Learning 3D Semantic Scene Graphs from 3D Indoor Reconstructions

Supplementary Material

This document supplements our paper with additional details, visualization and results. Section 1 contains more thorough statistics and characteristics regarding the 3DSSG dataset such as a visualization of the WordNet hierarchy, 2D graphs as well as the annotation interfaces. Section 2 gives additional information about the proposed method as well as more graph prediction results while Section 3 focuses on retrieval.

1. 3DSSG Dataset

Statistics In this paragraph, we present further data statistics. Fig. 1 and 2 show the number of relationships per 3D scan and object instance. The corresponding histograms for object attributes are in Fig. 3 and 4. Fig. 11, 12, 10, 13 show the most frequent object, predicate, attribute and affordance occurrences extracted from our ground truth graphs. Fig. 14 highlight some of the most common semantic connections present in the dataset. Further, a few example of object instances and the annotated attributes can be found in Fig. 5. These statistics show that the scene graphs in 3DSSG are not only semantically rich but also very dense.

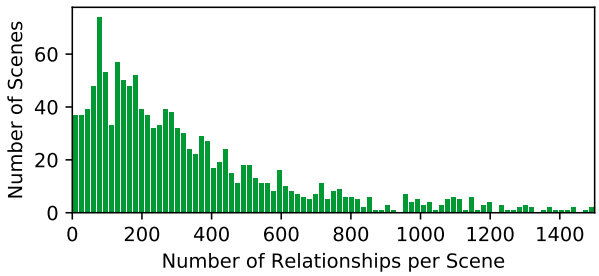

Figure 1. Histogram of scenes in 3DSSG and corresponding number of relationships

WordNet Graphs Fig. 17 shows a graphical visualization of the WordNet hierarchy of classes, which we use to extract the per-node hierarchy of labels c. Colored nodes show class labels from the annotation set, while white nodes are abstract representations that are not part of the original label set. For an instance annotated as `chair`, the hierarchical label c would be $c = \{$`chair, seat, furniture, ..., entity`$\}$.

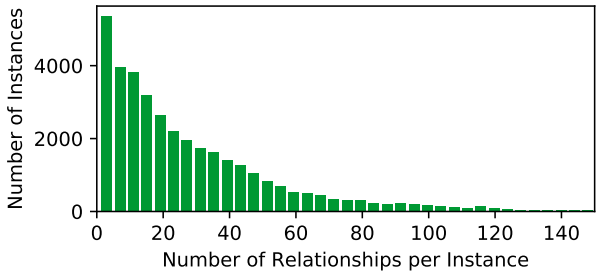

Figure 2. Histogram of object instances in 3DSSG and corresponding number of relationships

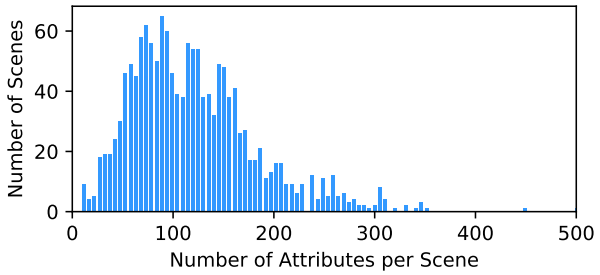

Figure 3. Histogram of scenes in 3DSSG and corresponding number of attributes

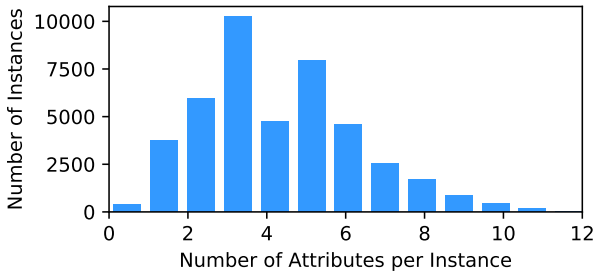

Figure 4. Histogram of object instances in 3DSSG and corresponding number of attributes

2D Graphs: Depth and Mask Fig. 18 illustrates 2D scene graphs of the 3DSSG dataset, which are obtained via rendering the 3D scene. We show that, while 2D scene graph datasets currently available [1] only have bounding box annotations, we also provide depth and dense seman-

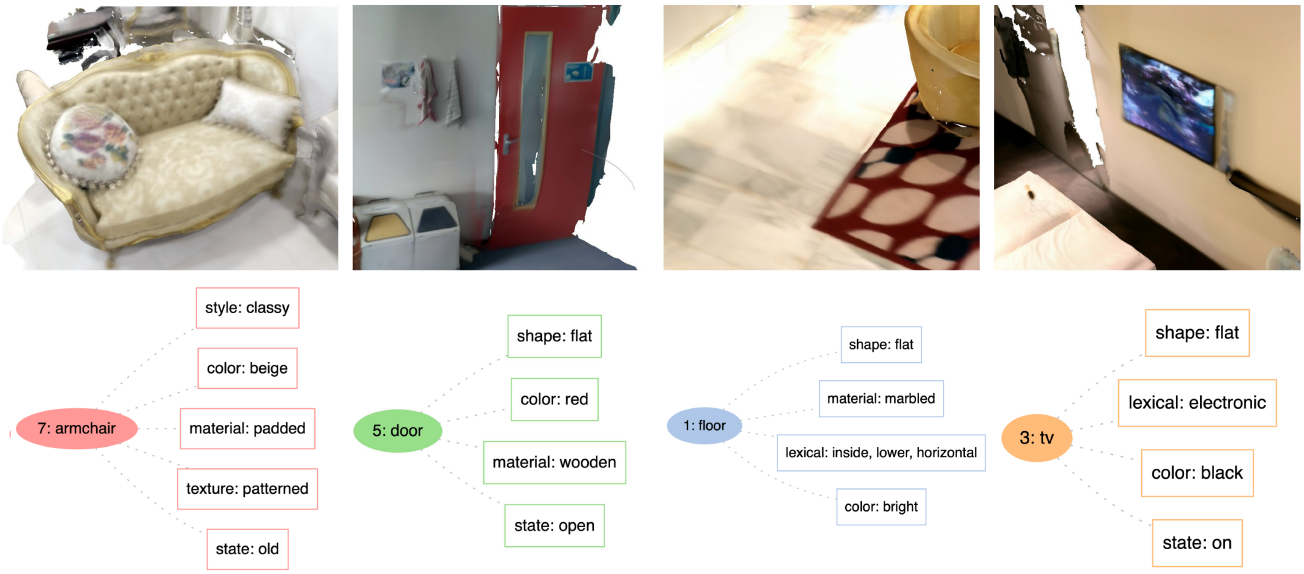

Figure 5. Example object instances (top) and their corresponding attributes (below).

tic instance masks, which we believe are relevant for the future, to explore alternative ways for the graph prediction and other underlying tasks.

State Changes While static instance attributes such as the color or material of an object do not change, dynamic instance attributes (*e.g.* `on / off`, `open / closed`) can change over time. Interestingly, these state properties are closely connected to human interaction and could potentially give information about activities that might have happened in a particular 3D space (see Fig. 6 for examples).

Annotation interfaces Fig. 7, 8 and 9 are screenshots of the user interfaces we used for the annotation, namely attribute and support relationships (for binary and semantic annotation and verification). Please note that semantic support annotations are done after the binary annotation, since they build upon the ground truth support pairs.

2. Graph Prediction

Implementation details For the feature extraction of nodes and edges, we adopt two standard PointNet architectures. The input points to ObjPointNet have three channels (x, y, z), while the RelPointNet inputs have four $(x, y, z, \mathcal{M}_{i,j})$. The size of the final features (per node and per edge) is 256. For the baseline model, the object and relationship predictors have namely three fully connected layers followed by batch norm and relu. The GCN encompasses $l = 5$ layers, where $g_1(\cdot)$ and $g_2(\cdot)$ are composed of a linear layer followed by a relu activation. The class prediction MLPs consists of 3 linear layers with batch normalization and relu activation. We set $\lambda_{obj} = 0.1$. For the per-class binary classification loss, α_t is set to 0.25. We use an Adam optimizer with a learning rate of 10^{-4}.

Data Processing Since our provided scene graphs are extremely dense (see statistics) a pre-processing and filtering of the ground truth graph data was required. We split the original graphs into smaller subgraphs of 4–9 nodes. We further consider only a subset of the relationships. Similarly, object instances with uncommon classes are filtered. In summary, in our experiments we use 160 different object classes and 26 relationships. For reproducibility our splits are made publicly available.

Qualitative Results Fig. 15 and 16 show more qualitative semantic scene graph results using our proposed network architecture. Please note that Fig. 16 shows rendered 2D subgraphs. Miss-classifications are reasonable (`desk` vs. `computer desk`, `object` vs. `toilet brush` or `picture` vs. `tv`). `Bright green` edges are correctly predicted relationships between two notes; `dark green` edges are partially correct (a subset of the edges is correctly predicted and the rest is either missing or miss-classified), `bright blue` edges are false positives (and often semantically correct), `red` edges are completely miss-classified, while `gray` edges are missing in the prediction (predicted as `none` while there exists an edge in the ground truth). In all edge and node predictions we show the prediction together with the ground truth data in brackets *e.g.* `shower`

Figure 6. 3 example scenes at two different time steps where human activity possibly have changed object states. *Left:* someone might have used the toilet (toilet seat is `down / up`), *Center:* someone might has cleaned this room (`desk` and `floor` are `messy / tidy`), *Right:* someone might have slept in the bed (`bed` is `tidy / messy`).

`curtain (GT: curtain)` in Fig. 6 (main paper) and Fig. 15 and 16.

Class-Agnostic Instance Segmentation – the input of our graph prediction network – is taken directly from the dense 3D ground truth instance segmentation from 3RScan. Since we only use the segmentation information (and not the class labels itself) we decided to name it *class-agnostic* instance segmentation. In theory, every 3D geometric segmentation method that is able to segment separate instances could be used as the input of our method.

3. Scene Retrieval

In Tbl. 3 of the main paper 3D-3D scene retrieval results are reported using ground truth graphs. While this gives us an upper bound for the scene retrieval task using 3D semantic scene graphs it also detects semantic changes. This interesting side effect is visualized in Fig. 19. Since only the parts of the graphs (nodes and edges) that could not be matched in the retrieval are visualized, it is easy to identify changes. In the upper example: 1.) the chair (8) pushed in a direction away from the bed (not `close by` anymore) and 2.) a pillow (20) was moved closer to pillow (22) and pillow (21). In the lower example 1.) a bag was added (`close by` cushion (8) and cushion (10)) and 2.) the purple cushion (10) was moved from couch (3) to couch (2). In these terms, the amount of change in a scene is the reverse of its ground truth similarity.

References

[1] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision (IJCV)*, 2017. 1

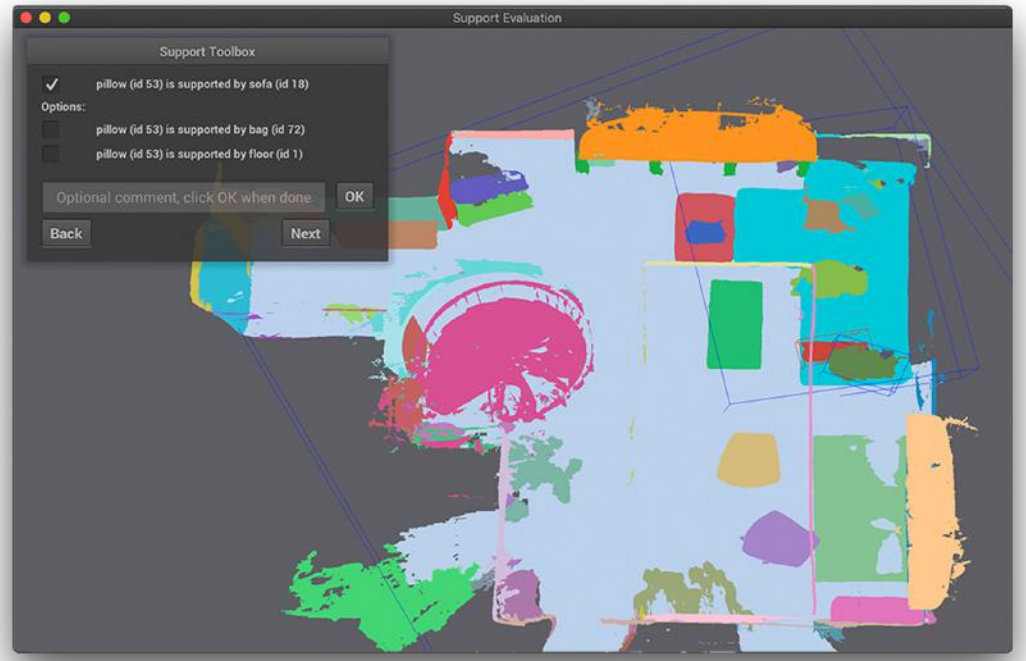

Figure 7. Interface for binary support annotation

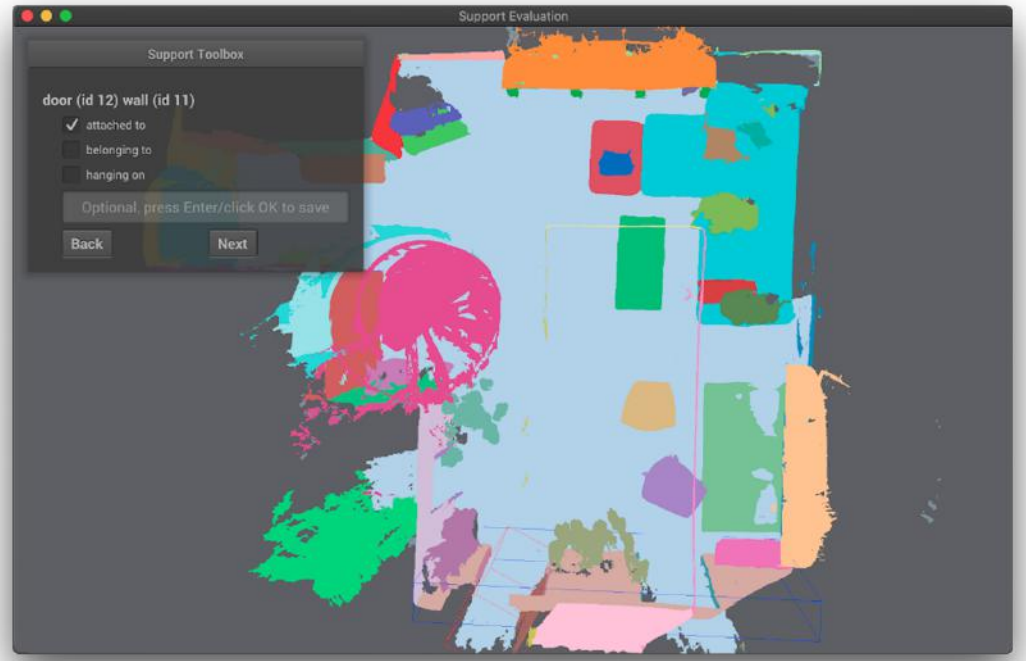

Figure 8. Interface for semantic relationship annotation

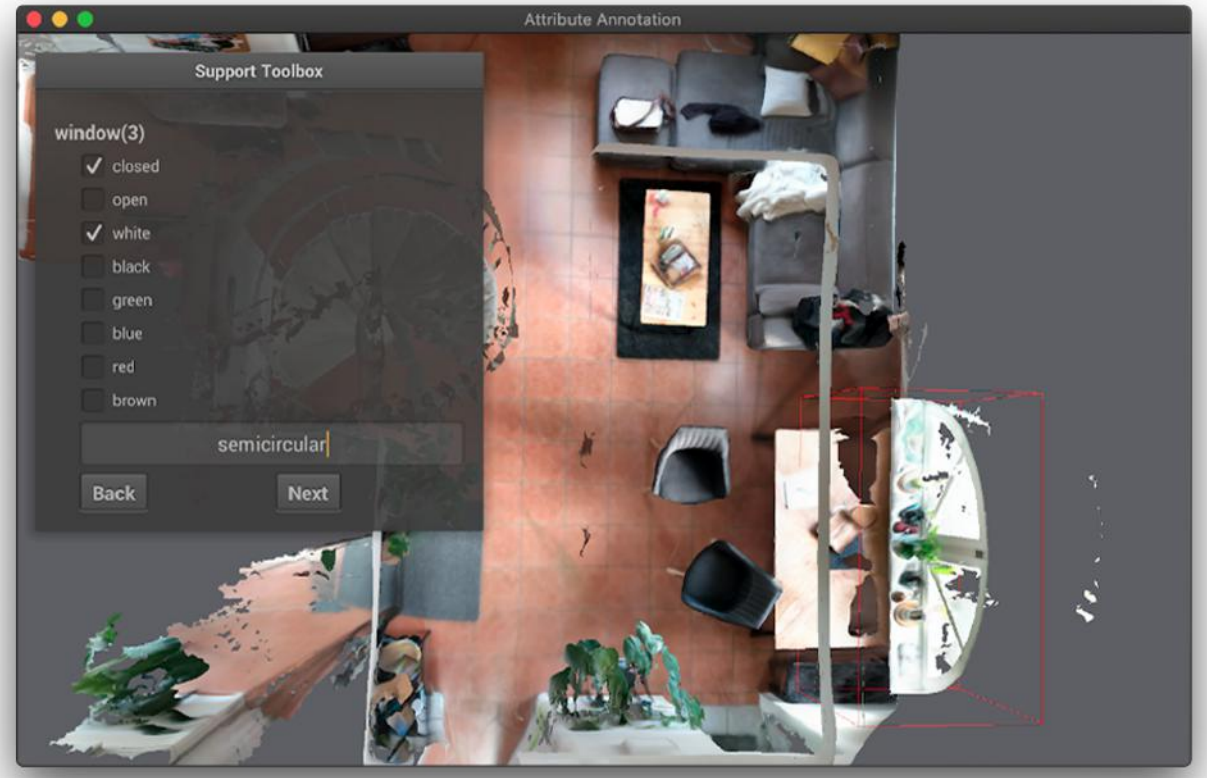

Figure 9. Annotation Interface for attribute annotation

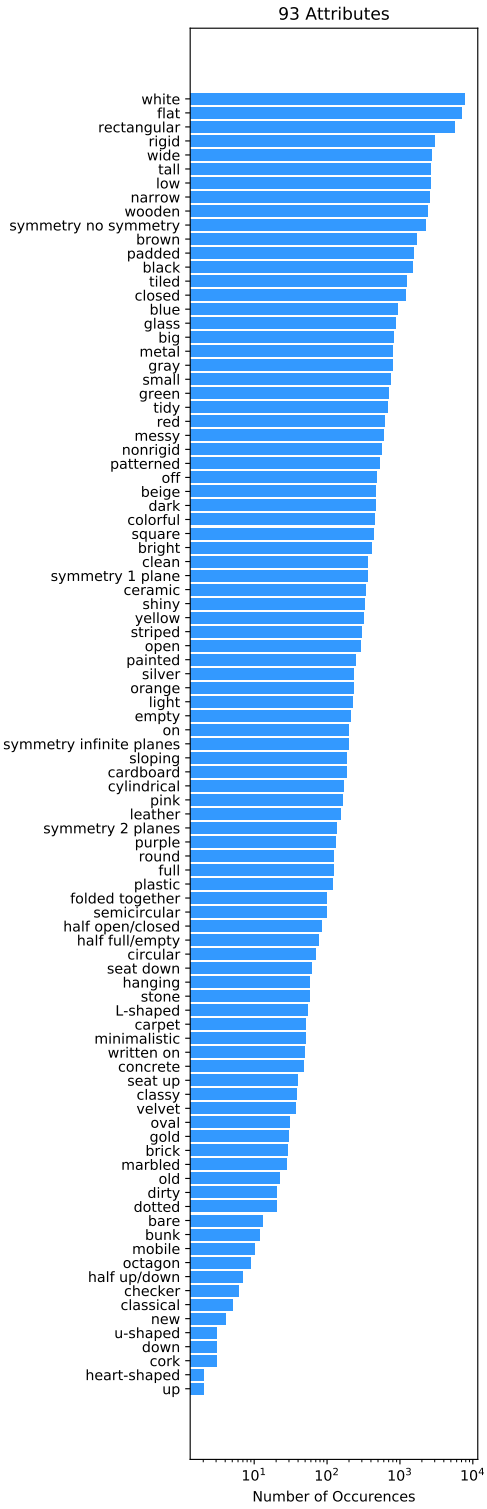

Figure 10. Attributes in the 3DSSG dataset, sorted by occurrence, presented in logarithmic scale

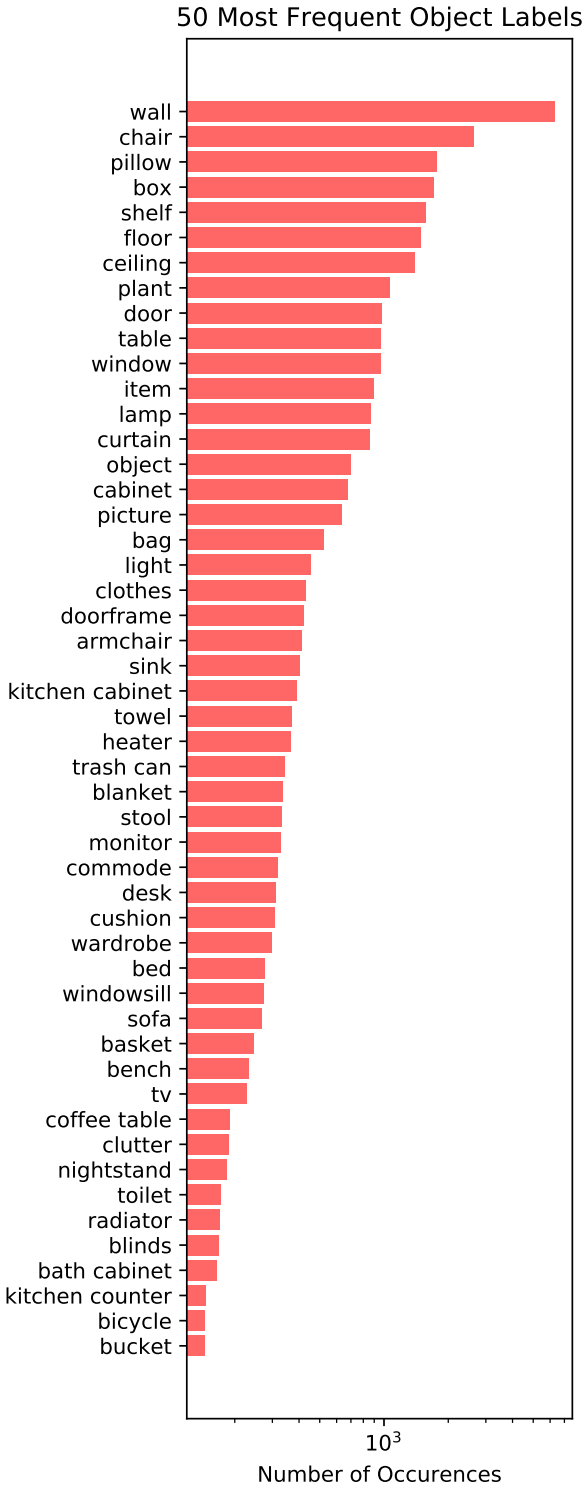

Figure 11. Most frequent (Top-50) object classes used for the training of the Scene Graph Prediction Network, sorted by occurrence, presented in logarithmic scale

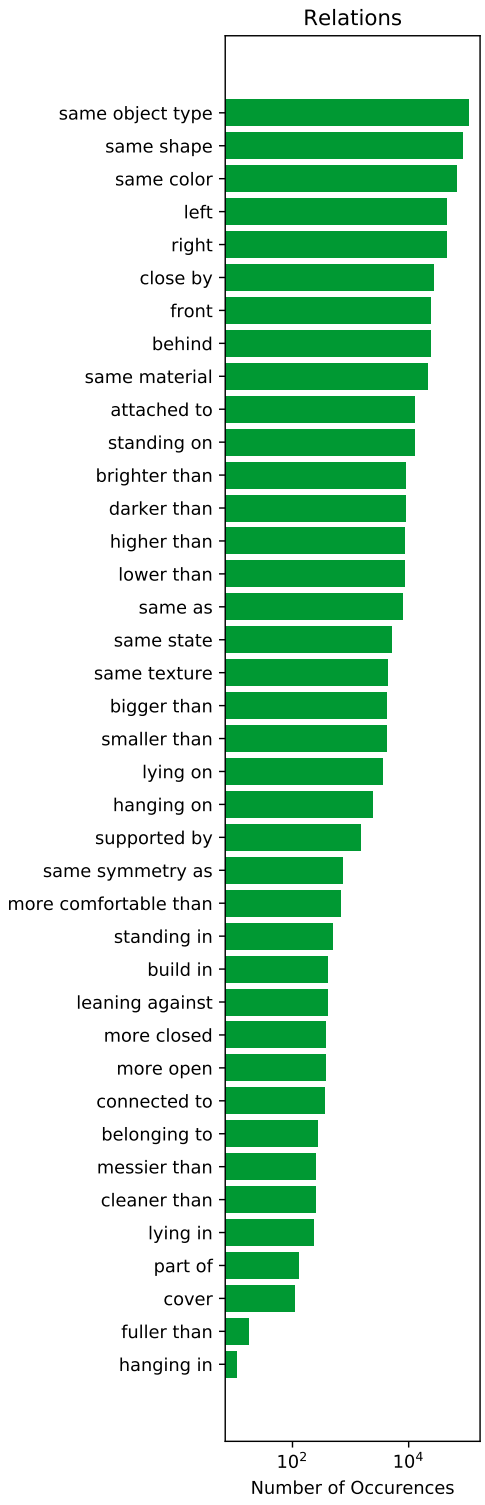

Figure 12. Predicate classes, sorted by occurrence, presented in logarithmic scale

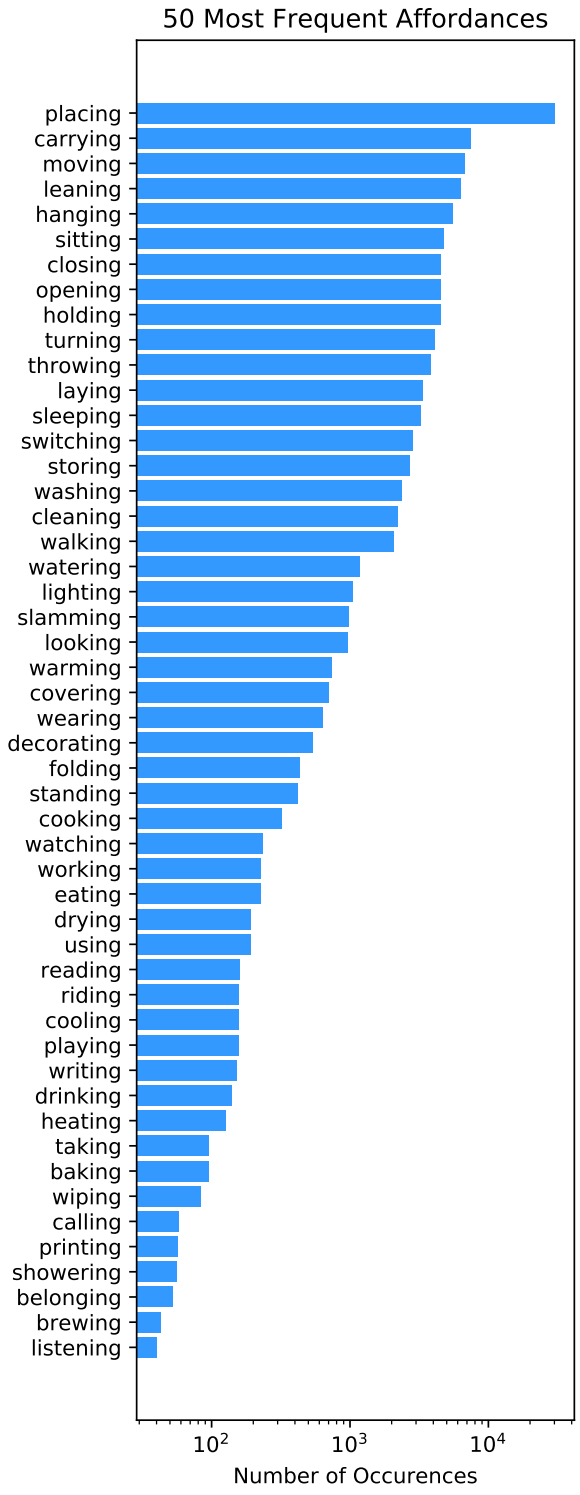

Figure 13. Simplified affordances in the 3DSSG dataset, sorted by occurrence, presented in logarithmic scale. Please note that for visualization purposes nouns and prepositions are removed such that *e.g.* `hanging in` or `hanging on` are combined into `hanging`.

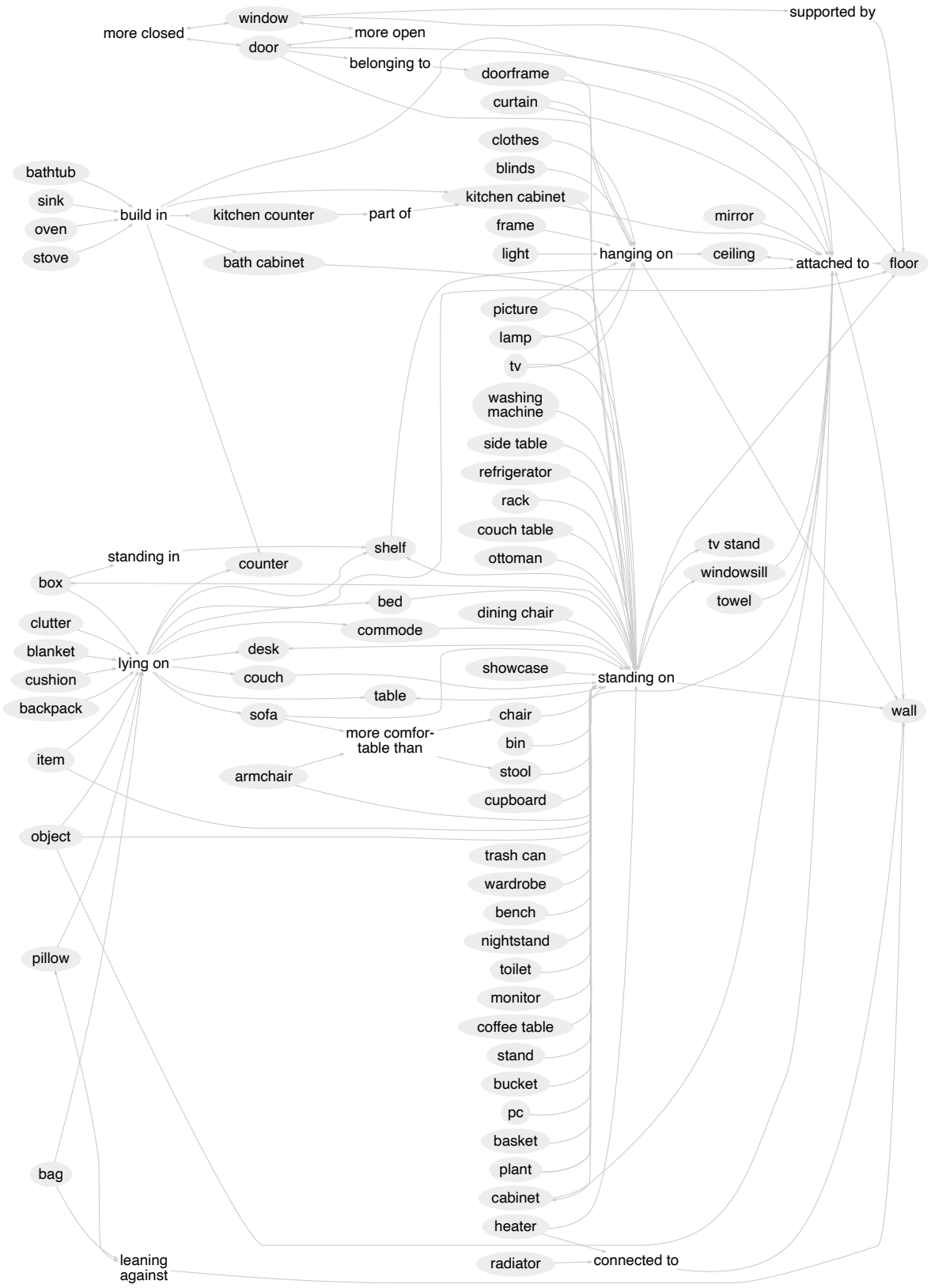

Figure 14. Simplified most frequent semantic `(object, predicate, subject)` tuples with more than 50 occurrences in the 3DSSG dataset. Please note that for simplification purposes proximity relationships and most comparative relationships are filtered in this graph.

Figure 15. Qualitative results of our scene graph prediction model (best viewed in the digital file).

Figure 16. Qualitative results of our scene graph prediction model rendered to 2D (best viewed in the digital file).

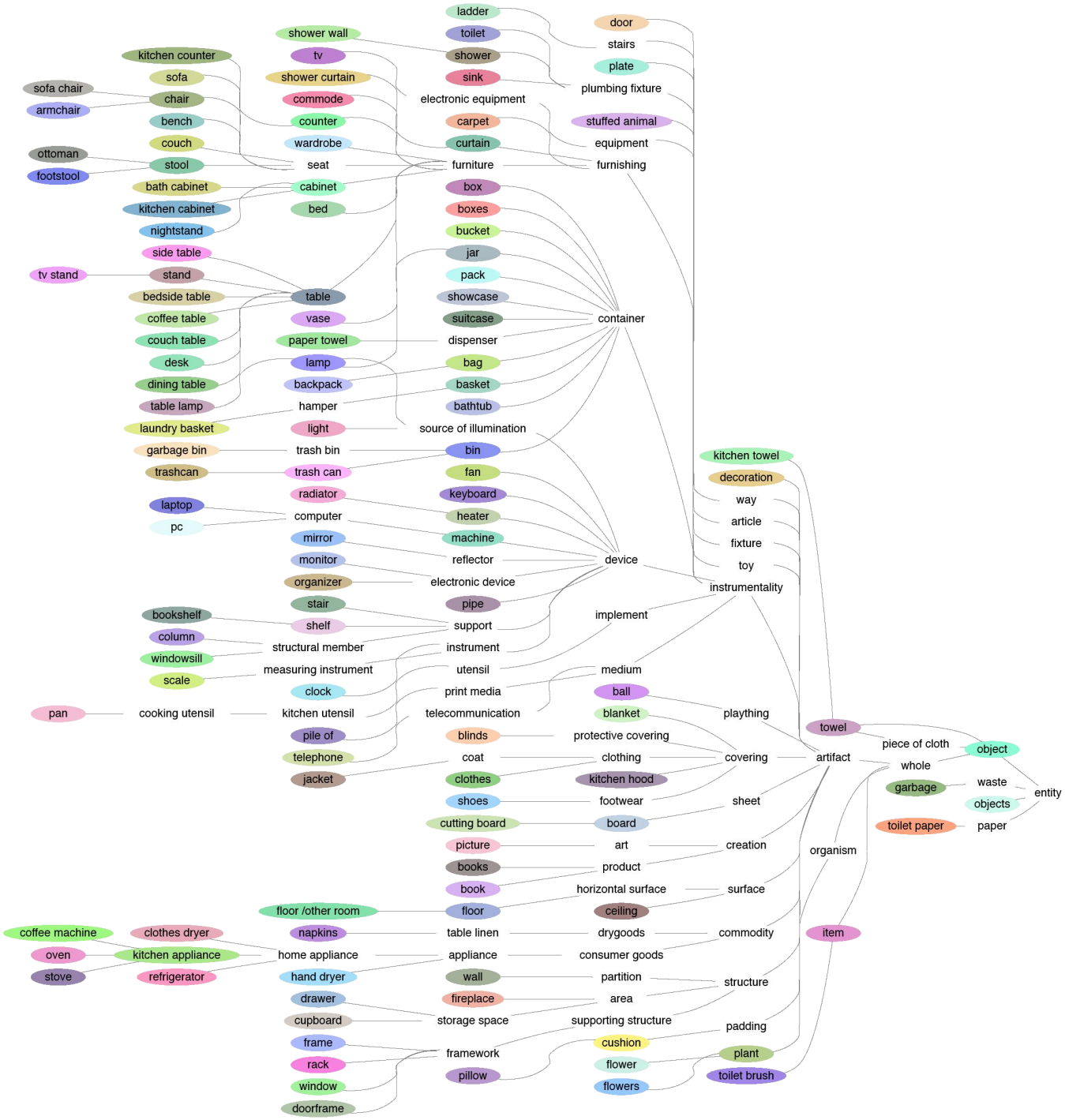

Figure 17. Graphical visualization of the hypernym / lexical relationships on a (bigger) subset of classes from 3RScan (see Fig. 3 in main paper).

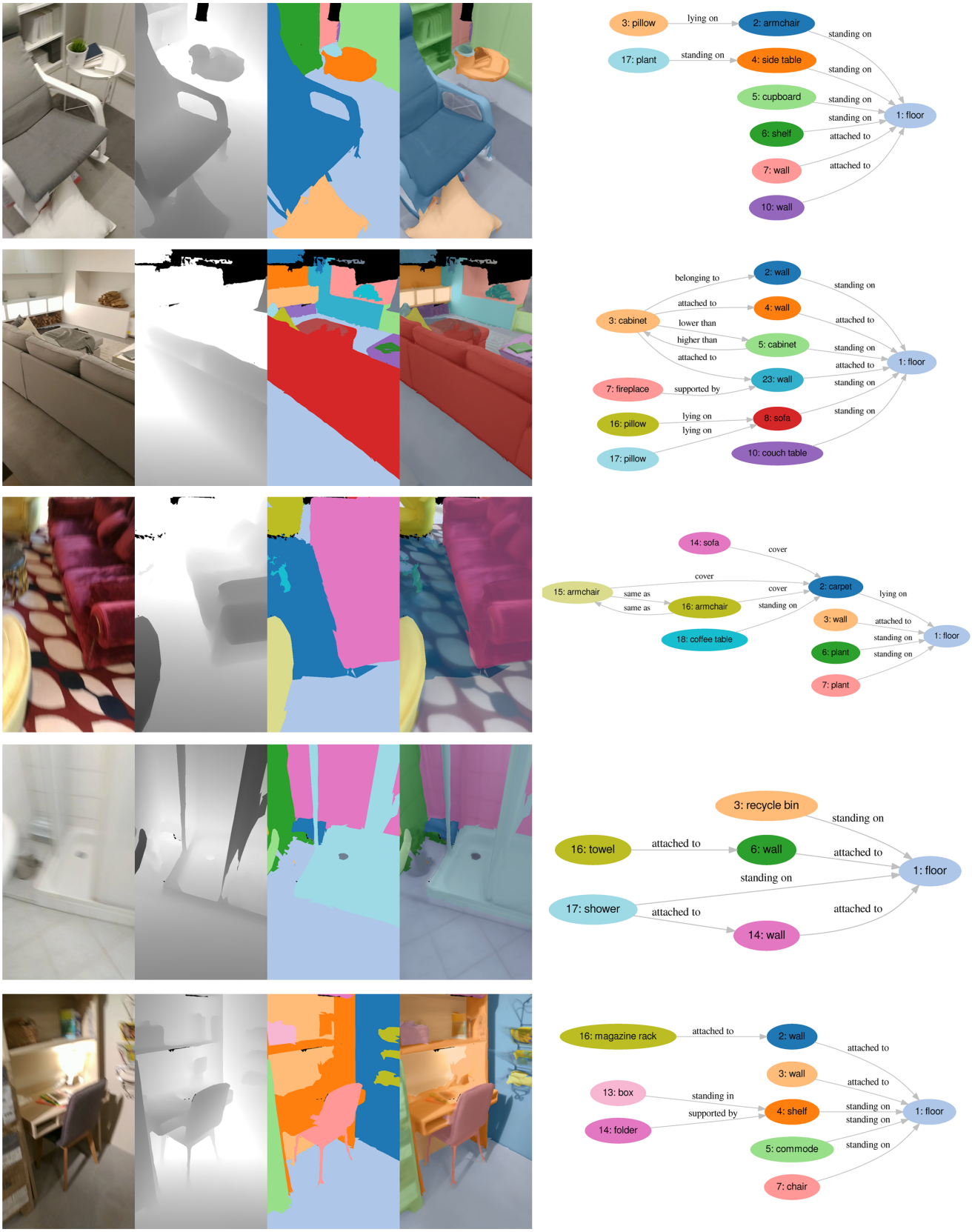

Figure 18. Rendered 2D graphs with a small subset of the relationships from our newly created 3D semantic scene graph dataset 3DSSG . From left to right: RGB image, rendered depth, rendered dense semantic instance segmentation, dense semantic instance segmentation on textured model, 2D semantic scene graph.

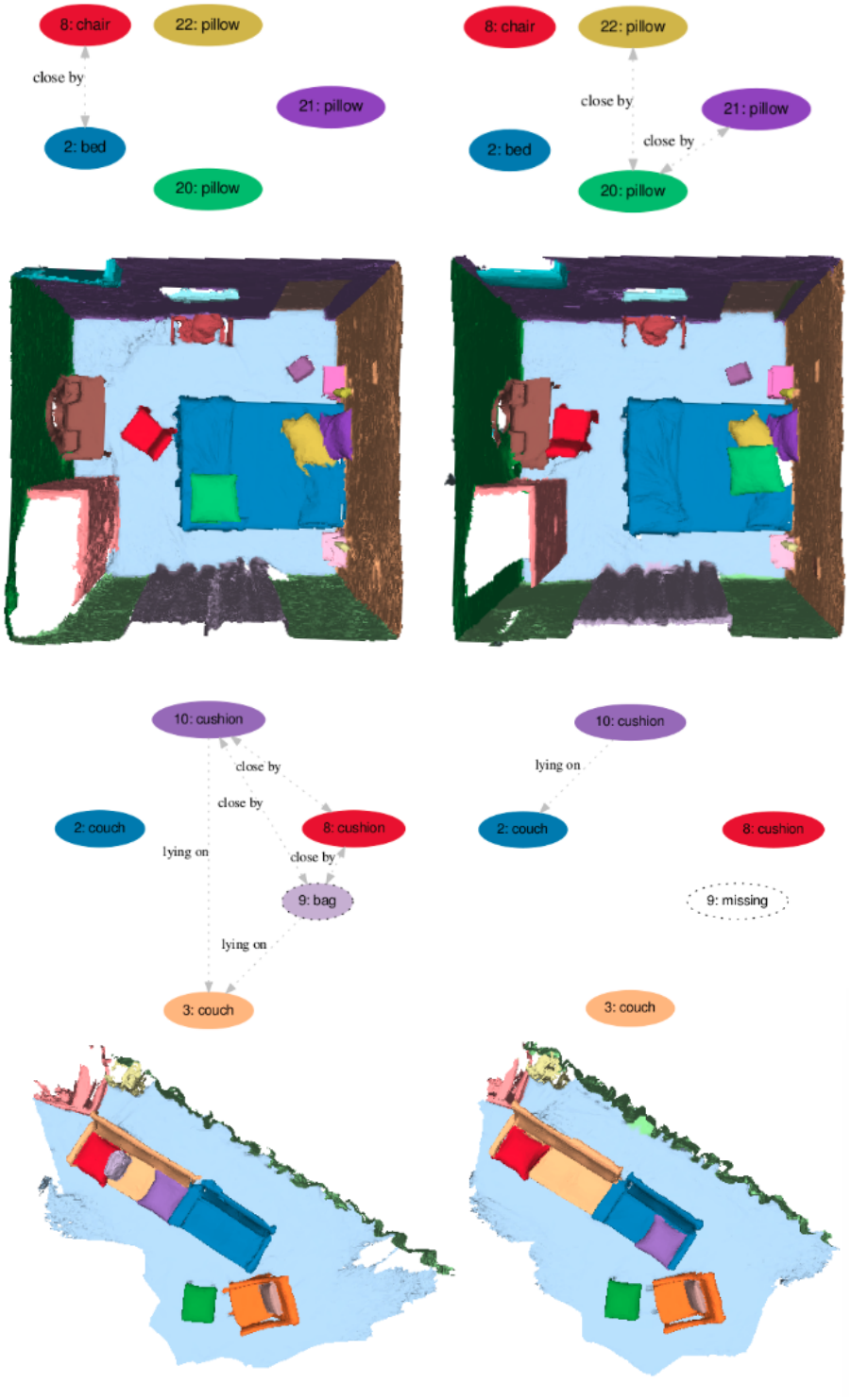

Figure 19. A byproduct of our scene retrieval is semantic change detection: Changed (added or removed) relationships and involved objects on two example scenes. Since we only show changes, all relationships *e.g.* between pillows and the bed are not visualized.