TECHNISCHE UNIVERSITÄT WIEN

AUTOMATION & CONTROL INSTITUTE
INSTITUT FÜR AUTOMATISIERUNGS- & REGELUNGSTECHNIK

Model Optimisation and Comparison for Improved Change Detection in Autonomous Systems

BACHELORARBEIT

Conducted in partial fulfillment of the requirements for the degree of a

Bachelor of Science (BSc)

supervised by

Ao.Univ.Prof. Dipl.-Ing. Dr.techn. M. Vincze

submitted at the

TU Wien
Faculty of Electrical Engineering and Information Technology
Automation and Control Institute

by
Florian Pfleiderer
Brandmayergasse 6/19, 1050 Wien

Wien, im Juni 2024

Vision for Robotics Group
A-1040 Wien, Gusshausstr. 27, Internet: http://www.acin.tuwien.ac.at

Abstract

This thesis addresses the challenge of improving change detection systems for indoor mobile robots. Although significant advances have been made in recent years, the use of home assistance robots without direct supervision remains a topic of contention. Models to improve such applications are investigated, using the existing tidy-up pipeline from Langer, Patten, and Vincze [1] as comparison. The pipeline achieves an average precision of 81%, but has unsolved problems with occluded objects, the quality of the reconstruction and the definition of the search areas for the robot. The model The Change You Want To See - 3D [2] showed potential solve the problems caused by the search area definition and the reconstruction, and is evaluated further. By applying CYWS-3D directly to camera data, the reconstruction step of the existing pipeline and the associated loss of information can be skipped and the model can be used as an indicator for objects missed by the tidy-up-pipeline. The evaluation on the ObChange data set shows that the new model achieves an accuracy of 65% in scenes with objects close to the camera, which appear larger in the field of view. In one scenario, this value exceeds the 64% of the existing tidy-up pipeline, but the result is overshadowed by a low average recall of 26%. Furthermore, CYWS-3D needs further work on the detection of small objects, where only a precision of 33% is achieved. This means that the overall precision drops sharply depending on the distribution of object sizes. The source code of the project can be found at the following link https://github.com/florianpfleiderer/CYWS3D-pipeline

Kurzzusammenfassung

In dieser Arbeit geht es um die Verbesserung von Systemen zur Erkennung von Veränderungen bei mobilen Robotern in Innenräumen. Obwohl in den letzten Jahren erhebliche Fortschritte erzielt wurden, bleibt der Einsatz von Assistenzrobotern im Haus ohne direkte Überwachung ein umstrittenes Thema. Es werden Modelle zur Verbesserung solcher Anwendungen untersucht, wobei die bestehende Tidy-Up-Pipeline von Langer, Patten, and Vincze [1] als Vergleich dient. Die Pipeline erreicht eine durchschnittliche Genauigkeit von 81%, hat aber ungelöste Probleme mit verdeckten Objekten, der Qualität der Rekonstruktion und der Definition der Suchbereiche für den Roboter. Das Modell The Change You Want To See - 3D [2] hat das Potenzial, die Probleme bei der Definition des Suchbereichs und der Rekonstruktion zu lösen, und wird in der Arbeit genauer untersucht. Durch die Anwendung von CYWS-3D direkt auf Kameradaten kann der Rekonstruktionsschritt der bestehenden Pipeline und der damit verbundene Informationsverlust übersprungen und das Modell als Indikator für die von der Tidy-Up-Pipeline übersehenen Objekte verwendet werden. Die Auswertung des ObChange-Datensatzes zeigt, dass das neue Modell in Szenen mit kameranahen Objekten, die dadurch im Sichtfeld größer erscheinen, eine Genauigkeit von 65% erreicht. In einem Szenario übertrifft dieser Wert die 64% der bestehenden Aufräum-Pipeline, aber das Ergebnis wird durch eine niedrige durchschnittliche Wiedererkennung von 26% überschattet. Darüber hinaus braucht CYWS-3D weitere Arbeit an der Erkennung kleiner Objekte, wo nur eine Genauigkeit von 33% erreicht wird. Dies bedeutet, dass die Gesamtgenauigkeit je nach Verteilung der Objektgrößen stark abnimmt. Der Quellcode des Projekts kann unter folgendem Link abgerufen werden: https://github.com/florianpfleiderer/CYWS3D-pipeline

Contents

List of Figures

List of Tables

1 Introduction

The concept of a domestic assistance robot has evolved from a futuristic fantasy to a clear reality due to rapid advances in robotics, making such systems available to consumers in the near future. The development of indoor robotics competitions, like WRS RoboCup@Home is driving progress and has become a crucial factor in the advancement of household robots [3]. Change detection is one of the critical tasks in autonomous systems, especially in indoor robotics. It involves identifying and responding to alterations in the environment during multiple visits at different timestamps.

This thesis is motivated by the need to improve the accuracy and efficiency of change detection systems on mobile autonomous systems. Robots, when used unsupervised at home, need to work flawlessly. Despite significant advancements in the field, existing models struggle with challenges such as occlusion, varying lighting conditions, and complex environments. These limitations result in missed detections or false positives, thereby compromising the overall effectiveness of the autonomous system [4].

1.1 Challenge

The task of change detection does not yet achieve sufficient results to be used on unsupervised robots in ordinary households. These robots need to work flawlessly in order operate reliably and safely in a indoor environments. Specifically, the issues of reconstruction accuracy, the detection of small objects and occlusion remain significant obstacles [5].

Langer, Patten, and Vincze [1] have proposed an autonomous system capable of achieving a precision of 81% and a recall of 64%. Their method achieves the task of change detection by combining scene reconstruction, object detection and object matching into an autonomous pipeline. Fig.1.1 demonstrates a model result of their method, serving as the groundwork for this thesis.

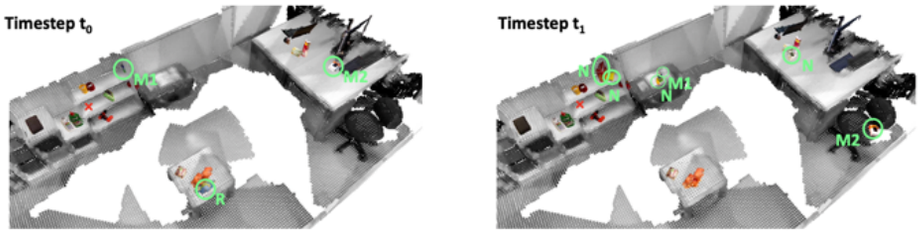


Figure 1.1: Results of the tidy-up-pipeline introduced by Langer, Patten, and Vincze [1]. M marks moved, N new and R removed objects at two different timestamps t_0 and t_1.

The proposed tidy-up-pipeline segments the reconstruction into local surfaces like tables, shelves and furniture seats to avoid having to scan an entire room at every visit. This

concept suffers from undetected objects at surface borders and struggles with curved or uneven spaces like the seat cushion of a sofa. The detection of small objects also struggles from the quality of the reconstructed point clouds and the handling of occlusions suffers from too few perspectives of the scenes. Furthermore, the system is designed to run on a mobile robot platform, which adds the constraint of limited computational power. The aim of this thesis is to find and evaluate a model that can improve the change detection capabilities of the tidy-up-pipeline by overcoming one of the remaining challenges, adhering to the constraints of mobile applications.

1.2 Contribution

The main contribution of this paper is the evaluation of The Change You Want To See - 3D (CYWS-3D) [2] in order to determine if the tidy-up-pipeline can be improved by integrating the new model. CYWS-3D takes two RGB-D frames from the camera stream looking at the same scene and detects changes between those. The frames can be from different perspectives to detect occluded objects or from the same viewpoint. The detected changes are used to indicate changes in the scene to the tidy-up-pipeline, which can re-scan the plane if objects where missed. Towards finding a solution to improve the existing pipeline, three papers presented at the International Conference on Computer Vision 2023 (ICCV23), are analysed to determine the best model for further evaluation. Sandström, Li, Van Gool, *et al.* [6] and Cao and Charette [7] tackle the problem of reconstructing the environment, while Sachdeva and Zisserman [2] proposed CYWS-3D, a change detection model that works on camera data directly. The choice fell towards the latter, which is evaluated to find the optimal configuration for indoor environments, before conducting an inference with the optimal settings on the *ObChange* dataset provided by Langer [8]. Detection results, illustrated in Fig. 1.2 are subsequently evaluated and compared against the performance reported by Langer, Patten, and Vincze [1] to empirically assess if the new model serves as a reliable indicator for identifying objects missed by the existing system.

CYWS-3D shows promising results on improving the restrictions of the tidy-up-pipeline regarding detecting missed objects on surface borders. Especially medium sized objects are detected at a precision of 65%. This number succeeds the relatively low average precision of 56% achieved by CYWS-3D, compared to the 81% achieved by the existing system. Looking closer at the kitchen environment of *ObChange*, it is evident that CYWS-3D outperformed the tidy-up-pipeline in this scene by one percentage point in precision. The scene is characterised by a low mean camera distance and therefore, the majority of the objects are being classified as medium sized or bigger. In contrast, the average recall of the tested model is 26%, which makes further work necessary to improve CYWS-3D. The code for this project can be found at: https://github.com/florianpfleiderer/CYWS3D-pipeline

1.3 Thesis Outline

The structure of this thesis is organised as follows: Chapter 2 provides a literature review of the theoretical constructs and methodologies relevant to change detection systems. Chapter 3 presents a comparative analysis of recent models and covers the criteria for the model selection process. Chapter 4 details the evaluation of the selected model, including

Figure 1.2: This figure shows results of CYWS-3D. The detections are used to indicate the existence of objects missed by the tidy-up-pipeline.

parameter overview and optimisation results. Chapter 5 compares the performance of the optimised model with the existing pipeline, analysing results and identifying pipeline weaknesses and problematic objects.

Through this structured approach, the thesis aims to provide a comprehensive solution to improve the system proposed by Langer, Patten, and Vincze [1], enhancing the capability to operate in complex and dynamic environments on mobile robot platforms.

2 Visual Perception Methods for Tidy Up Tasks

Methods like object recognition, scene differencing, robot localisation, scene reconstruction and object manipulation must work together seamlessly in order for robots to master everyday activities such as cleaning and tidying up [9]. The autonomous system presented by Langer, Patten, and Vincze [1] solves the problem of change detection by first dividing a room into surfaces, reconstructing and storing each surface, followed by detecting and categorising all detected objects into new, moved and removed objects. The location of detected and categorised objects are saved in a database. This chapter focuses on visual perception methods used in the existing system, as well as relevant theoretical concepts for this thesis. It also includes an overview of evaluation criteria for object recognition models, which will be used to compare the evaluated model to the existing system. The aim is to provide a theoretical basis for qualitative analysis and evaluation standards.

2.1 Object Detection

Object detection is a crucial area within computer vision, sharing its importance with image classification and semantic segmentation, aimed at locating and identifying objects within an image. This technology is essential for robotic applications related to indoor household robots like the Toyota Human Support Robot (HSR), on which the system by Langer, Patten, and Vincze [1] was tested.

In indoor applications, there are a number of challenges when it comes to recognising objects due to the narrow and often complex environments. Many objects in close proximity to each other in a scene can make it difficult to recognise individual objects without prior knowledge of the surroundings [10]. In addition, there are lighting-dependent factors such as strong fluctuations in light intensity and shadows, which can create new but incorrect contours in the images [11]. Metallic objects or mirrors produce reflections that create virtual images of objects elsewhere in the room and thus lead to incorrect results [12]. Another problem is occlusion, which is very common indoors. This requires different perspectives on the scenes, which is possible through advanced route planning. Another solution to this are multimodal sensor systems, which allow having different perspectives at all times [13].

Due to the rising number of appliances, there have been numerous recent advances in the field of object and change detection, with important developments listed below.

1. **Deep Learning:** Improved accuracy and efficiency in detecting objects by shifting from traditional methods to deep learning-based techniques [14].
2. **Synthetic Data:** Synthetic data for model training to increase the size of datasets and reduce the need for manual hand-labelling. [15].

3. **Efficient Algorithms:** Faster and more efficient detection models like Faster R-CNN, YOLO, and SSD. [16].

4. **Spatial Understanding:** Depth sensing and point cloud analysis are used for more accurate spatial understanding in indoor scenes [17].

2.2 Change Detection

Recognising changes involves identifying differences between scenes that were recorded at different times. It is important that the same scene is in focus, the camera position can change. This leads to two different approaches, 2-D change detection and 3-D change detection. In the former, the focus is on colour changes of pixels and changes to neighbouring pixels, which is due to the lack of depth information [2]. In 3-D change detection, technologies such as Light Detection and Ranging (LiDAR), point clouds and digital cave models (DEM) are used in addition to the RGB camera. This allows the advantages of existing depth data to be utilised. This three-dimensional approach can also be used if the camera position does not change and the advantages of depth data are still to be utilised [18].

For indoor applications, change detection is essential for activities like monitoring progresses and aiding autonomous navigation. Techniques have been developed to detect components of under-construction indoor partitions using 2D digital images, identifying elements like studs, insulation, and electrical outlets, thereby facilitating automated inspection and progress tracking in construction sites [19]. Similarly, indoor surveillance systems integrate 3D point cloud data with computer vision to detect and analyze changes, providing comprehensive monitoring of indoor spaces and detecting intruders or objects of interest [20].

2.3 Scene Reconstruction

Scene reconstruction deals with the generation of 3D models, such as point clouds, to create a digital model of the environment. The information about the environment can come from various data sources such as images, videos or LiDAR data. The technology is used in robotics, especially in indoor applications. Frequently used methods include Neural Radiance Fields (NeRF) and Simultaneous Localisation and Mapping (SLAM).

2.3.1 NeRF

NeRFs have revolutionized 3D scene representation and rendering by synthesizing photorealistic images from sparse views. NeRF operates by optimizing a neural network that maps spatial coordinates and viewing directions to color and density, enabling the rendering of continuous volumetric scenes. Recent advances include modular frameworks like NeRFstudio for efficient NeRF development, which balance speed and quality, and the introduction of methods like NerfingMVS that optimize NeRF for indoor multi-view stereo using depth priors to address the shape-radiance ambiguity, significantly enhancing depth quality in indoor settings. Evaluation of NeRF in indoor environments highlights its capability to produce detailed reconstructions, though challenges like accurate camera

pose estimation and optimization under complex lighting conditions persist, necessitating continued innovation for optimal performance [21], [22].

2.3.2 SLAM

SLAM is a computational process used in robotics and autonomous vehicles to construct or update a map of an unknown environment while simultaneously keeping track of the agent's location within it. SLAM integrates data from various sensors, such as cameras and LiDAR, to create a map of the environment and determine the location of the vehicle or robot relative to this map.

Recent advancements in SLAM have focused on enhancing accuracy, speed, and efficiency. Neural implicit functions for map representation have shown promising results, especially in dense visual SLAM, where hierarchical feature volumes facilitate map reconstruction and camera motion solving [23], [24]. Advances also include the development of efficient mapping systems like ESLAM, which combines neural radiance fields with SLAM for improved 3D reconstruction and camera localization [25].

Dense Neural RGB-D SLAM refers to the use of deep learning and neural networks to achieve dense mapping and localization with RGB-D data. Recent works have introduced methods that anchor scene features in a point cloud for iterative, data-driven generation, enhancing tracking and mapping accuracy while adapting the density of anchor points to the input's information density [6]. This approach not only improves the detail and accuracy of the generated map but also reduces memory and runtime requirements in less detailed regions.

2.4 Evaluation Metrics

In order to quantify and compare the results of machine vision models, the metrics *precision* and *recall* are used. The former reflects the accuracy of all positive predictions by measuring the proportion of correctly recognised objects (true positives) among all recognised objects. The second is used to show how well the result achieves relevant hits by measuring the proportion of objects recognised as true positives among all objects actually to be marked. The F1 score is the harmonic mean of *precision* and *recall* and combines these into a single metric if both values are to be weighted equally in the evaluation. Mean Average Precision (mAP) extends these metrics by averaging the precision values at different recall values across multiple classes and Intersection over Union (IoU) thresholds. This results in a comprehensive measure of the overall performance of the model in recognising objects of different sizes and shapes in different categories [26].

3 Comparative Analysis of Recent Models

Three papers, presented at the ICCV23, have the potential to improve the existing autonomous system. The models by Sandström, Li, Van Gool, *et al.* [6] and Cao and Charette [7] focus on reconstructing environments and can potentially replace the reconstruction done by *Voxblox* and *Elastic Fusion*, and Sachdeva and Zisserman [2] have presented CYWS-3D, which can be used to recognise changes directly from the RGB-D stream and thus provide hints for unrecognised objects. The next sections deal with open questions of the existing pipeline, outline the relevant decision criteria and explain each model in detail.

3.1 Remaining Challenges in Current Pipeline

The existing tidy-up-pipeline already achieves remarkable results with an average precision of 81% and a recall of 64%. These numbers already make it a robust model, but it leaves behind open problems regarding the quality of reconstruction, occlusion and the introduced surface-concept. Unsupervised use is not possible before significant enhancements to the open problems are made. This section provides an overview of those areas.

3.1.1 Reconstruction

A major open challenge in the tidy-up-pipeline concerns the accuracy of the reconstruction of the environment. The pipeline suffers from inaccurate localisation of the camera position, which is important for creating the point clouds. The current solution, in which the odometry data is used to improve the localisation of the camera position in ElasticFusion [27], works well but there is still a significant offset in a number of scenes compared to the RGB-D data, as seen in Fig. 3.1. The images show the annotated camera stream,

Figure 3.1: There is a mismatch between the position of the reconstructed objects and their actual position in the original scene. The bounding boxes show the position of the ground truth objects in ObChange, the first and third image show the displacement.

where the ground truth data is stored in the point clouds. The robot position is the same for the RGB image and the projected point cloud, which highlights the misalignment of

the point clouds. In addition, further improvements are needed to effectively reconstruct environments with few geometric or visual features.

The quality of the reconstructed point cloud is also in need of improvement. Particularly detailed environments such as densely packed shelves or objects touching each other are only recognised partially or even incorrectly. The reconstruction quality affects the recognition of small objects such as pens or screwdrivers. It can happen that these objects are filtered out as noise or as part of the reconstructed surface and are therefore no longer available for the mapping task. Fig. 3.2 shows the reconstructed marker to the left of the mug, suffering from the poor quality.

Figure 3.2: The reconstruction quality of *ElasticFusion* lacks fine details, as seen when looking at the marker to the left of the mug. This makes it difficult to detect small objects.

3.1.2 Search Space

Furthermore, the definition of the search spaces has room for improvement. The surface concept developed by Langer, Patten, and Vincze [1] creates a convex hull around each detected plane in the scene, which is used as the search space for objects. It works well for flat, large surfaces such as tables or shelves, but problems quickly become apparent. One of these concerns are seating cushions of sofas or other curved areas. Objects lying below the theoretical straight surface are wrongly being added to the surface and therefore not recognised. Objects that are placed at the edge of a surface, not having all their points within the convex hull are also at risk of being excluded as noise and are therefore not recognised.

3.1.3 Occlusion

Obscured objects are also problematic, as ElasticFusion [27] uses data from a single perspective of a journey through the environment it wants to reconstruct. This means that objects that are obscured by larger objects cannot be fully reconstructed and therefore cannot be used for further comparison. The partial reconstruction leads to difficulties in matching the object to other scenes containing fully reconstructed instances.

3.2 Criteria for Model Selection

The selection criteria outlined below should aid choosing one paper for further evaluation.

1. **Integration:** The new model should either be able to provide information about whether a missed object could be in the field of view, or provide data that can be integrated into the pipeline.

2. **Performance:** The tested systems should be able to run on a mobile robot platform and not be limited to GPU use only. This can be achieved through optimising the implementation strategy or through using a resource-efficient, small model.

3. **Benefit to existing Pipeline:** The new model should be a useful extension to the pipeline and improve one of the open problem areas. It does not have to replace the entire pipeline, but should support and improve a processing step such as reconstruction, object recognition or scene differencing.

3.3 Model Overview

This section outlines the method and discusses integration benefits and drawbacks of three Models presented at the ICCV23. The goal is to identify one model for further evaluation and testing with the dataset provided by Langer [8], which was used for the evaluation of the existing method described by Langer, Patten, and Vincze [1].

3.3.1 CYWS-3D

The paper titled *The Change You Want to See (Now in 3D)* by Sachdeva and Zisserman [2] from the University of Oxford's Visual Geometry Group addresses the challenge of detecting changes in 3D scenes using pairs of images taken from different viewpoints and times as shown in Fig. 3.3.

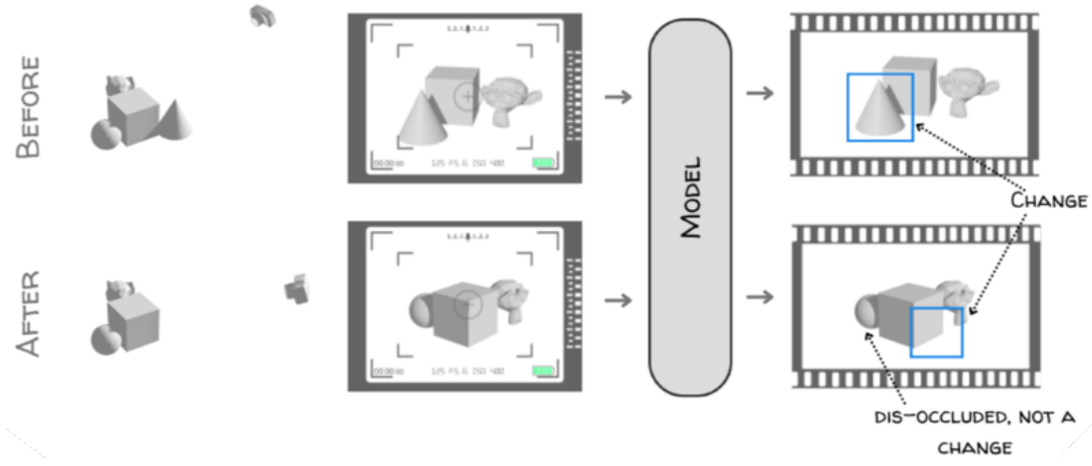


Figure 3.3: This illustrates the pipeline for the CYWS-3D [2] model. The cone is removed and gets detected as a change. Objects that are hidden from one perspective should be ignored as it is not known if they have changed.

This task is complicated by occlusions, viewpoint shifts, and the absence of large-scale, real-world training datasets. Their method employs a class-agnostic change detection model trained on synthetic data, capable of working with real-world images. The approach involves registering the two images in 3D space and then identifying differences, leveraging self-supervised techniques and synthetic training to generalize across various scenes. The model operates directly on RGB images, avoiding the need for detailed camera or depth information, but has the option of providing depth, camera intrinsic and extrinsic information to improve the model accuracy. Results show that the model achieves an average precision of 0.82 on their synthetic KC-3D dataset, which consists of 90955 image pairs for training and 4548 pairs for testing - indicating its robustness and adaptability to different scenarios [2]. The paper also discusses the limitations of the model, such as its dependency on reliable registration and the potential for further improvement with better depth estimation methods.

CYWS-3D is promising for improving the object mapping system of Langer, Patten, and Vincze [1], because the loss of information due to the reconstruction can be reduced. On the one hand, the new model works directly on RGB and depth images and can therefore recognise the changes directly from the robot's camera image and communicate them to the pipeline. If an object is subsequently not reconstructed, CYWS-3D still provides the information that there should have been an object and the reconstruction can be repeated. On the other hand, the model of Sachdeva and Zisserman [2] handles complex scene changes from different perspectives. This can be used to leverage additional viewpoints that were not used for the reconstruction using ElasticFusion [27]. The generalised training approach with synthetic data also makes it possible to generally recognise objects that are not part of the YCB object set [28]. Despite the model not being computationally light, it does run on devices without a GPU and it offers implementation flexibility, because it will not be necessary to compare changes in every single frame recorded by the robot. The implementation is focused on analysing predefined positions with a change in perspective, where it is possible. This only requires the pipeline to be adapted to save the RGB-D information at specific locations and use CYWS-3D to compare two instances looking at the same location from different angles.

3.3.2 PointSLAM

The method of Sandström, Li, Van Gool, *et al.* [6] is a new method for dense SLAM using monocular RGB-D data. In their approach, the density of the point cloud is dynamically adapted to the input data and thus, the mapping accuracy adapts to the density of information, as can be seen in Fig. 3.4. In contrast, other SLAM methods use a fixed grid for the scene representation, which leads to a lot of redundant information in the point cloud and consequently, to higher demands on the hardware. The goal of Point-SLAM is to further increase memory and runtime efficiency compared to other dense SLAM methods by implementing a flexible and more efficient feature anchoring.

The method uses multi-layer perceptrons (MLPs) for decoding geometric and color information from the neural point cloud, leading to accurate and detailed scene rendering. Evaluations on the Replica [29], TUM-RGBD [30] and Scan-Net [31] datasets datasets indicate that Point-SLAM performs better than comparable dense neural RGBD SLAM methods. Their results in tracking, mapping and rendering accuracy confirm its effectiveness in dynamic scene reconstruction.

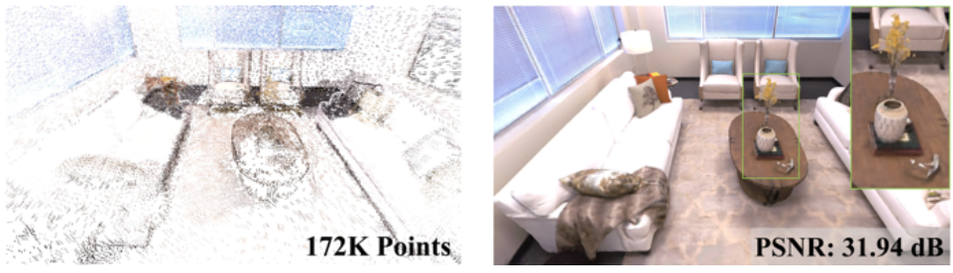


Figure 3.4: The left image shows the dense mapping, where PointSLAM predicts lots of details [6]. This increases the quality of reconstruction for critical areas containing lots of information, as seen in the right image.

Point-SLAM is not optimal for improving the autonomous object mapping in open-world environments described by Langer, Patten, and Vincze [1] for two reasons. Firstly, the method focuses on static, complete environment detection of entire rooms and does not have the option of reconstructing individual areas separately in the event of a change. This is in contrast to the semantic segmentation and surface adaptation used in the current method and leads to difficulties in integration. Secondly, despite the improvements in efficiency, the neural point cloud based approach of Point-SLAM has significant computational demands, making it less suitable for the real-time processing requirements of autonomous robot applications.

3.3.3 SceneRF

Cao and Charette [7]'s method called SceneRF, is a self-supervised method for monocular 3D scene reconstruction using radiation fields. It leverages advances in the field of Neural Radiance Fields (NeRF) and introduces a novel probabilistic sampling strategy and a spherical U-mesh to explicitly optimise depth and efficiently process large scenes. The model can generate new depth views from a single input image and merge multiple views into a 3D scene as shown in Fig. 3.5. The approach achieves better results than existing methods when reconstructing scenes, both on indoor (BundleFusion) and outdoor (SemanticKITTI) datasets [32], [33]. The paper also discusses the technical challenges, such as feature compression in spherical mapping and the need for improved inference time. SceneRF could be used to aid single plane reconstruction,

The model, with its reliance on advanced techniques like Neural Radiance Fields (NeRF) and complex probabilistic sampling, might face challenges in real-time deployment due to its computational intensity and resource demands. Additionally, the model faces challenges when it comes to reconstructing fine details, which can be a problem in cluttered and variable environments typical of the use case for indoor household robots. This requirement may limit its applicability and may also be problematic for the detection of small objects. Despite being an advancement in using NeRF for reconstruction, the overall quality when it comes to fine details, paired with the computational demands will be limiting factors when it comes to enhancing the existing pipeline.

Figure 3.5: SceneRF [7] works using a single image as input for reconstruction, but lacks enough fine details as seen in the second image. Merging multiple views into a single scene offers minor improvements.

3.4 Comparison Results

The choice for further evaluation fell towards the paper presented by Sachdeva and Zisserman [2] due to two main reasons. Firstly, the proposed method offers a new way of enhancing the pipeline, leveraging the availability of RGB-D data and overcoming the reconstruction problems. Secondly, the models simplicity, only needing to RGB-D input frames, offers implementation flexibility in terms of when to run the model. The plan is to use its advantages when comparing scenes from different perspectives and therefore, there is no need to deploy it on every frame recorded.

Regarding the model by Sandström, Li, Van Gool, *et al.* [6], which promises good reconstruction results using a dense SLAM algorithm, the computational demand is not suited for a mobile robot platform. Furthermore, the surface concept introduced by Langer, Patten, and Vincze [1] looks to be drastically different to the output data of Point-SLAM, which focuses on reconstruction of a complete room.

Lastly, SceneRF, presented by Cao and Charette [7], which uses Neural Radiance Fields to enhance Depth estimation and scene reconstruction still lacks overall reconstruction quality, especially when it comes to fines details or cluttered scenes. This, accompanied by long inference times leads to the decision of not investigation this model further.

4 Optimisation of CYWS-3D

The optimisation of the model by Sachdeva and Zisserman [2] to indoor dataframes is essential for enhancing model performance. By analyzing frames from the RGB streams collected during five navigations through various rooms, where notable perspective differences are present, the model's improvement potential is assessed. This chapter focuses on evaluating the optimal configuration on the *ObChange Dataset*[8] as the initial step.

4.1 Parameter Overview

The CYWS-3D model, originally trained with synthetic data, includes six adjustable parameters listed in Tab. 4.1 to evaluate the best configuration for indoor applications. To identify the optimal parameter settings, an inference was conducted using all parameter combinations across the extracted frames, running fives times for each setting to reduce statistical outliers. This approach is done to evaluate the model itself, hoping to identify trends to increase the performance indoors.

Parameter	**Tested Configurations**
bbox_areas	200, 300, 400, 500
keep_matching_bboxes	true, false
minimum_confidence_threshold	0.2, 0.25, 0.3, 0.35, 0.4
max_predictions_to_display	4, 5, 6
registration_strategies	2d, 3d
depth	true, false

Table 4.1: Tested Configurations for CYWS-3D

4.1.1 Bounding Box Area

The parameter is designed to filter out bounding boxes based on their area, as complex indoor scenes may contain numerous small changes like a slightly moved computer cable, which we do not want to identify as a changed Object. It takes an array of bounding boxes and a threshold area as inputs. The function iterates over each bounding box, calculates its area using the Shapely library, and removes those with an area smaller than the specified threshold. The remaining bounding boxes are then compiled into a new array, ensuring that only bounding boxes with sufficient size are retained. This helps in eliminating small, potentially irrelevant bounding boxes from further analysis.

The precision of the model increases with larger defined areas, but this is because the ground truth data suffers from reconstruction inaccuracies caused by the robot pose estimation in ElasticFusion [27], as explained in Section 3.1.1. The larger bounding boxes have a bigger IoU with the slightly moved ground truth and thus, better precision. The

Figure 4.1: The images on the right side show the results of using a minimum bounding box area of 500, where CYWS-3D dismisses the fruits as being minor changes.

expectation was that the precision will drop due to smaller objects not being recognised as seen in Fig. 4.1 on the right side, but the dominant effect is due to the shifted ground truth.

4.1.2 Matching Boxes Algorithm

The algorithm filters bounding boxes from two sets of predictions, ensuring each bounding box in one image has a corresponding match in the other image. Initially, bounding boxes with confidence scores below a specified threshold are removed, as explained in Section 4.1.3. The centers of the remaining high-confidence bounding boxes are then calculated for both images. These center points are mapped between the coordinate spaces of the two images using predefined transformation functions. The algorithm then iteratively checks if each transformed center point lies within any bounding box in the other image. If a match is found, the algorithm calculates the distance between the center points and retains the bounding box with the minimum distance. This process is repeated for both sets of bounding boxes, ensuring only corresponding pairs are kept. The result is two sets of filtered bounding boxes, one for each image, containing only those that have corresponding matches in the other image. This approach effectively ensures that changes detected between images are validated by having a corresponding bounding box in both images. Fig. 4.2 shows a result of CYWS-3D with the algorithm skipped.

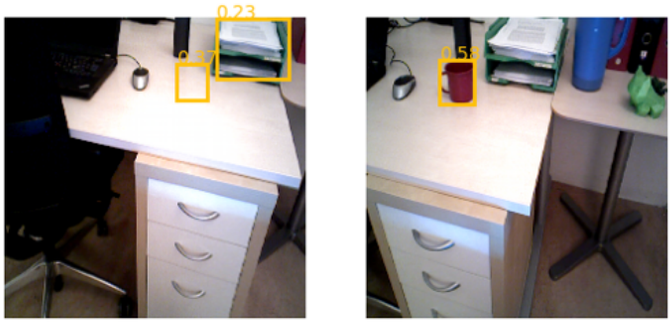


Figure 4.2: The *keep-matching-boxes* algorithm ensures detections always have corresponding bounding boxes in both images. This figure shows a bounding box with a confidence score of 0.23 not having a match in the second image, which would be dismissed by the algorithm.

4.1.3 Minimum Confidence Threshold

The parameter is used to filter out bounding boxes with low confidence scores. This function iterates through each bounding box and its corresponding confidence score, retaining only those boxes whose scores exceed the specified confidence threshold. By doing so, it ensures that only high-confidence predictions are considered for further processing, improving the reliability and accuracy of the detected objects. This parameter is crucial for eliminating uncertain detections and focusing on the most confident results, which is particularly important in applications where precision is prioritised.

4.1.4 Number of Predictions

The parameter limits the number of predictions shown by sorting bounding boxes based on their confidence scores and displaying only the top ones. This parameter is set to five because confidence scores beyond the fifth prediction consistently fall below 0.2, indicating lower reliability. Having a confidence threshold of 0.2 is significant as lower scores lead to false positives. This approach ensures that only high-confidence detections are considered, which is crucial for maintaining robust and reliable object detection in various applications [34].

4.1.5 Depth Registration

This parameter determines whether depth estimation is necessary for aligning feature maps. While 2D homography methods suffice for planar scenes, our case requires 3D registration due to differing perspectives in image pairs. This choice ensures accurate feature warping and alignment, enhancing the model's robustness and accuracy in indoor applications where perspective changes are frequent.

The *depth* parameter dictates whether the model estimates depth or uses ground truth depth. This flexibility allows the model to either generate depth maps via a monocular estimator or utilize provided depth data, enhancing accuracy in feature alignment and change detection, especially in complex 3D scenes.

4.2 Optimisation Results

The resulting parameters are shown in Tab. 4.2, and where evaluated by comparing the mean-average-precision and recall values for all parameter combinations.

Parameter	**Optimal Configuration**
bbox_areas	400
keep_matching_bboxes	false
minimum_confidence_threshold	0.2
max_predictions_to_display	5
registration_strategies	3d
depth	true

Table 4.2: Optimal Configuration for CYWS-3D

The average precision is weighted more heavily, as the predictions should indicate a missed object with high precision. Producing lots of false positives due to high recall values would not benefit the pipeline.

The metric *max_predictions_to_display* was evaluated separately, as mentioned in Section 4.1.4. The metric registration strategy was chosen to be 3d, because the model is intended to be used in scenes with perspective differences and setting this value to 2d resulted in a drop of 30% in overall precision.

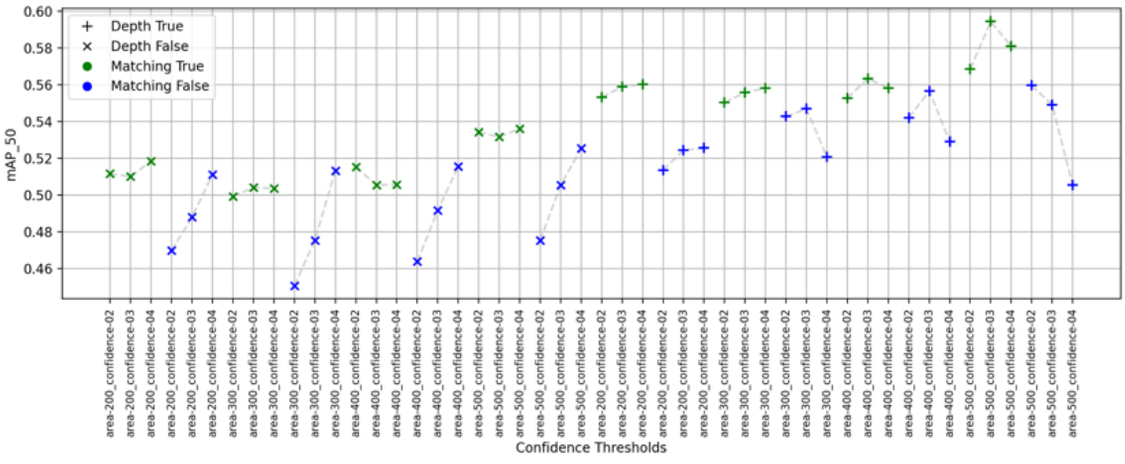


Figure 4.3: This shows the average precision for all possible parameter combinations at an IoU threshold of 0.5. The different bounding box areas and confidence thresholds are mentioned on the x-axis. The differences between using the camera depth compared to estimating the depth are highlighted by different markers.

The Mean-Average-Precision (mAP) is evaluated at an Intersection-over-Union (IoU) of 0.5, as the exact location of the object is less of an importance than the knowledge of whether there is an object in the scene or not. Furthermore, as mentioned in Section 3.1.1, the ground truth suffers from inaccuracies. It is evident in Fig. 4.3, that the performance increases when depth information is available and the depth estimation can be skipped. The choice of bounding box area fell towards using 400, as it offers the best trade-off between precision and recall increase. Despite being higher on average, the results indicate a decrease in performance for the highest confidence threshold of 0.4 when using the original depth information. The model predicts false positives with higher accuracy and the choice fell towards using the baseline setting of 0.2 because the recall significantly drops without a notable increase in precision for the chosen area and depth setting. Furthermore, the model surprisingly performs better when setting *keep_matching_bboxes* to *false*. The algorithm - as explained in Section 4.1.2 - deletes unmatched bounding boxes, but the model frequently draws correct bounding boxes on new objects, but misses the the corresponding box in the image taken, where the object should have been. These *true positives* would then be removed by the algorithm despite being correct.

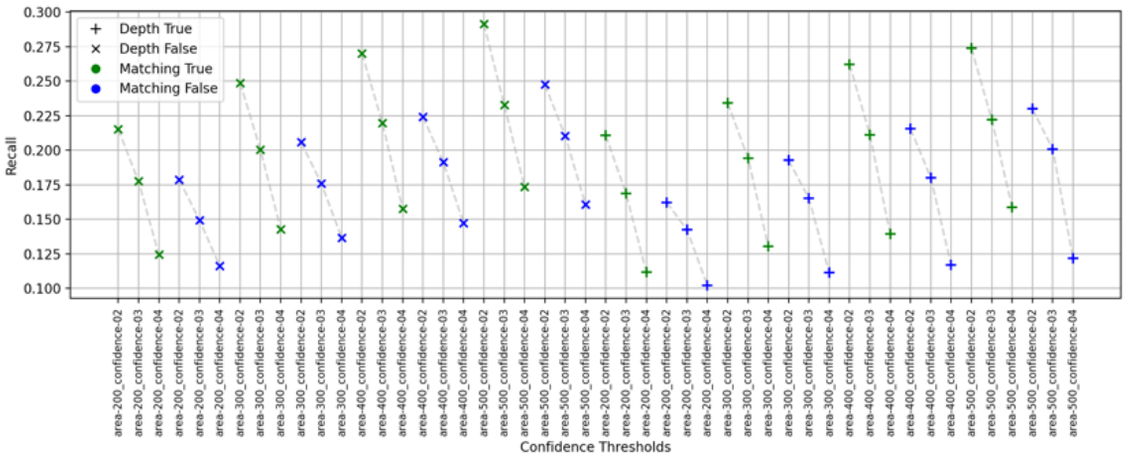


Figure 4.4: It is evident, that the recall drops drastically when using higher minimum confidence thresholds. The threshold levels are given on the x-axis, combined with different bounding box sizes.

The recall values, illustrated in Fig. 4.4, show a clear reduction when choosing confidence values 0.3 and 0.4, which is as expected. Moreover, recall values are lower when using the depth information from the camera stream and skipping the depth estimation process, as fewer objects get predicted more precisely. Another factor, which decreases the recall values, is the matching boxes algorithm as true positives are removed from the set of predicted bounding boxes. Using the original depth information provided by the camera, combined with skipping the matching-boxes algorithm still achieved the fourth highest recall value of 0.26 and therefore, it was chosen for the comparison to the existing pipeline.

5 Comparison to Existing Pipeline

To compare the model by Langer, Patten, and Vincze [1] with Sachdeva and Zisserman [2], the latter is evaluated on all frames where a change in perspective is present. If no significant perspective change is possible to derive from the camera stream, as it is the case for the *Kitchen* and partly the *Office* scene, the comparison is done using frames from similar camera poses. For better comparison, precision and recall are calculated from the data provided in the paper by Langer, Patten, and Vincze [1]. It introduces two metrics used for evaluation: *#Correctly categorized* and *#Correctly categorized in detected Objects*, the former comparing the results to the actual ground truth objects, including Objects that were not detected due to reconstruction errors. The metric *#Correctly categorized* is used for the calculations, as CYWS-3D also compares against the original ground truth due to working directly on the camera stream.

It is necessary to achieve high precision to improve the existing pipeline, because CYWS-3D should make predictions that are worth reconstructing the scenes again. To minimise the annotation error in the ground truth lowering the results, precision and recall are calculated at an IoU of 0.5.

5.1 Results & Analysis

The model from Sachdeva and Zisserman [2] achieves significantly lower results than the existing pipeline. One problem is the low recall value, which can be partly explained by the fact that many of the images are taken from a distance of around one metre and many household objects therefore appear very small in the field of view. Sachdeva and Zisserman [2] do not mention the recall in their paper, along with no mentions of the handling of small objects. The precision is also lower than that of the pipeline. One reason for this is the training with synthetic data. It allows a generalisation, but applied to the scenes from the dataset provided by *ObChange*, the tested model does not generalise well enough. The following table lists all the results for the individual rooms, the settings were taken from Tab. 4.2.

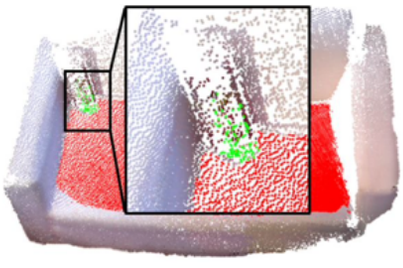
Room	**Precision**		**Recall**		**F1-Score**	
	Base	**CYWS-3D**	**Base**	**CYWS-3D**	**Base**	**CYWS-3D**
Big Room	0.81	0.55	0.67	0.24	0.74	0.33
Small Room	0.91	0.56	0.63	0.23	0.74	0.33
Living Area	0.88	0.53	0.69	0.26	0.78	0.35
Office Desk	0.76	0.55	0.76	0.26	0.76	0.35
Kitchen Counter	0.64	0.61	0.41	0.31	0.50	0.41
Overall Score	0.81	0.56	0.64	0.26	0.72	0.35

Table 5.1: Comparison of Precision, Recall, and F1-Score by Room

In the kitchen, the results are significantly better and the models achieve comparable results. This is because the objects in the kitchen are closer to the camera and the majority of objects are considered medium-sized objects, which are objects of size 32 x 32 or larger [35]. This is in contrast to all other scenes, where the average distance to the relevant objects is bigger and therefore, the objects appear smaller in the field of view.

5.1.1 Pipeline Weaknesses

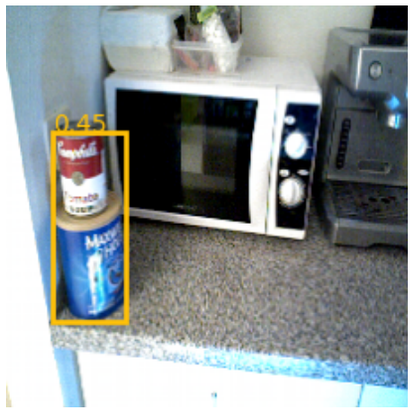
The pipeline proposed by Langer, Patten, and Vincze [1] has weaknesses regarding objects on curved surfaces and objects placed close to the edge of a surface. Objects placed on the edge of a surface are problematic for the reconstruction, as only few points lie within the convex hull of the surface and the are likely to be classified as noise. CYWS-3D does not face similar issues as it works on RGB-D data. In Fig. 5.1, two examples of objects missed by the pipeline are provided with their detection results using CYWS-3D.



(a) Reconstructed PointCloud [1]

(b) Detection using CYWS-3D

(c) Partially Reconstructed Object [1]



(d) Detection using CYWS-3D

Figure 5.1: The existing Pipeline could not reconstruct the chips can visible in (a), and the stacked cans in (c) due to the objects being partly outside the search space and not fully reconstructed. CYWS-3D has no such constraints.

It is worth noting that in the second detection, the new model was not able to differ between the objects due to the geometrical similarities. Objects on curved surfaces, like

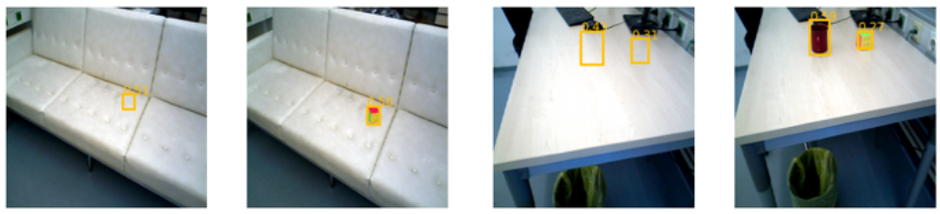
cups lying on a sofa in the *LivingArea* Dataset, also lead to a faulty reconstruction, because they are estimated to be part of the surface and therefore not reconstructed properly. CYWS-3D does not rely on detecting surfaces and therefore, it be used in both scenarios mentioned in this section to aid the detection of the existing pipeline.

5.1.2 Problematic Objects

A consistently missed object was the unsolved Rubik's cube, despite it having clear geometric features available from the depth information of the camera stream. The model presented by Sachdeva and Zisserman [2] works class agnostic due to the training on synthetic data, but nevertheless, the results did not show great generalisation for all objects. The model was able to detect the unsolved cube only on one location, which is shown in Fig. 5.2(c). This was likely possible because there was no other change in the

(a) Unsolved Rubik's Cube (b) Unsolved Rubik's Cube

(c) Only Unsolved & Detected Cube (d) Unsolved Rubik's Cube



(e) Solved Rubik's Cube (f) Solved Rubik's Cube

Figure 5.2: The unsolved Rubik's cube was mostly missed by CYWS-3D, despite having a simple geometric form that should aid the detection with available depth information. The solved cube was detected at a higher rate, examples are shown in (e) and (f).

image center and the cube was placed on a clear, flat surface. It is important to note here, that the cube was not detected at all when using the *registration_strategy=2d*, as the clear geometric form, which is visible through the ground truth depth aided detection. The issue does not apply to the solved Rubik's cube, which was detected similar to the mean-

average-precision of the model. The two screwdrivers were also missed in most detections and Fig. 5.2(b) shows an example were both the Rubik's cube and the screwdrivers were placed in a scene, but both were missed. It is interesting to note that this combination of objects was never guessed correctly over all inference runs.

The results also showed, that the new model has difficulties in scenarios where an object is replaced by another with similar geometries, which can be explained by the model relying more on depth information than comparison by colors. Fig. 5.3 shows an example of this problem together with an example of the same objects, but the new object is moved instead of replaced.

(a) (b)

Figure 5.3: CYWS-3D has problems detecting a change for replaced objects with similar geometries. This is because the model uses the estimated or the original depth maps to predict corresponding points between the images. (b) shows the results, if the new object is placed nearby.

As the model is trained class-agnostic, it fails to recognise objects appearing in clusters as such, and rather predicts one larger object. This can be explained by no constraints on what the size, colours or geometric features of an object could be and thus, a fruit basket, a couple boxes standing closely together and touching, or two stacked cans of similar proportions can be detected as one object instead of multiple. An example of this problem is shown in Fig. 5.4.

Figure 5.4: CYWS-3D had problems detecting single objects that are part of a cluster. Due to the class agnostic training process, the model generalises too much and is not irritated by the unusual geometric features of the cluster. The distance makes it even more difficult to predict multiple objects, as the cluster appears not bigger than other household objects.

5.1.3 Annotation Error

The ground truth in Langer [8] is annotated directly onto the local surface reconstruction by *Elastic Fusion* [27]. The solution of blending odometry and camera poses works better than just using the camera but is still inaccurate. This results in the ground truth bounding boxes, which are drawn onto the selected camera frames, being misaligned, as described in Section 3.1.1. The results are lower IoU values, leading to lower mean-average-precision. Increasing the bounding box areas helped to some extend, but as long as the localisation accuracy and thus, the point cloud accuracy does not increase, re-annotation by hand would be necessary to correct this error.

5.1.4 Small Objects

The model was able to detect the marker in a few cases, but the confidence score never surpassed a threshold of 0.25 in both images. In Fig. 5.5, a false positive result was also detected with similar certainty. The same applies to the screwdrivers.

Figure 5.5: Small Objects are predicted with a confidence score of around 0.25, rarely higher. CYWS-3D does make errors around this threshold as seem in the left image, but the mean average precision is still higher due the the reduction in recall when using higher confidence thresholds.

Detecting small objects, defined as objects below a size of 64x64 pixels [35], remains an open challenge for further work and the evaluation showed a reduction of 35% in average precision and almost 50% in recall for small objects, underlining the challenges that the proposed new model has regarding the detection of small objects. Fig. 5.6 shows this discrepancy for a sub-sample of the *ObChange* dataset containing medium and small objects.

The performance on small objects and the results in the kitchen, where the average distance between the camera and the objects is much lower, suggest the for CYWS-3D to work better, the robot should move as close as possible to all surfaces it visits.

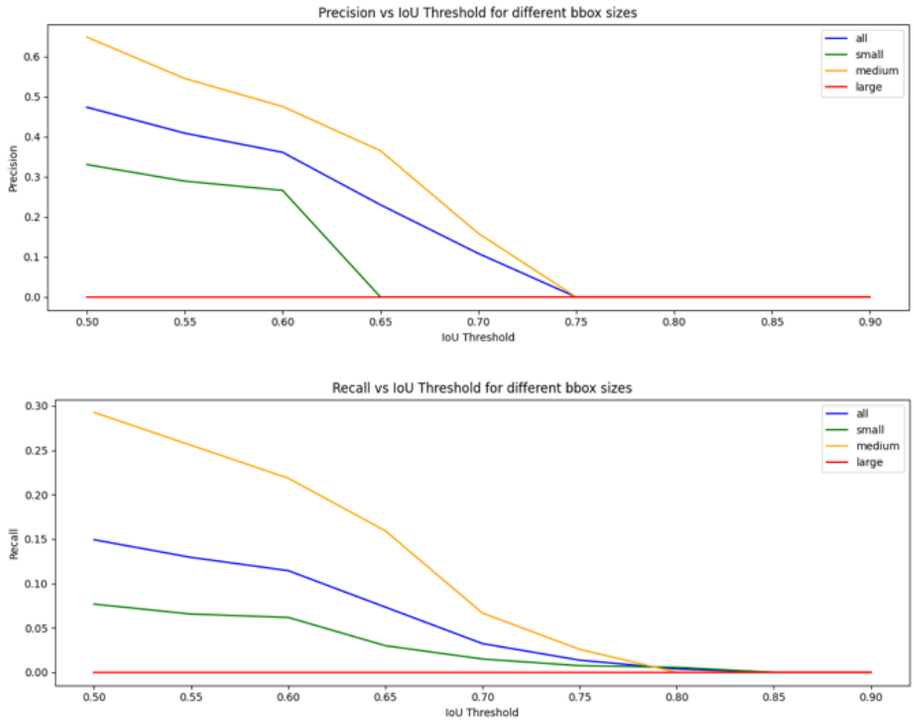


Figure 5.6: The results of a sub-sample of ObChange, comparing medium to small bounding boxes, show a 35% reduction in precision and a 50% reduction in recall compared to all detections in the sub-sample.

5.1.5 Original Depth Information

Another aspect needing further evaluation are the results achieved with using the depth information available from the camera stream, plotted in Fig. 5.7. The excepted trend would have been to see an increase in precision with rising confidence thresholds, as pictured in Fig. 4.3 for data points corresponding to using the depth estimation, but the average precision drops when using 0.4 as a confidence threshold, the effect increasing with bigger bounding box size.

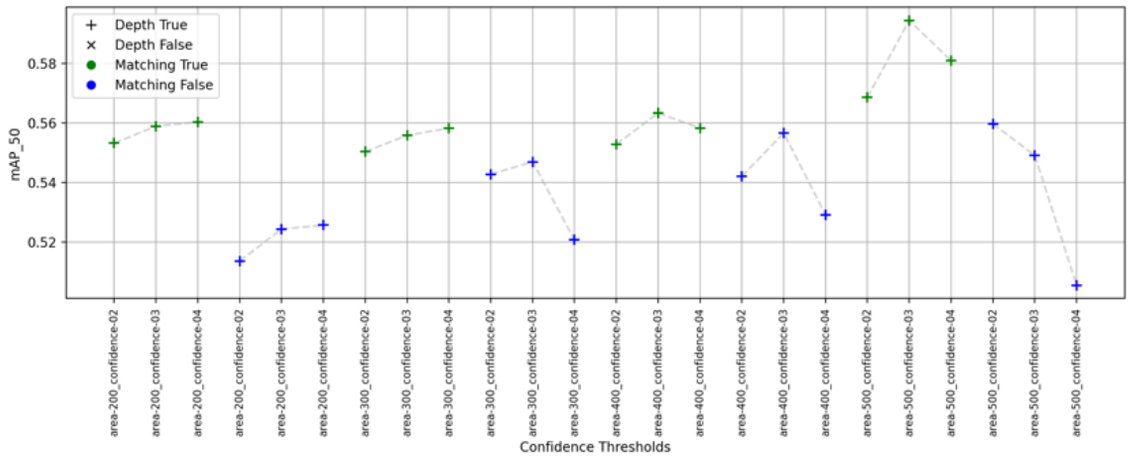


Figure 5.7: The drop in precision for high confidence thresholds remains an open challenge for investigation. It seems the matching-boxes algorithm does not influence the behaviour.

5.2 Discussions

The results indicate that CYWS-3D needs further training and evaluation to be used efficiently as an addition to the tidy-up-pipeline. Without further training, the new model could be used to evaluate all surface boundaries in order to detect objects that may have been missed in the reconstruction. Furthermore, it can make estimations about objects on curved surfaces, as CYWS-3D is not restricted to detecting objects on flat surfaces. These use cases still suffer from poor recall values, as CYWS-3D misses 3 out of 4 Objects on average. These results can partly be explained due to being trained on synthetic data, which was not not specifically targeted on indoor environments. One option for a following thesis would be to train the model on YCB-Objects in household scenes and re-evaluate on ObChange. The idea of using RGB-D data to skip the reconstruction process and increase overall detection rate by giving hints on changes that are missed by the existing pipeline is still a viable option that could be achieved with further training and examining the remaining challenges of CYWS-3D. The easy integration into the existing pipeline due to only needing to calculate positions that look at the same locations from different perspectives during the robots path through the room looks worth investigating through further work on this matter.

6 Conclusion

The aim of this work was to improve the indoor change detection system presented by Langer, Patten, and Vincze [1]. For this purpose, three models, each of which had the potential to solve one of the open problems of the existing system, were evaluated to select one model for further analysis and comparison. The motivation for this work arose from the open problems addressed by the authors of Langer, Patten, and Vincze [1], in combination with promising work presented at the ICCV23 that could deal with the remaining challenges. Concrete problems that still needed to be solved were hidden and small objects, the quality of the reconstruction and the problems caused by the introduced surface concept.

The model The Change You Want To See - 3D by Sachdeva and Zisserman [2] was chosen for further analysis, as it can work directly on the camera data and therefore had the potential to overcome the problems regarding the detailed reconstruction of the scenes. In particular, it was hoped that objects that were not fully reconstructed could be identified. The model was evaluated to test the various parameter combinations in order to evaluate the best configuration for interior scenes. The process included defining the minimum size of the recognised objects, finding the optimal minimum confidence threshold and examining the use of the original depth data.

Two use cases emerged, despite a number of shortcomings. Firstly, CYWS-3D does not suffer from the weaknesses associated with the surface concept and therefore it can be used to make a suggestion for missed objects on curved or sloping surfaces. Secondly, also related to the concept mentioned above, the model does not have any restrictions on objects being inside a convex hull of the extracted planes. It can therefore be used to check the edges of the limited search surfaces of the tidy-up-pipeline to identify partially reconstructed objects.

The results were evaluated on the basis of average precision and recall. It was possible to select parameters with which an increase in performance could be achieved. Medium-sized objects in particular are recognised significantly better than small objects, achieving an average precision of 65% and a recall of 30%, compared to 33% and 15% for small objects. This stands out in scenes where the robot moves close to the objects, and CYWS-3D outperformed the tidy-up-pipeline in the kitchen scene of *ObChange* by one percentage point in precision. The direct use of camera data proves to be a solution for reducing the number of missed objects in the two scenarios mentioned above, as long as the objects being searched for can be classified as medium-sized and are therefore recognised well enough.

Despite these results, the weaknesses of CYWS-3D dominate and further improvements to the model itself are necessary. It continues to have difficulties recognising small objects, which are only recognised with a very low confidence value around 25%. The low average recall of 26% indicates that the model cannot generalise well enough to the YCB objects in indoor scenes and better training is required. Additionally , the integration of a better system to reconstruct the scenes remains an open challenge of the existing system, as this

was not part of the improvement area to which CYWS-3D is applied.

Future work on CYWS-3D can address the issues the model has in indoor scenarios. More specific training could lead to an increase in recall for indoor household objects and including real images to the existing synthetic data could further advance the performance. Furthermore, the algorithm that deletes bounding boxes without a corresponding box in the previous frame can be further developed, as it deletes too many correct boxes. A following thesis could tackle the task of retraining CYWS-3D for indoor usage and examine potential improvements to the algorithm. Regarding the existing pipeline, another thesis could focus on testing more advanced reconstruction methods such as Point-SLAM, as this thesis did focus on change detection improvements through the direct use of the camera information available.

In summary, it can be said that two aspects were found in which the model of Sachdeva and Zisserman [2] can provide an indication of unrecognised objects and thus improve the pipeline. However, low values of average precision and recall predominate, as well as poor recognition of small objects, so that the model does not mark a significant improvement of the existing pipeline without further work.

Bibliography

[1] E. Langer, T. Patten, and M. Vincze, „Where does it belong? autonomous object mapping in open-world settings,“ *Frontiers in Robotics and AI*, vol. 9, 2022, ISSN: 2296-9144. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2022.828732.

[2] R. Sachdeva and A. Zisserman, „The change you want to see (now in 3d),“ in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[3] M. Asada and O. von Stryk, „Scientific and technological challenges in robocup,“ *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, no. 1, pp. 441–471, 2020. eprint: https://doi.org/10.1146/annurev-control-100719-064806. [Online]. Available: https://doi.org/10.1146/annurev-control-100719-064806.

[4] J. Ni, K. Shen, Y. Chen, and S. X. Yang, „An improved ssd-like deep network-based object detection method for indoor scenes,“ *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–15, 2023.

[5] Y. Ye, X. Ma, X. Zhou, G. Bao, W. Wan, and S. Cai, „Dynamic and real-time object detection based on deep learning for home service robots,“ *Sensors*, vol. 23, no. 23, 2023, ISSN: 1424-8220. [Online]. Available: https://www.mdpi.com/1424-8220/23/23/9482.

[6] E. Sandström, Y. Li, L. Van Gool, and M. R. Oswald, „Point-slam: Dense neural point cloud-based slam,“ in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[7] A.-Q. Cao and R. de Charette, „Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields,“ in *ICCV*, 2023.

[8] E. Langer, *Obchange dataset*, 2022.

[9] T. Kang, D. Song, J.-B. Yi, *et al.*, „Team tidyboy at the wrs 2020: A modular software framework for home service robots,“ *Advanced Robotics*, vol. 36, no. 17-18, pp. 836–849, 2022. eprint: https://doi.org/10.1080/01691864.2022.2111229. [Online]. Available: https://doi.org/10.1080/01691864.2022.2111229.

[10] H. Yan, G. Shen, R. Zetik, E. Malz, S. Jovanoska, and R. S. Thomä, „Stationary symmetric object detection in unknown indoor environments,“ in *2011 Loughborough Antennas & Propagation Conference*, 2011, pp. 1–5.

[11] M. Choi, G.-Y. Kim, and H.-I. Choi, „Robust object detection from indoor environmental factors,“ *Journal of the Korea Society of Computer and Information*, vol. 15, pp. 41–46, 2010.

[12] D. Park and Y. Park, „Identifying reflected images from object detector in indoor environment utilizing depth information," *IEEE Robotics and Automation Letters*, vol. 6, pp. 635–642, 2021.

[13] C. Chen, Z. Liang, H. Liu, and X. Liu, „Spatial and semantic information enhancement for indoor 3d object detection," *Int. Arab J. Inf. Technol.*, vol. 20, pp. 831–839, 2023.

[14] X. Wu, D. Sahoo, and S. Hoi, „Recent advances in deep learning for object detection," *ArXiv*, vol. abs/1908.03673, 2019.

[15] G. Georgakis, A. Mousavian, A. Berg, and J. Kosecka, „Synthesizing training data for object detection in indoor scenes," *ArXiv*, vol. abs/1702.07836, 2017.

[16] S. M. Irfan, M. M. Islam, M. S. Mia, M. Islam, S. Islam, and T. Islam, „Advancements in object detection and tracking algorithms: An overview of recent progress," *EPRA International Journal of Research & Development (IJRD)*, 2023.

[17] X. Wei, Y. Ren, and M. Huang, „Research on computer vision in object detection," vol. 12625, pp. 1262537 –1262537–5, 2023.

[18] R. Qin, J. Tian, and P. Reinartz, „3d change detection – approaches and applications," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 122, pp. 41–56, 2016.

[19] H. Hamledari, B. McCabe, and S. Davari, „Automated computer vision-based detection of components of under-construction indoor partitions," *Automation in Construction*, vol. 74, pp. 78–94, 2017.

[20] M. J. Gómez, F. García, D. Martín, A. de la Escalera, and J. M. Armingol, „Intelligent surveillance of indoor environments based on computer vision and 3d point cloud fusion," *Expert Systems with Applications*, vol. 42, no. 21, pp. 8156–8171, 2015, ISSN: 0957-4174. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0957417415004285`.

[21] M. Tancik, E. Weber, E. Ng, *et al.*, „Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH 2023 Conference Proceedings*, ser. SIGGRAPH '23, New York, NY, USA: Association for Computing Machinery, 2023, ISBN: 9798400701597. [Online]. Available: `https://doi.org/10.1145/3588432.3591516`.

[22] Y. Wei, S. Liu, J. Zhou, and J. Lu, „Depth-guided optimization of neural radiance fields for indoor multi-view stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 835–10 849, 2023.

[23] H. Li, X. Gu, W. Yuan, L. Yang, Z. Dong, and P. Tan, „Dense rgb slam with neural implicit maps," *ArXiv*, vol. abs/2301.08930, 2023.

[24] Z. Zhu, S. Peng, V. Larsson, *et al.*, „Nicer-slam: Neural implicit scene encoding for rgb slam," *ArXiv*, vol. abs/2302.03594, 2023.

[25] M. M. Johari, C. Carta, and F. Fleuret, „Eslam: Efficient dense slam system based on hybrid representation of signed distance fields," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17 408–17 419, 2022.

[26] R. Padilla, S. L. Netto, and E. A. B. da Silva, „A survey on performance metrics for object-detection algorithms," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2020, pp. 237–242.

[27] T. Whelan, S. Leutenegger, R. Salas-Moreno, B. Glocker, and A. Davison, „Elasticfusion: Dense slam without a pose graph," Robotics: Science and Systems, 2015.

[28] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, „The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 International Conference on Advanced Robotics (ICAR)*, 2015, pp. 510–517.

[29] J. Straub, T. Whelan, L. Ma, *et al.*, „The replica dataset: A digital replica of indoor spaces," *CoRR*, vol. abs/1906.05797, 2019. arXiv: 1906.05797. [Online]. Available: http://arxiv.org/abs/1906.05797.

[30] J. Sturm, W. Burgard, and D. Cremers, „Evaluating egomotion and structure-from-motion approaches using the tum rgb-d benchmark," in *Proc. of the Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RJS International Conference on Intelligent Robot Systems (IROS)*, vol. 13, 2012.

[31] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai, *Scannet++: A high-fidelity dataset of 3d indoor scenes*, 2023. arXiv: 2308.11417 [cs.CV].

[32] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, „Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface re-integration," *ACM Transactions on Graphics*, vol. 36, p. 1, Jul. 2017.

[33] J. Behley, M. Garbade, A. Milioto, *et al.*, „A dataset for semantic segmentation of point cloud sequences," *CoRR*, vol. abs/1904.01416, 2019. arXiv: 1904.01416. [Online]. Available: http://arxiv.org/abs/1904.01416.

[34] S. Wenkel, K. Alhazmi, T. Liiv, S. Alrshoud, and M. Simon, „Confidence score: The forgotten dimension of object detection performance evaluation," *Sensors*, vol. 21, no. 13, 2021, ISSN: 1424-8220. [Online]. Available: https://www.mdpi.com/1424-8220/21/13/4350.

[35] C. Peng, M. Zhu, H. Ren, and M. Emam, „Small object detection method based on weighted feature fusion and csma attention module," *Electronics*, vol. 11, no. 16, 2022, ISSN: 2079-9292. [Online]. Available: https://www.mdpi.com/2079-9292/11/16/2546.

Erklärung

Hiermit erkläre ich, dass die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt wurde. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder in ähnlicher Form in anderen Prüfungsverfahren vorgelegt.

Wien, im Juni 2024

Florian Pfleiderer