Spatial Commonsense Graph for Object Localisation in Partial Scenes Supplementary Materials

Francesco Giuliari[1,2] Geri Skenderi[3] Marco Cristani[1,3] Yiming Wang[1,4] Alessio Del Bue[1]
[1]Istituto Italiano di Tecnologia (IIT) [2]University of Genoa [3]University of Verona
[4]Fondazione Bruno Kessler (FBK)

1. Introduction

In this document, we present additional and complementary details as referred in the main paper. In particular, we detail the statistics about the proposed dataset of partial scenes in Section 2, and more qualitative results for our localisation approach in Section 3. Finally, we discuss the potential societal impact of our research in Section 4.

2. Dataset

We provide statistics regarding the geometric arrangement of the objects, such as the number of nodes and their class over all scene graphs. We also present illustrative figures of the partially reconstructed scenes to demonstrate how our dataset is constructed starting from the original ScanNet dataset.

2.1. Statistics

Table 1: Statistics regarding the number of nodes in the Spatial Commonsense Graph in the train/test split.

	# nodes				
	Mean	Std	Min	Max	Med
Train	59.54	30.06	3	177	55
Test	62.17	32.44	3	174	56

Total number of nodes in each scene. Table 1 presents some key statistics regarding the number of nodes (both *Object* nodes and *Concept* nodes) in our constructed *Spatial Commonsense Graph (SCG)* obtained from both the train and test split. We can observe that there is a very large variance in the number of nodes that compose the SCGs, with the smallest SCG having 3 nodes, and the largest one having about 170 nodes. This high variance is indeed a positive aspect as we can test small to wider environments populated by different numbers of objects. The varying number of objects is due to the process we use for generating each partial scene as described in section 5.1 of the main paper,

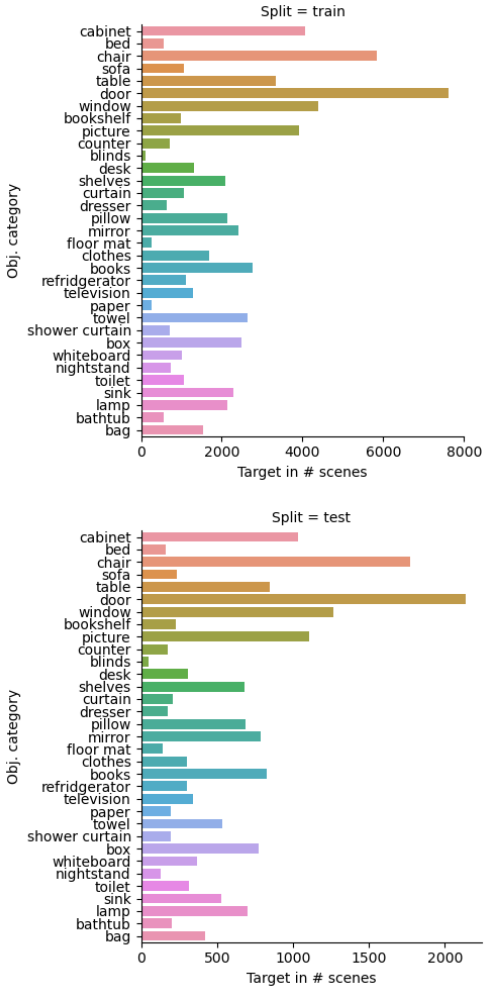

Figure 1: Numbers of scenes where an object appears as a target for the localisation. **Top:** Train set. **Bottom:** Test set.

i.e. by integrating RGB-D sequences of varying length. With a smaller number of RGB-D images, the reconstructed scene covers a smaller area with fewer objects, and vice versa.

Table 2: Statistics on the number of nodes in the Spatial Commonsense Graphs by node type, for both dataset partitions.

	# Object nodes					# AtLocation nodes					# UsedFor nodes				
	Mean	Std	Min	Max	Med	Mean	Std	Min	Max	Med	Mean	Std	Min	Max	Med
Train	12.77	8.58	3	74	11	22.10	9.67	0	52	22	24.67	14.64	0	86	22
Test	13.40	9.98	3	78	11	22.84	9.64	0	48	22	25.93	15.69	0	83	22

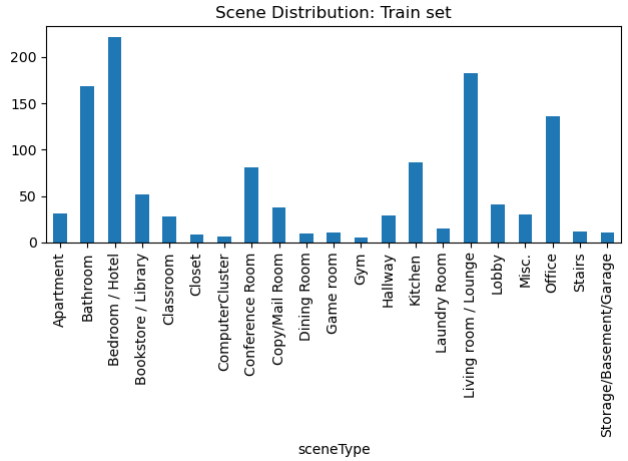

Figure 2: Room type distribution on the training set of ScanNet. The dataset mostly consists of bedrooms, bathrooms and living rooms. Some room types like closet and gym only have a few instances.

Another factor that contributes to the variance in the number of nodes is the addition of Concept nodes, whose number varies based on the relationship confidence queried from ConceptNet, as explained in the paper.

Node types in each scene. In Table 2 we distinguish the nodes in the SCG by their type, and we present the statistics for each one obtained from both the train and test split. Each type of node is defined by the type of edges that the node is linked to, where Object nodes are linked by Proximity Edges, AtLocation nodes are linked by AtLocation edges, and UsedFor nodes are linked by UsedFor edges. We observe that on average we have about 4 times more Concept nodes than the number of Object nodes. However, we notice that such a ratio does not scale to larger SCGs. The reason is that, while the Object nodes are duplicated for each object instance, the same is not true for the Concept nodes: for each Concept, only one such node exists in the graph, and multiple object nodes can be connected to it. Larger SCGs have many object nodes describing the same class, so the number of Concept Nodes does not increase linearly w.r.t. the number of object nodes. This behaviour can be observed in Fig. 4, where multiple Object nodes of the class "chair" are connected to the same Concept node.

Distribution of target objects. Fig. 1 shows the number of partial scenes where we estimate the position of the target object category, i.e. where an instance of that object category is in the unknown part of the scene. We can see that most target objects are of categories that tend to be present in all indoor environments, e.g. doors, windows, cabinets, chairs, and pictures. This type of class imbalance is also due to the room type imbalance in the original ScanNet dataset, as shown in Fig. 2 for the training split of ScanNet. Most of the reconstructed scenes are bedrooms, bathrooms, or living rooms, while other room types like closet or gym only appear a few times in the whole dataset. As such, objects that appear mostly in rooms of the minor categories will also appear less frequently as a target for our localisation task.

Geometrical arrangement of the objects. Table 3 shows the statistics on the geometrical arrangement of the objects in our dataset. While there is not much variance in the object's elevation (defined on the Z axis), the variance of the object position the horizontal plane, i.e. the (X, Y) plane, is large. This indicates that in an indoor environment, the main localisation challenge lies in finding the correct position on the (X, Y) plane.

Table 4 reports the statistics on the distance between the objects in the partial scenes computed on the *(X,Y)* plane. The high variance in the object position is reflected directly on the pairwise distances. This suggests that predicting the pairwise object distances stands for a similar difficulty as directly predicting the object position, but can better generalise to different reference systems.

Overall, these statistics show that our dataset of partial reconstructions contains very diverse scenes, with considerable variability regarding both object composition and their geometrical arrangement. Achieving a high Localisation Success Rate (LSR) on this dataset means that the method can generalise well in terms of both aspects described above.

2.2. Examples of partial scenes

Partial scenes of multiple levels of completeness. Fig. 3 shows three examples of partially reconstructed scenes. To obtain the partial reconstructions, we make use of the RGB-D sequences in ScanNet which are used to reconstruct the complete scene. From the full sequence, we extract a set of subsequences of different lengths, starting from the sequence with only the first frame, to the one containing all the frames. With these sub-sequences, the extracted Point Cloud Data (PCD) tends to cover a localised area of the scene, instead of having sparse reconstruction scattered around the whole scene. This allows us to simulate the use cases where a

Table 3: Statistics of the 3D positions of objects in our dataset of partial scenes for both dataset partitions. The X,Y plane is the floor of the room.

	X					Y					Z				
	Mean	Std	Min	Max	Med	Mean	Std	Min	Max	Med	Mean	Std	Min	Max	Med
Train	3.63	2.20	0.01	15.27	3.34	3.24	2.06	0.01	18.06	3.04	0.86	0.44	0.03	4.20	0.76
Test	3.40	1.92	0.06	11.95	3.18	3.27	1.99	0.01	10.84	2.93	0.83	0.42	0.04	3.09	0.73

Table 4: Statistics on the distances between objects for both dataset partitions

	Pairwise distances				
	Mean	Std	Min	Max	Med
Train	2.57	1.54	0.01	15.57	2.30
Test	2.57	1.49	0.05	10.03	2.32

visually enabled device visits only a limited part of the scene with the purpose of localising a target object in the unknown part.

Spatial Commonsense Graph from partial scenes. Fig. 4 shows a Spatial Commonsense Graph that is related to localising a sofa. The target node representing the sofa is highlighted in red, the object nodes are highlighted in green and the concept nodes are highlighted in pink. The edges' colours describe the relationship type, with *proximity* edges in black, the *AtLocation* in orange, and the *UsedFor* in blue. For the proximity edges we also show the pairwise distance between the objects.

We can see that some Concept nodes are connected to more than one object node, indicating a common usage or location, e.g. *sleeping* for both *sofa* and *pillow*, or *seat* for both *chair* and sofa.

This example demonstrates how much information can be added in the scene graph composed of only 5 object nodes, by integrating commonsense knowledge with 27 Concept Nodes. Note that the only criterion that we apply when retrieving Concept nodes from ConceptNet is to retain nodes with a weight score above a certain threshold (relation weight > 1). This explains why some Concept nodes may seem not closely related to our task, e.g. sofa *AtLocation* neighbour's house, chair *AtLocation* furniture_store.

3. Qualitative results

In this section, we show more qualitative results on the localisation with partial scenes. In particular, we show with real examples how the Localisation Module converts from pair-wise distance predictions to the position of the target object. Moreover, we show a comparison where we localise an object in a scene with different levels of scene completeness.

Additional qualitative results. Fig. 5 shows additional successful localisation with our **SCG Object Localiser**. On the left, we show the coloured reconstruction of the complete scene. On the right, we display the position predicted by our method, given a partial observation of the scene, highlighted with a yellow background. For all of the four examples, our approach was able to successfully estimate the position of the target object in the unseen part of the scene.

Demonstration of the Localisation Module. Fig. 6 demonstrates with an example how the pairwise distances predicted by our *Proximity Prediction Network* are converted by our *Localisation Module* to a single position in the unknown space. We define a cost function that is built with the pairwise distances of the Proximity edges, Eq. 5 of the main paper. The predicted distances between the target object and each observed object in the scene are visualised as a circle centred on each observed object, as shown at the left in Fig. 6, while the value of the defined cost function for all positions in the scene is visualised at the right in Fig. 6. The most yellow area indicates the lowest cost and the bluest the highest cost. The position where the object is most probably located is at the position with the lowest cost, i.e. the most yellow position. From the demonstrated cases in Fig. 6, we also observe that the predicted distances can be noisy with a certain degree of error. Methods with a high noise gain like Linear Least Squares, which are often used for multilateration, cannot be employed in this scenario. Differently, our approach can better tolerate erroneous distance measures.

The cases in the second and fourth rows show the presence of multiple local minima for solving the minimisation problem. Methods that search for a local minimum, e.g. gradient descent, non-linear Least Squares, may fail to converge to the correct solution due to a bad initialisation. Instead, our localisation module first divides the space into a coarse grid, where the cost of each cell is calculated. The one with the lowest cost is used as the starting point to initialise the solver for the minimisation method. This improves the chances of converging to the global minimum.

Localisation at different levels of scenes completeness. Figures 7 and 8 show examples of localisation at different scene completeness levels. In the first case (Fig. 7) we show the localisation of a sink in an apartment, while in the second case (Fig. 8) we show the localisation of a chair in a classroom. In general, as the scenes become more complete (left to right), the predicted position gets closer to the ground-truth position. The qualitative results coincide with the quantitative results presented in Fig. 5 of the main paper, i.e. the localisation error decreases and the Localisation

Success Rate (LSR) increases with the scene completeness.

Interestingly, where there are multiple instances of the target object category, i.e. the case in Fig. 8, we can see how the **SCG-OL** localises different instances of a chair based on the completeness of the scene. When the scene is mostly unobserved, i.e. the top-left case, our method places the chair behind a table in front of a whiteboard and manages to locate it correctly. In the top-right case with a more complete scene, **SCG-OL** correctly localises another chair on the side of the table. This is because the chair that was located in top-left case is now part of the SCG, thus not a valid target. In the bottom-left case, the previously predicted chairs are now part of the SCG, therefore no longer a valid target for the localisation. The network predicted a new position at the head of the table, although plausible, is considered a failure as there isn't a chair in the vicinity. In the bottom-right case, the model correctly predicts the most plausible position for a chair is between the two tables.

4. Ethical Discussion

Our new dataset has been built on top of the ScanNet dataset. Since ScanNet does not contain any human subject, by proxy, neither does our dataset of partial reconstructions.

Moreover, we proposed a novel graph modelling that enrich spatial scene representation with commonsense knowledge. Such formulation has a broader impact on the research community and foster methods for perception tasks that require spatial representation learning. In this paper, we demonstrated its effectiveness in terms of inferring the position of objects in unknown scenes, which by itself introduces potentials to advance applications, such as localisation service or suggestive layout design.

The proposed graph formulation aims to understand how we as humans model the arrangement of objects in rooms, and thus to learn a layout "profile". We note that this profile has been learned on thousands of different scenarios and therefore is too broad and generic to be used to negatively target specific individuals, races, or groups.

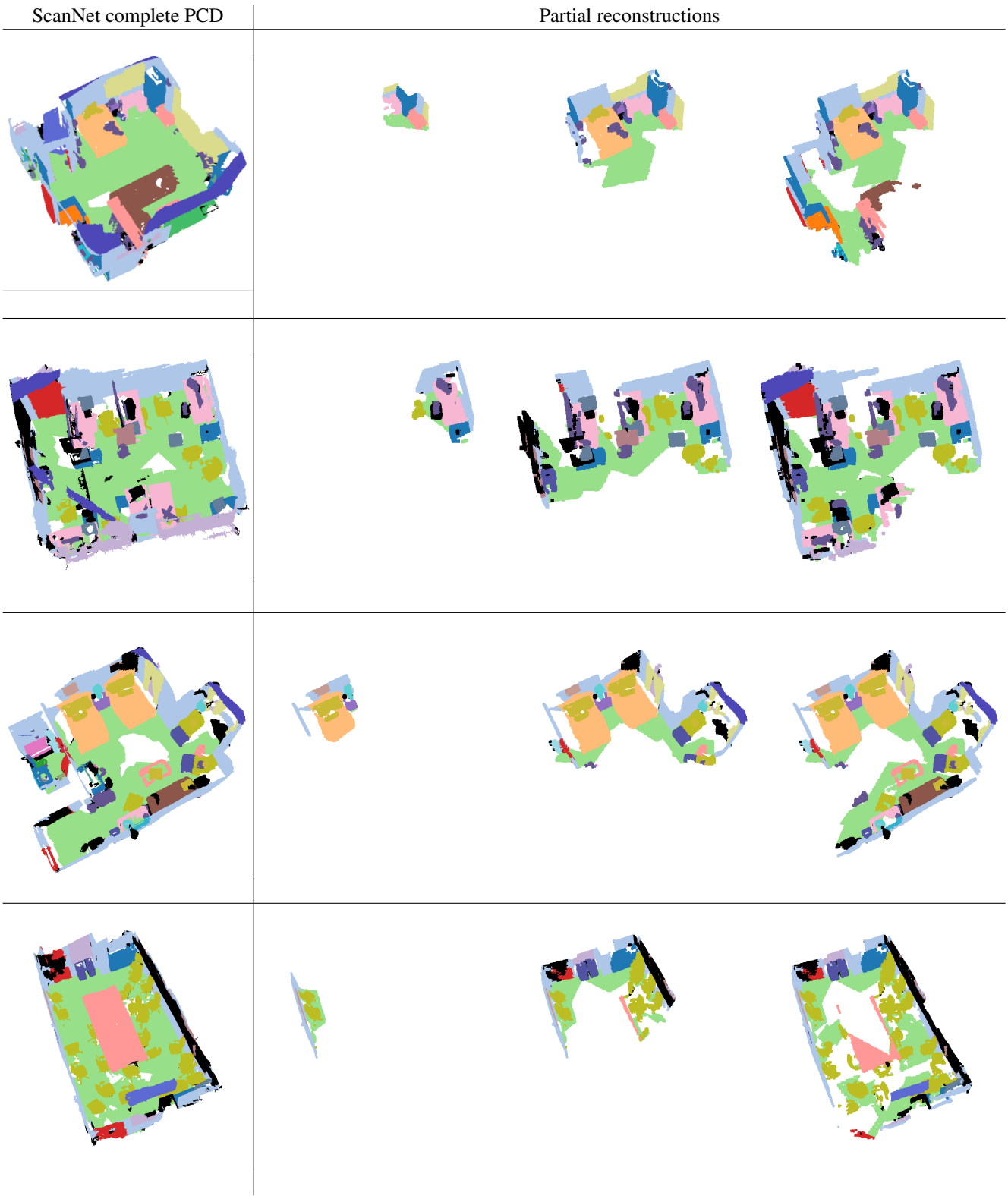

Figure 3: Examples of partially reconstructed scenes in our dataset. **Left:** The complete PCD of the room, semantically annotated with each colour indicating an object class. **Right:** Three partial reconstructions, obtained using a subset of the RGB-D sequence of increasing length (from left to right).

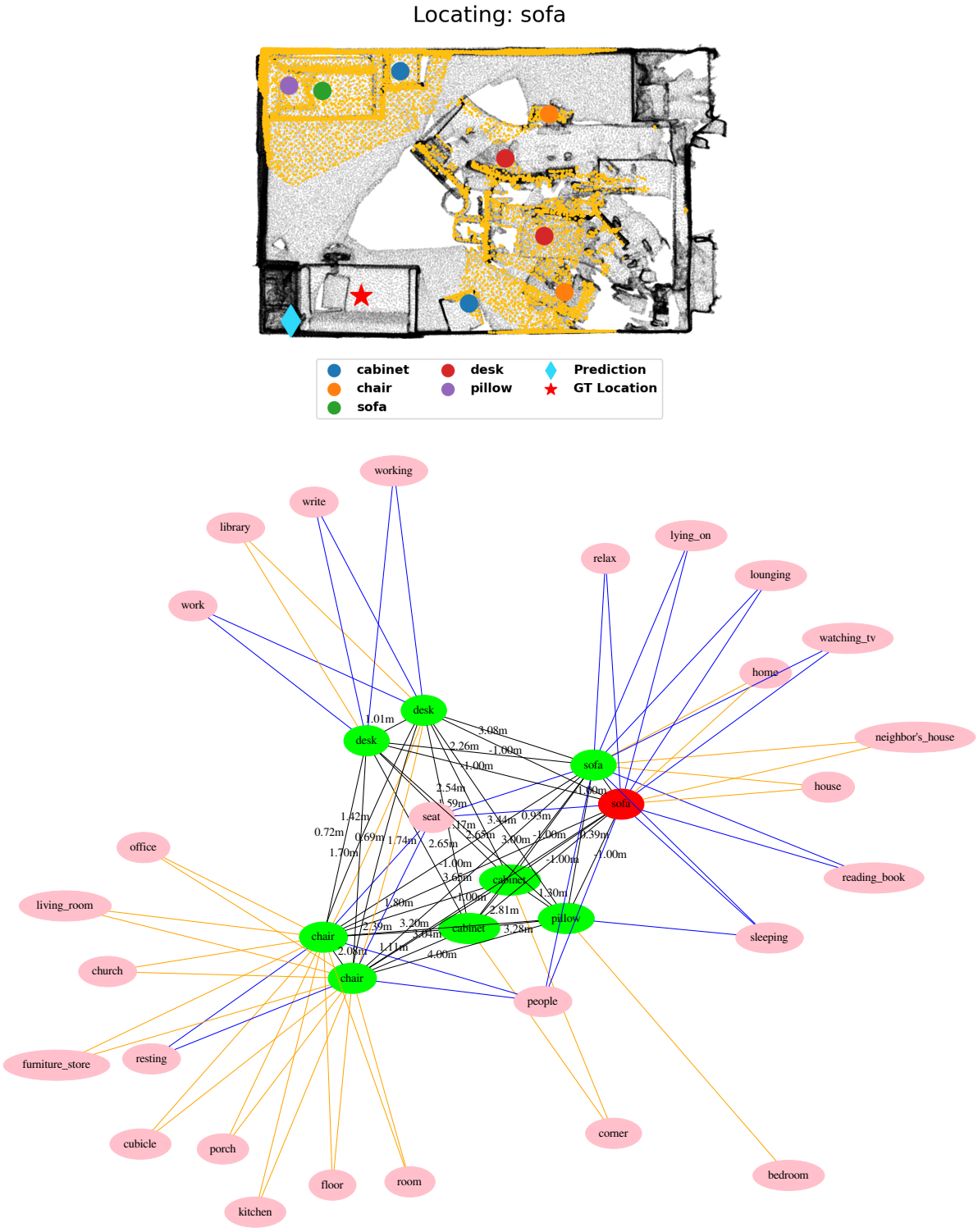

Figure 4: Example of the Spatial Commonsense Graph: **Top** - Image of the partial scene with highlighted the objects in the room. **Bottom** - Spatial Commonsense Graph for the top image. The target object is represented by the red node, the scene objects are the green nodes, and the concept nodes have a pink background. The colour of the edge distinguish the relationship type: orange are *AtLocation* edges, blue are *UsedFor* edges, and black are *Proximity* edges.

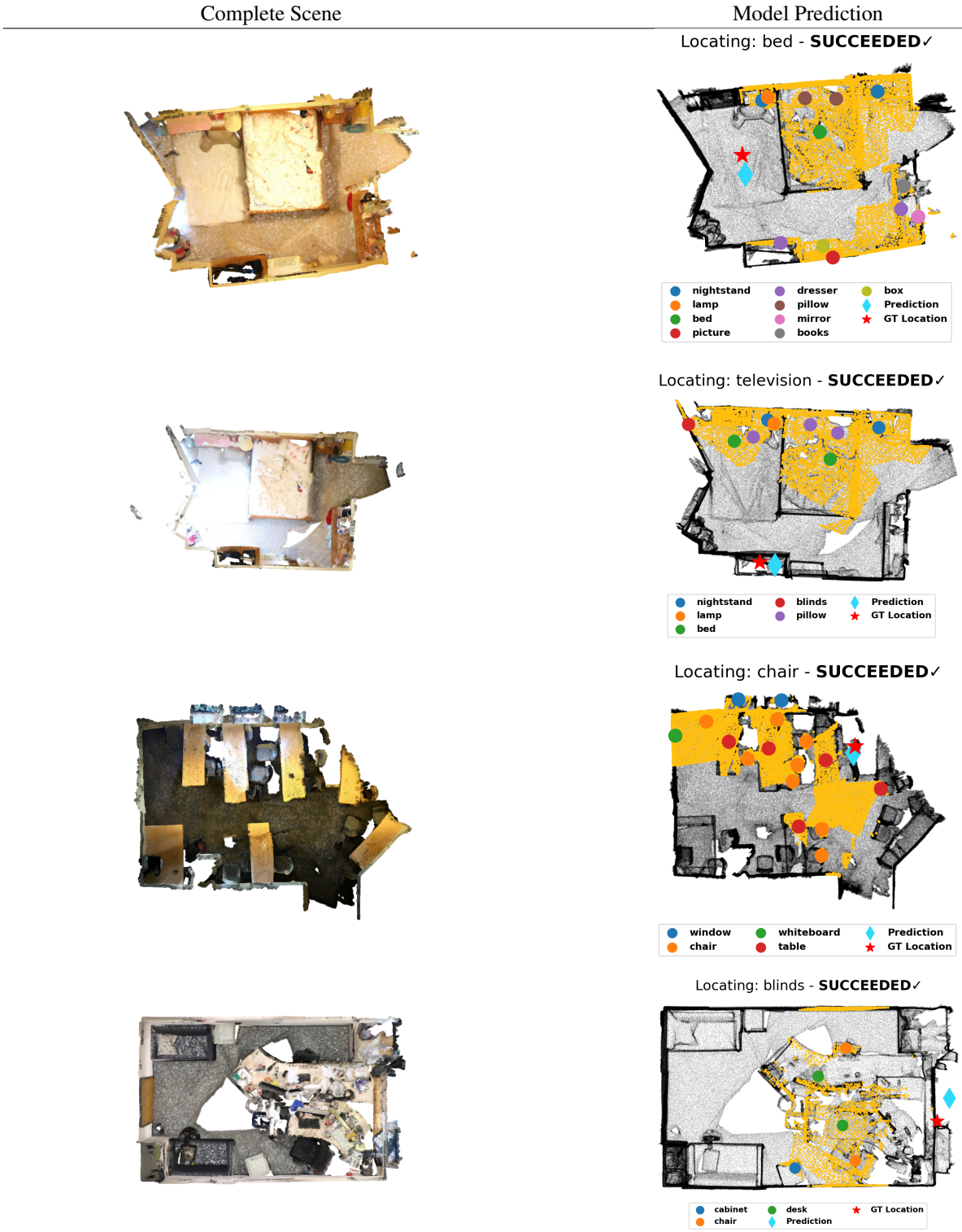

Figure 5: Successful localisation cases. The **left** column shows the complete scene from ScanNet. The **right** column shows the object nodes in the SCG and the position predicted by our **SCG Object Localiser** for the target object. The yellow areas indicate the visible part of the scene. Coloured dots show the objects in the SCG. The cyan diamond indicates the predicted position and, the red start is the ground-truth position of the target instance closest to the predicted position.

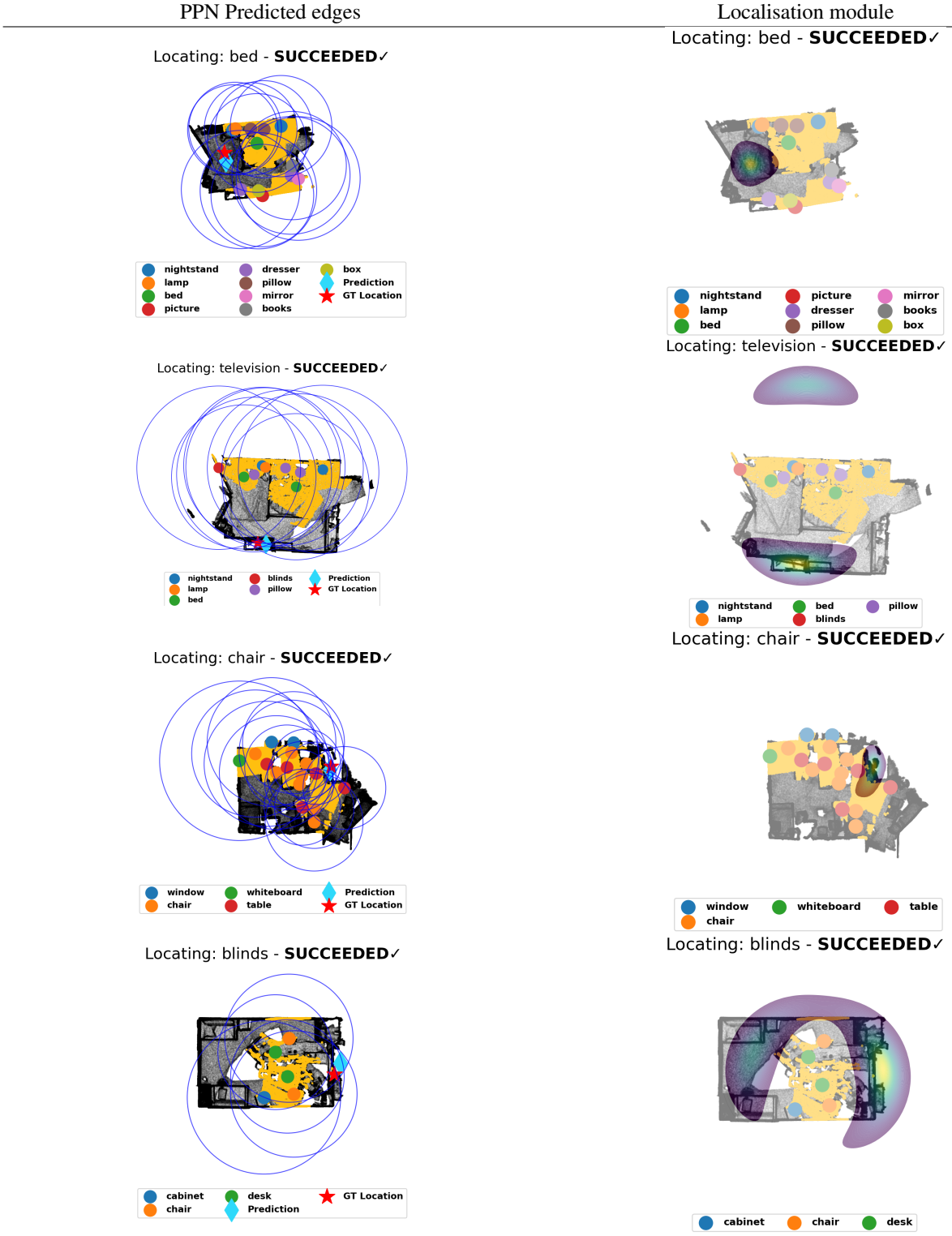

Figure 6: Effect of the Localisation Module, on the same examples as fig. 5. The **left** column shows the edges predicted by our *Proximity Prediction Network*, show in blue. The **right** column shows the cost defined in the *Localisation Module*, the areas where most edges overlap have a lower cost and are displayed in yellow, while the areas with higher cost are displayed in blue. Blank areas have a cost above the threshold set for the visualisation.

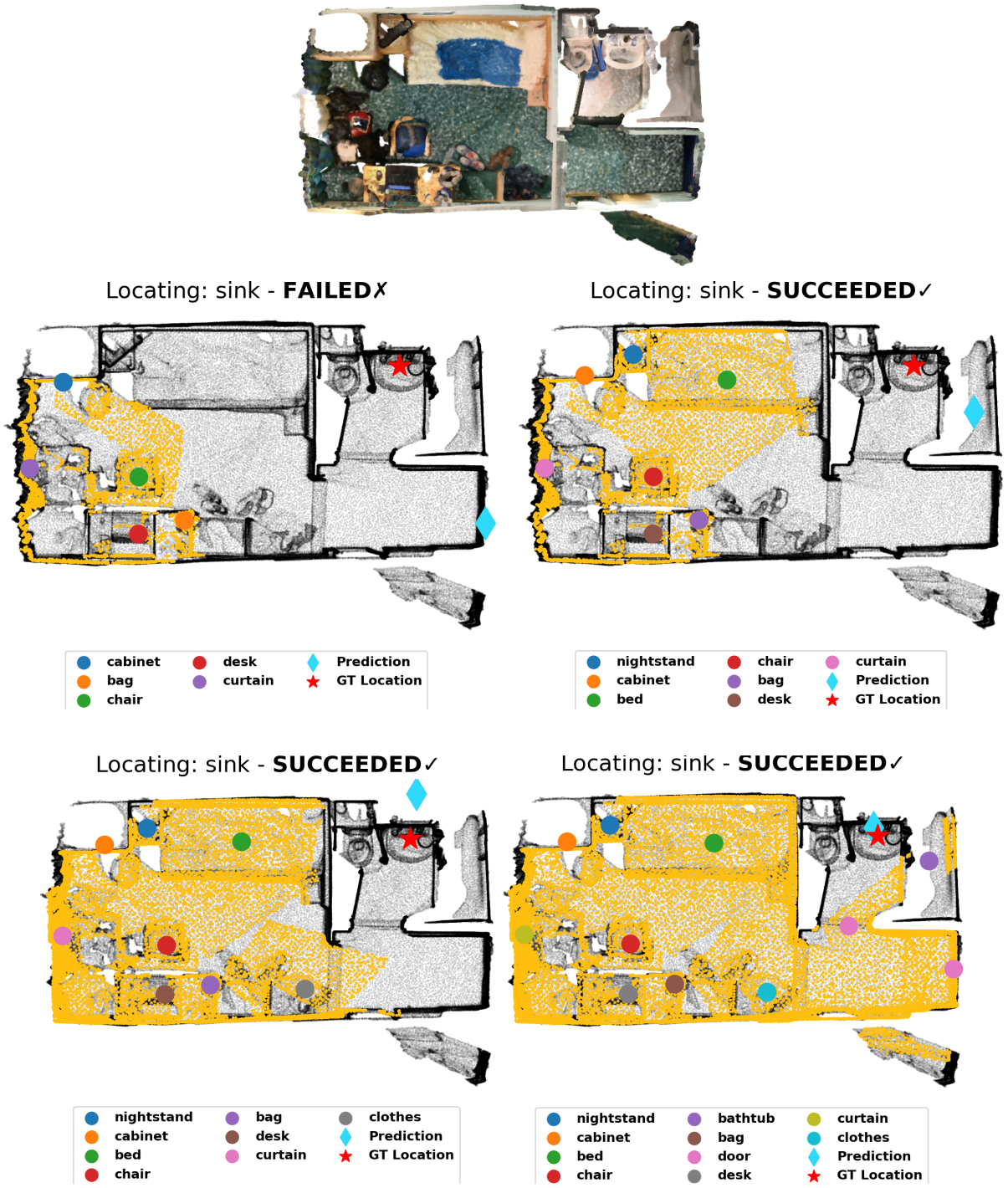

Figure 7: **Top** Complete scene of a small apartment. **Middle and Bottom** Localisation of the sink at different completeness levels. The localisation accuracy increases as the scene becomes more and more complete.

Figure 8: **Top** Complete scene of a classroom. **Middle and Bottom** Localisation of a chair at different completeness levels. Note that the red star indicates the ground-truth instance *closest* to the prediction. As the scene becomes more and more complete, the method is able to correctly adapt to changes in the SCG.