines the features in each relative pose refinement layer. Each layer predicts offsets for translation and rotation based on the extracted features. These offsets are subsequently added to the outputs of the previous layer to generate the outputs of the current layer. This process can be expressed as follows:

$$\Delta \mathbf{q}^i = \mathrm{Head}_r^i(f_i), \quad \Delta \mathbf{t}^i = \mathrm{Head}_t^i(f_i), \tag{5}$$

where $\Delta \mathbf{q}^i$ and $\Delta \mathbf{t}^i$ are the rotation and translation offsets of layer $i$, while $\mathrm{Head}_r^i$ and $\mathrm{Head}_t^i$ denote rotation and translation heads of layer $i$, respectively. $f_i$ is the pose feature of the layer $i$. The rotation quaternion $\mathbf{q}^i$ and the translation vector $\mathbf{t}^i$ in the layer $i$ are updated to:

$$\mathbf{q}^i = \mathbf{q}^{i-1} + \Delta \mathbf{q}^i, \quad \mathbf{t}^i = \mathbf{t}^{i-1} + \Delta \mathbf{t}^i. \tag{6}$$

The total relative pose loss of the network $\mathcal{L}_N$ is given by:

$$\mathcal{L}_N(\mathbf{p}) = \mathcal{L}_R(\mathbf{p}^c) + \sum_{i=1}^{N} \mathcal{L}_R(\mathbf{p}^i), \tag{7}$$

where $\mathcal{L}_R(\mathbf{p}^c)$, as defined in Equation (3), represents the coarse relative pose loss, $\mathcal{L}_R(\mathbf{p}^i)$ denotes the relative pose loss for each refinement layer, and $N$ is the number of relative pose refinement layers in the network.

The hierarchical coarse-to-fine pose refinement network iteratively refines the coarse relative camera pose with each layer, progressively approximating the ground truth relative pose. Therefore, this coarse-to-fine strategy enables the IRPRN to achieve more accurate camera pose estimation.

### Multi-Pair Training Strategy (MPTS)

The one-to-one matching strategy often leads the model to focus on image pairs with small relative camera pose transformations. However, in many real-world scenarios with large relative camera pose transformations, the model may struggle to estimate relative pose accurately. Moreover, a single matched image pair provides information about the target object from only two viewpoints. In contrast, incorporating information from more viewpoints allows the model to estimate relative pose more precisely.

Motivated by these findings, we propose the Multi-Pair Training Strategy (MPTS). For each image in the novel set, we select the top $K$ images that are most similar to it from the training set during the matching process, thereby forming $K$ matched image pairs.

The $K$ matched image pairs are employed as simultaneous inputs to train the model. The total loss $\mathcal{L}$ is defined as:

$$\mathcal{L}(\mathbf{p}) = \sum_{i=1}^{K} \frac{2(K+1-i)}{K(K+1)} \mathcal{L}_N(\mathbf{p}_i), \tag{8}$$

where $\mathcal{L}_N(\mathbf{p}_i)$, derived from Equation (7), represents the IRPRN's relative pose loss of the $i$-th matched image pair.

With $K$ matched image pairs, the model can explore a larger space of relative pose transformations. This strategy enables a more comprehensive understanding of the target object's structure and details in 3D space, which supports accurate estimation of relative camera pose even in scenarios with large relative pose transformations. This approach significantly enhances the accuracy of relative camera pose estimation while improving the ability of the model to handle challenging scenarios effectively.

## Experiments

### Settings

**Datasets and Evaluation Metrics** We evaluate our method on two datasets. The MAD dataset (Zhou et al. 2023) contains 20 categories of LEGO toys with synthetic anomalies such as surface burrs, discolored stains, and missing components. This dataset is constructed by Blender and the Ldraw (LEGO parts library). The RAD dataset (Zhou et al. 2024) is a real-world collection of 13 categories, including kitchenware, toys, and daily necessities. The anomalies in this dataset include surface scratches, missing parts, discolored stains, and squeezed deformations.

Following previous work, we adopt the Area Under the Receiver Operating Characteristic curve (AUROC) as the primary evaluation metric. $\mathrm{AUROC_P}$ and $\mathrm{AUROC_I}$ evaluate pixel-level anomaly localization and image-level anomaly detection, respectively. Additionally, to mitigate the scoring bias caused by varying anomaly sizes, we adopt the Area Under the Per-Region Overlap (AUPRO), a widely recognized metric for evaluating pixel-level anomaly localization.

| Category | CFA | CFlow | UniAD | FAVAE | OmniAD | SplatPose | Ours |
|---|---|---|---|---|---|---|---|
| Gorilla | 41.8 | 69.2 | 56.6 | 46.8 | **93.6** | 91.7 | 88.3 |
| Unicorn | 85.6 | 82.3 | 73.0 | 68.3 | 94.0 | **97.9** | 96.4 |
| Mallard | 36.6 | 74.9 | 70.0 | 33.6 | 84.7 | 97.4 | **97.5** |
| Turtle | 58.3 | 51.0 | 50.2 | 82.8 | 95.6 | 97.2 | **98.3** |
| Whale | 77.7 | 57.0 | 75.5 | 62.5 | 82.5 | 95.4 | **98.3** |
| Bird | 78.4 | 75.6 | 74.7 | 73.3 | 92.4 | **94.0** | 92.1 |
| Owl | 74.0 | 76.5 | 65.3 | 62.5 | 88.2 | 86.8 | **89.7** |
| Sabertooth | 64.2 | 71.3 | 61.2 | 82.4 | 95.7 | 95.2 | **96.2** |
| Swan | 66.7 | 67.4 | 57.5 | 50.6 | 86.5 | **93.0** | 89.1 |
| Sheep | 86.5 | 80.9 | 70.4 | 74.9 | 90.1 | **96.7** | 92.6 |
| Pig | 66.7 | 72.1 | 54.6 | 52.5 | 88.3 | 96.1 | **96.5** |
| Zalika | 52.1 | 66.9 | 50.5 | 34.6 | 88.2 | **89.9** | 89.2 |
| Phoenix | 65.9 | 64.4 | 55.4 | 65.2 | 82.3 | 84.2 | **85.1** |
| Elephant | 71.7 | 70.1 | 59.3 | 49.1 | 92.5 | **94.7** | 94.5 |
| Parrot | 69.8 | 67.9 | 53.4 | 46.1 | 97.0 | 96.1 | **98.0** |
| Cat | 68.2 | 65.8 | 53.1 | 53.2 | 84.9 | 82.4 | **88.7** |
| Scorpion | 91.4 | 79.5 | 69.5 | 66.9 | 91.5 | **99.2** | 93.7 |
| Obesobeso | 80.6 | 80.0 | 67.7 | 58.2 | 97.1 | 95.7 | **97.3** |
| Bear | 78.7 | 81.4 | 65.1 | 52.8 | 98.8 | 98.9 | **99.1** |
| Puppy | 53.7 | 71.4 | 55.6 | 43.5 | 93.5 | **96.1** | 94.3 |
| Mean | 68.2 | 71.3 | 62.2 | 58.0 | 90.9 | **93.9** | 93.8 |

Table 1: Anomaly detection results on the MAD dataset. Best results are highlighted in bold, and second-best results are underlined.

| Metrics | OmniAD | SplatPose | Ours |
|---|---|---|---|
| $\text{AUROC}_\text{P}$ | 98.4 | **99.5** | 99.2 |
| AUPRO | 86.6 | 95.8 | **96.2** |

Table 2: Anomaly localization results averaged across all categories on the MAD dataset.

**Implementation** The 3DGS model for novel view synthesis is trained on each category of the MAD and RAD datasets for 30,000 iterations, with image resolutions of 800×800 and 1291×721, respectively. In the RVSS, three random perturbation poses are generated for each camera pose in the training set, using hyperparameters $\mu_r = \mu_t = 0$, $\sigma_r = 0.02$, and $\sigma_t = 0.1$. In the MPTS, we select the top three most similar images for each image in the novel set. The IRPRN consists of three layers and is trained for 35 epochs, with an initial learning rate of 0.001, which decays to 0.0001 after 20 epochs. The batch size is set to 32. During inference, we utilize a pre-trained EfficientNet-B4 (Tan and Le 2019) model as the feature extractor. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

## Comparison with State-of-the-art Methods

We compare the proposed RPE-PAD with several advanced 2D image and pose-agnostic anomaly detection methods, including CFlow (Gudovskiy, Ishizaka, and Kozuka 2022), CFA (Lee, Lee, and Song 2022), FAVAE (Dehaene and Eline 2020), UniAD (You et al. 2022), OmniposeAD (abbreviated as OmniAD in tables) (Zhou et al. 2023) and SplatPose (Kruse et al. 2024).

**Anomaly Detection and Localization on MAD** Table 1 reports the image-level anomaly detection results on the MAD dataset. Our method achieves the best performance in 11 out of 20 categories. Compared to OmniposeAD, our RPE-PAD improves the average $\text{AUROC}_\text{I}$ by 2.9%. Additionally, our method exhibits competitive performance compared to the state-of-the-art SplatPose.

For pixel-level anomaly localization on the MAD dataset, as shown in Table 2, our method outperforms OmniposeAD in average $\text{AUROC}_\text{P}$ and AUPRO across all categories. Our RPE-PAD exhibits competitive performance compared to SplatPose and achieves the highest AUPRO score, indicating outstanding accuracy in anomaly localization.

As illustrated in Figure 4, our method exhibits closer alignment with ground truth in anomaly localization compared to the other two methods, demonstrating superior accuracy and practical effectiveness in the PAD task. Additional qualitative results of our RPE-PAD are provided in Appendix.

| Category | CFA | CFlow | UniAD | FAVAE | OmniAD | SplatPose | Ours |
|---|---|---|---|---|---|---|---|
| Binderclip | 61.5 | 68.1 | 53.6 | 66.2 | 48.0 | 52.1 | **72.1** |
| Bowl | 58.5 | 66.5 | 52.2 | 20.1 | 37.0 | 80.6 | **84.2** |
| Box | 70.5 | 75.8 | 52.9 | 51.3 | 62.0 | 55.5 | **84.9** |
| Can | 73.0 | **77.3** | 51.9 | 39.5 | 46.3 | 37.7 | 69.6 |
| Charger | 61.4 | **80.4** | 53.3 | 35.3 | 51.5 | 57.4 | 71.3 |
| Cup1 | 60.1 | 72.2 | 55.3 | 34.6 | 48.5 | **81.2** | 77.6 |
| Cup2 | 64.1 | 65.3 | 54.1 | 24.8 | 54.4 | 48.1 | **65.9** |
| Glue | 56.5 | 66.4 | 50.5 | 46.1 | 62.6 | 42.3 | **81.5** |
| Phonecase | 73.8 | 66.1 | 51.7 | 32.3 | 53.4 | 32.7 | **82.7** |
| Rubberduck | 62.9 | 57.6 | 55.2 | 33.2 | 33.0 | 26.9 | **69.1** |
| Spoon | 74.0 | 82.2 | 50.0 | 53.1 | 37.3 | 70.9 | **88.3** |
| Spraybottle | 59.3 | 60.4 | 50.1 | 44.8 | 51.6 | 60.2 | **86.9** |
| Tennisball | 83.5 | 79.4 | 51.3 | 49.3 | 37.6 | **91.6** | 71.4 |
| Mean | 66.1 | 70.6 | 52.5 | 40.8 | 47.9 | 56.7 | **77.8** |

Table 3: Anomaly detection results on the RAD dataset.

**Anomaly Detection and Localization on RAD** To evaluate the performance of our method in real-world scenarios, we conduct experiments on the RAD dataset. As reported in Table 3, our method outperforms other approaches in most categories and achieves the highest average $\text{AUROC}_\text{I}$ across all categories. Our RPE-PAD achieves state-of-the-art performance on this dataset, with an average $\text{AUROC}_\text{I}$ of 77.8%. The performance decline of OmniposeAD and SplatPose on this dataset is significant, showing their limited applicability in real-world scenarios.

These results confirm the strong adaptability and robustness of our approach in addressing complex and challenging real-world scenarios. Additional experimental results on the RAD dataset are available in Appendix.

## Inference Time on MAD

Our RPE-PAD significantly reduces inference time in comparison to methods based on query-specific framework, such as OmniposeAD and SplatPose, as detailed in Table 4.

The inference process is divided into three stages, as illustrated in Figure 1(a). In the pose estimation stage, our method adopts a query-independent framework, allowing simultaneous pose estimation on all query images. Therefore, our method consumes significantly less time at this

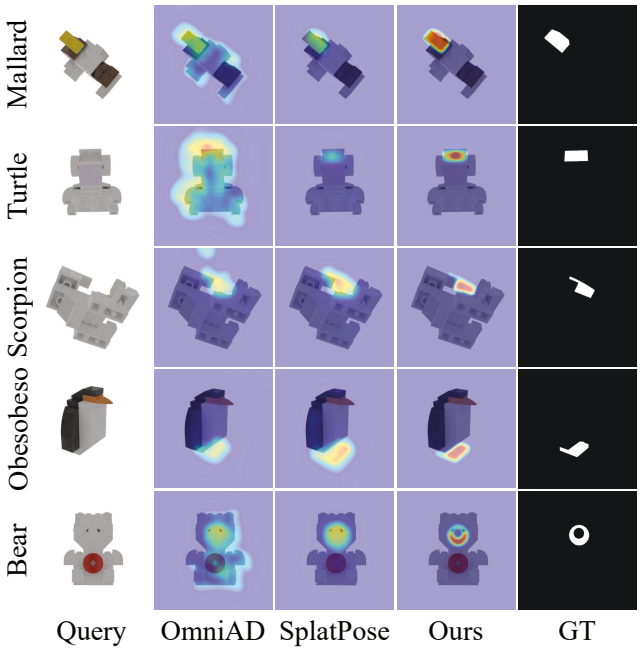



Figure 4: Visualization of anomaly localization results on the MAD dataset.

| Methods | Inference Time(s) | | | |
|---|---|---|---|---|
| | Pose Estimation | Rendering | Detection | Total |
| OmniAD | 55.047 | 3.277 | **0.003** | 58.327 |
| SplatPose | 4.095 | 0.005 | **0.003** | 4.103 |
| Ours | **0.006** | **0.002** | **0.003** | **0.011** |

Table 4: Inference time comparison of our method, SplatPose, and OmniAD. Inference is divided into three stages: pose estimation, rendering and detection. We compare the time consumption of each stage and the total inference.

| Metrics | OmniAD | SplatPose | Ours |
|---|---|---|---|
| $AUROC_I$ | 60.8 | 60.9 | **93.8** |
| $AUROC_P$ | 91.3 | 91.0 | **99.2** |
| AUPRO | 69.5 | 68.5 | **96.2** |

Table 5: Generalization performance comparison among our method, OmniAD, and SplatPose.

stage compared to the other two methods based on a query-specific framework. In the rendering stage, our method renders normal images more efficiently by utilizing the advanced 3DGS model, outperforming OmniposeAD, which relies on the slower NeRF model. The time gap between our method and SplatPose can be attributed to differences in 3DGS model settings. In the detection stage, all methods have identical time cost due to the same feature extractor.

Considering the acceleration in the pose estimation and rendering phases, our method is 5302 $\times$ faster than OmniposeAD and 373 $\times$ faster than SplatPose in total inference time. These results show that our method achieves both advanced performance and the fastest inference speed.

| # | Settings | | | $AUROC_P$ | $AUROC_I$ |
|---|---|---|---|---|---|
| | RVSS | IRPRN | MPTS | | |
| 1 | - | - | - | 92.8 | 62.5 |
| 2 | ✓ | - | - | 96.6 | 77.8 |
| 3 | ✓ | ✓ | - | 98.0 | 87.3 |
| 4 | ✓ | - | ✓ | 98.1 | 89.4 |
| 5 | ✓ | ✓ | ✓ | **99.2** | **93.8** |

Table 6: Ablation studies on the MAD dataset.

## Generalization on MAD

To demonstrate the generalization improvement of our method, we fix the model parameters during inference and apply them to all query images for OmniposeAD and SplatPose. For a fair comparison, we conduct multiple training runs for each method and select the model parameters that yield the best overall performance. In this setting, inference is performed using fixed parameters without any per-image iterative optimization.

As shown in Table 5, our method maintains stable performance under the generalization constraint. This stability can be attributed to the query-independent framework of RPE-PAD, which eliminates the necessity of re-optimization for each query image. In contrast, the performance of the other two methods based on the query-specific framework drops significantly, indicating their limited generalization ability. Overall, RPE-PAD achieves the best performance under a strict generalization constraint, demonstrating superior adaptability and robustness.

## Ablation Studies

We verify the effectiveness of each component on the MAD dataset. As shown in Table 6, all the proposed components significantly improve the overall performance of the model. By enriching the gallery with diverse viewpoints, the RVSS facilitates accurate relative pose estimation and leads to significant performance improvement. The IRPRN refines the relative camera pose through a coarse-to-fine strategy and achieves more accurate relative pose estimation. Additionally, by simultaneously leveraging multiple image pairs, the MPTS enables the model to explore a larger space of relative pose transformations and exhibits robust performance in challenging scenarios. Additional ablation studies are provided in Appendix.

# Conclusion

In this paper, we propose a novel framework for pose-agnostic anomaly detection using relative pose estimation to address the limitation of generalization within existing methods. The RVSS augments training data by adding Gaussian perturbations to original poses and rendering the corresponding views. We leverage multiple image pairs to train the proposed IRPRN, thus expanding the relative camera pose transformation space and improving pose estimation accuracy. Extensive experiments conducted on the MAD and RAD datasets demonstrate that our method delivers outstanding performance, significantly accelerates inference speed, and exhibits high generalization capability.

## Acknowledgments

This work was supported in part by Grant NSFC/RGC N_CUHK 415/19, the National Natural Science Foundation of China (62271203, 62202173, 62371190, 61971234), Grant ITF ITS/173/22FP, Grant RGC 14300219, 14302920, 14301121, CUHK Direct Grant for Research.

## References

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD – A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9592–9600.

Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; and Liu, Z. 2020. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11030–11039.

Cheng, K.; Long, X.; Yang, K.; Yao, Y.; Yin, W.; Ma, Y.; Wang, W.; and Chen, X. 2024. GaussianPro: 3D gaussian splatting with progressive propagation. In *Forty-first International Conference on Machine Learning*.

Cho, Y.; Eum, S.; Im, J.; Ali, Z.; Choo, H.; and Park, U. 2023. Deep photo-geometric loss for relative camera pose estimation. *IEEE Access*, 11: 130319–130328.

Dai, S.; Wu, Y.; Li, X.; and Xue, X. 2024. Generating and reweighting dense contrastive patterns for unsupervised anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1454–1462.

Dehaene, D.; and Eline, P. 2020. Anomaly localization by modeling perceptual features. *arXiv preprint arXiv:2008.05369*.

En, S.; Lechervy, A.; and Jurie, F. 2018. RPNet: An end-to-end network for relative camera pose estimation. In *Proceedings of the European Conference on Computer Vision Workshops*.

Gao, X.; Yang, Z.; Zhao, Y.; Sun, Y.; Jin, X.; and Zou, C. 2024. A general implicit framework for fast NeRF composition and rendering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1833–1841.

Gudovskiy, D.; Ishizaka, S.; and Kozuka, K. 2022. CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 98–107.

Jiang, B.; Xie, Y.; Li, J.; Li, N.; Chen, B.; and Xia, S.-T. 2024. IGSPAD: Inverting 3D gaussian splatting for pose-agnostic anomaly detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 10229–10237.

Kerbl, B.; Kopanas, G.; Leimkuhler, T.; and Drettakis, G. 2023. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4): 1–14.

Kruse, M.; Rudolph, M.; Woiwode, D.; and Rosenhahn, B. 2024. SplatPose & detect: Pose-agnostic 3D anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 3950–3960.

Lee, S.; Lee, S.; and Song, B. C. 2022. CFA: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10: 78446–78454.

Leng, J.; Tan, M.; Gao, X.; Lu, W.; and Xu, Z. 2022. Anomaly warning: Learning and memorizing future semantic patterns for unsupervised ex-ante potential anomaly prediction. In *Proceedings of the 30th ACM International Conference on Multimedia*, 6746–6754.

Liu, J.; Wu, K.; Nie, Q.; Chen, Y.; Gao, B.-B.; Liu, Y.; Wang, J.; Wang, C.; and Zheng, F. 2024a. Unsupervised continual anomaly detection with contrastively-learned prompt. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 3639–3647.

Liu, X.; Wang, J.; Leng, B.; and Zhang, S. 2024b. Dual-modeling decouple distillation for unsupervised anomaly detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5035–5044.

Melekhov, I.; Ylioinas, J.; Kannala, J.; and Rahtu, E. 2017. Relative camera pose estimation using convolutional neural networks. In *Advanced Concepts for Intelligent Vision Systems*, 675–687.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Niemeyer, M.; Barron, J. T.; Mildenhall, B.; Sajjadi, M. S.; Geiger, A.; and Radwan, N. 2022. RegNeRF: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5480–5490.

Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-NeRF: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10318–10327.

Rajendran, P. K.; Mishra, S.; Vecchietti, L. F.; and Har, D. 2022. RelMobNet: End-to-end relative camera pose estimation using a robust two-stage training. In *Proceedings of the European Conference on Computer Vision Workshops*, 238–252.

Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; and Zhou, X. 2021. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8922–8931.

Sun, X.; Lee, J. C.; Rho, D.; Ko, J. H.; Ali, U.; and Park, E. 2024. F-3DGS: Factorized coordinates and representations for 3D gaussian splatting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7957–7965.

Tan, M.; and Le, Q. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 6105–6114.

Von Stumberg, L.; Wenzel, P.; Yang, N.; and Cremers, D. 2020. LM-Reloc: Levenberg-Marquardt based direct visual relocalization. In *2020 International Conference on 3D Vision*, 968–977.

Yan, X.; Zhang, H.; Xu, X.; Hu, X.; and Heng, P.-A. 2021. Learning semantic context from normal samples for unsupervised anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3110–3118.

Yang, C.; Liu, Y.; and Zell, A. 2020. RCPNet: Deep-learning based relative camera pose estimation for UAVs. In *2020 International Conference on Unmanned Aircraft Systems*, 1085–1092.

Yang, Z.; Gao, X.; Zhou, W.; Jiao, S.; Zhang, Y.; and Jin, X. 2024. Deformable 3D gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20331–20341.

Yen-Chen, L.; Florence, P.; Barron, J. T.; Rodriguez, A.; Isola, P.; and Lin, T.-Y. 2021. iNeRF: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1323–1330.

You, Z.; Cui, L.; Shen, Y.; Yang, K.; Lu, X.; Zheng, Y.; and Le, X. 2022. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35: 4571–4584.

Yu, J.; Zheng, Y.; Wang, X.; Li, W.; Wu, Y.; Zhao, R.; and Wu, L. 2021. FastFlow: Unsupervised anomaly detection and localization via 2D normalizing flows. *arXiv preprint arXiv:2111.07677*.

Zhang, W.; Zhang, L.; Hu, P.; Ma, L.; Zhuge, Y.; and Lu, H. 2025. Bootstraping clustering of gaussians for view-consistent 3D scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10166–10175.

Zhou, K.; Cao, Y.; Kim, T.; Zhao, H.; Dong, H.; Ting, K. M.; and Zhu, Y. 2024. RAD: A dataset and benchmark for real-life anomaly detection with robotic observations. *arXiv preprint arXiv:2410.00713*.

Zhou, Q.; Li, W.; Jiang, L.; Wang, G.; Zhou, G.; Zhang, S.; and Zhao, H. 2023. PAD: A dataset and benchmark for pose-agnostic anomaly detection. *Advances in Neural Information Processing Systems*, 36: 44558–44571.