SPRING

GRANT AGREEMENT N. 871245

Deliverable D2.2

Semantics-based Localisation in Realistic Environments

Due Date: 31/08/2021
Main Author: CVUT
Contributors: APHP
Dissemination: Public Deliverable

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

DOCUMENT FACTSHEET

Deliverable no.	D2.2: Semantics-based Localisation in Realistic Environments
Responsible Partner	CVUT
Work Package	WP2: Environment Mapping, Self-localisation and Simulation
Task	T2.2: Audio-Visual Data Simulator
Version & Date	VFinal, 31/08/2021
Dissemination level	[X] PU (public) [ ] CO (confidential)

CONTRIBUTORS AND HISTORY

Version	**Editor**	**Date**	**Change Log**
1	CVUT	14/08/2021	First Draft
Final	CVUT	31/08/2021	Final Draft after reviewer comments

APPROVALS

Authors/editors	CVUT
Task Leader	CVUT
WP Leader	CVUT

D2.2: Semantics-based Localisation in Realistic Environments

Leader: CVUT
Contributors: HWU

August 2021

Contents

1 Introduction

The research project **H2020 SPRING** has a wide scope of objectives in different research fields. This deliverable presents progress made mainly towards two *specific objectives* **SpO-1.1**: *Performing self-localisation and tracking in cluttered and populated spaces*, and **SpO-1.3**: *Augmenting the 3D geometric maps with semantic information learned from associations between images (colour and depth) and natural language queries*. These specific objectives are part of the first of three *strategic objectives*. The strategic objective (**StO-1**) aims **to enable robust robot perception in complex, unstructured and populated environments**. We have presented our advances towards these objectives in our previous deliverable **D2.1** *Visual-based localisation in Realistic Environments* [1]. The presented work provided state-of-the-Art results in task of self-localisation within a realistic environment. The presented work builds on **InLoc** [2] and produces very accurate localisation in stable visual conditions. In other words it is highly reliant overall visual similarity between localisation map and query images. In order to improve the robustness we present an improvement based on taking into account semantic information of the scene as a final re-ranking step between candidate results. The advantage of adding semantic information is twofold. Firstly we believe it improves robustness to viewpoint, seasonal changes or general lighting conditions. Secondly it enriches localisation map with semantic information that can be further used in the human-robot interaction (**HRI**).

2 Semantics Extraction

To extend the visual localisation algorithm into semantic localisation, we need to introduce the semantic information on both the query side and the map (database) side. The former is done via a standard State-of-the-Art semantic segmentation tool like Detectron2 [3] or YOLACT [4] and is discussed in the while the latter requires semantic segmentation of 3D data which is a far more challenging task. There are some pioneering works like [5], [6] but most of the effort is currently driven by research in the field of autonomous driving, hence the focus is aimed towards traffic-specific sensors like LiDARs [7]. The LiDAR produces quite sparse pointclouds, especially in comparison to our high accuracy model. We could not achieve a reasonable quality of annotation using automated techniques and thus we annotated the 3D model manually. Manual processing carries two benefits, first it provides reliable ground-truth labels of the objects which can be potentially useful as benchmark or training data for future automation, and secondly, it allowed us to closely inspect the model and helped us understand what kind of objects can be encountered in a specific environment like hospital. Figure 1 shows some of the objects encountered in the model.

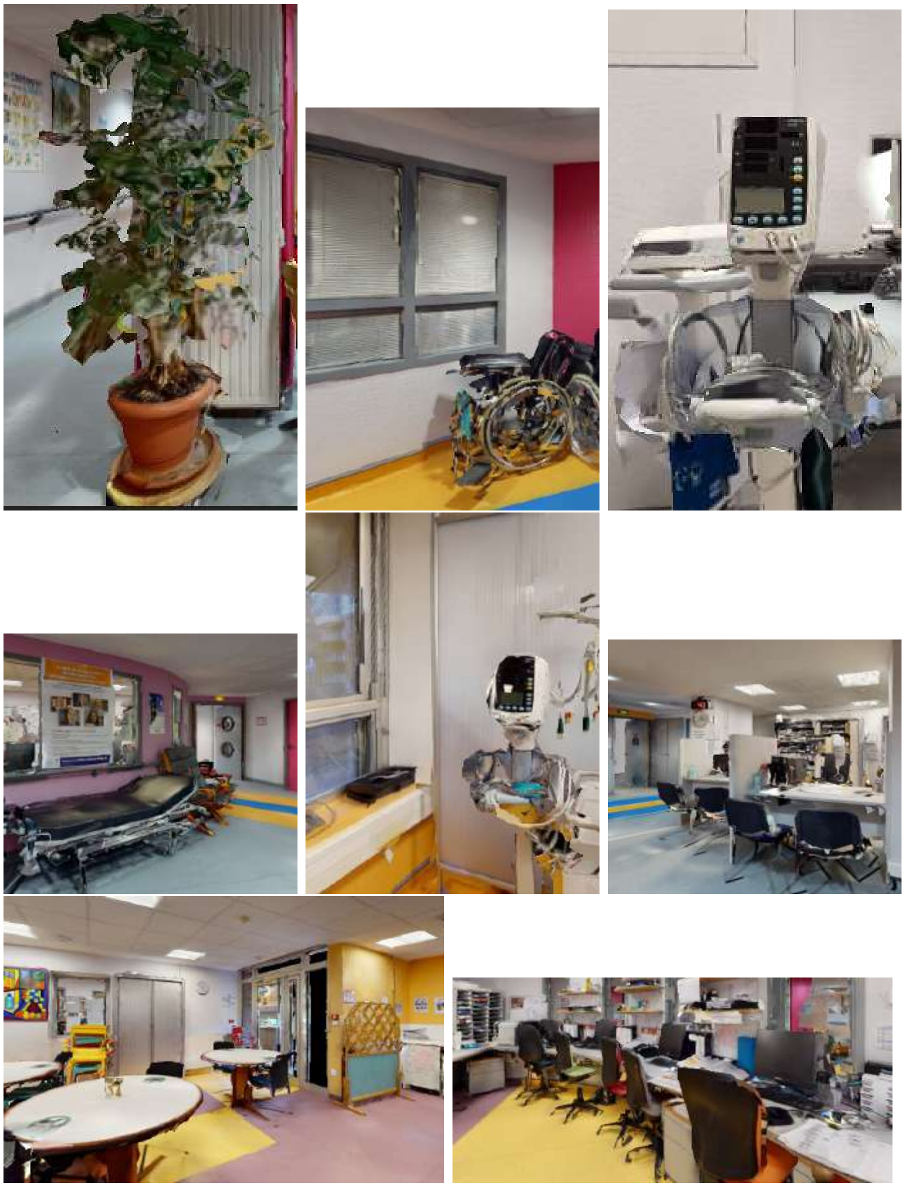

Figure 1: Examples of objects found within the Broca model.

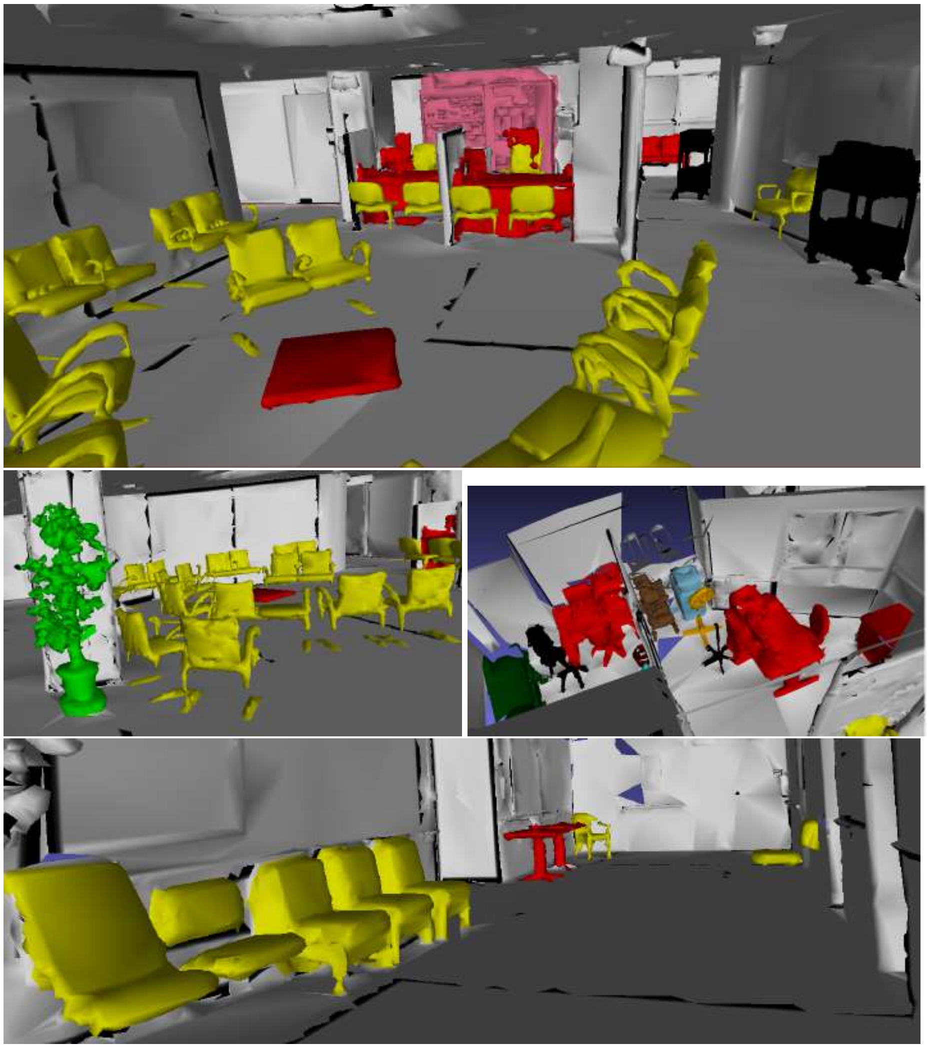

Figure 2: Examples of ground truth semantic annotation appended to Broca model. **Yellow** - Chairs, **Red** - Tables and Desks, **Green** - Plants

2.1 3D Semantic Segmentation of Broca Hospital

Contrary to our initial assumption, we discovered that the 3D model of the hospital is semantically very similar to a general office building with plenty of standard furniture like chairs and office desks and only a handful of specialized medical equipment like a wheelchair, medical bed or other specialized medical measurement or monitoring devices. This can be seen on Figures 1 and 2. We also observed that the medical equipment is mostly present in the check-up rooms where the actual patient examination is usually conducted. Even though the hospital model is very accurate and contains a lot of details, it is still a mesh model and it also contains some artifacts and it may happen that neighbouring objects are blended into a single object. We observed this behaviour especially around office desks, which were often inseparable from office chairs or smaller objects on the desk. A mesh model consists of triangular surfaces and this representation makes the separation extremely difficult. Therefore, we decided to leave the office desk as a single object with all items on it and around it. Full list of objects is available as a part of the semantic labeling package and the link can be found in the Software and data Chapter.

2.2 Image Semantic Segmentation using YOLACT

For the subtask of Image semantic segmentation, we use YOLACT [4]. Core advantage of YOLACT compared to other segmentation tools is its speed. The authors claim that it can run at a frequency around 30 to 40 frames per second, which is usable for real-time applications. We used an of-the-shelf ResNet50 [8] architecture model, pretrained on Microsoft's COCO dataset [9]. Figure 3 show some segmentation results on the Broca dataset images. As can be seen, YOLACT is able to detect most of the common objects relatively reliably. However, when YOLACT encounters an object which is not part of COCO dataset, we get completely false results. An example of such behaviour can be seen on Figure 4 where a wheelchair is detected as a bicycle.

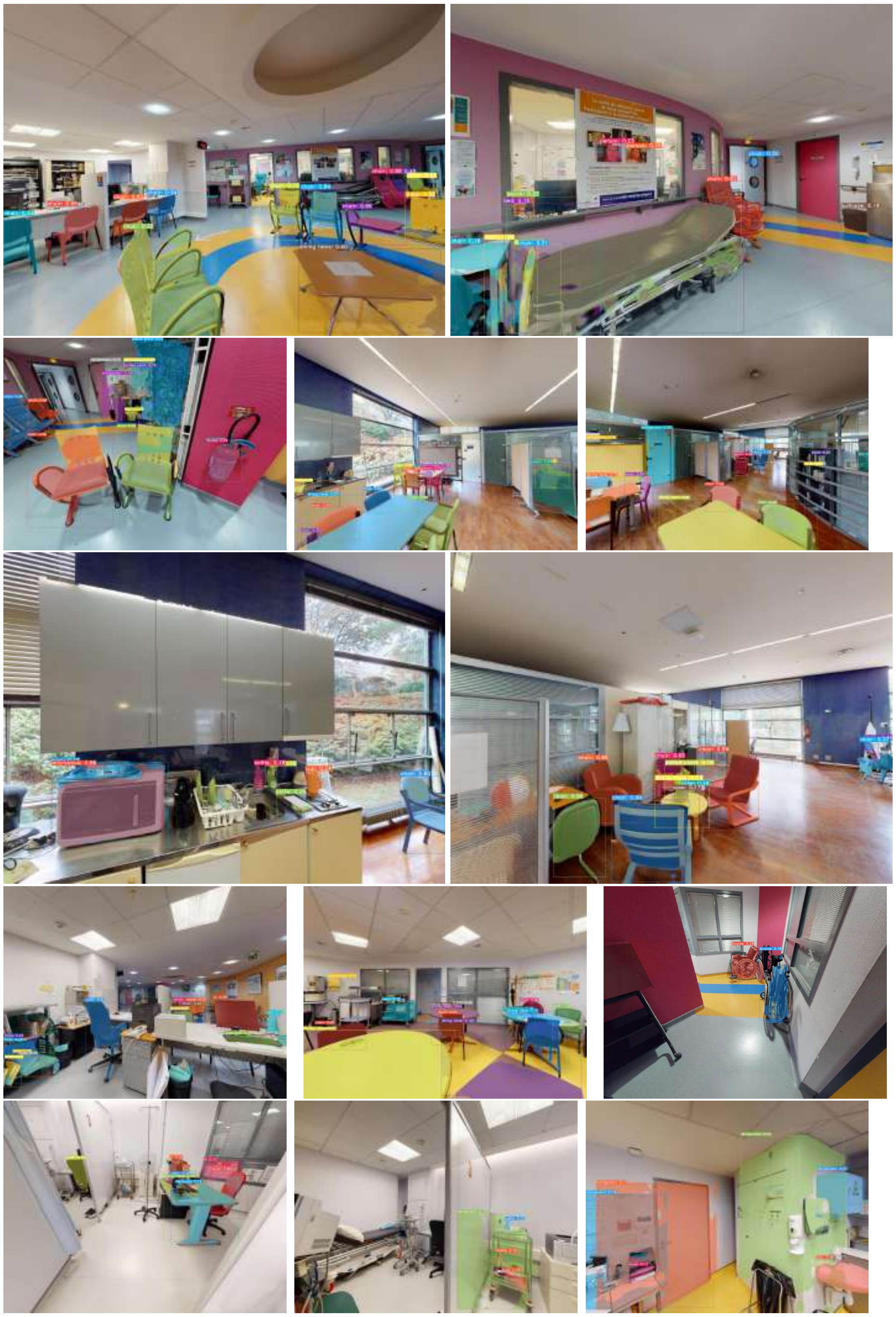

Figure 3: YOLACT semantic segmentation results of the Broca dataset cutouts.

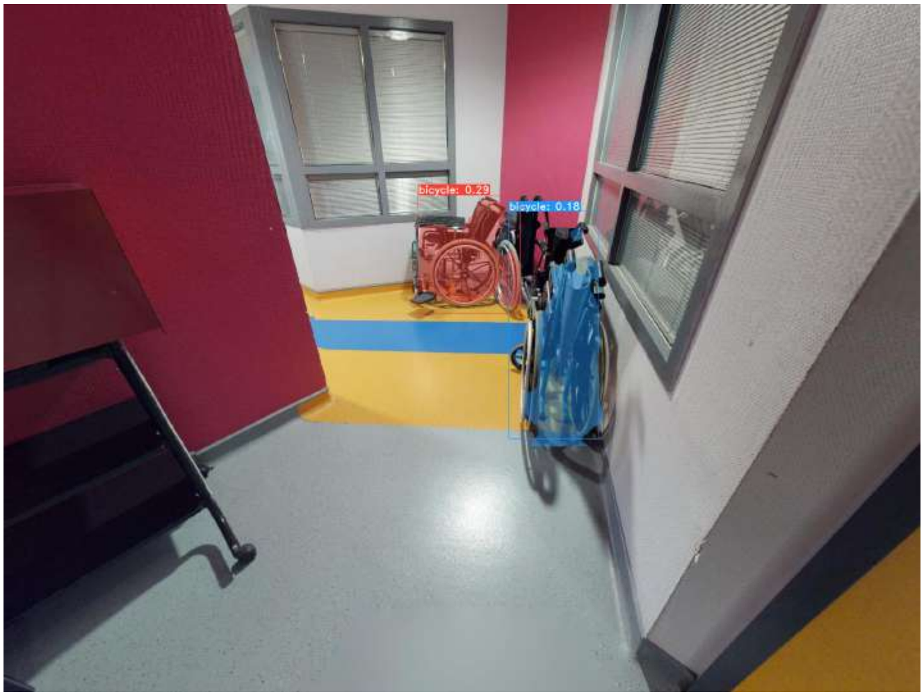

Figure 4: Visual example of YOLACT results observing an unknown class. The image contains almost a canonical view over multiple wheelchairs which are falsely classified as Bicycles.

3 Self-Localisation from Images

The Self-localisation from image data is still an open research problem. Even though we can achieve satisfactory results in a still environment as we published in our previous deliverable D2.1 purely on visual data, we can expect that the localisation precision will drop as more obstacles, occlusions will be introduced. Therefore, we work on improvements of the visual-based localisation with the main goal to robustify the localisation process rather than improve in the still environment.

3.1 Visual-Based Localisation

To provide a better understanding of the proposed modifications to the localisation algorithm, we briefly recap the visual-based localisation algorithm proposed in D2.1 [1]. The localisation algorithm can be divided into multiple consecutive steps. The process starts with **Image retrieval** where only a small subset of the most relevant database images is selected and further processed. This is followed by the extraction of feature points from the images and **establishing matches between the feature points** of the query image and the feature points of the database images. These feature matches are then pruned by **geometric verification** where we select only those in line with a geometrical transformation (We test for two hommographies). The selected tentative matches are then used to **estimate the relative pose** of the query image with respect to the database image. Given the known position of the database image within the localisation map, we reconstruct a candidate pose with respect to the localisation map. Final step is **pose verification** by comparing the visual similarity between the query image and rendering of the localisation map model from the candidate pose, we select the pose which can reproduce the query image best. Figure 5 shows the overview of the localisation process.

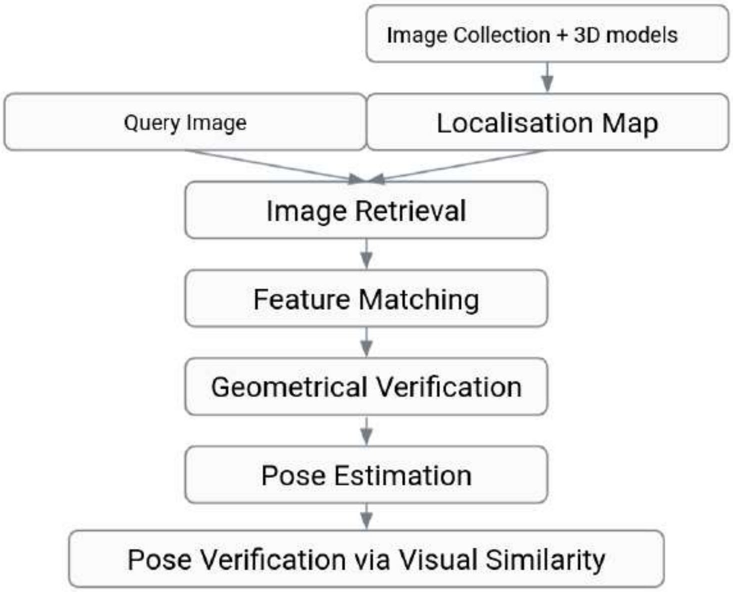

Figure 5: An overview of localisation pipeline according to InLoc [2]

.

3.2 Semantic-Based Localisation

Our initial goal was to follow the work of J. Schönberger et al. from 2018 - Semantic visual localization [10]. It relies on the availability of precise depth information for query images and comparison of semantic information in 3D space. The precision is key because the relative pose is estimated based on finding the best alignment between the semantically annotated depth of the query image and the localisation map. However, ARI robot does not provide depth information for cameras other than those used for collision avoidance, which have very limited scope of vision, mainly observing the floor in front of the robot. Even though there are some works on depth estimation from a single image, for example, the work by C. Godard et al.: Unsupervised monocular depth estimation with left-right consistency [11], the results are not comparable with real depth sensorical data. Therefore, this approach is not suitable for the SPRING project. We decided to modify our existing localisation pipeline by extending it by one more step where we compare the semantic similarity between the query and candidate pose render.

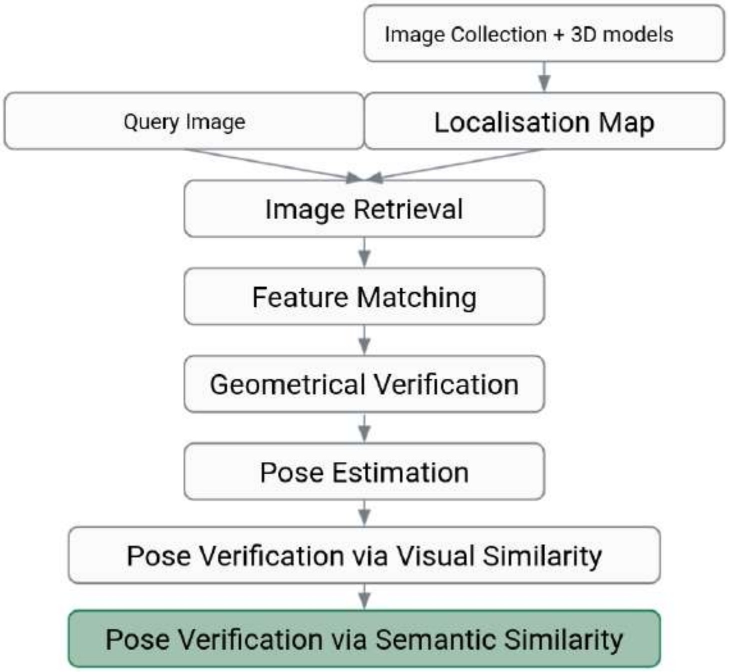
Image Collection + 3D models
Query Image
Localisation Map
Image Retrieval
Feature Matching
Geometrical Verification
Pose Estimation
Pose Verification via Visual Similarity
Pose Verification via Semantic Similarity

Figure 6: An overview of localisation pipeline according to InLoc [2] extended by semantic information.

4 Software and data

The results of described work consist of software tools and developed modules and processed data. Both are publicly available on the CVUT - SPRING repository which is physically stored on the servers of Czech Technical University. We have full control over the data storage and its accessibility and therefore we can guarantee four years of continuous support after the project is terminated as required by European Commission. The repository can be accessed at the web address:

`https://data.ciirc.cvut.cz/public/projects/2020SPRING/SPRING_D22_Semantic_Based_Localisation/` The repository is formed as a standard folder tree structure with two main folders in the top level:

`./segmentation_data`

and

`./software_modules`

4.1 Segementation data

The segmented data are available in multiple formats suitable for various visualization frameworks.

The list of annotated objects is available in PDF file at: `./broca_model_annotated/CLASSES.pdf`

The `./broca_model_annotated/README.txt` file contains additional information about the organisational structure of the segmentation data.

4.2 Software modules

We publish an improved version of the Visitor Module from D2.1 [1] which a simple agent walking within the 3D model of the hospital. It is built on top of AI Habitat framework developed by Facebook Research [12]. This version is enhanced with the semantic information of the newly labeled data. The module is available at `./visitor_module_with_semantics/` An example output from the visitor module can be seen in Figure 7.

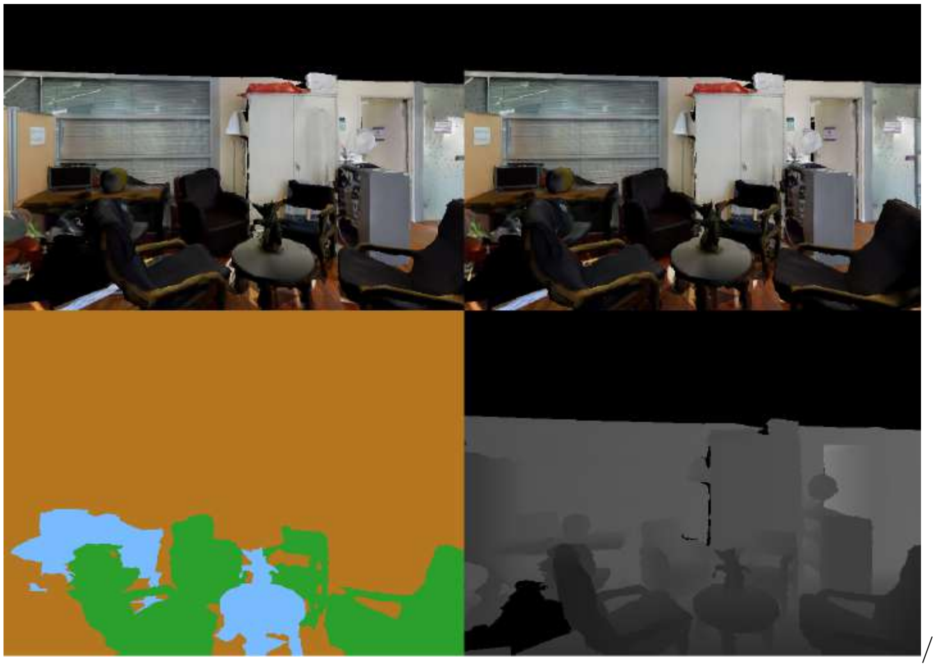

Figure 7: Visitor module in the Living lab model: **Top row** - a stereo pair of cameras, **Bottom left** - semantic segmentation of the scene (**Green** - chairs, **Blue** - tables, **Brown** - background), **Bottom right** - depth information of the scene

5 Conclusion

We have made a great progress in the introduction of semantic information into the localisation map. We have manually annotated the Broca hospital with semantic labels and thus prepared a source of reliable training data for adapting current segmentation models to the hospital environment. We proposed a modification to the current localisation algorithm by adding a final layer which compares the semantic similarity between the map and query image.

6 Future work

We see semantic localisation as a promising tool to enhance the quality of localisation in challenging scenarios with strong occlusions or less reliable localisation map.

Understanding the semantic context of the image will be important not only to improve the localisation quality, but also to identify parts of the image that we need to mask out and avoid using for localisation, as it is likely to provide no or even misleading information about the location of the robot. This in particular is relevant for humans, smaller movable objects like their belongings and furniture like chairs which often move and it is impossible to capture such information within the localisation map.

We also have to retrain the YOLACT to better adapt to the hospital environment with a particular focus on the medical equipment.

References

[1] CVUT. Visual-based localisation in realistic environments. SPRING Public Deliverable 2.1, 2021.

[2] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018.

[3] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[4] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9157–9166, 2019.

[5] Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. Segcloud: Semantic segmentation of 3d point clouds. In *2017 international conference on 3D vision (3DV)*, pages 537–547. IEEE, 2017.

[6] Timo Hackel, Jan D Wegner, and Konrad Schindler. Fast semantic segmentation of 3d point clouds with strongly varying density. *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences*, 3:177–184, 2016.

[7] Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4977–4987, 2021.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[10] Johannes L Schönberger, Marc Pollefeys, Andreas Geiger, and Torsten Sattler. Semantic visual localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6896–6906, 2018.

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition, pages 270–279, 2017.

[12] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.