SplatPose & Detect: Pose-Agnostic 3D Anomaly Detection

Mathis Kruse, Marco Rudolph, Dominik Woiwode, Bodo Rosenhahn
Institute for Information Processing, Leibniz University Hannover
{kruse, rudolph, woiwode, rosenhahn}@tnt.uni-hannover.de

Abstract

Detecting anomalies in images has become a well-explored problem in both academia and industry. State-of-the-art algorithms are able to detect defects in increasingly difficult settings and data modalities. However, most current methods are not suited to address 3D objects captured from differing poses. While solutions using Neural Radiance Fields (NeRFs) have been proposed, they suffer from excessive computation requirements, which hinder real-world usability. For this reason, we propose the novel 3D Gaussian splatting-based framework SplatPose which, given multi-view images of a 3D object, accurately estimates the pose of unseen views in a differentiable manner, and detects anomalies in them. We achieve state-of-the-art results in both training and inference speed, and detection performance, even when using less training data than competing methods. We thoroughly evaluate our framework using the recently proposed Pose-agnostic Anomaly Detection benchmark and its multi-pose anomaly detection (MAD) data set.

1. Introduction

In the industrial manufacturing of products, small errors and slight deviations from the norm inevitably lead to some defective products. To ensure both quality and effectiveness of the production process, the detection of defects or irregularities becomes necessary. Traditionally, this is done by a human supervisor trained in recognizing the special kinds of defects that may appear. In most cases, this is far more costly and prone to human error than fully automated solutions. Data of different kinds of defects are however scarcely available. This unavailability of anomalous samples leads standard classification approaches to underperform in these settings. More recently anomaly detection algorithms have shifted to train on normal data only [2], which is usually more widely available.

While state-of-the-art methods already perform very well on image data [2, 30], more challenging benchmarks have emerged [4, 5], with some including pose informa-
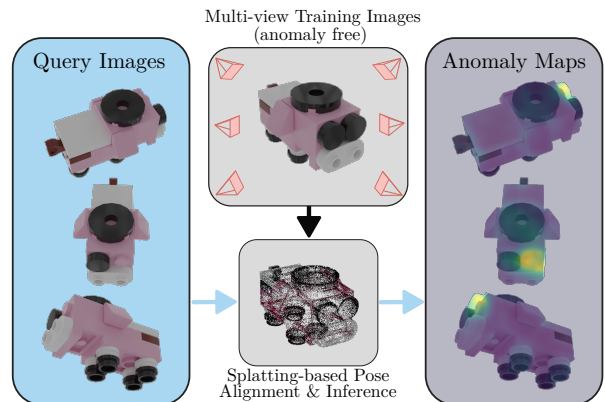

Figure 1. Example of SplatPose. A cloud representation is built from multi-view training images. During inference, query images with unknown poses are aligned and an anomaly map localizes their defects within the 3D object, irrespective of pose.

tion [47]. For this exact task, the MAD data set has been proposed [47], consisting of different views of 20 custom-built LEGO® figures and including anomalies such as missing bricks, discolorations, and artifacts caused by excess material. As standard anomaly detection approaches do not differentiate between poses, they underperform heavily in this setting [47]. In practice, objects in production may be randomly placed or rotated, while cameras are not adjustable as freely. This makes the MAD data set a good benchmark for the anomaly detection community to verify the robustness of their methods.

OmniposeAD has been proposed in conjunction with the MAD data set. It captures the anomaly-free volume density within a NeRF and aligns test images with unknown poses using the iNeRF framework [21, 47]. Despite working in principle, the high computational costs make it impractical for real-world usage. Also, NeRFs are known to struggle in sparse-view settings [25], limiting their applicability.

Recent progress on novel-view synthesis using 3D Gaussian Splatting (3DGS) showed that the volume density of complex three-dimensional objects can be encoded into a cloud of three-dimensional Gaussians using multi-view im-

ages and their respective camera poses [17]. Advances in this field enabled practitioners to generate realistic unseen views of complex objects while still achieving high frame rates suitable for real-time execution [17].

In this work, we propose *SplatPose* to tackle 3D pose-agnostic anomaly detection using 3DGS. The explicit 3D point cloud representation combined with the efficient rasterization results in up to **55** times faster training and **13** times faster inference times than other top competitors. We apply transformations to the 3D point cloud to perform pose estimation in a differentiable manner, letting us render defect-free images of arbitrary pose. Irrespective of a query object's pose, anomalies can be detected by matching features between the aligned rendering and query image, as shown in Fig. 1. SplatPose improves the anomaly detection performance on MAD, reaching a new state-of-the-art in both detection and segmentation. Even when using only 60% of training data, we are able to outperform all other methods, while they utilize the entire training set. Our contributions may be summarized as follows:

- We propose *SplatPose* for pose-agnostic anomaly detection, using 3D Gaussian Splatting to perform pose estimation in a differentiable manner.
- We achieve new state-of-the-art results on pose-agnostic anomaly detection.
- The proposed method is significantly more resource-efficient, both in terms of speed and required data, than other top competitors.
- Code is available on GitHub[1].

2. Related Work

As this paper deals with performing anomaly detection on multi-view images using novel-view synthesis techniques, a rough overview is given for both fields.

2.1. Industrial Anomaly Detection

In semi-supervised anomaly detection, also known as one-class classification or novelty detection, the task is to differentiate between anomalous and normal samples, while only learning from normal samples.

One line of work uses generative models like generative adversarial networks (GANs) or Autoencoders, which, when trained only on normal data, fail to reconstruct anomalous regions, thus enabling anomaly detection [2, 26, 34]. Another line of work tries to model the problem using density estimation, by assigning high likelihoods to the normal samples they are trained on, while anomalies receive much lower likelihoods. Leveraging complex pre-trained feature embeddings, and assuming them to be distributed as multivariate Gaussians, the Mahalanobis distance has been used for this density estimation [7, 28]. In

order to forego this distributional assumption, normalizing flows have been used, as they can map an arbitrarily complex feature distribution into a tractable Gaussian distribution [10, 29, 31, 32]. A proxy for the density can also be constructed with the nearest-neighbor distance, which is measured between a test sample and all normal features gathered from the training set and saved within a memory bank [6, 19, 30, 41]. Many of these methods need to choose meaningful features from pre-trained networks, which in itself is not trivial [13]. Student-Teacher architectures have also seen success [3, 8, 33, 40], as the student only learns to mimic the teacher on normal data and the regression error is also used as a proxy for density. Lastly, anomaly detection can also be learned by constructing synthetic anomalies and casting the problem into a supervised classification problem [20, 22, 46]. However, synthetic anomalies introduce a bias that may cause detection to fail on unseen anomalies.

With current methods excelling at detecting anomalies in high-quality images, anomaly detection data sets have shifted their attention to more difficult problems [4, 5]. The MAD data set serves as another example of this, as images no longer constrain themselves to fixed viewpoints, making reasoning about the object's pose crucial [47].

2.2. Novel-View Synthesis

In computer vision, 3D reconstruction is the task of estimating the three-dimensional representation of an object or a scene given several 2D images [11]. More specifically, one may want to recover the camera parameters, their poses and a sparse (or dense) 3D scene geometry, i.e. in the form of a 3D point cloud or mesh structures. Traditionally, *structure-from-motion (SfM)* algorithms have been extensively studied in the literature [35, 37] for extracting precise camera poses and coarse 3D representations of increasingly complex scenes. These mostly leverage feature-matching techniques and geometric constraints between neighboring views. With the results from SfM as input, more dense and precise 3D models of the object are generated using *multi-view stereo (MVS)* pipelines [36]. Novel views can then be extracted from these representations.

More recently, research on learning-based novel-view synthesis has increased in exposure, due to the improvements in neural rendering techniques such as NeRFs [24], which encode the volumetric density within the weights of a multilayer perceptron (MLP). Other approaches, namely 3DGS [17], encode the scene as a 3D point cloud of 3D Gaussian distributions, and have achieved similar photo-realistic view synthesis of increasingly bigger and more complex scenes while yielding a more explicit scene representation. These methods raised the state-of-the-art in view synthesis and spawned lots of follow-up research [1, 21, 25, 27, 44, 45]. Since they are crucial to our method, we will explain them in more detail in the following sections.

---

[1] https://github.com/m-kruse98/SplatPose

2.2.1 NeRF and iNeRF

Given a set of input views with corresponding camera poses, NeRFs [24] optimize a MLP with parameters Θ, to represent a continuous volumetric scene function. This function maps from 5D coordinates, containing the 3D spatial location $\boldsymbol{x}$ and the 2D viewing direction $\boldsymbol{d}$, to the emitted color $\boldsymbol{c}$ and volume density σ, i.e. $f_\Theta : (\boldsymbol{x}, \boldsymbol{d}) \mapsto (\boldsymbol{c}, \sigma)$. In order to render a pixel, NeRFs march a ray $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d}$ from a starting point $\boldsymbol{o}$ in the direction $\boldsymbol{d}$ with $t \in [t_n, t_f]$. The color $\boldsymbol{C}(\boldsymbol{r})$ emitted by this ray is calculated as

$$\boldsymbol{C}(\boldsymbol{r}) = \int_{t_n}^{t_f} \mathcal{T}(t) \cdot \sigma(\boldsymbol{r}(t)) \cdot \boldsymbol{c}(\boldsymbol{r}(t), \boldsymbol{d}) \, dt, \qquad (1)$$

where

$$\mathcal{T}(t) = \exp\left(-\int_{t_n}^{t} \sigma(\boldsymbol{r}(s)) ds\right) \qquad (2)$$

accumulates the probability of a ray marching from t_n to t without hitting another particle, and $\sigma(\boldsymbol{x})$ denotes the probability of a ray ending at a particle at location $\boldsymbol{x}$. To estimate this integral numerically, points are sampled along these rays, and the MLP predicts color $\boldsymbol{c}$ and density σ, optimized via the photometric loss, i.e. the total squared error between rendering and ground truth image.

Some of the downsides of NeRFs include their very high computational costs with training times of several hours per scene, and their performance problems in sparse-view settings [24, 25]. Further improvements have been proposed to address issues such as poor rendering quality [1, 25] or high inference times [45], among others.

By being differentiable, NeRFs can be used for pose estimation of novel views by "*inverting*" the trained NeRF f_Θ, spawning the iNeRF framework [21]. The unknown camera pose $\hat{T}$ of an image I, is recovered by keeping Θ frozen and optimizing the problem

$$\hat{T} = \operatorname*{argmin}_{T \in SE(3)} \mathcal{L}(T|I, \Theta), \qquad (3)$$

where $\mathcal{L}$ again measures the photometric loss between I and the image rendered using the camera pose transformation T, with T sampled from the group of euclidean motions $SE(3)$. Both NeRF and iNeRFs are sensitive to hyperparameters to ensure convergence at a reasonable speed. The computational intensity of iNeRF remains an open problem.

2.2.2 3D Gaussian Splatting

By encoding the scene using a 3D point cloud of Gaussians, 3DGS [17] is able to synthesize images of very high quality, while being efficient enough to render in real-time. Each 3D Gaussian in this point cloud is characterized by its center position μ and a covariance matrix Σ, which is constructed using a scaling matrix S and a rotation matrix R, utilizing the equation $\Sigma = RSS^T R^T$. Further parameters include an opacity factor α and a color component c, which is modeled using spherical harmonics (SHs), that act as a view-direction-dependent coloring of the Gaussian sphere.

Initially, given a set of multi-view images, their camera poses and an initial point cloud are estimated using SfM pipelines [35]. During optimization, novel views are repeatedly rendered using an efficient differentiable tile-based rasterizer and compared to their ground truth image. Since the rendering process is differentiable, the parameters of the Gaussians can be optimized using the mean squared error as well as the structural similarity index measure (SSIM) [42].

The rendering process is, just as with NeRFs, done along rays analogously to Eqs. (1) and (2). In the case for 3DGS, the densities σ and color values c are not sampled by querying an MLP, but rather saved explicitly within the Gaussian representations. Rasterization, efficient sorting and α-blending Gaussians at the queried pixels enables fast and high-quality image synthesis [17]. During optimization, the cloud of Gaussians is controlled and optimized in an adaptive rule-based manner. Some strategies include pruning away any points with negligibly small opacity values, proposing new points in sparsely populated areas, and splitting high-variance Gaussians into two of lower variance.

2.2.3 OmniposeAD

Since our work is inspired by OmniposeAD [47] (in this paper shortened to OmniAD), we describe its pipeline in the following. To become pose-agnostic, OmniAD trains a NeRF, just as described by Mildenhall *et al*. [24], and subsequently uses iNeRF to perform pose estimation.

The iNeRF framework relies on an initial pose, to start optimization. In line with the author's reference implementation, the "*Local Feature Matching with Transformers (LoFTR)*" framework [38] is used to collect feature matches between the test sample and every training sample. The training sample with the most matches is chosen as the initial coarse pose. Afterward, pose estimation for OmniAD follows the iNeRF [21] framework. Repeatedly rendering images with a NeRF becomes a major bottleneck for fast inference and computational requirements.

Lastly, with sufficiently good pose estimation, OmniAD generates an image using the trained NeRF at the estimated pose. As the NeRF is trained only on defect-free data, it is void of any anomalies, apart from any NeRF-related rendering mistakes. The synthesized image is then compared to the original query sample, by gathering a pyramid of features from a pre-trained neural network for both images and calculating their mean squared error at every pixel. Aggregating these pixel-wise score maps is done for image-wise anomaly detection.

3. Method

In the pose-agnostic anomaly detection setting, defect-free multi-view images and their respective poses are provided during training. At inference time, the task is to infer whether a test sample contains any anomalies without having access to its camera pose. We aim to solve this task using *SplatPose*, which is visualized in Fig. 2. We take inspiration from the OmniAD pipeline as described in Sec. 2.2.3, which solves the problem by coarsely estimating a test sample's pose from all training examples and then refining it with a previously optimized NeRF model.

Instead of using NeRFs, we resort to modeling the defect-free objects with 3DGS and estimate a fine pose by transforming the learned 3D point cloud to match a test image.

3.1. Parametrizing Pose Transformations

In order to align our rendered image to the unknown pose of a query image, we need to estimate its camera pose. Instead of transforming the camera pose directly, we keep the camera fixed to its coarse estimate and transform the entire scene, i.e. the entire 3D point cloud. This simulates camera movement, allowing us to efficiently align the image rendered by SplatPose to the query image.

Concretely, a transformation $T \in SE(3)$ is applied to the position of every Gaussian, as well as to their intrinsic rotation R. All possible transformations from the Euclidean group $SE(3)$ may be modeled using a screw axis $S = (\omega, v) \in \mathbb{R}^6$ along which to rotate and a distance $\theta \in \mathbb{R}$ to translate along said axis [23]. According to Rodrigues' formula for rotations [11, 23], sampling from the group of 3D rotations $SO(3)$ is done by sampling such a scalar θ and a unit vector $\omega \in \mathbb{R}^3$ and calculating

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

where $[\omega]$ is the skew-symmetric 3×3 matrix representation of ω. The full transformation $T \in SE(3)$ is then parametrized as

$$T = e^{[S]\theta} = \begin{bmatrix} e^{[\omega]\theta} & G(\theta)\,v \\ 0 & 1 \end{bmatrix}, \text{ where} \quad (5)$$

$$G(\theta) = I\theta + (1 - \cos\theta)\,[\omega] + (\theta - \sin\theta)\,[\omega]^2, \quad (6)$$

which can be directly applied to our 3D point cloud [23]. Thereby, any gradients flowing to the 3D point cloud of Gaussians can also be propagated back to the parametrization of our transformation T, which is given by the parameters $\omega, v \in \mathbb{R}^3$ and $\theta \in \mathbb{R}$.
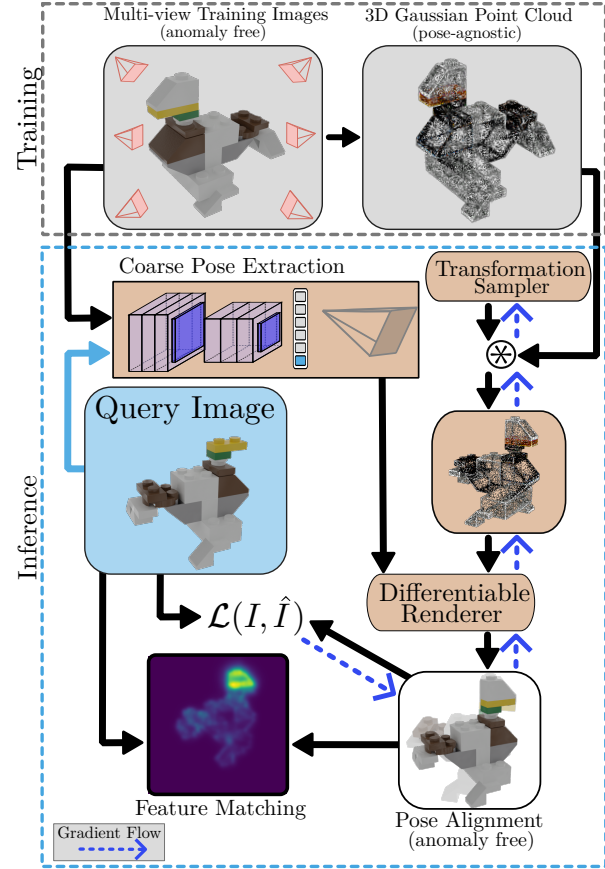

Figure 2. Overview of our pipeline. Multi-view training images are represented in a 3D point cloud of Gaussians. The unknown camera pose of a query image is first coarsely estimated and then iteratively refined by applying a pose transformation on the 3D point cloud before the differentiable renderer. The final anomaly-free rendering is then compared to the original test image to perform pixel-wise comparison for anomaly detection.

3.2. Anomaly Detection with SplatPose

Since NeRFs are too resource-intensive for practical use [24, 47], SplatPose uses 3DGS to encode the multi-view images of the objects within a 3D point cloud of Gaussians. Apart from following the standard representations proposed by Kerbl *et al.* [17], the spherical harmonics, which represent the colors, are restricted to a degree of zero. This restriction still suffices for good image rendering while saving compute power and keeping the color invariant to the rotations that are applied during pose estimation.

With the ability to synthesize high-quality views of arbitrary poses using our 3D point cloud, the pose of a test query image I_q still remains unknown. To recover an initial coarse pose, SplatPose sticks to OmniAD's proposed method using LoFTR, which also lets us better estimate the impact of replacing both NeRFs and iNeRF with our 3DGS-based framework.

	Feature Emb.		Memory Banks		Student-Teacher			Normal. Flow		Synthetic		View Synthesis	
	2D Image-based without Poses											3D Object-based	
Category	PaDiM [7]	Mahal. [28]	CFA [19]	PatchCore [30]	RD4AD [8]	AST [33]	STFPM [40]	CFlow [10]	CS-Flow [32]	SimpleNet [22]	DRÆM [46]	OmniAD [47]	SplatPose (ours)
Gorilla	46.9	32.4	41.8	66.8	51.9	28.3	65.3	69.2	41.9	40.9	58.9	**93.6**	91.7 ± 1.1
Unicorn	81.0	86.4	85.6	92.4	67.7	87.0	79.6	82.3	85.5	88.7	70.4	94.0	**97.9 ± 1.1**
Mallard	14.8	89.1	36.6	59.3	54.4	48.6	42.2	74.9	36.8	43.4	34.5	84.7	**97.4 ± 0.5**
Turtle	54.7	21.2	58.3	87.0	82.6	28.1	64.4	51.0	56.0	62.9	18.4	95.6	**97.2 ± 0.7**
Whale	75.7	35.0	77.7	86.0	64.1	26.6	64.1	57.0	47.4	78.5	65.8	82.5	**95.4 ± 3.0**
Bird	55.6	79.7	78.4	82.9	59.8	87.1	52.4	75.6	84.9	76.3	69.1	92.4	**94.0 ± 1.2**
Owl	75.8	67.0	74.0	72.9	69.0	81.8	72.7	76.5	71.9	74.1	67.2	**88.2**	86.8 ± 0.9
Sabertooth	65.4	75.8	64.2	76.6	69.9	82.2	56.0	71.3	73.9	69.6	68.6	**95.7**	95.2 ± 1.5
Swan	59.7	70.5	66.7	75.2	60.5	83.0	53.6	67.4	65.9	66.2	59.7	86.5	**93.0 ± 0.7**
Sheep	69.1	82.0	86.5	89.4	77.7	96.0	56.5	80.9	84.5	81.1	59.5	90.1	**96.7 ± 0.1**
Pig	50.4	61.5	66.7	85.7	59.0	72.9	50.6	72.1	68.2	65.9	64.4	88.3	**96.1 ± 1.9**
Zalika	39.0	63.0	52.1	68.2	49.1	68.1	53.7	66.9	47.9	62.4	51.7	88.2	**89.9 ± 0.7**
Phoenix	61.1	63.7	65.9	71.4	62.0	74.8	56.7	64.4	63.7	62.5	53.1	82.3	**84.2 ± 0.3**
Elephant	48.9	70.3	71.7	78.6	62.3	86.5	61.7	70.1	71.0	71.7	62.5	92.5	**94.7 ± 0.9**
Parrot	53.9	64.1	69.8	78.0	47.8	76.0	61.1	67.9	59.4	66.2	62.3	**97.0**	96.1 ± 1.1
Cat	53.6	54.2	68.2	78.7	54.6	70.2	52.2	65.8	61.0	64.5	61.3	**84.9**	82.4 ± 1.3
Scorpion	80.4	78.4	91.4	82.1	69.7	90.1	68.9	79.5	76.9	80.5	83.7	91.5	**99.2 ± 0.1**
Obesobeso	68.8	69.7	80.6	89.5	77.6	86.3	60.8	80.0	77.4	81.0	73.9	**97.1**	95.7 ± 0.7
Bear	65.5	74.4	78.7	84.2	67.5	87.4	60.7	81.4	75.3	79.4	76.1	98.8	**98.9 ± 0.2**
Puppy	42.9	63.3	53.7	65.6	60.3	67.9	56.7	71.4	64.1	61.0	57.4	93.5	**96.1 ± 0.9**
mean	58.2	65.1	68.4	78.5	63.4	71.4	59.5	71.3	65.7	68.8	60.9	90.9	**93.9 ± 0.2**

Table 1. AUROC ($\uparrow$) for image-level anomaly detection performance on MAD. We run our approach $n = 5$ times and report the standard deviation and mark the best result per class in **bold** and the runner-up underlined.

With an initial coarse pose fixed, transformations T from the group of 3D motions are sampled as described in Sec. 3.1, and applied to each Gaussian in the 3D point cloud. In each time step i, using the transformation T_i, a new sample I_i is generated, which may be compared to the query image I_q. For this, we again stick to the original 3DGS framework and use the mean absolute error $\mathcal{L}_1$ and SSIM [42] to optimize the pose with the loss

$$\mathcal{L} = (1-\lambda)\mathcal{L}_1 + \lambda\left(1 - \mathcal{L}_{SSIM}\right). \qquad (7)$$

In the standard 3DGS setting, the loss is propagated solely back to the 3D point cloud. By modifying the cloud using our transformations T, the gradients instead flow back to the pose transformation sampler, making the entire pose estimation process differentiable. Contrary to iNeRF, no region-of-interest sampling is needed and the entire scene can be used for optimization.

After k steps of pose estimation, a rendering I_k, which has the same object pose as I_q, but without any anomalies, is generated. Lastly, similar to OmniAD, an anomaly score map is constructed by pixel-wise comparison of extracted pre-trained features on both I_q and I_k, again using the mean squared error.

4. Experiments

We evaluate the anomaly detection, segmentation, and pose estimation performance of SplatPose empirically on the recently proposed pose-agnostic anomaly detection benchmark [47].

4.1. Evaluation Protocol

As is standard in anomaly detection settings, only anomaly-free data is available at training time. For evaluation, both defective and normal samples are given. The anomaly detection performance is measured by producing an anomaly score for each sample, separating normal data and anomalies. We calculate the area under the ROC curve (AUROC) using this score. Similarly, as ground truth masks of the anomaly positions are available, the pixel-wise performance can be evaluated by calculating the AUROC across all pixels. In practice, however, all kinds of anomalies should be detected with equal importance regardless of their size. To treat smaller anomalies with the same weight as larger ones, AUPRO is a well-established metric [2]. It assigns the same weight to each of the connected components within the anomaly masks, making it more difficult to reach high scores and we report it for all models reproduced by us.

4.2. Data Sets

We test our method on the Multi-pose Anomaly Detection (MAD) data set proposed by Zhou *et al.* [47], which introduced the challenge of performing pose-agnostic detection in images. To this end, the MAD-Sim data set contains images of 20 different kinds of LEGO® toys of different complexities in terms of shape and texture. All of these were generated synthetically using Blender. At training time, the camera poses for all images are given, whereas they are not available at test time. A few examples of anomalies can be found in Fig. 1. This benchmark is motivated

| Metric | Feat.Emb. | Memory banks | | Student-Teacher | | | Normal. Flow | | Synthetic | | View Synthesis | |
	PaDiM [7]	CFA [19]	PatchCore [30]	RD4AD [8]	AST [33]	STFPM [40]	CFlow [10]	CS-Flow [32]	SimpleNet [22]	DRÆM [46]	OmniAD [47]	SplatPose (ours)
AUROC (↑)	90.7	89.4	74.9	91.3	68.1	89.3	90.8	71.7	89.7	58.0	<u>98.4</u>	**99.5** ± 0.01
AUPRO (↑)	77.0	-	83.7	79.2	68.1	-	-	27.1	77.4	-	<u>86.6</u>	**95.8** ± 0.03

Table 2. Results for anomaly segmentation averaged across all categories in MAD. We run our approach $n = 5$ times and mark the best result in **bold** and the runner-up <u>underlined</u>. Comprehensive tables are in the supplementary in Sec. A.1.

by all major anomaly detection data sets featuring tightly aligned poses, which may interfere with detecting partially occluded anomalies on more complex-shaped objects.

To verify our pose estimation more thoroughly, we also use the 360° synthetic scenes provided by NeRF [24], which are already divided into train and test splits. These also contain synthetically generated multi-view images but are not constrained to simple LEGO® toys, as they also include more complex objects with finer details.

In our experiments, we find that for both data sets, the poses in the training split are uniformly distributed in a sphere around the object. For MAD, the test poses are sampled randomly within that sphere without any well-defined distribution, signifying the need for the pose estimation to be able to generalize to previously unseen poses. The test poses in the NeRF data set follow the same uniform distribution as the train set and instead describe one circular orbit around the objects.

4.3. Implementation

We base our implementation on the reference given by OmniAD[2] [47]. Implementations for Gaussian Splatting are taken from the original project by Kerbl *et al.*[3][17], which we modify to allow for the insertion of pose transformations into the differentiable rendering process.

As for the modules replicated from OmniAD, coarse pose estimation is done using a pre-trained LoFTR [38] with a ResNet [12] backbone, while the image feature matching is done using a pre-trained EfficientNet-b4 [39]. Both are pre-trained on ImageNet [9]. All images are also resized to 400×400 pixels before processing.

We train our 3D Gaussian model for $30,000$ iterations while keeping everything to the author-given hyperparameters, except for the spherical harmonics [17]. The pose estimation is optimized using the Adam optimizer [18] with a learning rate of 0.001 and momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For the loss in Eq. (7), we use $\lambda = 0.2$. The pose estimation iterates for $k = 175$ steps, which we find to yield good estimations while still allowing for quick inference times.
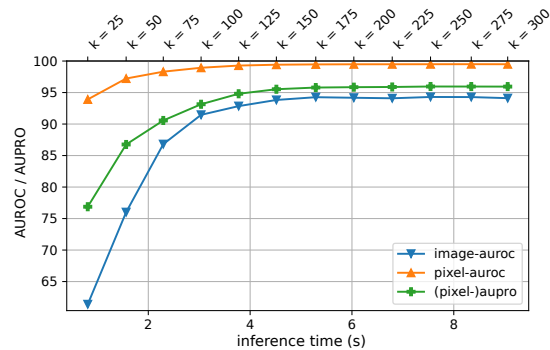

Figure 3. Influence on all detection metrics and inference speed, when changing the number of pose estimation steps k from 25 to 300, with the performance saturating around $k = 175$. Since OmniAD has inference times magnitudes larger than SplatPose, we do not include it in this experiment.

4.4. Anomaly Detection Results

Detailed results for image-wise anomaly detection on MAD are given in Tab. 1. We have taken numbers for competitors from Zhou *et al.* [47] and also reproduced additional state-of-the-art models ourselves. Our method confidently reaches a new state-of-the-art performance of 93.9%, which is three points higher than the next best competitor OmniAD. It should also be noted, that all methods that do not incorporate any pose information, i.e. the ones not synthesis-based, lack behind heavily in performance, with PatchCore [30] reaching the highest AUROC of only 78.5%. We observe, that SplatPose especially outperforms OmniAD, whenever the pose estimation is more accurate. This intuition will be verified quantitatively in Sec. 4.6.

We furthermore also evaluate the anomaly segmentation performance by looking at the pixel-wise AUROC and AUPRO in Tab. 2. Again SplatPose clearly outperforms all competitors with a pixel-wise AUROC of 99.5% and an AUPRO of 95.8. Some qualitative intuition of the superior performance of SplatPose will later be visualized in Fig. 5.

4.5. Computational Impact

One of the major shortcomings of our main competitor OmniAD is its high training and inference times, which we overcome with our approach. Therefore, we evaluated training and inference times across our experiments in Tab. 3. For a fair comparison, all experiments are executed on the same machines, utilizing a single NVIDIA RTX 3090.

---

[2] https://github.com/EricLee0224/PAD
[3] https://github.com/graphdeco-inria/gaussian-splatting

Method	training ↓ $(h:min:sec)_{\pm(min:sec)}$	inference ↓ $(min:sec)_{\pm(sec)}$
OmniAD	$04{:}33{:}43_{\pm 04{:}52}$	$01{:}06_{\pm 00}$
SplatPose	$\mathbf{00{:}04{:}54}_{\pm 00{:}18}$	$\mathbf{00{:}05}_{\pm 00}$

Table 3. Both training and inference speed measured on the MAD data set for OmniAD [47] and our approach. One standard deviation is given across the 20 different classes. Best results in **bold**.

Since both our and the OmniAD pipeline perform the same coarse pose estimation, we only measure the inference times caused by the iterative fine pose estimation and the subsequent feature matching. We closely match the computational requirements reported by Zhou *et al.* for OmniAD [47]. As our approach takes far fewer optimization steps both when training the cloud of Gaussians and when fine-tuning the estimated pose, we can confidently beat their computational requirements. In total, it results in roughly **55** times faster training and **13** times faster inference when using our method. Furthermore, our 3DGS-based training always converged to good solutions, while NeRF struggled to converge in roughly 15% of training runs due to poor ray sampling. SplatPose's robustness and speed advantage allow for both quick prototyping and on-the-fly anomaly detection in real production scenarios.

Furthermore, we run an ablation on the number of steps used for pose estimation k in Fig. 3. Starting with $k = 25$, we observe an image-wise AUROC of 61% and inference times of roughly one second. While inference time linearly increases with the number of steps, the detection performance improves drastically and starts to saturate after $k = 175$, which we suggest to be a good compromise between quick inference and precise detection. Higher values of k do not substantially improve the performance, as the pose is already very precisely estimated.

4.6. Pose Estimation Results

Despite the focus of our work being on anomaly detection, we evaluate the pose estimation process more thoroughly and compare the performance of our method and iNeRF.

As commonly done in pose estimation literature, we compare the location of the object whose pose we estimate (here the camera, as we only have access to ground truth camera poses) to the ground truth position using the Euclidean distance [14].

For the estimated rotation matrices, we first convert all rotation matrices into their representation as unit quaternions. For two quaternions q_1 and q_2, we can then quantify the difference between two rotations, by measuring the rotation needed to go from q_1 to q_2, resulting in the formula

$$\Phi(q_1, q_2) = 2 \arccos\left(|q_1 \cdot q_2|\right), \tag{8}$$

with · denoting the dot product of two vectors. Φ defines a metric on the group of 3D rotations $SO(3)$ and takes values in the ranges of $[0, \pi]$ with lower scores denoting a higher similarity in rotation [15].
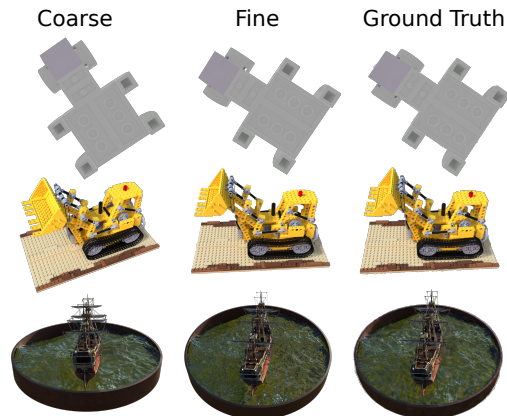

Figure 4. Examples of pose estimation using SplatPose. Starting from OmniAD's coarse pose [47], we refine it to match the ground truth. Examples are from MAD and the NeRF synthetic data.

A few estimated poses are visualized in Fig. 4. As no true ground truth poses are available for the MAD data set, we randomly split the training data set into a custom training and testing split to validate our methods. Quantitative results of the experiments are presented in Tab. 4, with the full table in the supplementary in Sec. A.2.

Method	Transl. Err. (total) ↓		Rot. Err. (rad) ↓	
	MAD	Synth	MAD	Synth
Coarse [47]	0.803	0.624	0.163	0.084
iNeRF [21]	0.179	0.055	0.056	0.007
SplatPose	**0.094**	**0.021**	**0.040**	**0.003**

Table 4. Pose Estimation Error measured as the Euclidean distance for Translation and using Φ from Eq. (8) for Rotation. Results are averaged across the MAD data set [47] and NeRF synthetic scenes [24]. Best results in **bold**.

We observe, that the coarse pose estimation step of OmniAD is already able to achieve good baseline scores. Still, both iNeRF and SplatPose can vastly outperform the rough estimation, with us performing the best out of all approaches, despite using fewer optimization steps and very short training times. According to the full results in the supplements, we outperform iNeRF by several magnitudes in most of the categories, with closer margins only occurring in a few cases. We argue that this is caused by us being able to optimize using the entire image and full 3D point cloud, while iNeRF is restricted to sampling the most important rays, which misses out on some details.
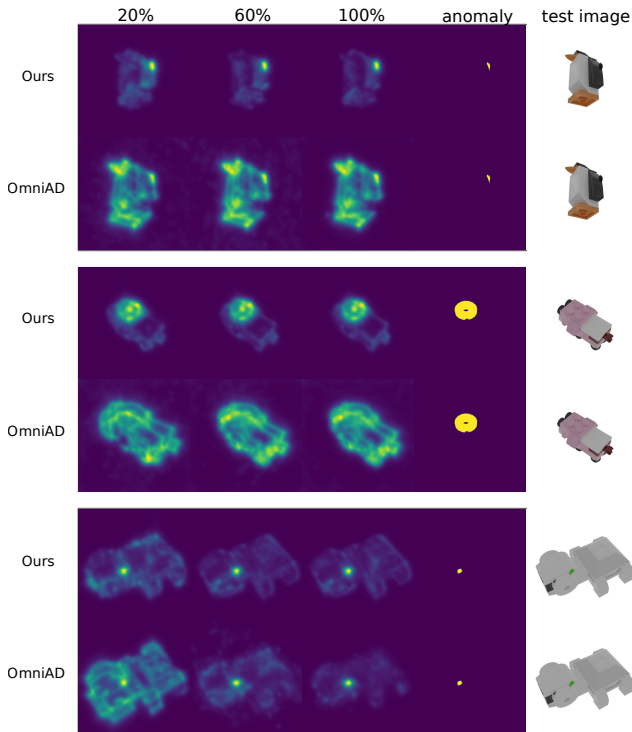
Figure 5. Quantitative comparison of performance for both OmniAD when using between 20% and 100% of the available training data. We show the anomaly maps achieved by feature matching for both methods. Best viewed in color.

%	Detection		Segmentation	
views	OmniAD	SplatPose	OmniAD	SplatPose
20	67.2	**77.6**	73.7	**86.5**
40	79.1	**89.3**	81.5	**93.3**
60	82.2	**92.2**	85.0	**95.0**
80	86.8	**93.1**	86.8	**95.5**
100	90.9	**93.9**	86.6	**95.8**

Table 5. Image-wise Anomaly Detection performance for the sparse-view setting measured as AUROC ($\uparrow$) and AUPRO ($\uparrow$) respectively on MAD. Best results in **bold**.

niAD is able to fit a good pose, artifacts and surrounding rendering mistakes are more frequent than with SplatPose.

The quantitative results for image-wise anomaly detection and segmentation under a sparse-view setting can be found in Tab. 5, with the full table in the supplement in Sec. A.3. We compare SplatPose to the other top competitor OmniAD, despite the high computational costs [47].

SplatPose decisively beats OmniAD for every step of view-sparsification, by margins of around ten points in the 20, 40 and 60% settings for both the image-wise AUROC and AUPRO. For lower levels of sparsity, SplatPose still outperforms OmniAD decisively. Again, we can optimize using the full views, while NeRF resorts to sampling rays and is known to struggle in sparse-view settings [25].

5. Conclusion

In this paper, we presented a novel pose-agnostic method for anomaly detection. Given multi-view images, we represent the object as a 3D point cloud of Gaussians, which is used for pose estimation, and to find anomalies in images without prior pose information. Our method beats all competitors at the detection task, while still being magnitudes faster for both training and inference time, making it more suited for deployment in production environments.

We would like to dedicate future work towards improving the coarse pose estimation and image feature comparison. Applying our findings to adjacent fields, such as pose estimation in humans [16, 43], strikes us as a promising next direction. Closing the gap between synthetic and real-world data would also require more work. Lastly, we want to investigate ways to include the three-dimensional point cloud information in existing two-dimensional approaches.

The quantitative comparison for the synthetic NeRF scenes from Tab. 4, again have SplatPose performing better. For both the NeRF's uniformly sampled test scenes and the randomly sampled ones from MAD, SplatPose achieves a stronger pose estimation performance. Therefore, it is not only able to better capture the 3D geometry of the scene for known camera poses but also its capability to generalize to previously unseen views beats that of iNeRF. This superiority in pose estimation also shows up in SplatPose beating out OmniAD at the anomaly detection tasks, since the rendered images are aligned more accurately to the test samples.

4.7. Results for Sparse-View Data

Each category in the MAD data set contains 210 different views for training. In practice, multi-view data can not be guaranteed to be this dense. Thus, we sample subsets of the original training splits, to simulate sparse-view settings.

Some selected visualizations of the anomaly maps for training with 20, 60, and 100% of the data are shown in Fig. 5. Here, our intuition is confirmed, that better pose estimation leads to more precise anomaly maps. OmniAD struggles most when the coarse pose is too far away, or when large parts of the object are missing. Even when Om-

References

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5460–5469. IEEE, 2022. 2, 3

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9584–9592, 2019. 1, 2, 5

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4182–4191. Computer Vision Foundation / IEEE, 2020. 2

[4] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *Int. J. Comput. Vis.*, 130(4):947–969, 2022. 1, 2

[5] Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2022, Volume 5: VISAPP, Online Streaming, February 6-8, 2022*, pages 202–213. SCITEPRESS, 2022. 1, 2

[6] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *CoRR*, abs/2005.02357, 2020. 2

[7] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges - Virtual Event, January 10-15, 2021, Proceedings, Part IV*, pages 475–489. Springer, 2020. 2, 5, 6, 1

[8] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9737–9746, 2022. 2, 5, 6, 1

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[10] Denis A. Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. CFLOW-AD: real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 1819–1828. IEEE, 2022. 2, 5, 6, 1

[11] Andrew Hartley and Andrew Zisserman. *Multiple view geometry in computer vision (2. ed.)*. Cambridge University Press, 2006. 2, 4

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 6

[13] Lars Heckler, Rebecca König, and Paul Bergmann. Exploring the importance of pretrained feature extractors for unsupervised anomaly detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 2917–2926. IEEE, 2023. 2

[14] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary R. Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Computer Vision - ACCV 2012 - 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I*, pages 548–562. Springer, 2012. 7

[15] Du Q. Huynh. Metrics for 3d rotations: Comparison and analysis. *J. Math. Imaging Vis.*, 35(2):155–164, 2009. 7

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014. 8

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. 2, 3, 4, 6

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[19] Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. CFA: coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10:78446–78454, 2022. 2, 5, 6, 1

[20] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9664–9674. Computer Vision Foundation / IEEE, 2021. 2

[21] Yen-Chen Lin, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2021, Prague, Czech Republic, September 27 - Oct. 1, 2021*, pages 1323–1330. IEEE, 2021. 1, 2, 3, 7

[22] 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 (CVPR)*, pages 20402–20411, 2023. 2, 5, 6, 1

[23] Kevin M. Lynch and Frank C. Park. *Modern Robotics - Mechanics, Planning, and Control*. Cambridge University Press, Cambridge, 2017. 4

[24] 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. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 405–421. Springer, 2020. 2, 3, 4, 6, 7, 1

[25] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5470–5480. IEEE, 2022. 1, 2, 3, 8

[26] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. OC-GAN: one-class novelty detection using gans with constrained latent representations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2898–2906. Computer Vision Foundation / IEEE, 2019. 2

[27] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10318–10327. Computer Vision Foundation / IEEE, 2021. 2

[28] Oliver Rippel, Patrick Mertens, and Dorit Merhof. Modeling the distribution of normal data in pre-trained deep features for anomaly detection. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pages 6726–6733. IEEE, 2020. 2, 5

[29] Bodo Rosenhahn and Christoph Hirche. Quantum normalizing flows for anomaly detection. *CoRR*, abs/2402.02866, 2024. 2

[30] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter V. Gehler. Towards total recall in industrial anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14298–14308. IEEE, 2022. 1, 2, 5, 6

[31] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 1906–1915. IEEE, 2021. 2

[32] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 1829–1838. IEEE, 2022. 2, 5, 6, 1

[33] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2023. 2, 5, 6, 1

[34] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Information Processing in Medical Imaging - 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*, pages 146–157. Springer, 2017. 2

[35] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4104–4113. IEEE Computer Society, 2016. 2, 3

[36] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, pages 501–518. Springer, 2016. 2

[37] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, 2006. 2

[38] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8922–8931. Computer Vision Foundation / IEEE, 2021. 3, 6

[39] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 6105–6114. PMLR, 2019. 6

[40] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 306. BMVA Press, 2021. 2, 5, 6, 1

[41] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 8032–8041. IEEE, 2023. 2

[42] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. 3, 5

[43] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11179–11188. IEEE, 2021. 8

[44] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 2

[45] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5732–5741. IEEE, 2021. 2, 3

[46] Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. Dræm - A discriminatively trained reconstruction embedding for surface anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8310–8319. IEEE, 2021. 2, 5, 6, 1

[47] 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. In *NeurIPS*, 2023. 1, 2, 3, 4, 5, 6, 7, 8