

**humantech**

## **D4.3 – Wearable-user localization algorithm**



This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement n° 101058236. This document reflects only the author's view, and the EU Commission is not responsible for any use that may be made of the information it contains.



#### D4.3 – Wearable-user localization algorithm

<b>Project Title</b>	Human-Centred Technologies for a Safer and Greener European Construction Industry.
<b>Project Acronym</b>	HumanTech
<b>Grant Agreement No</b>	101058236
<b>Instrument</b>	Research & Innovation Action
<b>Topic</b>	HORIZON-CL4-2021-TWIN-TRANSITION-01-12
<b>Start Date of Project</b>	June 1, 2022
<b>Duration of Project</b>	36 months

<b>Name of the Deliverable</b>	Wearable-user localization algorithm
<b>Number of the Deliverable</b>	D4.3 (D18)
<b>Related WP Number and Name</b>	WP4 Wearable Technologies for Construction
<b>Related Task Number and Name</b>	T4.3 Wearable camera digital twin localization
<b>Deliverable Dissemination Level</b>	PU
<b>Deliverable Due Date</b>	31.08.2024
<b>Deliverable Submission Date</b>	30.08.2024
<b>Task Leader/Main Author</b>	Suresh Guttikonda (DFKI)
<b>Contributing Partners</b>	Markus Miezal (SCT), Yaxu Xie (DFKI), Bruno Mirbach (DFKI), Mahdi Chamseddine (DFKI)
<b>Reviewer(s)</b>	Bruno Mirbach (DFKI), Jason Rambach (DFKI)

**Keywords**

Localization, BIM model, scene graph, visual-inertial tracking

## Revisions

Version	Submission date	Comments	Author
V1.0	30.08.2024	Submitted Version	Suresh Guttikonda et al.
...			
...			
...			

## Disclaimer

This document is provided with no warranties whatsoever, including any warranty of merchantability, non-infringement, fitness for any particular purpose, or any other warranty with respect to any information, result, proposal, specification, or sample contained or referred to herein. Any liability, including liability for infringement of any proprietary rights, regarding the use of this document or any information contained herein is disclaimed. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted by or in connection with this document. This document is subject to change without notice. HumanTech has been financed with support from the European Commission. This document reflects only the view of the author(s) and the European



## Acronyms and definitions

Acronym	Meaning
BCF	BIM Collaboration Format
BIM	Building Information Model
BIMxD	Extended Dynamic BIM
BSN	Body Sensor Network
IMU	Inertial Measurement Unit
LiDAR	Light Detection and Ranging
SLAM	Simultaneous Localization and Mapping
XR	Extended Reality

### **Abstract**

This document summarizes the research results accomplished within Task 4.3. The target of this task has been to develop and assess methods to localize a camera in the digital twin of a construction site from a data stream consisting of video images and IMU data of a body worn sensor system, as the one described in D4.1.

Numerous approaches to extract relevant 3D information from a visual body sensor network and their alignment to a BIM model have been considered within Task 4.3. This deliverable reports the results to two main directions of research. One direction combines the detection of fiducial markers with a visual-inertial tracking. The second research direction focused on the localization without the support of markers

To evaluate the developed methods a laboratory space at DFKI has been equipped with markers provided by ZHAW and a ground truth optical sensor system. The lab space has been scanned and a BIM model been generated with the support of RPTU. A data set has been recorded with the body sensor network with integrated camera described in D4.1 as well as with an Aria Camera system with integrated visual-inertial SLAM.

SCT implemented marker supported camera and body pose tracking algorithm on mobile processing platform and evaluated the accuracy in terms of KPI 4.3. The full body pose tracking server moreover as basis for the Intelligent exoskeleton with intention prediction prototype reported in D4.2.

Regarding the second, marker-less, localization, DFKI has followed the direction of recently published existing works [5] on scene graph alignment. DFKI has developed a method to generate from a BIM model a global scene graph using techniques from the Scan2BIM pipeline developed in WP3. In parallel, a novel alignment method for pairs of scene graphs that works also for incomplete and partial overlap has been developed and meanwhile been published. The most challenging module in the pipeline of a scene graph-based localization turned out to be the generation of a local scene graph from the video stream of the body camera. Identified issues and first results conclude this deliverable.



## The HumanTech project

The European construction industry faces three major challenges: increase the safety and wellbeing of its workforce, improve its productivity, and become greener, making efficient use of resources.

To address these challenges, HumanTech proposes to develop **human-centred cutting-edge technologies** such as wearables for workers' safety and support and robots that can harmoniously coexist with human workers while contributing to the ecological transition of the sector.

**HumanTech aims to achieve major advances in cutting-edge technologies that will enable a safe, rewarding, and digital work environment for a new generation of highly skilled construction workers and engineers.**

These advances will include:

- **Robotic devices equipped with vision and intelligence** that allow them to navigate autonomously and safely in highly unstructured environments, collaborate with humans and dynamically update a semantic digital twin of the construction site in which they are.
- **Smart, unobtrusive workers protection and support equipment.** From exoskeletons activated by body sensors for posture and strain to wearable cameras and XR glasses that provide real-time workers' location and guidance for them to perform their tasks efficiently and accurately.
- An entirely new breed of **Dynamic Semantic Digital Twins (DSDTs) of construction sites** that simulate in detail the current state of a construction site at the geometric and semantic level, based on an extended Building Information Modelling (BIM) formulation that contains all relevant structural and semantic dimensions (BIMxD). BIMxDs will act as a common reference for all human workers, engineers, and autonomous machines.

The **HumanTech consortium** is formed by 22 organisations — leading research institutes and universities, innovative hi-tech SMEs, and large enterprises, construction groups and a construction SME representative — from 10 countries, bringing expertise in 11 different disciplines. The consortium is led by the German Research Center for Artificial Intelligence's Augmented Vision department.



# Contents

1	Introduction.....	8
2	Lab-setup at DFKI for localization testing.....	10
2.1	Final marker setup and integration.....	11
2.2	Scanning and BIM model.....	11
3	Localization based on marker detection.....	13
3.1	Evaluation setup and calibration.....	13
3.2	Evaluation.....	14
3.3	Conclusion.....	17
4	Localization based on scene graphs.....	18
4.1	Motivation and related work.....	18
4.2	Scene graph representation of 3D-scenes.....	19
4.2.1	Node attributes.....	19
4.2.2	Edge relationships.....	19
4.3	Point cloud and scene graph alignment.....	21
4.3.1	Problem definition.....	21
4.3.2	Scene graph and point encoding.....	22
4.3.3	Scene graph alignment and point cloud registration.....	23
4.3.4	Results on open-source datasets.....	23
4.4	Global scene graph generation of a BIM model.....	25
4.4.1	3D global scene graph from BIM-model of DFKI lab.....	26
4.5	Local scene graph generation.....	28
4.5.1	Hardware specifications and recording profile.....	30
4.5.2	Aria machine perception services.....	31
4.5.3	Semi-dense 3D semantic segmentation.....	32
5	Conclusion.....	34
6	References.....	36



# 1 Introduction

This document summarizes the research results accomplished within task T4.3. The target of the research is to develop and assess methods to localize a camera in the digital twin of a construction site from a data stream consisting of video images and IMU data of a body worn sensor system, as the one described in deliverable D4.1.

Localization means thereby detecting and tracking the 3D position and the 3D pose of a camera with respect to the coordinate system of a BIM model of a building. This localization functionality has several applications. The localization information can help a worker localizing himself on a large construction site with respect to the digital twin, to automatically assign the localization information to captured visual data and issues reported via BCF to the BIMxD platform. A precise 6D localization is also required for rendering of BIM content in XR glass visualization, as those developed in task T4.4.

There are already established and robust visual-inertial SLAM algorithms which fuse inertial information with sparse visual feature tracking thus providing the localization of the camera in a sparse local map (see e.g. [1] and references therein). However, to localize the camera unambiguously with respect to the digital twin of a building requires the alignment of this local map to the BIM reference system using, e.g. some known landmarks in the building. A localization only based on a sparse local map may be unstable due to sparsity of data and is moreover prone to ambiguities due to repetitive structures such as corners and walls.

Most existing works localize the camera using a pre-collected image database or within a large-scale pre-built 3D model [2]. However, these representations of the environment are costly in terms of storage and maintenance. In contrast, indoor environments including most commercial real estate such as warehouses, offices and apartments already possess a floorplan. F3loc [3] propose to localize the camera with respect to a given floorplan. Also, in this approach repetitive structures cause ambiguity in the localization, which is eliminated to a certain extent by using image sequences.

Alternatively, representations like Situational Graphs (S-Graphs) [3] [4] have been shown to be efficient in representing the scene in hierarchical structure with structural and topological constraints between the elements in the scene. These graphs enable the robots to understand and navigate using high-level abstractions (such as chairs, tables, and walls) and the inter-connections between them (such as set of walls from a room or corridor).



Kimera [1] and Hydra [4] approaches studied real-time performance and adapting hierarchical representations for robot spatial perception systems using RGB-D camera inputs. Kimera [1] provides a novel metric-semantic hierarchical representation of the environment; however, the generation of the scene graph is not real-time. Hydra [4], built-upon Kimera framework, overcomes the drawback. Mono-Hydra [5] tries to achieve real-time spatial perception with a monocular camera and an IMU setup. Additionally, it leverages a set of deep learning neural networks to predict the depth [6], [7] and semantics [8]. Although, this approach shows a potential for generating scene graph using monocular depth prediction and hierarchical metric-semantic mesh generation, the approaches fail to perform hierarchical loop closure optimization, accurate room detection, with temporal consistency.

The research of T4.3 has followed two main directions. One direction is the development of a real-time localization method which combines a visual-inertial tracking of the visual BSN developed in T4.1 (see D4.1) with the detection of markers. The second research direction focused on the localization without support of markers using scene graph alignment.

To evaluate the developed methods a complex hallway at DFKI has been equipped with standardized markers provided by ZHAW and a ground truth optical sensor system. The lab space has been scanned and a BIM model has been generated with the support of RPTU. This test setup is presented in section 2.

Section 3 presents the marker supported visual-inertial camera and body pose tracking algorithm SCT has developed and deployed on a mobile body-worn processing platform. A data set has been recorded with the BSN with integrated camera described in D4.1 and the accuracy of both the camera and body pose been evaluated on these data. The full body pose tracking serves moreover as a basis for the Intelligent exoskeleton with intention prediction prototype reported in D4.2.

Regarding the second, marker-less, localization technique, DFKI has followed the direction of recently published existing works [5] on scene graph alignment. DFKI has developed a concept to build a global scene graph from a 3D point cloud of a BIM model and a local scene graph from the RGB-IMU streams of the visual BSN, followed by matching for scene alignment and localization. Section 4 presents the investigated approaches and results.

## 2 Lab-setup at DFKI for localization testing

A sufficiently complex hallway in DFKI was chosen for localization testing. It features a long corridor with a meeting area in the center. For BIM-model generation and marker based tracking, sheets with multiple markers with known dimensions have been placed on the walls. Each sheet contained a Chilitag marker [9], two reference points for surveyors and a marker for a laser range finder for BIM-model generation. This standardized marker system has been developed in a previous project under the lead of ZHAW and meanwhile been approved as European pre-standard. For details see deliverable D3.1.

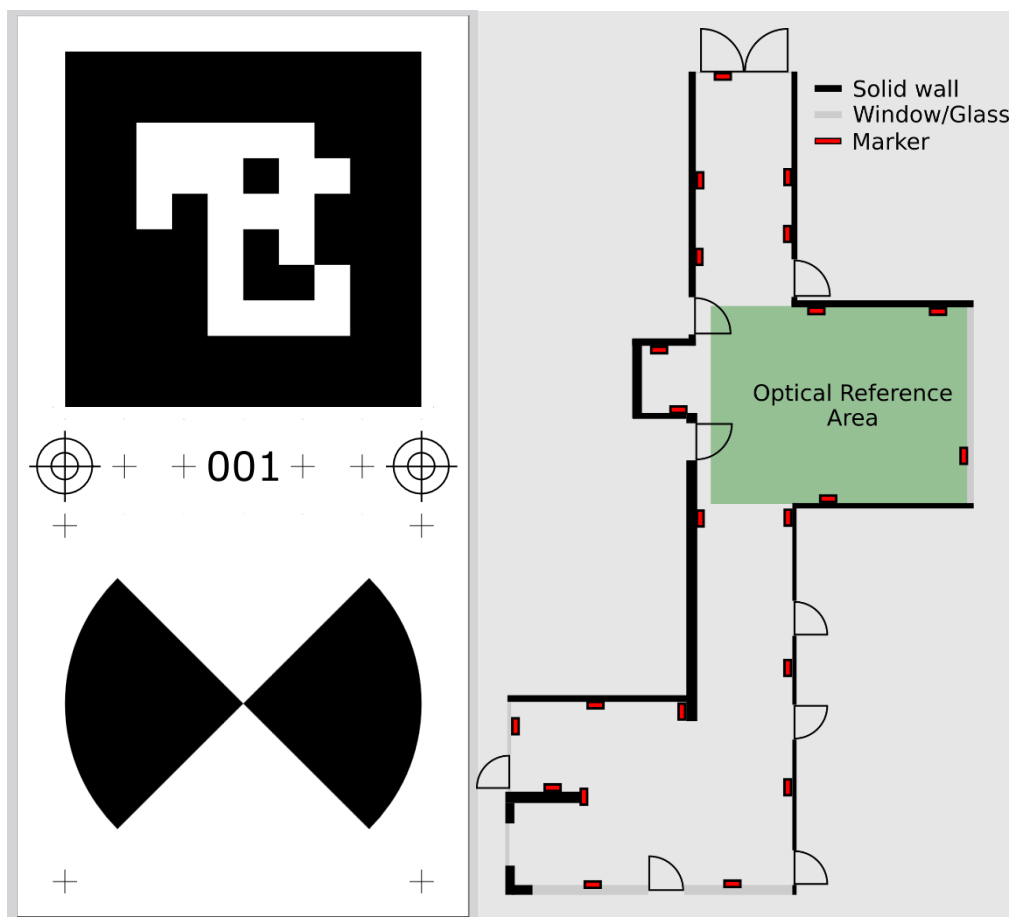


Figure 1: Left: marker used in experiment; Right: floorplan of experimental area

In addition to the markers, an optical reference system has been setup. To relate the optical reference system to the markers, reflective markers have been placed on the surveyor points of each marker sheet. The z-axis was assumed to be gravity. During the calibration process, the optical reference systems vertical axis was aligned with gravity using spirit levels.

### 2.1 Final marker setup and integration

For the final marker layout for HumanTech, Apriltags [10] and a QR code are placed on the marker sheets as well. ZHAW measures each marker and creates a marker world which can be retrieved at a certain endpoint in JSON format from the Catenda-Hub. For April- or Chilitags, the four marker corners in BIM model coordinates are provided. If, however, neither April- nor Chilitags are supported, the QR code holds the URL to an endpoint, where the location of the surveyor points can be retrieved.

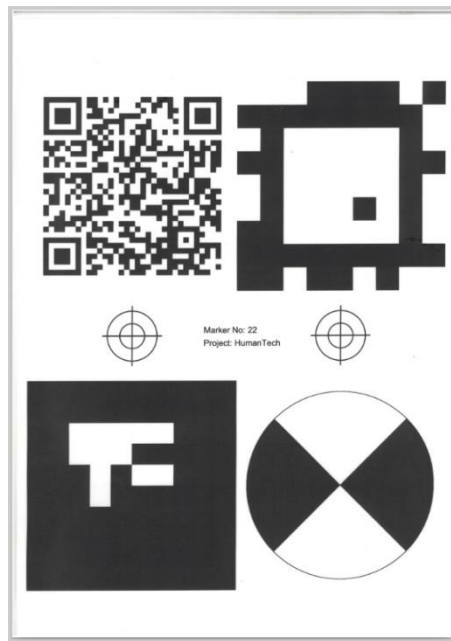
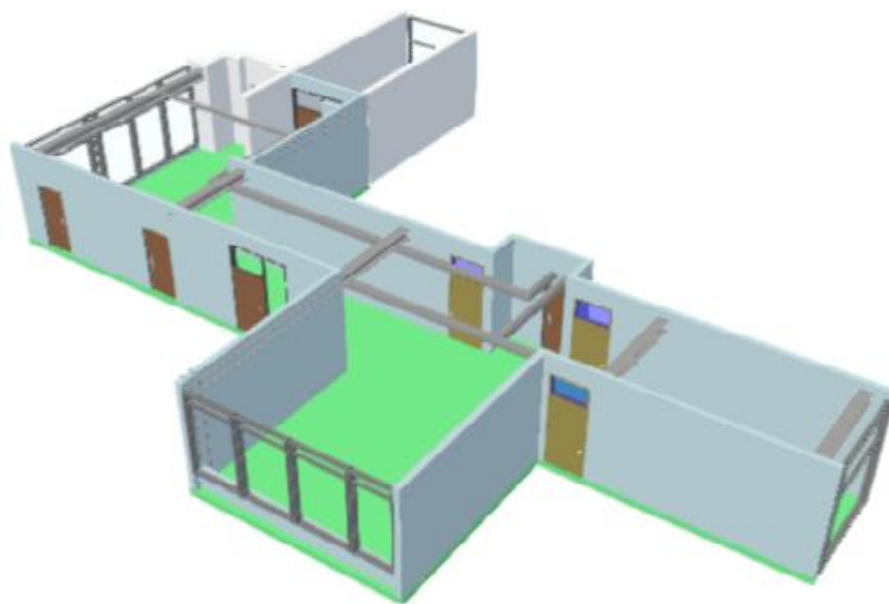


Figure 2: Final HumanTech marker layout with additional Apriltag and QR-code.

### 2.2 Scanning and BIM model

After being equipped with the markers, the hallway has been scanned with the support of RPTU with a Leica BLK370 scanner. RPTU has then also generated from the point cloud a BIM model visualized below in **Error! Reference source not found.** The 3D position of the markers with respect to the BIM model has been determined from the position of laser range finder tags in the scans.



*Figure 3: BIM model of generated from the scan of the DFKI hallway*

### **3 Localization based on marker detection**

The localization cannot rely solely on marker detection, since the number of markers and effort to survey them is too large given the scale of the intended BIM models. Visual SLAM systems, however, can solve this problem but are computationally expensive. One solution is to combine the inertial body tracking with the visual marker tracking. Since the inertial tracking system is capable of estimating translation it can fill the gaps between marker detections, thus reducing the number of required markers.

Typically, a marker detection algorithm returns certain features on the detected marker. In case of *Chilitags* and *Apriltags*, these are the 2D point, i.e. the pixel coordinates of each corner. Together with the associated 3D points provided by ZHAW, the resulting 2D/3D correspondences can be directly utilized by the tracking, given that the camera position on the workers body is known.

Besides allowing localization with respect to the BIM model, it mitigates weaknesses of the inertial tracking system. Inertial sensors usually contain magnetometers which can be easily disturbed by any magnetic field or ferro-magnetic materials, which influence the measurement. These local disturbances result in different magnetic references among the sensors and hence in errors. Consequently, the magnetometer measurements are omitted which in turn leads to an inevitable global yaw drift. The information from the markers ensures consistent yaw measurements within the BIM model.

#### **3.1 Evaluation setup and calibration**

The inertial system (an XSens Awinda) has been set up to run at 60Hz and trigger the Ricoh camera at 30Hz and the optical reference system at 60Hz. One inertial sensor has been placed on the camera forming the visual inertial unit. Although all involved devices were synchronized, the recording was manually started, and some devices do not start on the first trigger signal directly, thus delaying the first measurement. This issue requires to calibrate an integer frame offset between all involved system based on the magnitude of the rotational velocity of the CamIMU device.

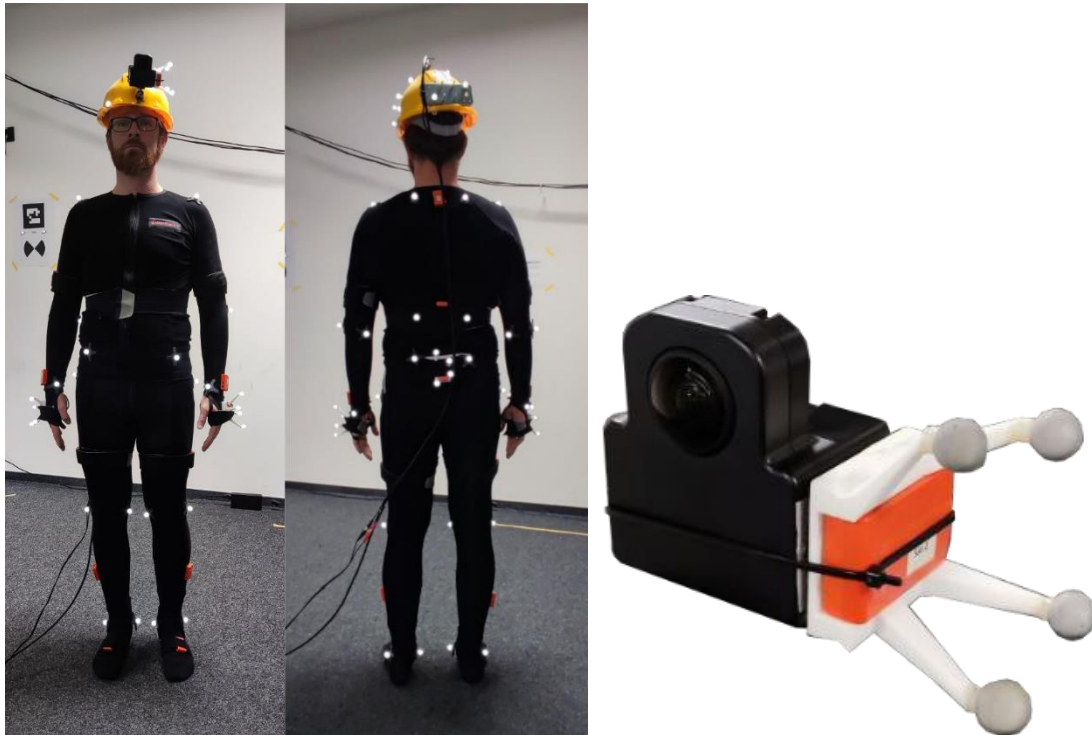


Figure 4: Left: Test subject with marker protocol and CamIMU unit on the helmet; Right: CamIMU in detail.

A subject wears the sensor network, the CamIMU unit on the helmet and was equipped with reflective markers on certain bony landmarks. Additionally, marker clusters which allow to retrieve a full 6D pose, have been attached to the CamIMU and the HandIMUs. These contraptions required a handeye calibration as well, so that the IMU orientation is directly reflected by the optical reference system.

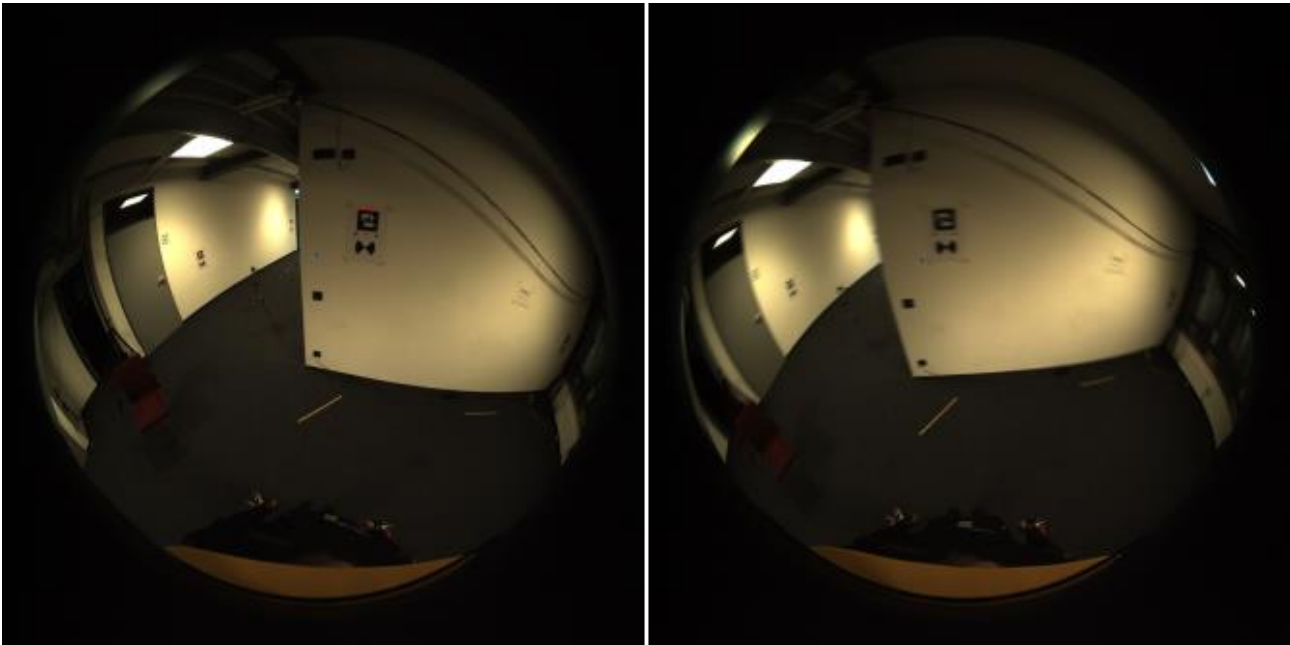
The optical system was calibrated so the z-axis is aligned with gravity with a spirit level. This, together with the axes exposed by the reflective markers on the surveyor points (cmp. Figure 1), enable to find the marker positions within the optical reference system.

### 3.2 Evaluation

For evaluation a generic walking sequence of one minute length is used. Traversing a BIM model is the most challenging sequence, since markers will be out of sight and localization depends on the translation estimation quality of the inertial system over a certain time interval.

The focus for evaluation lies on the pose error of the CamIMU device with respect to the Optitrack system.

During the selected sequence, although the usable area is comparably small with over six comparably large markers (15cm edge length) distributed on all four walls, only 41% of the time a marker is detected. To a large extent this is due to motion blur, which prevented the computer vision operations to detect proper rectangles. Please see Figure 5 below for an example.



*Figure 5: Example images for marker detection. Left: proper marker detection (even small marker to the left is detected). Right: Motion blur prevents detection. Notice the origin is still drawn onto the image.*

Figure 6 below show the error plot of the sequence both for marker detection and the tracking approach. The marker detection has gaps as mentioned before and if the marker is detected, position and rotation errors frequently exceed 1m and 20° respectively. This is due to the world scale. The position of the camera is calculated via the 2D/3D correspondences of the 4 corners of the markers. They only cover a small area of the image and even though subpixel precision is targeted, errors in the 2D points are expected. Since the origin is comparably far away from the camera position (11m on the direct line) small errors in the 2D points have a large impact on the overall position. That is especially the case, when only a single marker is detected. To a large extent the error may also result from the fact that due to internal issues the markers had to be removed and replaced after they had been surveyed. This manual process has a large potential of

inaccuracy. Thus, the system is assumed to perform better in a properly calibrated environment.

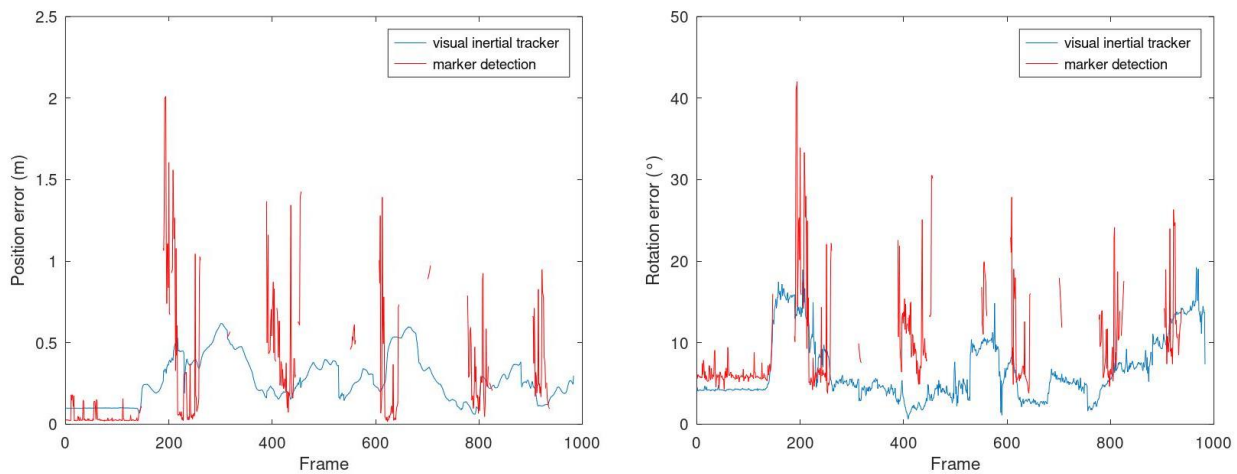


Figure 6: Position and Rotation error for the walking sequence. The marker detections have gaps and are exposed to larger errors than the fusion approach.

Especially position has sometimes larger changes in error which come with a sliding, whenever a marker is detected. When body measurements are imprecise and the orientation estimate is only slightly off, the quality of the translation estimate declines between marker detections. Once a marker comes into view again, the tracker tends to jump towards this new information. The tracking results also yield KPI values. First KPI 4.02 which measures the accuracy of the human pose tracking. For this, the center point between the optical markers for each joint is calculated and the mean absolute difference to the tracked joint center is taken. The average over the whole sequence and all joints is sought performance indicator. The most recent value is 22cm.

Also, KPI 4.03 is measured, which tracks the availability of the pose tracking in ms. Since the inertial body tracking is driven by the frequency of the IMU data, the latency is supposed to be below 20ms.

Another KPI 4.04 is the precision of the camera pose in terms of position and orientation. This KPI is met, if the percentage of frames, where both, translation **and** rotation are below the threshold is below 30% and later below 10%. Currently, this is only the case in 55% of all frames. Note that translation is majorly contributing to the low percentage. In over 90% of all frames the goal toward rotation is met.

All KPIs are also presented in Table 1 below:

Table 1: KPIs calculated with the visual BSN

KPI n°	Name	KPI Index/Method	Target M36	Actual Value	Notes
K4.02	Accuracy of 3D human pose tracking	Mean per Joint relative Position Error (MPJRPE)	< 10 cm	22 cm	Calculated over 8 joints
K4.03	Real time usage of pose tracking	Latency of pose tracking smaller than [ms]	< 40 ms (>25fps)	< 20 ms	Value based on pose tracking with BSN only
K4.04	Accuracy of wearable camera pose (w.r.t BIM/3D Scan)	Localization accuracy (rotation, translation) and Tracking failure rate	$\epsilon = (15 \text{ cm}, 10^\circ)$  $p < 10\%$	$\epsilon = (15 \text{ cm}, 10^\circ)$  $p = 55\%$	Based on marker detection.  Computed error comprises calibration inaccuracies between reference system and camera

### 3.3 Conclusion

The visual inertial body sensor network can track the workers position if marker information is not present and can thus be implemented as a backup solution for localization within the BIM model. In this experiment the camera was hardware triggered by the IMU device which is not feasible for the pilots if IMU sensors are used that do not provide trigger outputs. The next target is therefore a system that handle a camera that runs asynchronous and time-displaced to the IMU network.

## 4 Localization based on scene graphs

### 4.1 Motivation and related work

Recent approaches [11] use a novel framework to localize robots leveraging not only geometry but also high-level hierarchical information from architectural plans. Where, the BIM information is modelled in the form of a graph, denoted as Architectural Graph (A-Graph) - provides a global scene representation, and the online Situational Graph (S-Graphs) [12] [13] is estimated using 3D LiDAR measurements – provides a local scene representation, that the robot builds as it navigated the environment. To localize a robot within the environment, refer Figure 7, a graph-matching algorithm is utilized by exploiting the hierarchical information from both graphs to provide the best match candidates finally resulting in informed (iS-Graphs) that fuses the information of both.

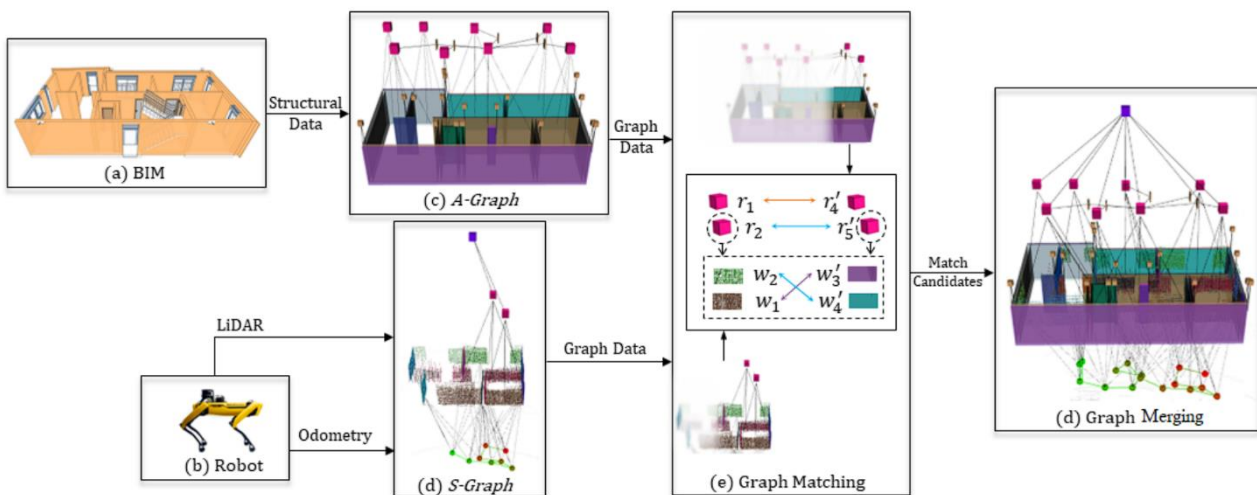


Figure 7: (a) An offline generated Architectural Graph (A-Graph) from a BIM model is matched and aligned to (b) online estimated Situational Graph (S-Graph). Which is then utilized by the robot to be localized w.r.t BIM [11].

However, for our use case, the Body Sensor Network (BSN) is equipped with 360-degree RGB device which is capable of recording RGB+IMU streams only. As described in section 4.5, we investigated methods to generate online scene graphs from RGB image streams only. Additionally, the BIM models in HumanTech are also accompanied by coloured point clouds (+ RGB images). These rich informational clues provided by RGB + Normal 3D points can further improve the scene graph generation process, as described in section 4.4. Furthermore, our novel point cloud and scene graph alignment technique is described in section 4.3 which achieved state-of-the-art performance with open-source datasets experiments.

## 4.2 Scene graph representation of 3D-scenes

Scene understanding is a cornerstone in many computer vision applications requiring perception, interaction and manipulation. Semantic Scene Graphs (SSGs) go beyond recognising individual entities (objects and stuff) by reasoning about the relationships among them. They also proved to be a valuable representation for complex scene understanding tasks, such as scene manipulation, task planning, and image captioning.

As illustrated in Figure 9, a semantic scene graph  $\mathcal{G}$  is a set of tuples  $(V, \mathcal{E})$  between nodes  $V$  and edges  $\mathcal{E}$ . Nodes represent specific 3D object instances in a 3D scan, refer Figure 8, where each node is assigned to either a single object category  $\mathcal{C}$  [14], [15] or by a hierarchy of classes  $c = (c_1, \dots, c_d)$  where  $c \in \mathcal{C}^d$ , and  $d$  can vary [16]. Additionally, to these object categories each node has a set of attributes  $\mathcal{A}$  that describe the `visual` and `physical` appearance of the object instance. The edges in graphs define semantic relationships (predicates) between the nodes such as `standing on`, `hanging on`, `more comfortable than`, `same material`.

### 4.2.1 Node attributes

Attributes of nodes are semantic labels that describe object instances. This includes static and dynamic properties, as well as affordances.

**Static Properties** includes visual object features such as the color, size, shape or texture but also physical properties e.g. the (non-)rigidity. Geometric data and class labels are utilized to identify the relative size of the object in comparison with other objects of the same category.

**Dynamic Properties** are states such as `open / close`, `full / empty` or `on / off`. The state category is defined to be class specific and can change over time.

**Affordances** are interaction possibilities or object functionalities of nodes of a specific object class e.g. a `chair` is for `sitting`. However, these are conditions on their state attribute: only a `closed door` can be opened.

### 4.2.2 Edge relationships

The relationships can be classifiable into: a) spatial / proximity relationships, b) support relations, and c) comparative relationships.

**Support Relationships** indicate the supporting structures of a scene. An instance can have multiple supports; walls are by default supported by the floor.

**Proximity Relationships** describe the spatial relationships (e.g. next to, in front of) with respect to a reference view.

**Comparative Relationships** are derived from comparison of attributes, e.g. bigger than, darker than, cleaner than, and same shape as.

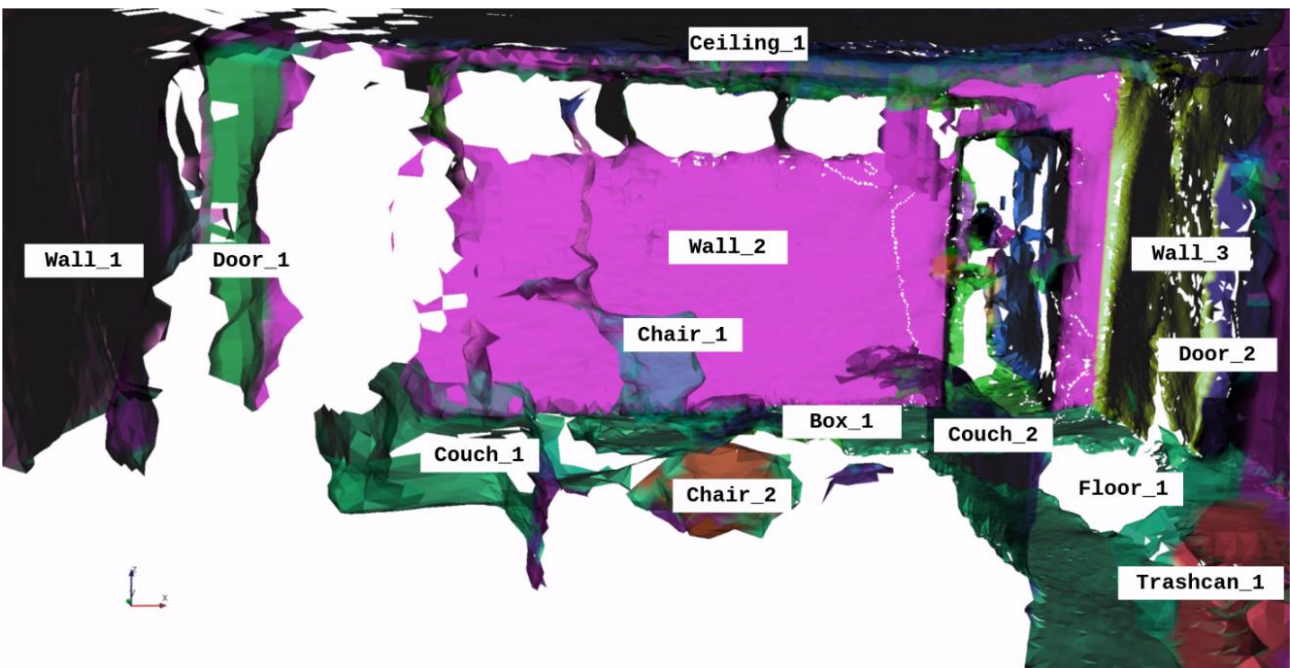


Figure 8: (top) Raw RGB image of DFKI Coffee room area scan recorded with Ricoh-3D device developed in T3.2 (see D3.2); (bottom) 3D pointcloud segmentation + instances extraction in T3.4. Note that the Ricoh-3D device provides spherical image, while the point cloud representation is perspective.

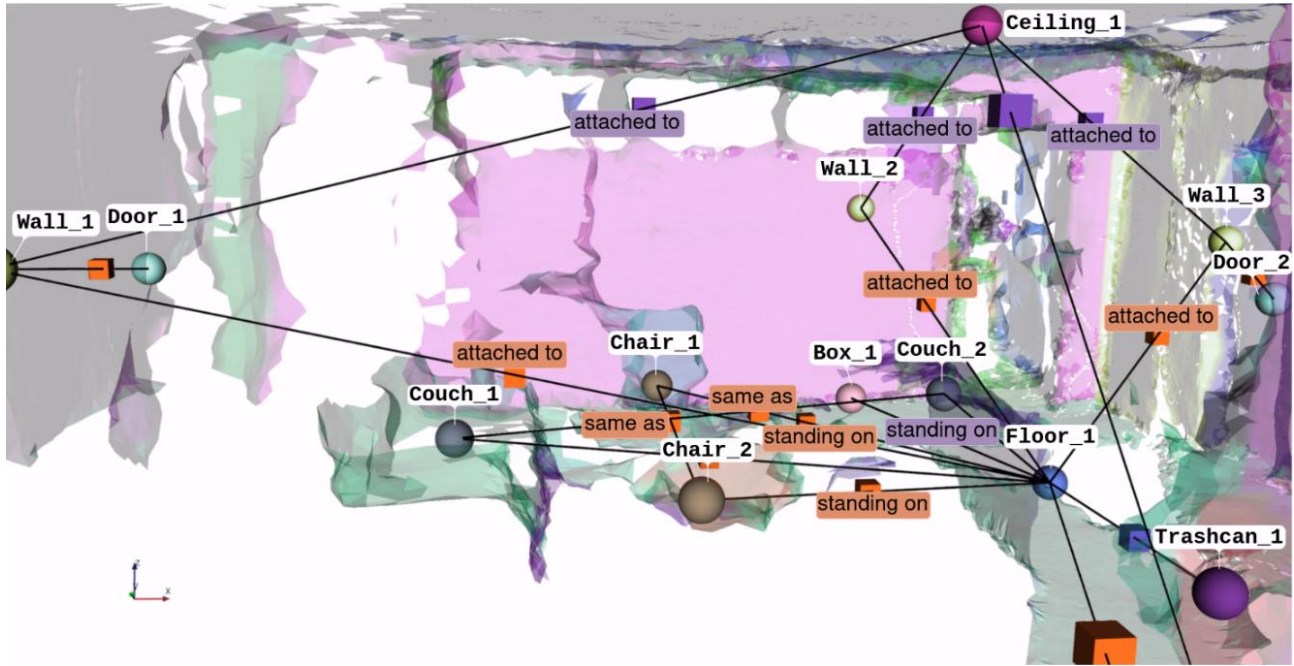


Figure 9: An example 3D Semantic Scene Graph manually annotated for DFKI Coffee room area 3D scan recorded with Ricoh-3D device developed in T3.2.

### 4.3 Point cloud and scene graph alignment

#### 4.3.1 Problem definition

For the accurate alignment of a global and local scene graph DFKI has developed a novel method, named SG-PGM [17], which is designed to work also for incomplete and partially overlapping scene graphs and incorporates also the corresponding point clouds for accurate alignment. This method is described in detail below.

Given a 3D scene graph with semantic node and edge attributes:  $\mathcal{G}' = (\mathcal{V}, \mathbf{A}, \mathbf{X}, \mathbf{E})$ , where it consists of a finite set of object nodes  $\mathcal{V} = \{v_1, \dots, v_M\}$ , an adjacency matrix  $\mathbf{A} \in \{0,1\}^{M \times M}$ , representing the edges between the nodes, a node feature matrix  $\mathbf{X} \in R^{M \times D}$  and an edge feature matrix  $\mathbf{E} \in R^{M \times M \times D}$ . Additionally, each 3D points of the corresponded point cloud  $\mathbf{P} = \{p_i \in R^3 \mid i = 1, \dots, N\}$  is assigned to one specific object node with point-to-object map  $O : \{1, \dots, N\} \rightarrow \{1, \dots, M\}$ . As illustrated in Figure 10, we formulated the graph matching problem as optimizing the following objective function, SG-PGM [17].

$$\arg \max_s f(\mathcal{S}; \mathcal{G}'_{src}, \mathcal{G}'_{ref})$$

in which  $S \in \{0,1\}^{M_{src} \times M_{ref}}$  is the binary permutation matrix that maps nodes between the source graph  $\mathcal{G}'_{src}$  and the reference graph  $\mathcal{G}'_{ref}$ .

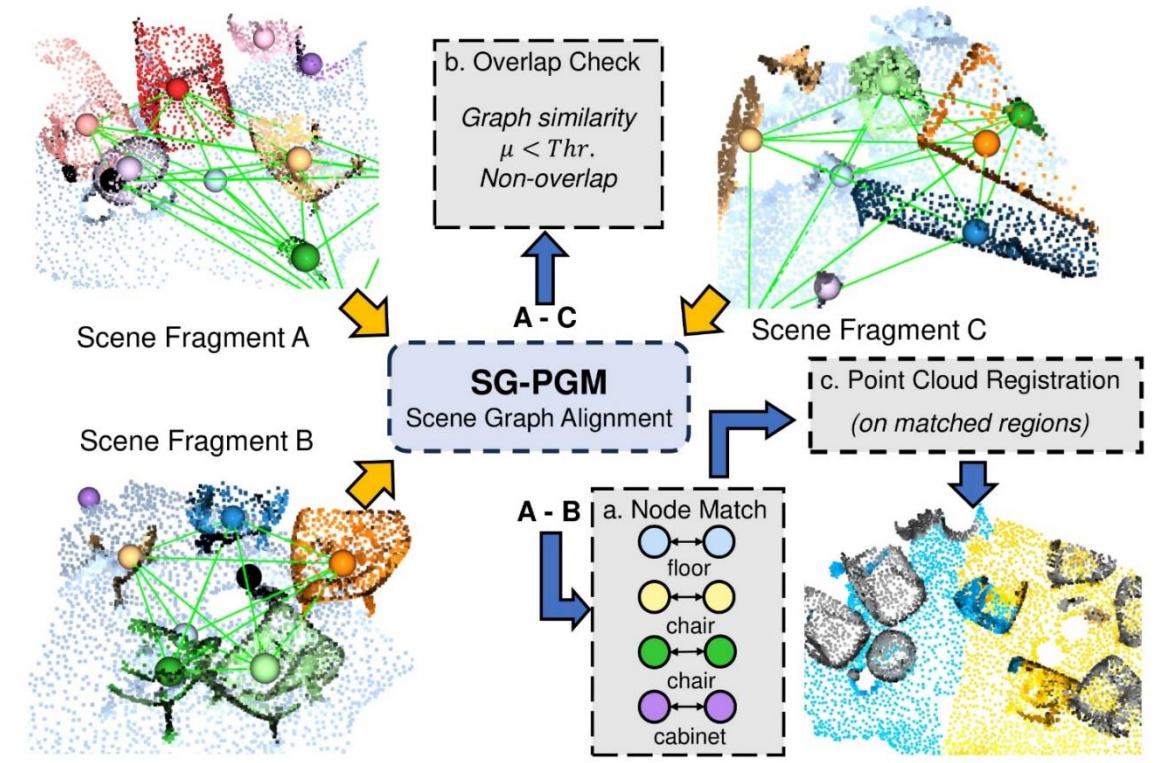


Figure 10 **SG-PGM** [17] partial graph matching for 3D scene graph alignment.

### 4.3.2 Scene graph and point encoding

As illustrated in Figure 11, our matching network first projects the semantic node features  $\mathbf{X}$  and semantic edge features  $\mathbf{E}$  of the source and reference graphs into the semantic scene graph embedding  $F_S \in R^{M \times d_s}$ . We then combine the point geometric embedding  $F_P \in R^{M \times d_p}$  of each object node from the point cloud encoder to form the fused embedding  $F_{S+P} \in R^{M \times (d_s+d_p)}$ .

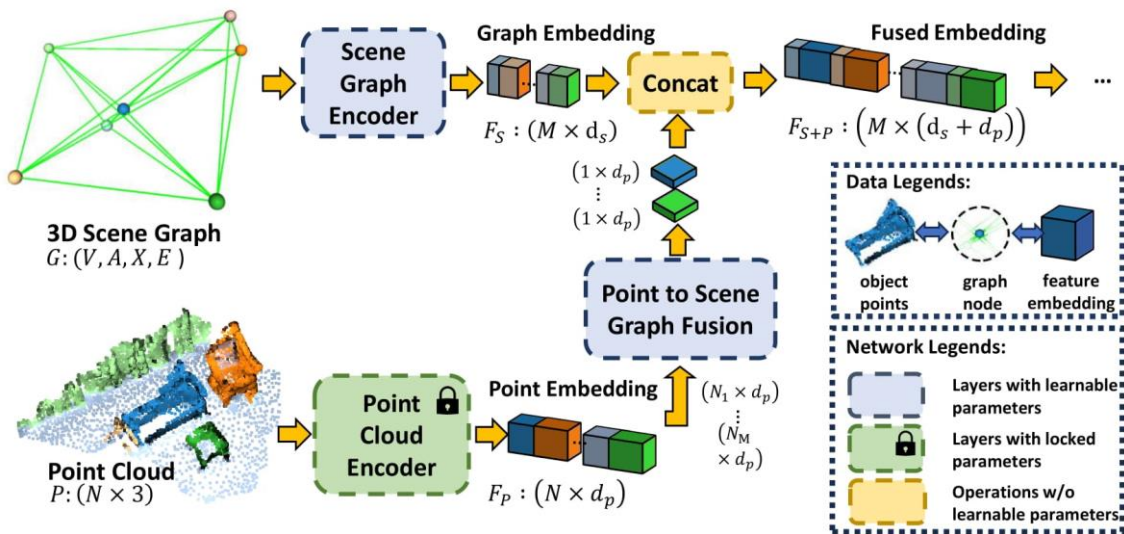


Figure 11 Point to Scene Graph Feature Fusion of one single point cloud and its associated 3D scene graph [17].

### 4.3.3 Scene graph alignment and point cloud registration

As illustrated in Figure 12, the fused embedding of the source and reference scene graph is taken by the AIS [18] module to provide a cost matrix that measures the pair-wise similarity. Then the joint scene graph and geometric node embedding  $F_{S+P}^{\text{ref}}$  and  $F_{S+P}^{\text{src}}$  are used to compute an affinity matrix  $\mathbf{A}$  by:

$$\mathbf{A} = F_{S+P}^{\text{ref}} \mathbf{W}_s + F_{S+P}^{\text{src}} \mathbf{W}_p$$

in which  $\mathbf{W}_s$  and  $\mathbf{W}_p$  are the learnable weights for computing the affinity of both nodes embedding. Then  $\mathbf{A}$  is normalized via instance normalization and processed by the Sinkhorn [19] operator with an additional row and column of zeros.

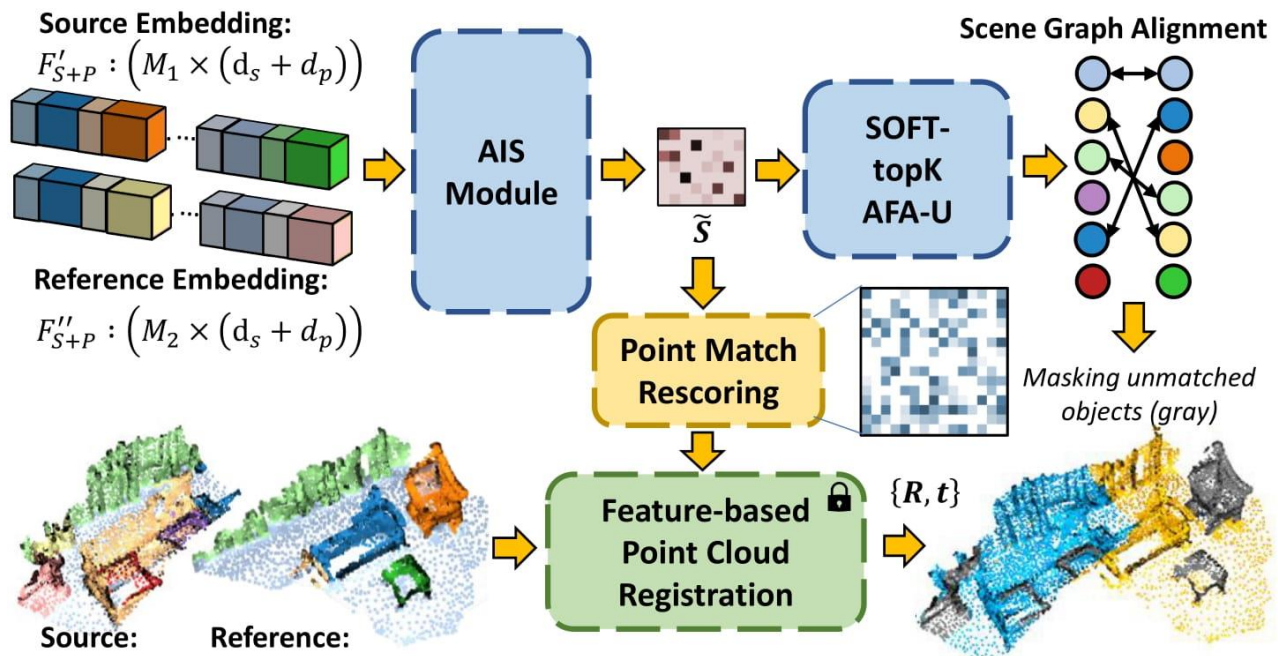


Figure 12 The alignment stage between the source and the reference scene graphs and the registration stage of point clouds with the guidance of Superpoint Matching Rescoring method [17].

### 4.3.4 Results on open-source datasets

For our experiments, we use 3RScan [15], [16] open-source dataset that provides multiple rescans of one scene with changes such as moved, removed, and deformed objects. Our training procedure takes 10 epochs with the ADAM optimizer and an initial learning rate of 0.0001, which decreases by 0.1 for every 4<sup>th</sup> epoch. Figure 13 provide some qualitative results by combing our SG-PGM method and GeoTransformer [20] for point cloud registration.

To summarize, our SG-PGM method [17] shows significant performance improvements on scene graph alignment, overlap-checking, point cloud registration and other downstream tasks. Specifically, the alignment accuracy is improved by 10~20% compared to existing state-of-the-art method, especially when transformation  $T \neq I_4$  exists between scene fragments. The rotation error is reduced by 50% and the translation by 24% on the point cloud registration tasks. Additionally, our scene graph alignment method remains decoupled from registration and robust to scene dynamics and noises.

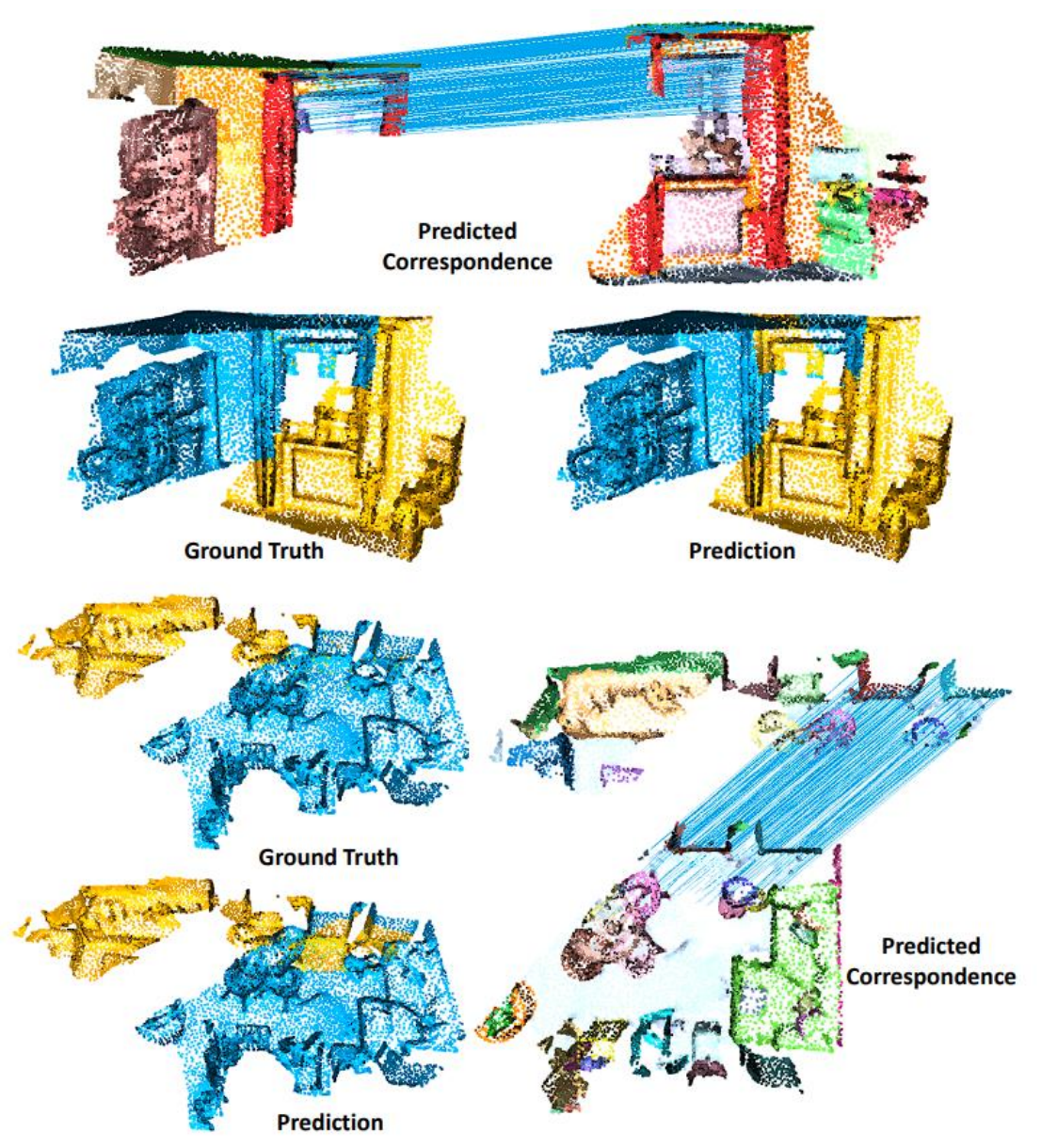


Figure 13 Point Cloud Registration using SG-PGM [17].

## 4.4 Global scene graph generation of a BIM model

In order to apply a scene graph matching methods like our method from section 4.3, we need to be able to extract scene graphs from a BIM reference model (global scene graph), as well as from the observed scene by the user (local scene graph). Our efforts for estimating these scene graphs are presented in the current and next section. We have a point cloud  $\mathbf{P} \in \mathbb{R}^{N \times 3}$  with  $N$  3D points (refer Figure 15(c)), which is synthetically generated from the BIM model (see D3.4), a set of class-agnostic instance masks  $I = \{I_1, \dots, I_M\}$  that associate the point cloud  $\mathbf{P}$  with  $M$  semantic instances [14]. The aim is to predict a 3D semantic scene graph as a directed graph  $\mathcal{G} = \{\mathcal{O}, \mathcal{R}\}$ . The set of objects  $\mathcal{O} = \{o_i\}_{i=1}^M$  are all named object instances that are specified by instance masks  $\mathcal{M}$ . Each edge  $r_{ij}$  in  $\mathcal{R}$  depicts the predicate in a relation triplet  $\langle \text{subject}, \text{predicate}, \text{object} \rangle$ , where the head node  $o_i$  of this edge is the subject and the tail node  $o_j$  is the object. To be specific,  $o_i$  indicates an object label from  $N_{obj}$  semantic class,  $r_{ij}$  is a predicate label from  $N_{rel}$  predicate classes.

The VL-SAT [21] 3D prediction model is employed as GNN-based scene graph prediction methods such as SGFN [22], which mainly consists of node encoder, edge encoder and scene graph reasoning modules.

**Node Encoder** Based on one class-agnostic instance mask  $M_i$  along with the input point cloud  $\mathbf{P}$ , we can extract the set of points  $\mathbf{P}_i$  that corresponds to one semantic instance. A simple PointNet [23] is employed to extract instance level features  $o_i^{3d} \in \mathbb{R}^D$  as node features for GNN-based scene graph reasoning.

**Edge Encoder** To encode the edge features for the GNN-based scene graph reasoning the difference between several attributed between the linked instances are calculated. For each instance, these attributes include the mean  $\mu$  and standard deviation  $\sigma$  of the 3D points, the size  $b = (b_x, b_y, b_z)$ , the volume  $v = b_x b_y b_z$ , and the maximum side length  $l = \max(b_x, b_y, b_z)$  of the bounding box. The edge features  $r_{ij}^{3d} \in \mathbb{R}^D$  are encoded by projecting the concatenated differences of these attributes between two instances, via multi-layer perceptron (MLP) layers, i.e.,  $r_{ij}^{3d} = \text{MLP} \left( \text{cat} \left( \mu_i - \mu_j, \sigma_i - \sigma_j, b_i - b_j, \ln \frac{l_i}{l_j}, \ln \frac{v_i}{v_j} \right) \right)$ , where the subscript  $i$  indicates the instance  $\mathbf{P}_i$  in the head node, and the  $j$  means the instance  $\mathbf{P}_j$  in the tail node.

### Scene Graph Reasoning

Using the Feature-wise Attention (FAT) module messages are passed between the nodes and edges, and then the updated node and edge features are retrieved. Each GNN module is paired with a Multi-Head Self Attention (MHSA) module, and they are repeated for  $T$  times to extract the final node and edge features  $\{o_i^{\widetilde{3d}}\}_{i=1,\dots,M}$  and  $\{r_{ij}^{\widetilde{3d}}\}_{i \neq j, i, j=1,\dots,M}$ , respectively. An object classifier and a predicate classifier are used to predict the elements of  $\{o_i, r_{ij}, o_j\}$  of each possible relation triplet from the final node and edge features to construct the semantic scene graph  $\mathcal{G} = \{\mathcal{O}, \mathcal{R}\}$ .

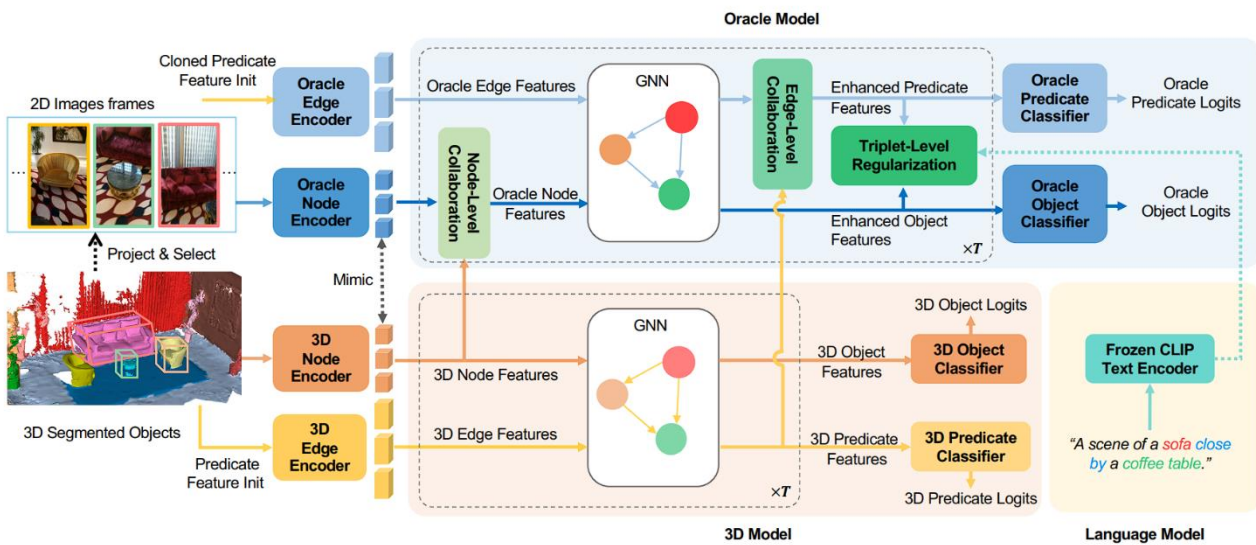


Figure 14 Visual-Linguistic Semantics Assisted Training (VL-SAT) for 3D scene graph prediction [21].

### Visual-Linguistic Semantics Assisted Training

The key idea is to achieve discriminative power from auxiliary learning a powerful multi-modal prediction model, refer Figure 14, that receives structural semantics from vision, and language, as well as the 3D geometry from the 3D prediction model. The multi-modal semantics are expected to be heterogeneously aligned with the 3D semantics as the node and edge levels, and the benefits from the oracle model can be efficiently absorbed by the 3D prediction model during the training process. As illustrated in Figure 14, during training, VL-SAT takes 2D and language semantics as extra inputs and helps 3D scene graph prediction with node- and edge-level collaboration and triplet-level regularization. In inference, VL-SAT only takes the 3D point cloud to predict reliable 3D scene graphs.

#### 4.4.1 3D global scene graph from BIM-model of DFKI lab

For experiments, we use the synthetic point cloud dataset generated in T3.4, refer Figure 15, using the BIM models generated from Scan-to-BIM pipeline T3.6 (for details see deliverable D3.6). We used a pre-trained VL-SAT network, which is end-to-end optimized

using AdamW optimizer with a batch size as 8. The network was trained for 100 epochs, and the base learning rate is set as 0.001 with a cosine annealing learning rate decay strategy.  $N_{obj} = 160$  and  $N_{rel} = 26$  in all of our experiments. The 2D image inputs are only used during the training stage. During the inference stage the top@1 class of both object and predicate are selected. Figure 16 demonstrate the output from VL-SAT method for 3D synthetic scan of DFKI hallway described in 2. The method successfully distinguishes some similar objects like door versus wall.

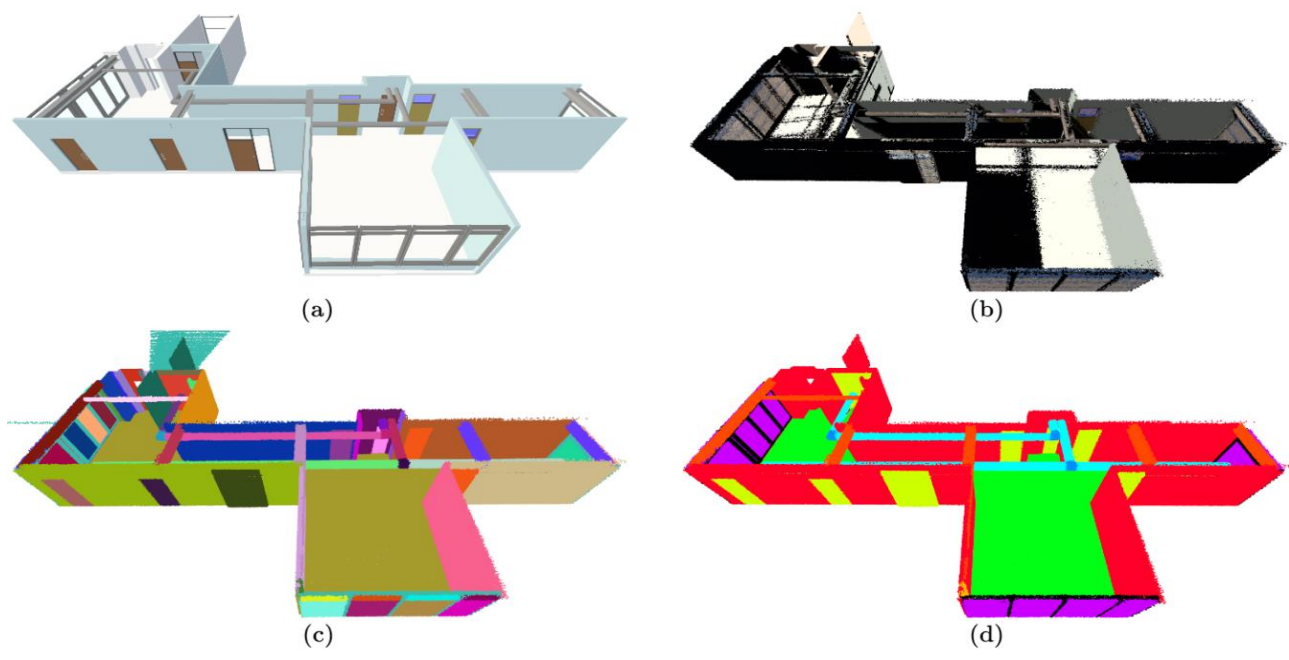


Figure 15 (a) BIM model of DFKI hallway obtained manually and/or from scan-to-bim pipeline, (b): Synthetic RGB point cloud generated using BIM model in UnrealEngine5, (c) and (d): Object instances and semantics ground truth annotations for generated synthetic scan in T3.4 .

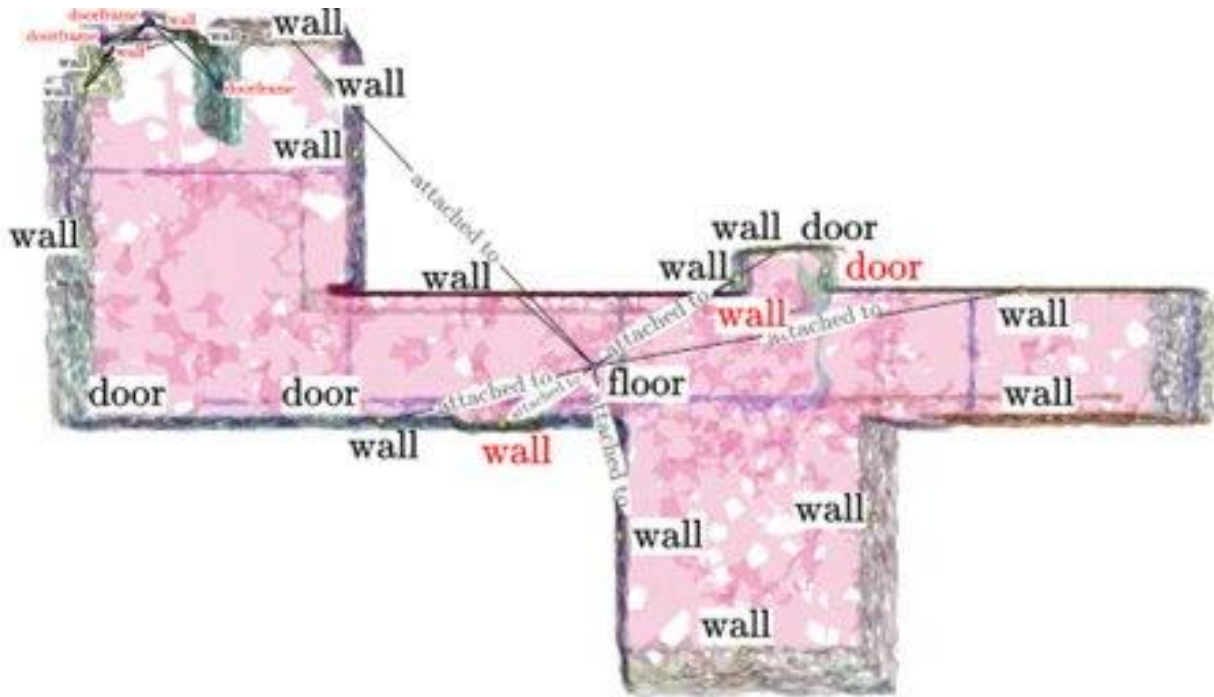


Figure 16 3D Semantic Scene Graph generated from DFKI hallway's synthetic 3D scan.

#### 4.5 Local scene graph generation

MonoSSG [24] framework, as shown in Figure 17, given a sequence of RGB images, can estimate a 3D semantic scene graph incrementally. The Incremental Entity Estimation (IEE) front end make use of the images to generate segmented sparse points. Those are merged into 3D entities and used to generate both an entity visibility graph and a neighbour graph. The Semantic Scene Graph Prediction (SSGP) network uses the entities and both graphs to estimate multiple scene graphs for the entire sequence and then fuse them into a consistent 3D SSG.

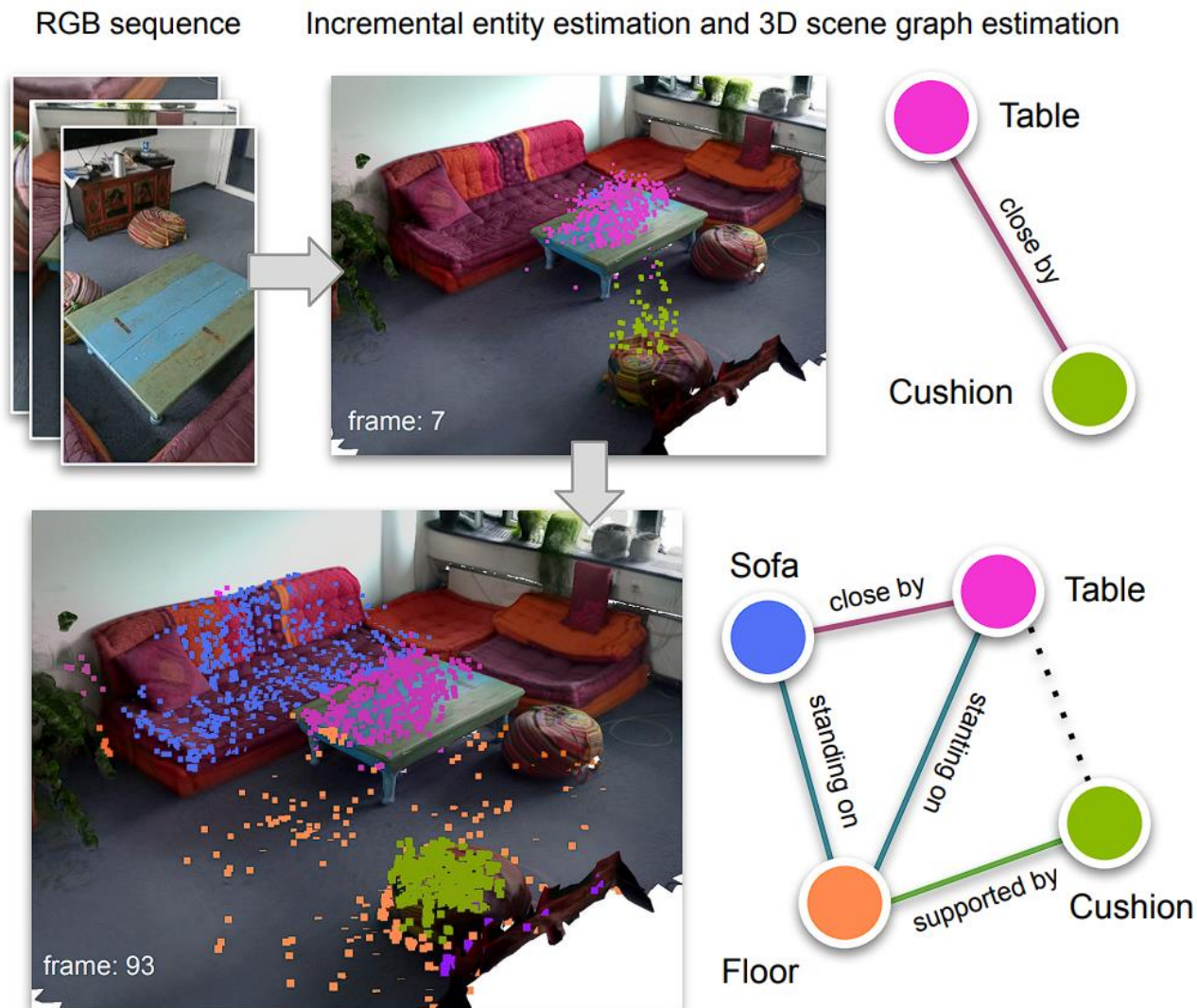


Figure 17 A scene graph incrementally estimated and fused into a global 3D scene graph [24].

However, we were not able to use this approach out-of-the-box for the HumanTech setup because the authors use the ground truth poses with respect to a world reference system to guide the scene reconstruction and to tackle severe image blur and jittery motion, refer Implementation details [24]. On the contrary, our goal is to obtain the poses from state estimation methods (such as SLAM approaches) which provide the poses together with point clouds from an RGB+IMU stream. Prior to implementing a visual-initial SLAM in the HumanTech BSN, an experimental device (Meta Aria glasses) with integrated SLAM capabilities has been used to assess the possibility of generating a local scene graph from a sparse point cloud. The hardware setup of the system is explained in the following section.

### 4.5.1 Hardware specifications and recording profile

As illustrated in Figure 18, Project Aria glasses [25] have five cameras (two Mono Scene, one RGB, and two Eye Tracking cameras) as well as non-visual sensors (two IMUs, magnetometer, barometer, GPS, Wi-Fi beacon, Bluetooth beacon and Microphones). Mono Scene Cameras are often used to support SLAM algorithms, but they can have other applications. All cameras, as well as the IMU, magnetometer, barometer and microphone are calibrated, and all sensor measurements are timestamped on a common clock at nanosecond resolution. The SLAM and RGB cameras have fisheye lenses to maximize the visible field of view.

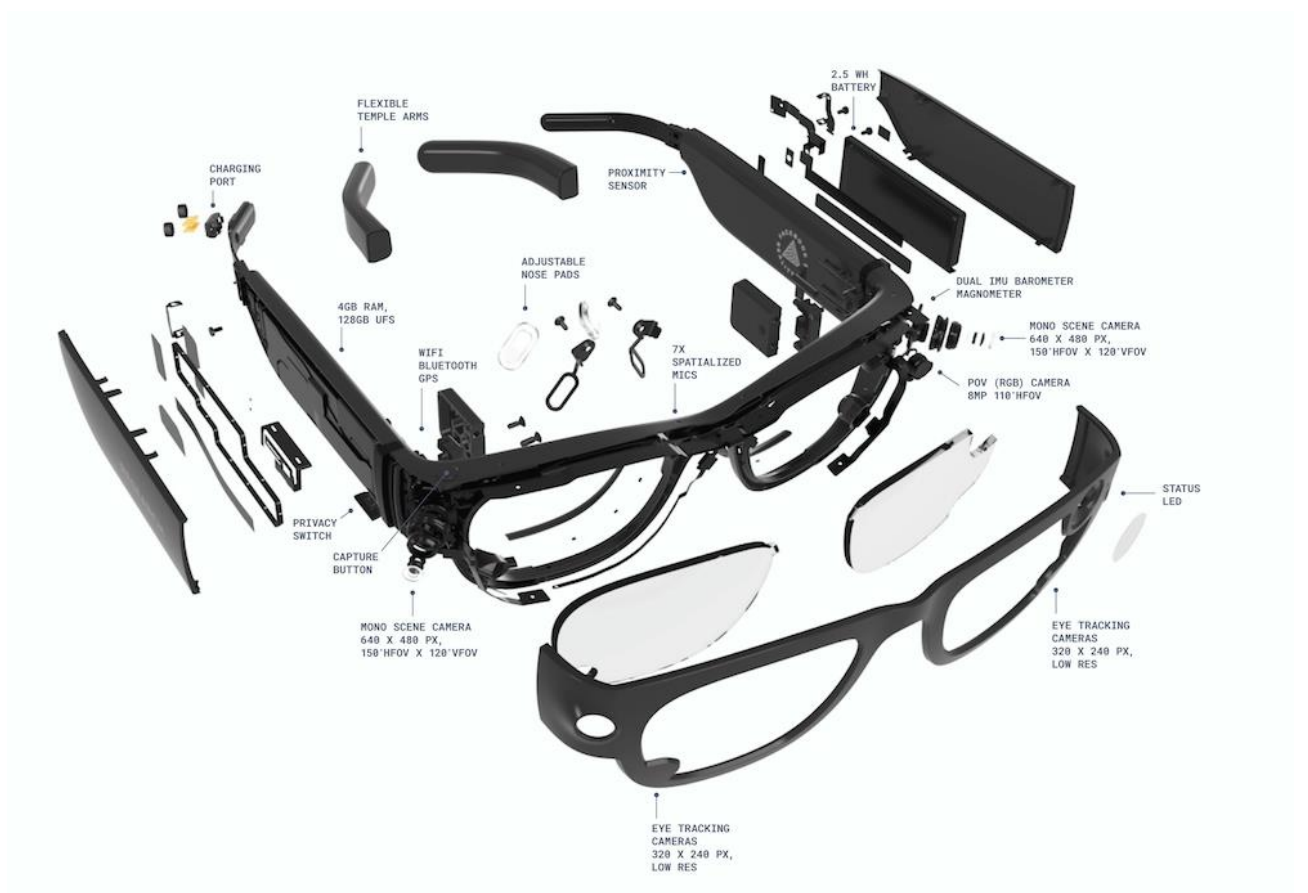


Figure 18 Project Aria Glasses Sensor Diagram [25].

Using Project Aria glasses we recorded the following data:

1. RGB Camera – 1408x1408 resolution @ 20 FPS and Auto Exposures is ON.
2. SLAM (Mono Scene Camera) – 640 x 480 resolution @ 20 FPS and Auto Exposures is ON.
3. GPS – 1 Hz
4. IMU1 & 2 – 1000 Hz



5. Magnetometer – 10 Hz
6. Barometer – 10 Hz
7. Wi-Fi – 10 seconds
8. Bluetooth – 10 seconds

Project Aria chose VRS as its data container because it is a file format designed to record and playback streams of XR sensor data and support huge file sizes. The VRS files contain streams of time-sorted records generated for each sensor, with one set of sensors per stream.

#### 4.5.2 Aria machine perception services

To accelerate research with Project Aria, several Spatial AI machine perception capabilities that help form the foundation for future Contextualized AI applications and analysis of egocentric data is provided. MPI provide SLAM services with following outputs:

1. 6DoF Trajectory
2. Semi-Dense Point Cloud
3. Online Sensor Calibration

**Open loop trajectory** is the high frequency (IMU rate, which is 1kHz) odometry estimation output by visual-inertial odometry (VIO), in an arbitrary odometry coordinate frame. The estimation includes pose and dynamics (translational and angular velocities). The open loop trajectory has good “relative” and “local” accuracy: the relative transformation between two poses is accurate when the time span between two frames is short (within a few minutes). However, the open loop trajectory has increased drift error accumulated over time spent and travel distance.

**Closed loop trajectory** is the high frequency (IMU rate, which is 1kHz) pose estimation output by mapping process, in an arbitrary gravity aligned world coordinate frame. The estimation includes poses and dynamics (translational and angular velocities). Closed loop trajectories, refer Figure 19, are fully bundle adjusted with detected loop closures, reducing the VIO drift which is present in the open loop trajectories. However, due to the loop closure correction, the “relative” and “local” trajectory accuracy within a short time space (i.e. seconds) might be worse compared to open loop trajectories.



Figure 19 Display of interactive visualization of the Project Aria VRS RGB frames along with MPS data – Closed loop trajectory (in green), estimated semi-dense global point cloud (in purple) of DFKI hallway, and RGB image with estimated 6DOF pose.

### 4.5.3 Semi-dense 3D semantic segmentation

As illustrated in Figure 20, the 3D semantic segmentation result of the method developed by DFKI within task T3.5. The trained model was able to properly segment main sections of the building more specifically, roofs and walls. However, other scene objects such as door, floor, table and chair are incorrectly segmented.

Due to the sparsity and the lack of texture of the point clouds, the quality of the semantic segmentation was not sufficient for the Scan-to-BIM pipeline to generate a proper BIM of the scan. Generating the BIM is a necessary step to retrieve the instances (BIM-to-Scan) since the instance labels are assigned knowing their respective object in the model. With the absence of instance segmentation, it is thus not possible to generate the scene graph.

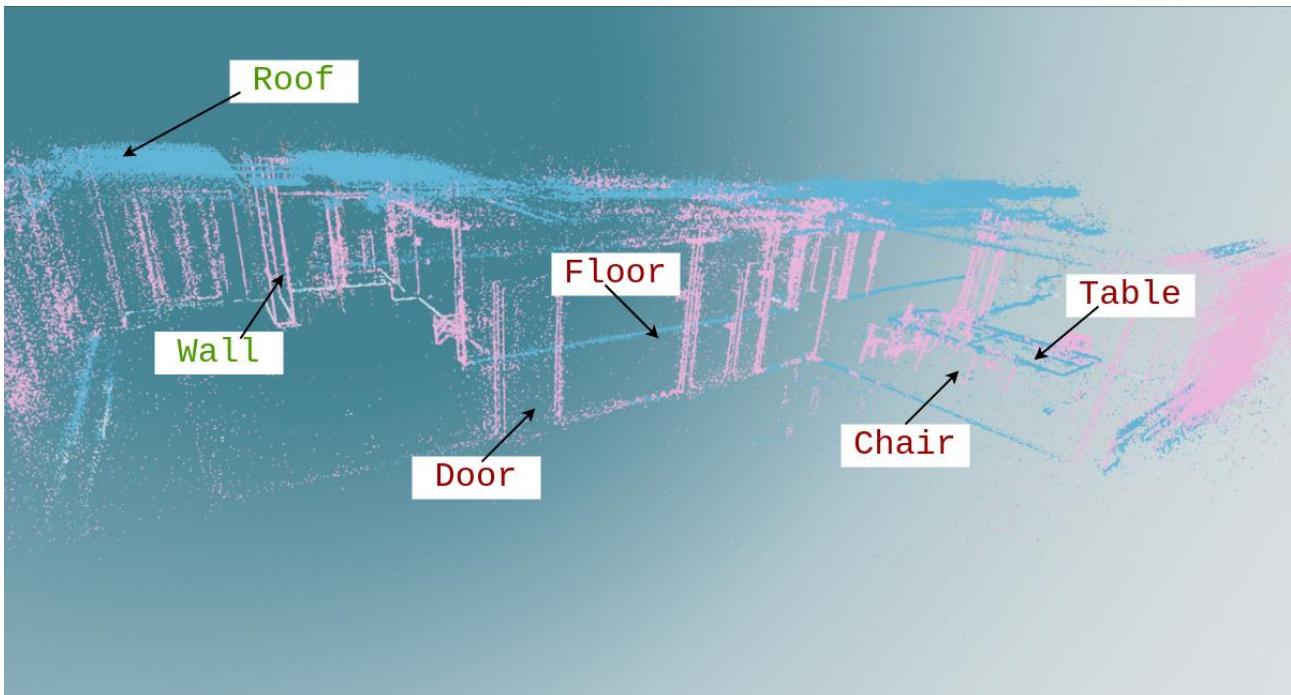


Figure 20 Semantic Segmentation of Project Aria VRS semi-dense global point cloud using T3.5.

## 5 Conclusion

This document summarizes the research results accomplished within Task 4.3. The target of the research is to develop and assess methods to localize a camera in the digital twin of a construction site from a data stream consisting of video images and IMU data of a body worn sensor system, as the one described in D4.1. The research and development in this task have been accomplished within two paths. While SCT has developed a marker supported visual-inertial tracking system, DFKI has concentrated on research on marker-less localization methods.

The SCT approach has been a tracking algorithm that fuses marker detection with IMU data in a common framework which provides both camera and body poses in real time. This algorithm has been ported on a mobile processing platform and evaluated the localization accuracy in terms of KPI 4.04 as well as the body pose accuracy in terms of KPI 4.02. The full body pose tracking server moreover as basis for the Intelligent exoskeleton with intention prediction prototype reported in D4.2.

To develop and evaluate the localization approaches, a sufficiently complex hallway at DFKI has been equipped with an optical reference system and markers. The markers were developed under the lead of ZHAW in a previous project and have meanwhile become a European standard and are used in HumanTech throughout different work packages and sensor systems (see D3.1). The space has then been scanned in collaboration with RPTU who have generated a BIM model.

DFKI has identified, implemented and assessed several approaches to extract useful 3D and semantic information from a video stream and to align them with a BIM model. The most promising approach is based on the alignment of scene graphs. Scene graphs are compact representation of both the 3D-geometry and semantics of building elements and their relationship to each other and thus less prone to ambiguity issues than purely geometrical approaches.

To apply this concept to the WP4 use-case of localizing a visual-internal BSN to a BIM model, several modules had to be developed. Based on a method developed in WP3 (see) to generate from a BIM model a point cloud with semantics object instances, DFKI has generated a global scene graph from a BIM model. In parallel, a novel alignment method for pairs of scene graphs that works also for incomplete and partial overlap has been developed and meanwhile been published [17].



### D4.3 – Wearable-user localization algorithm

---

The most challenging module in the pipeline of a scene graph-based localization turned out to be the generation of a local scene graph from the data stream of the visual-BSN. The bottleneck in the generation of a local scene graph is thereby the semantic segmentation of semi-dense point clouds a BSN can provide via visual-inertial SLAM. Here future research directions are clearly indicated. Possible directions are to adapt the dense point cloud segmentation developed in T3.6 to semi-dense point clouds or to conceive techniques to fuse SLAM methods with monocular depth estimation networks which recently emerged [26].

## 6 References

- [1] A. Rosinol, A. Violette, M. Abate, N. Hughes, Y. Chang, J. Shi, A. Gupta and M. Carlone, “Kimera: from SLAM to spatial perception with 3D dynamic scene,” *Intl. J. of Robotics Research*, vol. 40, no. 12-14, p. 1510–1546, 2021.
- [2] L. Liu, H. Li and Y. Dai, “Efficient global 2d-3d matching for camera localization in a large-scale 3d map,” in *In Proceedings of the IEEE International Conference on Computer Vision*, 2017.
- [3] C. Chen, R. Wang, C. Vogel and M. Pollefeys, “F3Loc: Fusion and Filtering for Floorplan Localization,” in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [4] N. Hughes, Y. Chang and L. Carlone, “Hydra: A real-time spatial perception system for 3D scene graph construction and optimization,” 2022.
- [5] U. V. B. L. Udugama, G. Vosselman and F. Nex, “Mono-hydra: Real-time 3D scene graph construction from monocular camera input with IMU,” 2023.
- [6] C. Y. Wu, J. Wang, M. Hall, U. Neumann and S. Su, “Toward practical monocular indoor depth estimation,” in *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.
- [7] N. Zhang, F. Nex, G. Vosselman and N. Kerle, “Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation,” in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [8] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao and D. Liu, “Deep high-resolution representation learning for visual recognition,” in *IEEE transactions on pattern analysis and machine intelligence*, 2020.
- [9] Q. Bonnard, S. Lemaignan, G. Zuffery, A. Mazzei, S. Cuendet, N. Li, A. Özgür and P. Dillenbourg, *Chilitags 2: Robust Fiducial Markers for Augmented Reality and Robotics*, CHILI, EPFL, Switzerland, 2013.

- [10] J. Wang and E. Olson, “AprilTag 2: Efficient and robust fiducial detection,” in *Proceedings of the {IEEE/RSJ} International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, South Korea, 2016.
- [11] M. Shaheer, J. A. Millan-Romera, H. Bavle, J. L. Sanchez-Lopez, J. Civera and H. Voos, “Graph-based Global Robot Simultaneous Localization and Mapping using Architectural Plans,” 2023.
- [12] H. Bavle, J. L. Sanchez-Lopez, M. Shaheer, J. Civera and H. Voos, “Situational graphs for robot navigation in structured indoor environments,” in *IEEE Robotics and Automation Letters*, 2022.
- [13] H. Bavle, J. L. Sanchez-Lopez, M. Shaheer, J. Civera and H. Voos, “S-graphs+: Real-time localization and mapping leveraging hierarchical representations,” in *IEEE Robotics and Automation Letters*, 2023.
- [14] I. Armeni, Z. Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik and S. Savarese, “3d scene graph: A structure for unified semantics, 3d space, and camera,” in *In Proceedings of the IEEE/CVF international conference on computer vision*, 2019.
- [15] J. Wald, A. Avetisyan, N. Navab, F. Tombari and M. Nießner, “Rio: 3d object instance re-localization in changing indoor environments,” in *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
- [16] J. Wald, H. Dharmo, N. Navab and F. Tombari, “Learning 3d semantic scene graphs from 3d indoor reconstructions,” in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [17] Y. Xie, A. Pagani and D. Stricker, “SG-PGM: Partial Graph Matching Network with Semantic Geometric Fusion for 3D Scene Graph Alignment and Its Downstream Tasks,” in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [18] K. Fu, S. Liu, X. Luo and M. Wang, “Robust point cloud registration framework based on deep graph matching,” in *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.



- [19] G. Mena, D. Belanger, S. Linderman and J. Snoek, “Learning latent permutations with gumbel-sinkhorn networks,” 2018.
- [20] Q. Zheng, H. Yu, C. Wang, Y. Guo, Y. Peng and K. Xu, “Geometric transformer for fast and robust point cloud registration,” in *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.
- [21] Z. Wang, B. Cheng, L. Zhao, D. Xu, Y. Tang and L. Sheng, “VI-sat: Visual-linguistic semantics assisted training for 3d semantic scene graph prediction in point cloud,” in *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.
- [22] S. C. Wu, J. Wald, K. Tateno, N. Navab and F. Tombari, “Scenegrphfusion: Incremental 3d scene graph prediction from rgb-d sequences,” in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- [23] C. R. Qi, H. Su, K. Mo and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.
- [24] S. C. Wu, K. Tateno, N. Navab and F. Tombari, “Incremental 3d semantic scene graph prediction from rgb sequences,” in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [25] K. Somasundaram, J. Dong, H. Tang, J. Straub, M. Yan, M. Goesele, J. J. Engel, R. De Nardi and R. Newcombe, “Project aria: A new tool for egocentric multi-modal ai research,” 2023.
- [26] L. Yang, B. Kang, Z. Huang, Z. Zhen, X. Xu, J. Feng and H. Zhao, “Depth Anything V2,” 2024.