A Unified and Hybrid Approach for Image-based Scene Change Detection and Pose-agnostic Object Anomaly Detection

by

Yizhe Liu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science in Systems Design Engineering

Waterloo, Ontario, Canada, 2024

© Yizhe Liu 2024

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Yizhe Liu is the sole author of all chapters except for Chapter 3. Chapter 3 is adapted from the published work "SplatPose+: Real-time Image-Based Pose-Agnostic 3D Anomaly Detection" [30], in which Yizhe Liu is the primary author, and Yan Song Hu, Yuhao Chen, John Zelek are the reviewers.

Abstract

Image-based Pose-Agnostic 3D Anomaly Detection is an important task that has emerged in industrial quality control. This is an object-central task that seeks to find anomalies from pose-known query images of a tested object given a set of reference images of a standard anomaly-free object. There is also a similar task: Image-based Scene Change Detection which focuses on the differences in a scene instead of an object. Image-based Scene Change Detection is a critical task in mapping and monitoring a scene, seeking to find the semantic changes in a scene described by two sets of images (reference and query) captured at different timestamps.

For those industrial detection tasks, image sensors are widely used for their ease of use and low cost to acquire. However, the most commonly used image sensors: RGB cameras are only capable of capturing 2D information in the form of an RGB image from a specific angle of an object or a scene. While, in the context of imaged-based anomaly detection and change detection, reference images and query images are often taken from different poses; and the poses of the query views can be unknown. As a result, reference images and query images cannot be compared easily. Recent learning-based methods, for example, OmniposeAD [66] and SplatPose [23], employ Novel View Synthesis (NVS) Methods, i.e., Neural Radiance Field [33] (NeRF) and Gaussian Splatting [20] to bridge the gap by simultaneously localizing the query image with respect to the reference images and synthesizing pseudo reference images for the query views for direct pixel-to-pixel comparison. However, these learning-based methods suffer from long localization overhead during the inference stage because inversed Neural Radiance Field methods, e.g., INeRF [64], can take hundreds of gradient descent steps to localize and refine the poses.

This paper introduces a hybrid approach SplatPose+ [30] that maintains both a learning-based model (Gaussian Splatting [20]) for NVS and a Structure-from-Motion (SfM) model (Hierarchical Localization [52]) for localization, which takes the advantage of the fast training and inference of 3D Gaussian Splatting [20] and the fast localization of Hierarchical Localization [52]. On the Image-based Pose-Agnostic 3D Anomaly Detection task, although our proposed pipeline requires the computation of an additional SfM model, it offers real-time inference speeds and faster training compared to SplatPose. Quality-wise, we achieve a new SOTA on the Pose-agnostic Anomaly Detection benchmark with the Multi-Pose Anomaly Detection (MAD-SIM) dataset. On the Image-based Scene Change Detection task, we achieve a higher Intersection over Union (IoU) than previous supervised methods on the binary change detection sub-task without environment variations. Moreover, we demonstrate the potential of combining SAM2 [44] with SplatPose+[30] to further refine the object-level change masks toward higher accuracy.

Acknowledgements

First, I want to thank Professor John Zelek for giving me guidance and suggestions from selecting the research topic to publishing part of it as an ECCV workshop paper. Professor John Zelek also funded me to attend amazing top-tier conferences in the field of Computer Vision, including CVPR and ECCV. These experiences open my eyes and enable me to get a better understanding of the frontier of the Computer Vision Research.

Second, I also would like to thank all the people who made this thesis possible, including my lab mates, Yuxiang Huang and Yan Song (Kevin) Hu. They helped me better understand Structure-from-Motion and solve related problems.

Last but not least, I want to thank my parents for supporting my living expenses and giving me emotional support when I am studying abroad.

Dedication

This thesis is dedicated to my parents.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements v

Dedication vi

List of Figures x

List of Tables xi

List of Abbreviations xiii

1 Introduction 1

1.1 Motivation 1

1.2 Problem Definition and Scope 2

1.3 Outline 5

1.4 Contributions 5

2 Review of Previous Methods **7**
2.1 Overview . . . 7
2.2 2D Image-based Methods . . . 8
2.3 3D Methods . . . 10
2.4 Conclusion and Motivation . . . 12

3 Real-time Image-Based Pose-Agnostic 3D Anomaly Detection **13**
3.1 Preface . . . 13
3.2 Introduction . . . 13
3.3 Related Work . . . 15
3.3.1 3D Reconstruction . . . 15
3.3.2 Novel View Synthesis (NVS) . . . 15
3.3.3 Image-Based Pose-Agnostic 3D Anomaly Detection . . . 16
3.4 Method . . . 16
3.4.1 Hybrid Representation . . . 17
3.4.2 Anomaly Score . . . 19
3.5 Evaluation . . . 20
3.5.1 Dataset . . . 20
3.5.2 Metrics . . . 20
3.5.3 Results . . . 21
3.5.4 Efficiency . . . 23
3.5.5 Robustness on Sparse-View Data . . . 23
3.6 Conclusion . . . 25

4 Image-Based 3D Scene Change Detection **27**
4.1 Introduction . . . 27
4.2 Related Methods . . . 28
4.2.1 Paired Image Difference Methods . . . 28

4.2.2 Open-set Image Difference Methods . . . 28
4.2.3 Object Tracking Methods . . . 29
4.3 Method . . . 29
4.4 Evaluation . . . 30
4.4.1 Dataset . . . 30
4.4.2 Limitations . . . 30
4.4.3 Metrics . . . 31
4.4.4 Results . . . 32
4.5 Mask Refinement . . . 32
4.5.1 Pipeline . . . 33
4.5.2 Results . . . 33
4.6 Conslusion . . . 34

5 Conclusion and Recommendation **39**
5.1 Conclusion . . . 39
5.2 Recommendation . . . 39

References **41**

List of Figures

3.1 SplatPose+ pipeline incorporates an SfM model for localizing query views and initialization for the 3D Gaussian Splatting [20] (3DGS) model. 3DGS model is trained only for Novel View Synthesis. 18

3.2 Defect Samples in MAD Dataset[66] 20

3.3 Quantitative Comparison for Anomaly Detection Performance on Sparse-View Reference Images 24

List of Tables

1.1 Samples of Anomalies 3

1.2 Image-based 3D Scene Change Detection Example: Predict change masks given reference and query image sequences. 4

3.1 Reconstruction efficiency and quality comparison. SHS: Spherical Harmonics Degree. 19

3.2 Comparison of NVS-based methods on MAD-SIM using Image-level AUROC (The higher the better ↑). The best result per class is **bold** and the runner-up is underlined. 22

3.3 Pixel-wise anomaly segmentation on MAD. The best result per class is **bold** and the runner-up is underlined. 22

3.4 Training and inference time on the MAD-SIM Dataset. 23

3.5 Qualitative comparison of predicted anomaly maps under 3 sparse-view settings. "Ours Psuedo Ref" is generated by SplatPose+ to help readers find the anomalies. 26

4.1 ChangeSim Dataset Samples (Reference Images are captured under normal visibility). 31

4.2 Qualitative binary Scene Change Detection (SCD) performance evaluated on ChangeSim [40] dataset without environment variations. 35

4.3 Binary SCD performance evaluated on ChangeSim [40] dataset without environment variations. (The higher the better, top results are bold.) 36

4.4 Qualitative comparison of SAM2 [44] assisted sub-mask refinement. Sub-masks are refined with respect to either query image or synthetic reference image. 37

4.5 SAM2 [44] assisted mask refinement IoU. (The higher the better, top results are bold.) 38
4.6 SAM2 [44] assisted final mask refinement. 38

List of Abbreviations

3DGS 3D Gaussian Splatting [20] x, 3, 5, 14, 16–19, 21, 23, 25, 30

CNN Convolutional Neural Network 9–11, 39

DSM Digital Surface Model 3, 4, 7, 8, 11

IoU Intersection over Union iv, 2, 31, 32

KNN k-nearest neighbors 11

MLP Multi-layer Perceptron 15

NeRF Neural Radiance Field [33] iv, 12, 14–16, 23

NVS Novel View Synthesis iv, xi, 4, 11, 14, 16, 17, 21, 22, 32, 39

PAD Pose-agnostic Anomaly Detection 1–3, 5, 29, 32

PCA Principal Component Analysis 8, 10

SCD Scene Change Detection xi, 1–6, 27–30, 32, 34–36, 39

SfM Structure-from-Motion iv, x, 3, 5, 14–19, 21, 23, 25, 28, 29

ViT Vision Transformers 4, 9, 10

Chapter 1

Introduction

1.1 Motivation

In the context of computer vision, various detection tasks focus on detecting objects or regions of interest, including anomalies and changed objects, in different representations, for example, images, videos, point clouds, and meshes. Some of the detection tasks, for example, face detection in images, are well-researched and have been applied to modern smartphones. While other tasks are under-researched and still require human interventions. These tasks are often tedious for humans to mark the targets manually. Moreover, human errors could impact the detection accuracy. For example, it is hard for humans to find all the differences in the "Spot the Difference" game, in which, a lot of tiny changes exist. As a result, the automation of these detection tasks is urgently demanded to facilitate important jobs in various industrial and engineering applications, including quality control, video surveillance, and autonomous vehicles. In this thesis, we will focus on bridging the gaps in two recently proposed industrial detection tasks: Image-based Pose-agnostic Anomaly Detection (PAD) and Scene Change Detection (SCD).

- Image-based PAD is a sub-task of anomaly detection that detects the anomalies in pose-unknown query images of a tested object given its anomaly-free reference images. This task is non-trivial because it is often impossible to find a pose-aligned refer image for comparison with a query image. Another difficulty in developing a PAD algorithm is that its applications, e.g., quality control on an assembly line, often require real-time performance. Therefore, an efficient PAD algorithm, ideally over 10 fps, is demanded for this use case. Besides the efficiency, AUROC and AUPRO

(defined in Eq. (3.1) and Eq. (3.3)) are two important metrics in evaluating the accuracy of a PAD algorithm, both measuring the performance of a binary classifier across different thresholds.

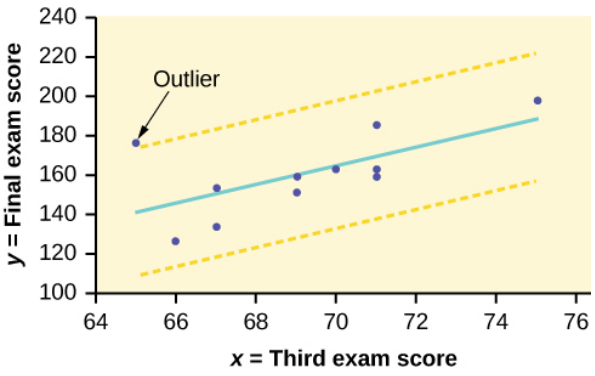
- Image-based SCD is a sub-task of change detection that differs between two captures of a scene at different times, e.g., detecting new or missing objects. This task is different from the PAD task in that the changes are object-level and multiple changed objects can occur. Moreover, SCD may involve semantic change where curtain objects still remain in the scene but are moved, rotated, or regrouped. However, the challenge in this task is similar to that in PAD where the two captures could involve images taken at unknown and unaligned perspectives, prohibiting the direct comparison of query and refer images. SCD algorithms are usually evaluated on Intersection over Union (IoU) metric, which evaluates the overlap ratio between the predicted change mask and the ground truth.

1.2 Problem Definition and Scope

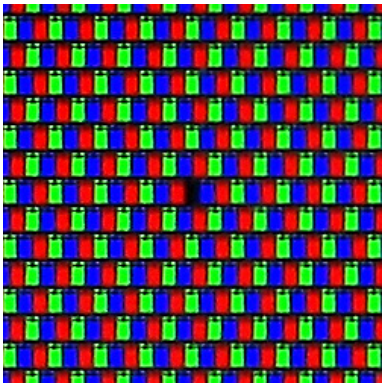
PAD

Anomalies are the parts of the data that do not follow the patterns of the majority of the data. Anomalies are often referred to as outliers, and we might use these words interchangeably. For example, as illustrated in Tab. 1.1, a data point that deviates from a linear distribution dataset can be an outlier in Statistics, and a dead pixel in an LCD display can be an anomaly in the electronics industry.

Anomalies detection is the process of identifying and locating the abnormal data that deviates from the standard or expected distribution. The task of anomaly detection can be traced back to as early as the nineteenth century by Edgeworth [17]. According to a survey on anomalies detection by Chandola, *et al.* [8], a substantial amount of research [46, 3, 4] was focused on statistical outlier detection before the 21st century. Since then, increasingly more methods have been developed for this task from statistical methods to deep learning methods. However, according to Chandola, *et al.* [8], more targeted methods are developed than generic methods for specific application domains. This is because there are emerging sub-tasks of anomaly detection that focus on locating outliers for different applications in different data modalities, e.g., from unstructured data (images, videos, tweets, ...) to structured data (credit card transactions, stock prices, and volumes, ...). Therefore, in this thesis, we limit the scope of the research to the PAD sub-task and the data type to multi-view image data.



Outliers in Statistics [25]



Anomaly (Dead Pixel) on a LCD Display [36]

Table 1.1: Samples of Anomalies

Image-based PAD is a task that compares a set of inward-facing reference images R comprised of multiple views of a standard object (without anomalies) with a set of query images Q to deduce if each query image contains anomalies. The reference set R is consistent. While the query set Q can be inconsistent, that is, these images do not necessarily come from the same object. Therefore, PAD algorithms are designed to predict anomalies in each query image separately. The main application for this task is to automatically detect faulty products from an assembly line. However, previous algorithms typically rely on time-consuming INeRF [64]-like algorithms to locate query images, which slow down the inference speed significantly.

For this task, the previous best method: SplatPose [23] inferencing only at 0.2 fps, cannot run in real-time. We will improve it to approximately 30 fps with even better accuracy by adopting an efficient Structure-from-Motion (SfM) framework: hloc [52] for localization and the initialization of 3D Gaussian Splatting [20] (3DGS).

SCD

Most scenes and areas are non-stationary, which means they are constantly changing as time goes on. As its name, change detection is a process designed to locate changes in a scene or at a location over time. This task usually operates on time series data, which can be as dense as daily tide levels at a beach or as sparse as two satellite images of an area taken at two different times. The data types can also vary depending on the applications, which can be images (2D), point clouds (3D), Digital Surface Model (DSM)s, and multiview

images (3D). For example, DSMs are mostly used in remote sensing, while images can be used for generic change detection.

Change detection was first developed as an important part of the research field of remote sensing [12, 43], which developed various statistical (for 2D representations), geometrical (for 3D representations) and deep-learning-based (for 2D & 3D representations) methods. Nowadays, there are new applications for change detection, for example, surveillance and autonomous driving that require the development of customized algorithms. In this thesis, we will focus our research on scene change detection that operates on 3D scenes described by multi-view images.

Image-based SCD is a task that compares 2 sequence of images R, Q taken at the same location but different timestamps t_0, t_1 as illustrated in Tab. 1.2, and identifies the changed objects. Each set of images is consistent, i.e., taken from the same scene, and can be inward-facing or outward-facing. For example, an inward-facing scene can be sampled by taking a video by going around a table and pointing the camera toward the table; while an outward-facing scene can be sampled by taking a 360° panoramic video toward all the walls in a room. The use cases for this task include tracking construction progress, updating street view maps, and monitoring critical areas.

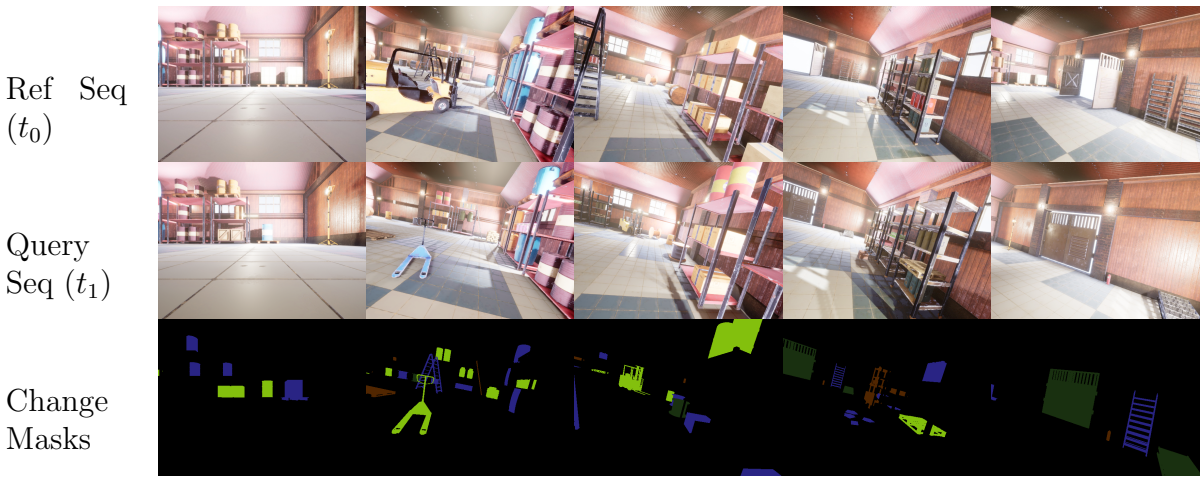


Table 1.2: Image-based 3D Scene Change Detection Example: Predict change masks given reference and query image sequences.

For the image-based SCD task, current methods are mostly supervised methods based on 2D vision models, e.g., Vision Transformers (ViT) [16]. These methods require a long training process and are likely to suffer precision loss when input images are captured from different perspectives. We bridge the gap by being the first method to apply Novel View

Synthesis (NVS) methods on the SCD task to align the views of input images, which, at the same time, reduces this task into a simple image differencing problem.

Overall, PAD and SCD play important roles in industrial applications and are labour-intensive if without automation. Because of the similarities in these two tasks, it is possible to develop an automated algorithm for both tasks. This thesis will focus on solving both tasks in a unified pipeline that employs hybrid 3D representations including a 3DGS model and a SfM model. The unification is achieved by transforming the problem into a generalized image differencing problem that only requires the reference image set R to be consistent.

1.3 Outline

- Chapter 2 contains a comprehensive review of previous anomaly detection and change detection methods. The latest methods of our specific research domains, i.e., PAD and SCD will be introduced in Chapter 3 and Chapter 4 respectively.
- Chapter 3 introduces a state-of-the-art real-time image-based PAD method and benchmarks the performance on the MAD-Sim [66] dataset.
- Chapter 4 explains how to adapt the method introducted in Chapter 3 on the SCD task. This chapter also demonstrates that it is possible to employ SAM2 [44] to refine the binary change masks.
- Chapter 5 concludes the thesis and discusses the future works that can be further researched based on this work.

1.4 Contributions

First, we introduce the very first real-time method for the PAD task, which also achieves higher accuracy and sparse-view robustness than previous methods. The ability to run in real-time allows this method to be truly applicable for quality control on an automatic assembly line that never stops.

Second, we adapt our method to solve the SCD tasks. Our adaption is revolutionary because previous learning-based SCD methods are mostly supervised. These methods employ heavy neural network structures that are slow to train and require a large amount

of training data. On the contrary, our method is unsupervised, that is, does not require training and supervision on the training data as well as labels. This improvement is crucial for the development of the SCD task because a large scene can contain thousands of images, in which, capturing the scene and annotating all images are costly.

Lastly, we show that a foundation model SAM2 [44] can be combined with our method to further refine the quality of the predicted change mask. Since SAM2 is pre-trained on universal object classes, no further training is required to integrate SAM2.

Chapter 2

Review of Previous Methods

2.1 Overview

Change Detection and Anomaly Detection are popular tasks, where various methods have been developed to cope with different types of scenes and use cases. On the high level, they can be classified by the dimensionality of the scene, i.e., 2D representations or 3D representations.

2D Change Detection and Anomaly Detection operate on scenes or objects described by 2D images. One of the very first use cases of 2D Change Detection is analyzing changes in satellite images for a Geographic Information System (GIS). GIS relies on 2D Change Detection algorithms to study changes in landforms, e.g., vegetation, water systems, and artificial landforms, which are important for researchers to understand the impacts of human activities or natural disasters on the macro-level. 2D Anomaly Detection algorithms are also important in industrial applications for quality control and workpiece inspection. MVTec AD: the MVTec Anomaly Detection dataset [5, 6] illustrated different kinds of anomalies in the different industries. For example, 2D Anomaly Detection algorithms can be used to detect broken threads on carpets in the textiles industry, cracks on nuts in the food processing industry, and broken capsules in the pharmaceutical industry.

3D Change Detection and Anomaly Detection operate on scenes or objects described by 3D representations, including stereo images, multi-view images, point clouds, DSM, and other explicit or implicit 3D representations. 3D representations contain richer information than 2D representations, which could facilitate Change Detection and Anomaly Detection under more complex environments with occlusions, rotations, or translations of

objects. However, running Change Detection and Anomaly Detection algorithms on 3D representations also needs more effort in collecting and processing 3D data as 3D data require special sensors to collect and are often magnitudes bigger in size than 2D images. With the development in 3D vision sensors, including stereo cameras and Light Detection and Ranging (LiDAR) sensors, 3D data can be acquired more easily at a cheaper cost. As a result, more methods and applications of 3D Change Detection and Anomaly Detection have emerged. For example, self-driving cars can use stereo-camera or LiDAR to detect moving pedestrians and other vehicles. While, in GIS, height-differencing is often applied to compare landscapes modeled by DSMs.

2.2 2D Image-based Methods

2D Change Detection and Anomaly Detection are popular tasks even before the recent developments and applications of deep learning methods in the field of Computer Vision. According to a 2004 literature review "Change Detection Techniques" by Lu, *et al.*, [12], image differencing, Principal Component Analysis (PCA), and post-classification comparison were the most common methods used for change detection. These methods are mostly unsupervised and are based on heuristics and statistics.

- Image differencing is the most straightforward algebraic method which simply subtracts the query and reference image pixel-by-pixel. Muchoney and Haack applied image differencing on "Change Detection for Monitoring Forest Defoliation" [34]. However, this method requires pixel-to-pixel alignment of the query and reference images and picking an appropriate threshold or applying a filter on the difference between those two images to get the final change mask.

- PCA is a popular statistical method that works on Change Detection or Anomaly Detection under the assumption that the underlying image or background are the main components, and the changes or anomalies are the minor components in the compressed space after performing PCA. According to Lu *et al.* [12], there are two ways to apply PCA: perform PCA on reference and query image together and analyze the minor components; or perform PCA separately and then subtract the query principle component image from the corresponding reference principle component image. Li and Yeh [24] applied PCA on finding a temporal change of remote sensing images for studying urban expansion. Meijer, *et al.* [32] applies PCA on unsupervised anomaly detection in sewer images. However, this method may fail when the changes

are drastic where the changes will change the distribution and become the new principle component. Moreover, this method also requires picking an appropriate threshold and determining which components to select as the major (background)/minor (changed/anomalous) components.

- Post-classification comparison is a classification/segmentation-based method that first segment both query and reference images into objects or classes, and then compare the segmented images region by region. Serra, *et al*. [37] applied this method to change detection for land cover and land use. A benefit of this method is that it can handle environmental changes. For example, although the colour of the sky can be blue in the reference image but gray in the query image, the sky will not be marked as changed as long as the sky region is classified correctly. The limitation is that this method highly relies on the accuracy of the segmentation model.

After deep learning became popular in the field of Computer Vision, more learning-based methods emerged to solve change detection and anomaly detection tasks. Learning-based methods can be divided into two groups: supervised and unsupervised. Popular supervised methods include Convolutional Neural Network (CNN) and ViT [16]; while unsupervised methods are commonly based on Autoencoder.

- CNN are one of the most popular network structures in deep-learning-based Computer Vision thanks to their efficiency, accuracy, and versatility in various tasks. CNN employs convolution and pooling operations to quickly extract spatial information while reducing the resolution of the input images. Because the number of learnable weights in the convolution kernels is far smaller than the number of pixels in input images and convolution can be efficiently executed in a highly parallel manner on modern graphic cards, CNN is memory-efficient and fast to train. C-3PO [63] applied different CNN structures, including VGG, [56], ResNet [19], and MobileNetV2 [50] to solve general change detection task. The limitation of CNN is that the CNN captures more spatial information in local neighbourhoods than in the global context because of the relatively small receptive field of convolution kernels. As a result, CNN may not be able to capture global changes/anomalies accurately.

- ViT is a modified Transformer [60] that operates on 16*16 pixel blocks instead of text inputs. It divides an image into 16*16 blocks and applies the attention mechanism on the sequence of flattened blocks, i.e., 256 or 768-dimensional vectors for grayscale or RGB images. ViT inherits the capability of the Transformer to extract information between any pair of data in the input sequence. Thus, ViT can learn more global

spatial information than CNN, while still being able to capture local spatial information in images. GeSCF [21] used pre-trained ViT to extract features, and adaptive thresholding function on the features to obtain the change mask. The main drawback of ViT is its slow training and inference speed due to its $O(n^2)$ complexity attention operation where n represents the sequence length.

- Autoencoder is similar to PCA, which are both designed to extract compact lower dimensional data from higher dimensional inputs. Similar to PCA, Autoencoder can solve the change detection and anomaly detection task by only learning the unchanged/normal distribution from the reference images, thus not being able to reconstruct changed/anomalous areas in query images. Therefore, the reconstruction loss indicates the deviation of the query image from the reference image distribution, i.e., a higher reconstruction loss indicates a higher chance of anomaly. However, unlike PCA, which is linear and deterministic, Autoencoder is nonlinear and learnable, which enables it to extract complex and nonlinear latent information in the data. Zimmerer, *et al.* [67] applied Autoencoder on unsupervised MRI image anomaly detection.

2.3 3D Methods

3D Change Detection and Anomaly Detection are generally more difficult than 2D image-based Change Detection and Anomaly Detection. According to Qin, *et al.* [43], there are three main challenges for 3D Change Detection: higher uncertainty in 3D representations, fusion of multi-modal data, and handling different viewing perspectives. 3D representations, especially point clouds, cannot describe a scene definitely because of the constraints on sampling resolutions and the zero-dimensionality property of points. As a result, thin objects might be missed and textures (2D features) cannot be captured in a point cloud. Moreover, 3D representations with different uncertainties and distributions need to be fused for change detection in certain scenarios, for example, self-driving cars. In which, stereo images from stereo cameras and point clouds from LiDARs are usually combined for the perception of moving cars and pedestrians. Lastly, 3D representations could be captured from different perspectives, which hinders the direct comparison of two scenes or objects. As a result, data alignment and/or registration are often required.

Because of the complexity in collecting and processing 3D data, 3D Change Detection and Anomaly Detection methods are generally developed later than 2D methods. Initially, rudimentary methods have been developed by comparing the geometry between two 3D

representations. Later, the development of machine learning models enables the fusion of geometrical information and other features, including RGB images and radar images. Lately, rendering or projection-based methods have been developed thanks to the advancement in NVS methods.

- Geometric comparison methods often use heuristics to compare two 3D representations, including DSMs and point clouds. Heuristics include height differencing, euclidean distance, and k-nearest neighbors (KNN). For example, height differencing is the simplest heuristic that is often used in comparing DSM where the height of the two surfaces are subtracted, filtered and thresholded to find changes between two surfaces. Height differencing is used widely in remote sensing research, for example, detecting building changes [15] and updating maps [53]. Euclidean distance is another popular metric, which can be applied on both DSMs and point clouds. Euclidean distances can be calculated between surface-to-surface, point-to-surface, or point-to-point when comparing a DSM to a DSM [42], a point cloud to a DSM [42], or a point cloud to a point cloud [7] to measure the similarity and difference between two 3D representations. KNN algorithm is often used together with the Euclidean distance to computer the average Euclidean distance between a point and its k-nearest neighbours. For example, Chai, *et al*. [7], proposed a KNN-based metric: "Density Adaptive Local Euclidean Distance", for 3D Change Detection. KNN can also be combined with other point features to compute a k-nearest-neighbour-distance-weighted features. For example, PointCore [65] employed such a strategy for 3D Anomaly Detection.

- 3D scenes can carry more information than the geometry, e.g., colour, which can assist in detecting changes and anomalies. Therefore, various methods have been developed to fuse multi-modal data for better accuracy. At an earlier stage, data are often simply fused. For example, Sasagawa [53] simply overlayed the height-differenced DSM change mask with the pixel-wise satellite image change mask to predict the changes in topographic maps. Later, machine learning models, including CNN and Transformers, enable early fusion of features and multi-modal data. For example, Liu, *et al*. [27], proposed a Transformer-based change detection model that fuses images and DSM data by learning a shared representation via cross-attention.

- Rendering or projection-based methods are designed to convert 3D representations back to 2D for comparison. One motivation is that 3D data are often bigger in size and more complex. A point cloud or a mesh of a scene may contain millions of points and faces, the running time and memory requirement for methods processing

3D representations directly could be unaffordable. The other motivation is to conduct change detection or anomaly detection between 2D and 3D representations, e.g., finding changes in an image to give a 3D reference model. In the work "3D change detection at street level using mobile laser scanning point clouds and terrestrial images" [41], terrestrial images were registered against a mobile laser scanned point cloud, and then the point cloud was projected onto each terrestrial image for comparison. However, the projection of the point cloud to 2D is not trivial because one needs to determine the occluded points that should not be rendered. Although the front and back faces can be determined when using a mesh as the 3D representation, mesh is expensive to construct and the projection of mesh onto 2D may not be photo-realistic from unseen perspectives. Later, the emergence of Neural Radiance Field [33] (NeRF) solved these issues by learning an implicit 3D model for high-fidelity rendering from unseen views. OmniposeAD [66] is a rendering-based 3D anomaly detection method that maintains a reference NeRF model of an anomaly-free reference object. For each query image, OmniposeAD first localizes it against the reference NeRF model and then renders synthetic reference images from the same perspective for comparison.

2.4 Conclusion and Motivation

As discussed in Sec. 2.2 and Sec. 2.3, 2D methods are generally simpler and faster than 3D methods. However, the limitation is also obvious: 2D methods cannot handle different perspectives easily. On the other hand, 3D methods and representations are more powerful but are more computationally expensive. As a result, we aim to combine the advantages of both 2D and 3D methods while minimizing the computation overhead by adopting the path of rendering-based 3D Change Detection and Anomaly Detection methods. This path essentially helps us reduce the comparison of two 3D scenes or objects down to the comparison of a pair of 2D images. We will explain the process in detail in the next chapter.

Chapter 3

Real-time Image-Based Pose-Agnostic 3D Anomaly Detection

3.1 Preface

This chapter is adapted from the paper "SplatPose+: Real-time Image-Based Pose-Agnostic 3D Anomaly Detection" [30] accepted at the ECCV 2024 VISION workshop.

3.2 Introduction

Inspection is a critical procedure in manufacturing to sort out defective products with manufacturing flaws such as missing parts, stains, burrs, and holes. Traditionally, humans are employed to visually inspect the products on assembly lines; however, because of the high cost and human errors, algorithms have been developed to automate anomaly detection by comparing photos of a standard product (reference images) and photos of a product to be tested (query images). Traditional approaches [39, 61, 14] rely on supervised training with both normal and defective samples. According to a survey on anomaly detection conducted by Liu, *et al*. [28], more recent anomaly detection methods have shifted towards unsupervised training, using only normal samples. The reason behind the shift towards unsupervised training is the dis-proportionality of positive and negative samples and the inability of supervised models to generalize on unseen anomalous samples. One challenge

in developing an unsupervised algorithm is that the model only sees an anomaly-free object from a set of predetermined reference views, while, in the testing stage, the object may have arbitrary orientations. Thus, anomaly detection algorithms need to be able to detect defects from images taken from any viewing angle, which is unknown and can be different from the reference views. This specific problem has been researched and named by Zhou, *et al.*, as the Pose-Agnostic 3D Anomaly Detection (PAD) problem[66].

Intuitively, this problem can be broken down into two sub-problems. First, the poses of query images are unknown. Second, no reference images are available for the query views. The first sub-problem can be solved by building a structure-from-motion model to localize query images. However, before the recent development of NVS methods, the second sub-problem is hard to solve. SfM models, though able to optimize a dense mesh for the scene, cannot render high-fidelity unseen views. Recent developments in learning-based NVS methods including NeRF and 3DGS enable direct pixel-to-pixel comparison by synthesizing pseudo reference images from the query view. Current state-of-the-art (SOTA) PAD methods rely on the same learning-based 3D representation for both sub-problems: localization and NVS. However, localization via learning-based 3D representation is costly because these methods need to back-propagate gradients from the image reconstruction loss to the camera poses, and refine those poses through gradient descent hundreds of times. As a result, current methods could take seconds, or even minutes, to detect anomalies in a query image.

Therefore, in this work, we propose an unsupervised method: SplatPose+, to solve the PAD task in real time. SplatPose+ uses a hybrid 3D representation, combining a feature-matching-based SfM model and a learning-based 3DGS model. This hybrid structure enables us to efficiently localize query views and synthesize high-fidelity anomaly-free pseudo reference images for direct pixel-to-pixel comparison. Efficiency-wise, SplatPose+ is 37.4% faster in training and 147 times faster in inference than the current SOTA: SplatPose evaluated on the MAD-Sim dataset. Accuracy-wise, SplatPose+ reaches the new SOTA on the MAD-Sim dataset as shown in Tab. 3.2 and Tab. 3.3. Moreover, SplatPose+ outperforms current methods given sparse-view training data. As illustrated in Fig. 3.3, when only using 40% of the training data, the anomaly detection and segmentation scores of SplatPose+ are still higher than the current SOTA trained on all reference images. To summarize, our contributions include:

- The first real-time image-based 3D pose-agnostic anomaly detection method.
- Achieves top efficiency and best anomaly detection and segmentation scores on the MAD-Sim Dataset.

- Robust anomaly detection performance given sparse-view training data.

3.3 Related Work

3.3.1 3D Reconstruction

In computer vision, 3D Reconstruction is a process that computes a 3D representation from a set of 2D images of a scene or an object from multiple viewing angles. Traditionally, SfM methods are developed to optimize a sparse model and extract poses from a set of images through feature extraction and feature matching. Colmap [54, 55] is a well-known SfM pipeline that uses SIFT [31] as the feature extractor and RANSAC [18] as the feature matcher. Later on, a more advanced and flexible SfM pipeline: Hierarchical Localization (hloc) [52] has been proposed to incorporate more advanced feature extractors (SuperPoint [13] and DISK [58]) and feature matchers (SuperGlue [62] and LightGlue [26]), which can generate better feature descriptors and more robust matches.

3.3.2 Novel View Synthesis (NVS)

Novel view synthesis is a task that synthesizes arbitrary views from a given set of images of a scene or an object. Although SfM methods can reconstruct dense 3D meshes, 3D meshes generally cannot generate high-fidelity images from unseen views. As a result, new classes of learning-based methods have been proposed for this task, including NeRF [33] and 3D Gaussian Splatting [20]

NeRF [33] represents a scene of an object using an implicit neural representation. NeRF deconstructs each image into individual pixels, with each pixel representing a unique ray that extends from the camera to the object in the scene. A set of 5D Points, consisting of coordinates (x, y, z) for the location and (θ, ϕ) for the viewing angle, are sampled along each ray as input for a Multi-layer Perceptron (MLP) that predicts the colour and density of each point. Finally, the colour and density values predicted along a ray are accumulated via volume metric rendering to calculate the pixel colour. NeRF is able to synthesize photorealistic novel views. However, NeRF is computationally expensive, taking hours to train for each scene. This is because it is an implicit representation, which requires millions of queries of the MLP to generate a single image.

3D Gaussian Splatting (3DGS) [20], unlike NeRF, represents a scene or an object using an explicit representation consisting of anisotropic 3D Gaussian colourized by spherical

harmonics. 3D Gaussians are initialized by randomization or a reference sparse point cloud produced by a SfM system. Rays are sampled in a similar fashion as in NeRF. However, the colour of a ray is accumulated by the view-dependent colour and density values from 3D Gaussians that the ray intersects. By using an explicit representation, no MLP queries are required. As a result, 3DGS can be trained within minutes and rendered in real-time.

3.3.3 Image-Based Pose-Agnostic 3D Anomaly Detection

OmniposeAD [66] is the first Pose-Agnostic 3D Anomaly Detection method proposed together with the MAD-Sim dataset. It utilizes NeRF to reconstruct pseudo reference and query image pairs for pixel-to-pixel comparison. A neural radiance field is trained using multiview images from the anomaly-free reference object. For each query image, the coarse query pose is determined by the pose of the reference image that has the most LoFTR [57] matches with the query image. INeRF [64] optimizes the coarse pose for the fine pose given the reference NeRF model. Then, NeRF synthesizes the pseudo reference image as if captured from the query view using the fine pose localized from each query image. After that, a pre-trained CNN is applied to extract features from the pseudo reference and query image pair. Finally, the anomaly scores are calculated by the L2 difference between the feature maps and the difference in RGB values. According to Zhou, *et al.*, OmniposeAD outperformed other non-NVS-based methods [66] at that time. However, OmniposeAD is slow to train (4.5 hours) and infer (1 minute).

SplatPose [23] is proposed to improve the efficiency and the anomaly detection accuracy of OmniposeAD [66]. A well-known issue for NeRF is the prohibitively high training and inference time. According to Kruse, *et al.* [23], SplatPose reduce the training and inference time by 2 magnitudes, which is achieved by replacing NeRF and INeRF with a pose-wise differentiable 3D Gaussian Splatting implementation. Moreover, the qualitative anomaly detection scores, and sparse-view robustness (defined in Sec. 3.5.5) are also improved thanks to the better NVS ability of 3DGS.

3.4 Method

For the pose-agnostic anomaly detection task, we define the reference set as $R = \{(r_1, T_1^R), (r_2, T_2^R), \cdots, (r_n, T_n^R)\}$ and the query set as $Q = \{(q_1, T_1^Q), (q_2, T_2^Q), \cdots, (q_m, T_m^Q)\}$ where the pose information T_i^R is known for the reference set and T_i^Q is unknown for the query set. The goal is to find image-level and pixel-level anomalies in Q while only trained on R.

During the training stage, we first build a SfM model with the groud truth poses $\{T_i^R | i = 1, \cdots, n\}$ using hloc [52] with the configuration of (NetVLAD [2] + Superpoint [13] + LightGlue [26]). Then, we use the sparse point cloud from the SfM model as the initialization to train a 3DGS model. During the testing stage, we first localize the query images against the SfM model to obtain the query poses. Then, the 3DGS model synthesizes pseudo reference images at those query poses. Lastly, a pre-trained CNN is employed to calculate the anomaly score.

The current SOTA: SplatPose uses the same 3DGS for both localization and NVS. In comparison, our method triangulates an additional SfM model for 3DGS initialization and localization, and the hybrid representation (3DGS and SfM) can train faster and detect anomalies better in real time. The pipeline is illustrated in Fig. 3.1.

3.4.1 Hybrid Representation

SfM Model

We employ hloc [52], which is the SOTA modularized visual localization pipeline, to build the SfM model. For the training stage, both images and poses are known for the reference views. We use NetVLAD [2], a lightweight but effective image-level feature descriptor for image retrieval, to build a reference database D^R. Then, overlapped reference image pairs are extracted from D^R. After that, we use SuperPoint [13] to extract local features and use LightGlue [26] to match features. We choose SuperPoint and LightGlue because SuperPoint can find more feature points than traditional methods, for example, SIFT, and LightGlue is the SOTA lightweight feature matcher. The matched features are then triangulated to build a reference sparse model M^s. For the inference stage, the poses for the query views are unknown. To localize query views, we also use NetVLAD for the image retrieval against the reference database D^R to find 15 most related reference images for each query image. NetVLAD helps speed up both training and inference by narrowing down the search space because we only need to match features between a fixed number of similar images instead of an exhaustive search. Similarly, SuperPoint features are extracted from matched reference and query image pairs. Then, LightGlue is employed to match features. Finally, the query views are localized against the sparse model M^s from the matched features by hloc.

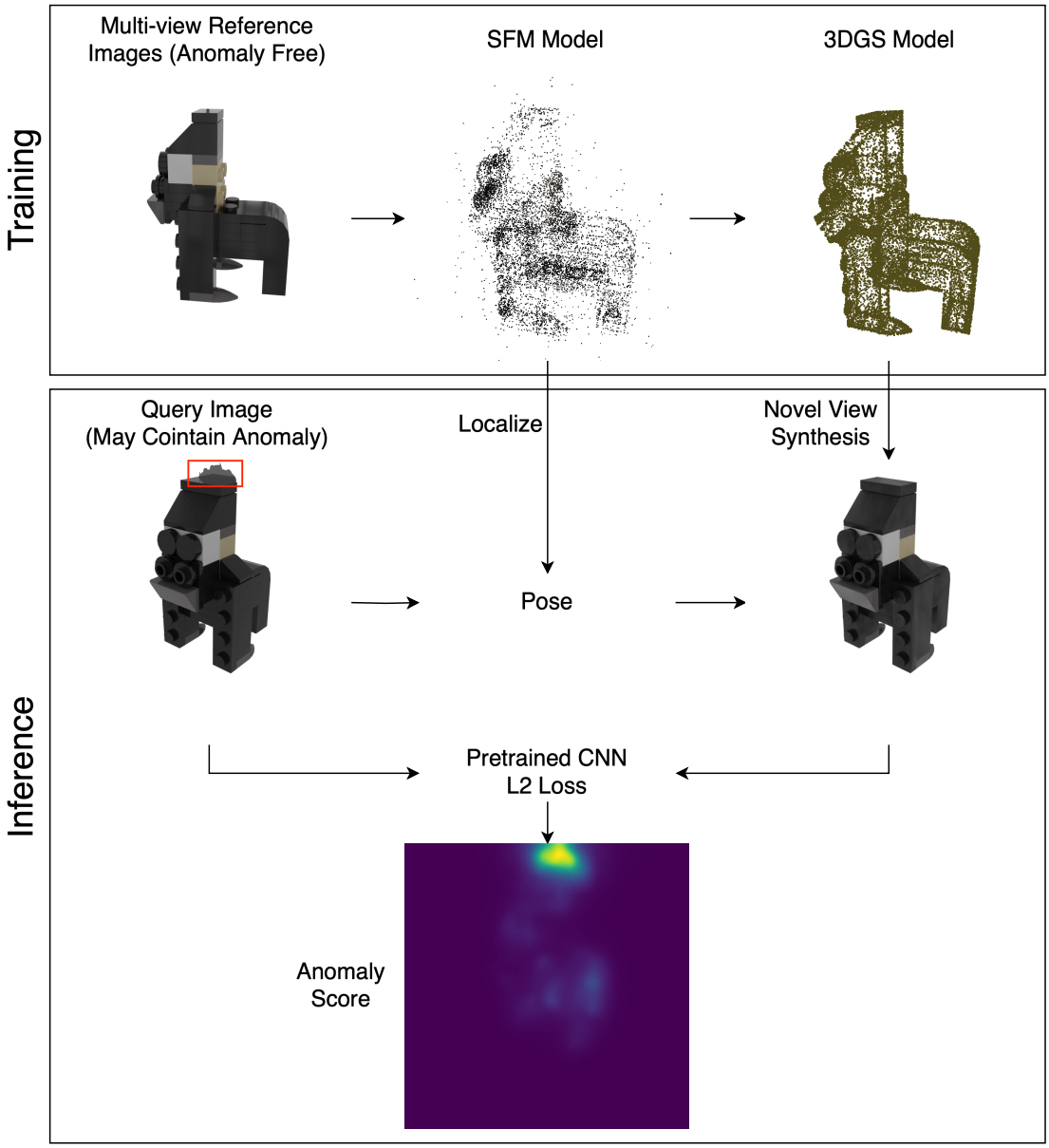


Figure 3.1: SplatPose+ pipeline incorporates an SfM model for localizing query views and initialization for the 3DGS model. 3DGS model is trained only for Novel View Synthesis.

Novel View Synthesis

We first train a 3DGS model and then use the poses localized by the SfM model M^s to reconstruct pseudo reference images at the query views. Unlike SplatPose, which trains a

3DGS model from random initialization, SplatPose+ trains a 3DGS model M^{gs} initialized by the sparse point cloud extracted from the sparse SfM model M^{s}, in which points are resided and concentrated on the true object surfaces. Moreover, SplatPose only uses RGB to colorize gausssians (maximum spherical harmonics degree: SHS = 0) in 3DGS for stable gradient descent during pose optimization. Since we no longer use 3DGS for localization, we are able to use the full spherical harmonics degree of 3 for better view-dependent colourization. Lastly, inspired by the experiments conducted in Compact3D [35], we reduce the densification frequency by setting the densification interval to 1000 because the SfM point cloud is already well-distributed on the object surface illustrated in Fig. 3.1. With good initialization, increasing the densification interval greatly improves the efficiency while causing marginal deterioration in reconstruction quality. Our experiments Tab. 3.1 show that with those three changes, our 3DGS model trains 2x faster than SplatPose and achieves comparable construction quality quantified by the peak signal-to-noise ratio (PSNR).

3DGS Setting	Initialization	Iters	Densification	SHS	Avg. PSNR	Train Time (s)
SplatPose [23]	Random	30k	100	0	42.13	294
SplatPose+	SfM	15k	1000	3	42.12	107

Table 3.1: Reconstruction efficiency and quality comparison. SHS: Spherical Harmonics Degree.

3.4.2 Anomaly Score

Given a query image q and the pseudo reference image $\hat{r}$ generated by 3DGS from the same view (paired image), we have successfully reduced the task to an image differencing problem. We used the same pre-trained 5-layer Convolutional Neural Network (CNN) used in OmniPoseAD [66], to extract 5 features $(f_1^{\hat{r}}, f_1^{q}), \cdots, (f_5^{\hat{r}}, f_5^{q})$ from the both images from layer 1 to layer 5 with decreasing spatial resolution. We use the same CNN as well as the L2 norm to calculate the anomaly map for a fair comparison. Pixel-wise L2 norms are calculated between the paired reference and query image $||q - \hat{r}||_2$ and all feature pairs $||f_1^{\hat{r}} - f_1^{\hat{r}}||_2, \cdots, ||f_5^{\hat{r}} - f_5^{\hat{r}}||_2$. All L2 norms are then resized to $224 * 224$ and then summed, Gaussian-smoothed and normalized to form the final pixel-wise anomaly score. Image anomaly scores are assigned by the maximum pixel anomaly score in each image.

3.5 Evaluation

3.5.1 Dataset

The Simulated Multi-pose Anomaly Detection (MAD-Sim) dataset is the only dataset for the PAD task proposed together by Zhou, *et al*. [66]. The MAD-Sim dataset, illustrated in Fig. 3.2, contains 3 types of defects (Burrs, Stains and Missing Parts) for 20 different synthetic Lego objects. This dataset mimics the standardized product defect detection scenario in the real world where products to be tested can have arbitrary orientation when passing a checkpoint with multiple fixed cameras. MAD-Sim is an object-central anomaly detection dataset. For each object, MAD-Sim provides

- 210 pose-known reference images (without defects) from different views.
- Approximately 300 pose-unknown query images from different views with different types of defects, and the masks of the defects.

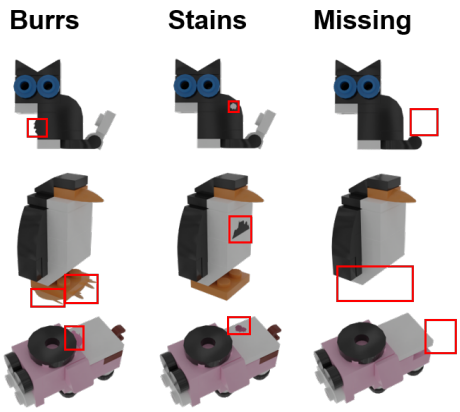


Figure 3.2: Defect Samples in MAD Dataset[66]

3.5.2 Metrics

We use the well-established metrics AUROC and AUPRO [5] to evaluate the anomaly detection and segmentation of our model. AUROC evaluates the performance of a classification

model for all thresholds by measuring the area under the receiver operating characteristic (ROC) curve. ROC curve plots true positive rates (R_{TP}) against false positive rates (R_{FP}) at all thresholds. A higher AUROC means the ROC curve bends further toward the top left, that is, the model classifies anomalies better by maintaining a higher R_{TP} and a lower R_{FP}

$$\text{AUROC} = \int R_{\text{TP}}\, dR_{\text{FP}} \tag{3.1}$$

AUPRO evaluates the performance of a segmentation model by measuring the area under the per-region-overlap (PRO) curve. PRO better describes the model's ability to segment anomalies of all sizes than intersection-over-union (IOU) because this metric calculates the overlap ratio per region, which weighs smaller anomalies the same as larger anomalies.

$$\text{PRO} = \frac{1}{N} \sum_{i=1}^{N} \frac{|P \cap C_i|}{|C_i|} \tag{3.2}$$

$$\text{AUPRO} = \int \text{PRO}\, dR_{\text{FP}} \tag{3.3}$$

where P represents pixels classified as anomalies, and C_i represents the set of pixels in the i-th connected anomalous region in the ground truth.

3.5.3 Results

We conduct all experiments under the unsupervised setting, where only anomaly-free reference images are used for training. For the SfM model, we use the default settings for NetVLAD, SuperPoint and LightGlue with num_matched = 15, that is, each query image is matched with 15 most similar reference images using the NetVLAD features. For the 3DGS model, compared to SplatPose, we decrease the densification frequency from once every 100 iterations to once every 1000 iterations and the total training iterations from 30000 to 15000.

Under these settings, our methods perform better than other NVS-based methods on image-level anomaly detection Tab. 3.2 as well as pixel-wise anomaly segmentation tasks Tab. 3.3 on the MAD-SIM dataset. SplatPose+ achieved a higher image-level AUROC for 18 out of 20 test objects. Overall, SplatPose+ achieved image-level AUROC of 96.4 and pixel-level AUPRO of 96.9 on the PAD task, the new SOTA on the MAD-SIM dataset.

Methods Objects	OmniposeAD [66]	SplatPose [23]	SplatPose+
Gorilla	93.6	91.7	**95.5**
Unicorn	94.0	97.9	**99.0**
Mallard	84.7	97.4	**98.3**
Turtle	95.6	97.2	**97.4**
Whale	82.5	95.4	**98.9**
Bird	92.4	94.0	**98.7**
Owl	**88.2**	86.8	87.9
Sabertooth	95.7	95.2	**98.3**
Swan	86.5	93.0	**95.7**
Sheep	90.1	96.7	**97.4**
Pig	88.3	96.1	**98.4**
Zalika	88.2	89.9	**93.9**
Phoenix	82.3	84.2	**90.4**
Elephant	92.5	94.7	**96.8**
Parrot	97.0	96.1	**98.9**
Cat	84.9	82.4	**89.4**
Scorpion	91.5	99.2	**99.6**
Obesobeso	**97.1**	95.7	96.8
Bear	98.8	98.9	**99.3**
Puppy	93.5	96.1	**98.0**
Avg.	90.9	93.9	**96.4**

Table 3.2: Comparison of NVS-based methods on MAD-SIM using Image-level AUROC (The higher the better ↑). The best result per class is **bold** and the runner-up is underlined.

Methods Metrics	OmniposeAD [66]	SplatPose [23]	SplatPose+
AUROC (↑)	98.4	99.5	**99.6**
AUPRO (↑)	86.6	95.8	**96.9**

Table 3.3: Pixel-wise anomaly segmentation on MAD. The best result per class is **bold** and the runner-up is underlined.

3.5.4 Efficiency

Inference efficiency is crucial in industrial anomaly detection applications because each product usually has less than a second to be inspected on an assembly line. The training time is relatively unimportant because it is considered as a one-time overhead. Previous methods OmniposeAD and SplatPose do not have real-time inference capability. Our method can not only infer in real-time but also train faster. For a fair comparison, all experiments are conducted on Ubuntu using a single NVIDIA RTX 3090. For the training stage, SplatPose+ takes an average of 76.7 seconds to build an SfM model and 107.2 seconds to train a 3DGS model. For the inference stage, SplatPose+ takes an average of 22 milliseconds on localization, 9.8 milliseconds on novel view synthesis, and 2.5 milliseconds on CNN feature extraction and anomaly score calculation. Overall, on average, the training time per object is only 184 seconds and the inference time is 34 milliseconds per query view. Equivalently, the inference can be run in real-time at approximately 30 fps.

Method	Training (h:min:sec) ↓	Inference (sec) ↓
OmniposeAD [66]	4:33:43	66
SplatPose [23]	0:04:54	5
SplatPose+	0:03:04	0.034

Table 3.4: Training and inference time on the MAD-SIM Dataset.

3.5.5 Robustness on Sparse-View Data

The MAD-Sim Dataset provides 210 evenly distributed reference views for each object. In practice, the number of reference views might be limited. Therefore, the anomaly detection performance on sparse-view data is crucial in real-world applications. As a result, we also evaluated our method on randomly sampled partial training data (20%, 40%, 60%, and 80%) to simulate sparse-view settings. Quantitatively, SplatPose+ achieves the highest AUROC and AUPRO scores given sparse-view training data under all settings. The difference is the most significant when using only 20% of the training data because learning-based 3D representations (NeRF and 3DGS) will be blurry given limited training views, and, as a result, localizing with a learning-based 3D representation will produce inaccurate poses. On the other hand, a SfM model can still localized query poses well given $20\% \cdot 210 = 42$ reference views. Qualitatively, as illustrated in Tab. 3.5, SplatPose+ predicts the cleanest anomaly map for all three kinds of defects in the MAD-Sim dataset under all sparse-view settings.

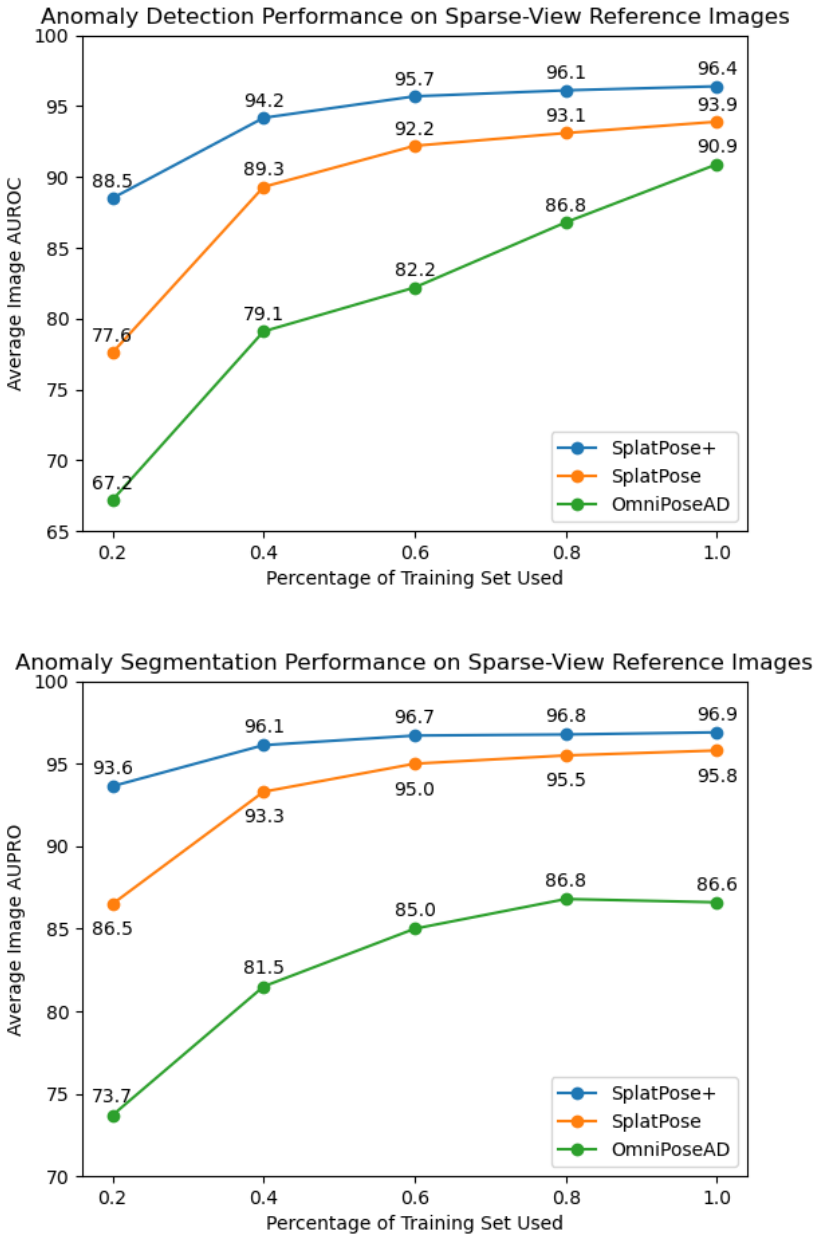


Figure 3.3: Quantitative Comparison for Anomaly Detection Performance on Sparse-View Reference Images

3.6 Conclusion

In this chapter, we present the first method that can infer in real-time for the Pose-agnostic 3D Anomaly Detection Task. We employ a hybrid representation comprised of a SfM model and a 3DGS model to encode a reference object. Given an un-posed query image, our method is able to localize it and find anomalies more accurately and 147 times faster than the previous SOTA. Moreover, our method outperforms current methods qualitatively and quantitatively under the sparse-view settings. Future works can possibly incorporate more efficient 3DGS implementations, for example, Compact3D [35], to further reduce the training and inference time.

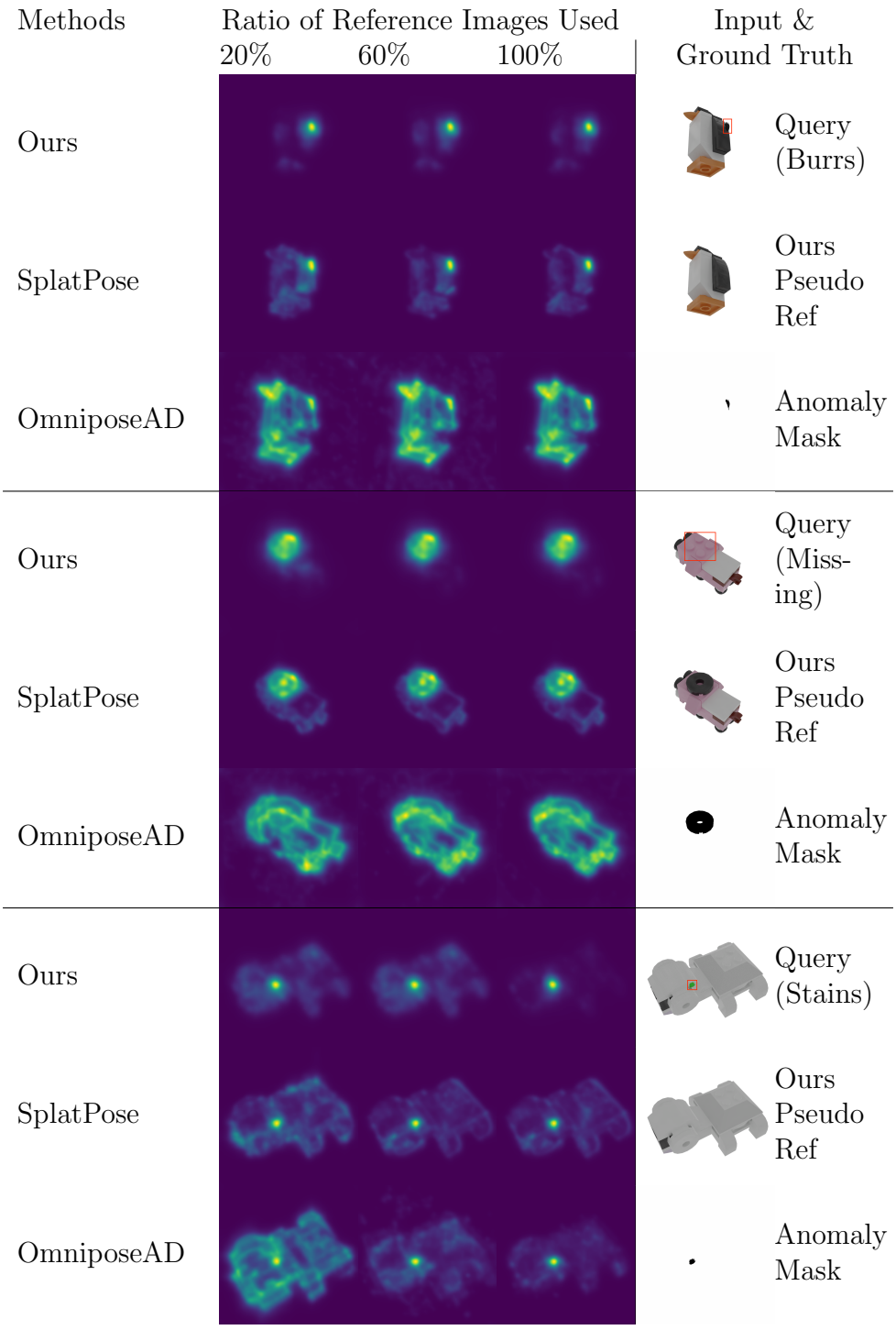


Table 3.5: Qualitative comparison of predicted anomaly maps under 3 sparse-view settings. "Ours Psuedo Ref" is generated by SplatPose++ to help readers find the anomalies.

Chapter 4

Image-Based 3D Scene Change Detection

4.1 Introduction

Scene Change Detection (SCD) is an important task that aims to find differences in a scene at different timestamps. Typical applications include locating vehicles and pedestrians entering an intersection for autonomous driving, monitoring a critical area of a factory for surveillance, and assisting scene understanding for future augmented reality (AR) applications. In the SCD task, a scene can be collected and represented in different formats, including images, point clouds, and meshes. To cope with different scene representations, different SCD algorithms have been developed to resolve each kind of scene. For example, Santos, *et al.* [51] employs Gaussian Mixture Model to cluster point clouds to conduct SCD on International Space Station Scenes are described by point clouds. However, more works, for example, CSCDNet [49], ChangeNet [59], and C-3PO [63] focus on image-based SCD, because images are generally easier and cheaper to collect. Moreover, the latest image-based methods, including GeSCF [21] and Zero-Shot Scene Change Detection [10], can take advantage of the foundation model to extract richer semantic information from images. For these reasons, this chapter will focus on the Image-based 3D Scene Change Detection task, where we have a set of reference images that are captured at time t_0 and a set of query images captured at t_1 of the same scene (with objects changed). The goal is to find changes between t_0 and t_1 as illustrated in Tab. 1.2. The challenge is that the reference and query images, though described in the same scene, can be captured at different unknown poses. Therefore, it is often hard to compare the reference images and query

images directly.

The task of Image-Based 3D Scene Change Detection is similar to Image-Based 3D Anomaly Detection in that both tasks seek to find differences between images. The major difference is that we aim to find the difference in a scene, which can be outward-facing, instead of an object, which is naturally inward-facing. Luckily, the method we introduced in Chapter 3: SplatPose+ [30] is built upon 3DGS [20], which unlike NeRF [33], can handle outward-facing scene well. Moreover, hloc [52], as a feature-based SfM framework, does not care about the orientation of the scenes. Therefore, we can adapt the SplatPose+ algorithm to the SCD task.

4.2 Related Methods

There are three main categories of methods that tackle the SCD task. The first class treats the SCD task as a pair image difference problem, where the query images and reference images that are captured at the same or similar poses are paired together. The second class operates on open-set images that can have significant differences in poses between the query images and reference images. The third class formulates the problem as an object-tracking problem instead of an image comparison problem.

4.2.1 Paired Image Difference Methods

These methods take paired query and reference images taken from the same or similar poses as input and typically use a neural network to find the difference in the paired image. ChangeNet [59] and C-3PO [63] use pre-paired images and convolution neural networks to infer a pixel-wise change map. C-3PO [63] also uses perspective transformation to correct slight viewpoint changes. Later, GeSCF [21] improves the accuracy by employing foundation models, specifically, the Vision Transformer (ViT) [16] encoder in Segment Anything (SAM) [22] to extract features and further refining predicted masked using SAM [22].

4.2.2 Open-set Image Difference Methods

These methods are similar to the paired image difference methods while releasing the similar-pose requirement on the paired image. CSCDNet [49] uses an encoder-decoder

structure to extract information from pose-different images. Later, Sachdeva and Zisserman proposed a method The "Change You Want to See (Now in 3D)" [47] that uses ViT to extract features from both images and then applies monocular depth estimation to lift and align features into 3D and re-render the feature map back to the same view for comparison.

4.2.3 Object Tracking Methods

These methods view a scene as a set of objects by utilizing image segmentation models and attempt to find the changed objects via tracking algorithms on the object level. Zero-Shot Scene Change Detection [10], the SOTA method in this category, utilizes SAM [22] to segment objects from a paired reference and query frames, and apply DEVA [9] to track the unchanged objects from one frame to another, assigning the unmatched object as changed object.

4.3 Method

For the scene change detection task, the inputs are two sequences of images of a scene captured from different perspectives at different times t_0 and t_1 as illustrated in Tab. 1.2. We define the reference sequence as $R = \{(r_1, T_1^R), (r_2, T_2^R), \cdots, (r_n, T_n^R)\}$ and the query sequence as $Q = \{(q_1, T_1^Q), (q_2, T_2^Q), \cdots, (q_m, T_m^Q)\}$ where the pose information T_i^R and T_i^Q are both unknown. The goal is to find pixel-level anomalies in Q while only trained on R.

We slightly modified the method we introduced in Sec. 3.4 to fit the SCD task.

- First, the pose information for the reference images is known in the PAD task but unknown in the SCD task. Therefore, in the PAD task, we only need to complete a partial sparse reconstruction that simply triangulates matched features into 3D to form a sparse point cloud. While, in the SCD task, we need to run a full structure-from-motion reconstruction on the reference images with pose estimation and bundle adjustment to obtain the reference poses T_i^R.
- Second, for the PAD task, the query set is not consistent, i.e., images do not necessarily come from the same object. As a result, we localize each query image individually. While in the SCD task, the query set is consistent, i.e., images come from the same scene. Therefore, when localizing Q with respect to the SfM model built on R, we also match features between query images and run bundle adjustment to acquire more accurate poses for the query set.

After that, we follow the same pipeline as in SplatPose+, where we train a 3DGS model on the reference images and their poses to generate synthetic reference images at all query perspectives. Then, the same pre-trained CNN is applied to each query image and its corresponding synthetic reference image pair to predict the change mask.

4.4 Evaluation

4.4.1 Dataset

ChangeSim [40] is the only dataset that fits the Image-based 3D Scene Change Detection task because this dataset provides a decent coverage (in hundreds) of images of each scene for building a 3D model and pixel-wise change masks are given for all query views. Here "decent coverage" requires enough multi-view images to capture all objects in a scene with no blind spots. Other datasets designed for general SCD task, for example, PCD [48] and VL-CMU-CD [1] only provide $1, 1$ and $3, 3$ reference images and query images respectively for each scene.

This dataset provides ten simulated scenes in industrial indoor environments. Each scene contains two routes traversing the scene. Each route consists of four sequences: a reference sequence of images captured in the scene following a simulated drone-flying trajectory with normal visibility, and three query sequences of images captured in the mutated scene following different simulated drone-flying trajectories with normal visibility, dusty air and low illumination respectively. The mutated scene includes four types of changes: new objects, missing objects, rotated objects and replaced objects. For all three query sequences, the pixel-wise change mask is provided as illustrated in Tab. 4.1.

4.4.2 Limitations

SplatPose+ [30] uses a binary pre-trained CNN to infer image differences. Therefore, we will only be able to perform binary change detection, i.e., changed or not changed. Moreover, the pre-trained CNN is not trained on the non-target environment change in the ChangSim dataset. Therefore, we will only work on the query sequences with normal visibility. Although it is possible to train environment-robust 5-class CNN to deduce the 5 classes: unchanged, new, missing, rotated and replaced, then the method will no longer be unsupervised because the CNN need to be trained on this dataset with the environment variation as well as the 5-class change labels.

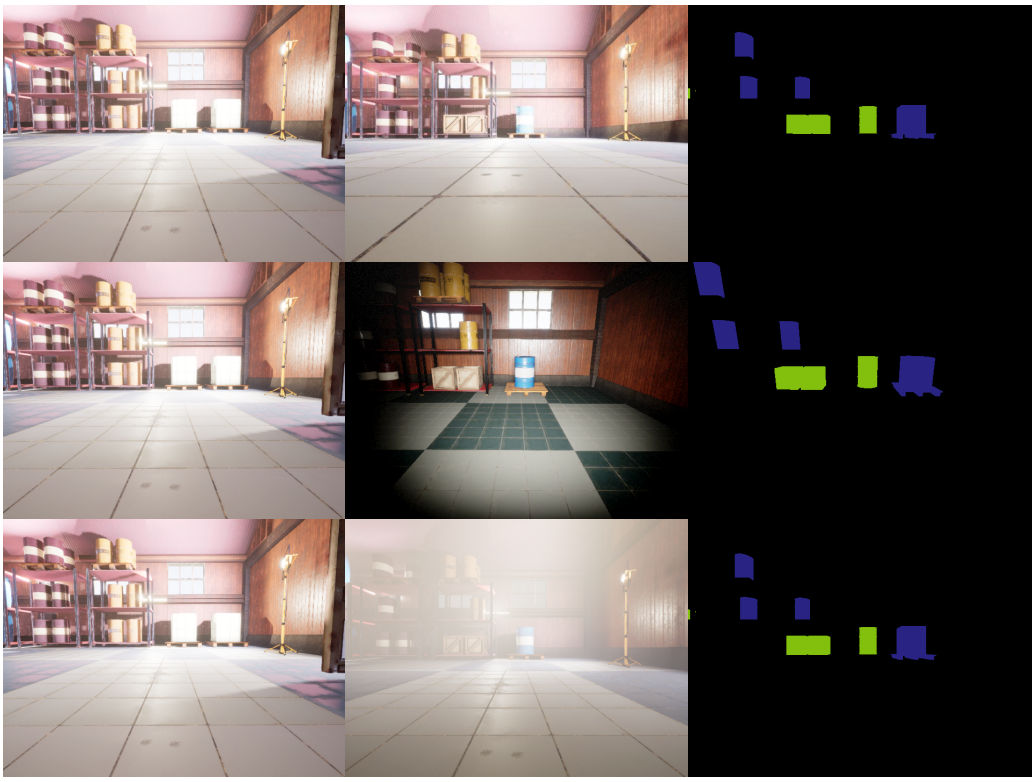
Query Visibility	Reference Image	Query Image	Change Mask
Normal			
Dark			
Dust			

Table 4.1: ChangeSim Dataset Samples (Reference Images are captured under normal visibility).

4.4.3 Metrics

We report the mean IoU metric as in the ChangeSim [40] paper. To generate a binary change mask, we set the anomaly score threshold to be 0.215, i.e., a pixel is classified as changed if its anomaly score is greater than 0.215. This value is selected empirically to maximize the F_1 score.

$$F_1 = \frac{2R_{TP}}{2R_{TP} + R_{FP} + R_{FN}} \tag{4.1}$$

where R_{TP}, R_{FP}, R_{FN} represents the true positive rate, false positive rate and false negative rate respectively.

4.4.4 Results

We conduct all experiments under an unsupervised setting, where we do not access any label during training. Note that this method may also be classified as a zero-shot method in terms of this task, as we do not train or access the training set of the ChangeSim [40] dataset. We use the same settings as in Chapter 3, with the only difference being that we bundle adjust the query sequence during localization as described in Sec. 4.3.

The quantitative comparison with ChangeNet [59] and CSCDNet [49] is shown in Tab. 4.2. Note that ChangeNet [59] and CSCDNet [49] operate on the pre-paired reference and query images illustrated in the top two rows of Tab. 4.2. It is noticeable that the poses of the first and second rows are different, e.g., the floor in the last column. This is because the pre-paired images are paired based on the closest L1 distance between the 7-D poses, according to Park, *et al*. [40]. Our method generates synthetic reference images via NVS to pair with query images, which more accurately mimics the query image poses as illustrated in the second and third row of Tab. 4.2. As a result, our model, though unsupervised, predicts more accurate masks than previous supervised methods by taking advantage of pixel-to-pixel comparison.

Quantitatively, our modified model achieves a higher IoU than previous supervised methods on the binary SCD task without environment variation illustrated in Tab. 4.3.

4.5 Mask Refinement

Inspired by previous methods described in the Sec. 4.2 that utilize SAM [22] for pose refinement, we also attempt to apply Segment Anything 2 (SAM2) [44] on our unsupervised masks for mask refinement. SAM2 is a new version of the foundation model SAM, which is the first foundation model that enables zero-shot object segmentation on various applications. SAM2 has recently proposed to segment objects in both images and videos faster and more accurately than SAM. SAM2 can be prompted by points, bounding boxes and coarse masks that describe the location of an object. Therefore, we can use SAM2 for mask refinement in the SCD task by breaking the change mask into multiple object-level sub-change masks and refining those sub-masks one by one. *Note that we only apply SAM2 on this task because the changes in this task (SCD) are object-level while the changes in the PAD task are not.*

4.5.1 Pipeline

- First, we find all connected components in each mask generated by the modified SplatPose+ algorithm and store each connected component as sub-masks. The reason is that SAM2 segments objects and refines masks on the object level.
- Second, we convert the sub-masks to pseudo logits by mapping the binary change map (0 for unchanged and 1 for changed) to (-10 for unchanged and 10 for changed) that matches the range of the logit mask prompt for SAM2.
- Third, we prompt the SAM2 model with each sub-mask represented by pseudo logits on both the query image and the synthetic reference image to obtain the most confident refined mask in the query domain and the reference domain respectively. *We run the SAM2 model twice on both images because when an object disappears, the change mask only describes the object in the reference frame; while when a new object appears, the change mask only describes the object in the query frame.*
- Finally, we combine the refined masks with two different strategies: pick the **higher-score** refined sub-mask between the reference and query domain, or simply apply the **union** of two masks.

4.5.2 Results

Qualitatively, SAM2 [44] is able to refine the coarse sub-masks to the accurate shape if the image domain (reference or query) for segmentation is correct, i.e., the mask describes an object in the corresponding image. For example, the first row of Tab. 4.4 demonstrates the mask of the disappeared boxes is correctly refined in the synthetic reference image domain but not in the query image domain because these boxes no longer exist in the query scene. On the contrary, the fourth row of Tab. 4.4 demonstrates the mask of the newly appeared forklift is correctly refined in the query image domain but not in the synthetic reference image domain.

Quantitatively, as illustrated in Tab. 4.5, we found applying both the union strategy and the higher-score strategy can improve the Changed IoU and Unchanged IoU metrics over the baseline (without mask refinement). The union strategy achieves a higher IoU on the changed part, while the higher-score strategy achieves a higher IoU on the unchanged part. The mean IoU for both strategies is similar.

Overall, as illustrated in Tab. 4.5 and Tab. 4.6, SAM2 [44] can help improve the SplatPose+ generated mask quality and accuracy on the binary SCD task by aggregating the higher-score refined sub-masks or taking the union of refined sub-masks.

4.6 Conslusion

In this chapter, we first explain how to adapt SplatPose+ [30] to solve the SCD task with little modifications. Then, we introduce a custom SAM2 [44]-based pipeline that can help refine the SplatPose+ generated change masks. Overall, our unsupervised methods achieve higher IoU scores than previous supervised methods on the binary SCD task evaluated on the ChangeSim[40] dataset without mask refinement. Moreover, we can further improve the IoU scores by 2% to 4% with SAM2-assisted mask refinement.

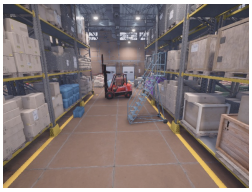
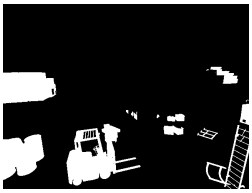
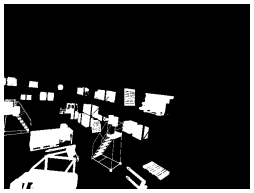
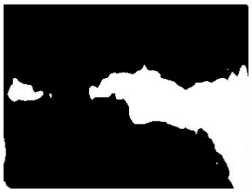
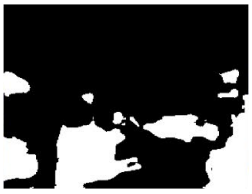
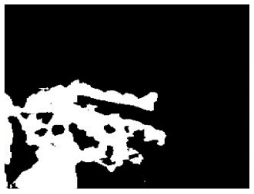
Paired Ref Img			
Query Img			
Syn Ref Img			
GT Mask			
ChangeNet Mask			
CSCDNet Mask			
Ours Mask			

Table 4.2: Qualitative binary SCD performance evaluated on ChangeSim [40] dataset without environment variations.

Methods	Unchanged IoU (%)	Changed IoU (%)	Mean IoU (%)
ChangeNet [59]	73.3	17.6	45.4
CSCDNet [49]	87.3	22.9	55.1
C-3PO [63]	90.4	28.8	59.6
Ours	**91.2**	**38.6**	**64.9**

Table 4.3: Binary SCD performance evaluated on ChangeSim [40] dataset without environment variations. (The higher the better, top results are bold.)

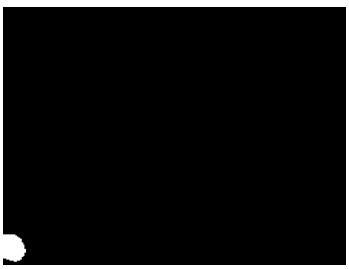
Input Sub-mask	Refined on Query Img	Refined on Syn Ref Img
		
		
		
		
		

Table 4.4: Qualitative comparison of SAM2 [44] assisted sub-mask refinement. Sub-masks are refined with respect to either query image or synthetic reference image.

Methods	Unchanged IoU (%)	Changed IoU (%)	Mean IoU (%)
Baseline	91.2	38.6	64.9
Union	93.1	**42.0**	**67.6**
Higher-score	**93.9**	41.1	67.5

Table 4.5: SAM2 [44] assisted mask refinement IoU. (The higher the better, top results are bold.)

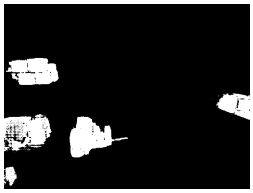
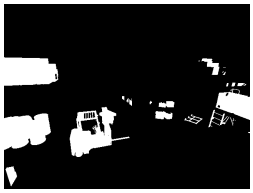
Input Mask	Higher Score	Union	Ground Truth
			

Table 4.6: SAM2 [44] assisted final mask refinement.

Chapter 5

Conclusion and Recommendation

5.1 Conclusion

In conclusion, we showcase that SplatPose+ [30] can not only achieve SOTA performance (speed and accuracy) on the Image-based Pose-agnostic Scene Change Detection task but is also able to solve the Image-based Scene Change Detection task with little modifications. We also demonstrate that SAM2 [44] can be combined with SplatPose+ to refine the SplatPose+ generated masks for higher accuracy in the SCD task.

5.2 Recommendation

In general, we do realize the scalability limitation of SplatPose+ [30]. The SFM method we use: hloc, could take hours to build a sparse model when there are thousands of images in a scene. Moreover, according to Liu, *et al.* [29], 3DGS may run out of memory and suffer from low inference speed because of the speed bottleneck in depth sorting when fitting a large-scale scene. To improve the scalability, later works may employ more efficient SFM methods, for example, GLOMAP [38], and 3DGS variants designed for larger scale scenes, for example, CityGaussian [29]. Moreover, SplatPose+ [30] cannot handle environment changes, e.g., lighting and visibility, which may limit its application in complex environments. It is possible for future works to incorporate environment embeddings in either the NVS part or the CNN part to bridge the gap.

Specifically for the image-based SCD task, this method can only classify binary changes, i.e., either changed or not changed. Future work can work on incorporating a pre-trained

semantic-aware model to classify semantic changes, including rotation and displacement of objects.

References

[1] Pablo F Alcantarilla, Simon Stent, Germán Ros, Roberto Arroyo, and Riccardo Gherardi. Street-view change detection with deconvolutional networks. *Auton. Robots*, 42(7):1301–1322, October 2018. 30

[2] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition, 2016. 17

[3] Vic Barnett. Outliers in statistical data. *John Wiley & Sons google schola*, 2:705–708, 1994. 2

[4] Richard J Beckman and R Dennis Cook. Outlier..... s. *Technometrics*, 25(2):119–149, 1983. 2

[5] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, January 2021. 7, 20

[6] 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*. SCITEPRESS - Science and Technology Publications, 2022. 7

[7] JX Chai, YS Zhang, Z Yang, and J Wu. 3d change detection of point clouds based on density adaptive local euclidean distance. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:523–530, 2022. 11

[8] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009. 2

[9] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation, 2023. 29

[10] Kyusik Cho, Dong Yeop Kim, and Euntai Kim. Zero-shot scene change detection, 2024. 27, 29

[11] Wen-Hsuan Chu and Kris M. Kitani. *Neural Batch Sampling with Reinforcement Learning for Semi-supervised Anomaly Detection*, page 751–766. Springer International Publishing, 2020.

[12] E. Brondízio D. Lu Corresponding author, P. Mausel and E. Moran. Change detection techniques. *International Journal of Remote Sensing*, 25(12):2365–2401, 2004. 4, 8

[13] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description, 2018. 15, 17

[14] Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022. 13

[15] GR Dini, Karsten Jacobsen, Franz Rottensteiner, M Al Rajhi, and Christian Heipke. 3d building change detection using high resolution stereo images and a gis database. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences; XXXIX-B7*, 39:299–304, 2012. 11

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 4, 9, 28

[17] Francis Ysidro Edgeworth. Xli. on discordant observations. *The london, edinburgh, and dublin philosophical magazine and journal of science*, 23(143):364–375, 1887. 2

[18] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, June 1981. 15

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 9

[20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. iv, x, xiii, 3, 15, 28

[21] Jaewoo Kim and Uehwan Kim. Towards generalizable scene change detection, 2024. 10, 27, 28

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 28, 29, 32

[23] Mathis Kruse, Marco Rudolph, Dominik Woiwode, and Bodo Rosenhahn. Splatpose & detect: Pose-agnostic 3d anomaly detection, 2024. iv, 3, 16, 19, 22, 23

[24] X. Li and A. G. O. Yeh. Principal component analysis of stacked multi-temporal images for the monitoring of rapid urban expansion in the pearl river delta. *International Journal of Remote Sensing*, 19(8):1501–1518, 1998. 8

[25] LibreTexts. *Introductory Statistics*. LibreTexts, 2023. 3

[26] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed, 2023. 15, 17

[27] Biyuan Liu, Huaixin Chen, Kun Li, and Michael Ying Yang. Transformer-based multimodal change detection with multitask consistency constraints. *Information Fusion*, 108:102358, August 2024. 11

[28] Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1):104–135, January 2024. 13

[29] Yang Liu, He Guan, Chuanchen Luo, Lue Fan, Naiyan Wang, Junran Peng, and Zhaoxiang Zhang. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians, 2024. 39

[30] Yizhe Liu, Yan Song Hu, Yuhao Chen, and John Zelek. Splatpose+: Real-time image-based pose-agnostic 3d anomaly detection, 2024. iii, iv, 13, 28, 30, 34, 39

[31] Jorge Martinez-Gil. Sift: An algorithm for extracting structural information from taxonomies, 2016. 15

[32] Dirk Meijer, Mitchell Kesteloo, and Arno Knobbe. Unsupervised anomaly detection in sewer images with a pca-based framework. In *International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI)*, pages 354–359, 2018. 8

[33] 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, 2020. iv, xiii, 12, 15, 28

[34] Douglas M Muchoney and Barry N Haack. Change detection for monitoring forest defoliation. *Photogrammetric engineering and remote sensing*, 60(10):1243–1252, 1994. 8

[35] KL Navaneet, Kossar Pourahmadi Meibodi, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Compact3d: Smaller and faster gaussian splatting with vector quantization, 2024. 19, 25

[36] Selçuk Oral. Lcd display dead pixel, 2011. 3

[37] X. Pons P. Serra and D. Saurí. Post-classification change detection with data from different sensors: Some accuracy considerations. *International Journal of Remote Sensing*, 24(16):3311–3340, 2003. 9

[38] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L. Schönberger. Global structure-from-motion revisited, 2024. 39

[39] Guansong Pang, Choubo Ding, Chunhua Shen, and Anton van den Hengel. Explainable deep few-shot anomaly detection with deviation networks, 2021. 13

[40] Jin-Man Park, Jae-Hyuk Jang, Sahng-Min Yoo, Sun-Kyung Lee, Ue-Hwan Kim, and Jong-Hwan Kim. Changesim: Towards end-to-end online scene change detection in industrial indoor environments, 2021. xi, 30, 31, 32, 34, 35, 36

[41] Rongjun Qin and Armin Gruen. 3d change detection at street level using mobile laser scanning point clouds and terrestrial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 90:23–35, 2014. 12

[42] Rongjun Qin, Armin Gruen, and Cive Fraser. Quality assessment of image matchers for dsm generation – a comparative study based on uav images, 2021. 11

[43] Rongjun Qin, Jiaojiao Tian, and Peter Reinartz. 3D change detection - Approaches and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 122:41–56, December 2016. 4, 10

[44] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric

Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. iv, xi, xii, 5, 6, 32, 33, 34, 37, 38, 39

[45] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection, 2022.

[46] PJ Rousseeuw and AM Leroy. Robust regression and outlier detection, 1987. 2

[47] Ragav Sachdeva and Andrew Zisserman. The change you want to see, 2022. 29

[48] Ken Sakurada and Takayuki Okatani. Change detection from a street image pair using cnn features and superpixel segmentation. In *British Machine Vision Conference*, 2015. 30

[49] Ken Sakurada, Mikiya Shibuya, and Weimin Wang. Weakly supervised silhouette-based semantic scene change detection, 2022. 27, 28, 32, 36

[50] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks, 2019. 9

[51] Jamie Santos, Holly Dinkel, Julia Di, Marina Moreira, Brian Coltin, Paulo Borges, Trey Smith, and Oleg Alexandrov. Unsupervised change detection for space habitats using 3d point clouds. In *AIAA SCITECH 2024 Forum*. American Institute of Aeronautics and Astronautics, January 2024. 27

[52] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale, 2019. iv, 3, 15, 17, 28

[53] A Sasagawa, K Watanabe, S Nakajima, K Koido, H Ohno, and H Fujimura. Automatic change detection based on pixel-change and dsm-change. *jiS*, 600:2, 2008. 11

[54] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 15

[55] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 15

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 9

[57] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers, 2021. 16

[58] Michał J. Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient, 2020. 15

[59] Ashley Varghese, Jayavardhana Gubbi, Akshaya Ramaswamy, and P Balamuralidhar. ChangeNet: A deep learning architecture for visual change detection. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 129–145. Springer International Publishing, Cham, 2019. 27, 28, 32, 36

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 9

[61] Qian Wan, Liang Gao, and Xinyu Li. Logit inducing with abnormality capturing for semi-supervised image anomaly detection. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12, 2022. 13

[62] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems, 2020. 15

[63] Guo-Hua Wang, Bin-Bin Gao, and Chengjie Wang. How to reduce change detection to semantic segmentation, 2023. 9, 27, 28, 36

[64] Lin Yen-Chen, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. Inerf: Inverting neural radiance fields for pose estimation, 2021. iv, 3, 16

[65] Baozhu Zhao, Qiwei Xiong, Xiaohan Zhang, Jingfeng Guo, Qi Liu, Xiaofen Xing, and Xiangmin Xu. Pointcore: Efficient unsupervised point cloud anomaly detector using local-global features, 2024. 11

[66] 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, 2023. iv, x, 5, 12, 14, 16, 19, 20, 22, 23

[67] David Zimmerer, Simon A. A. Kohl, Jens Petersen, Fabian Isensee, and Klaus H. Maier-Hein. Context-encoding variational autoencoder for unsupervised anomaly detection, 2018. 10