FROSS: Faster-than-Real-Time Online 3D Semantic Scene Graph Generation from RGB-D Images

Supplementary Material

6. Detailed Evaluation Metric

The evaluation procedure in this paper follows closely with Wu [35] to ensure a fair comparison. The only difference is the exclusion of the 'none' relationship category, as FROSS does not predict it. Wu [35] also provided results evaluated under this protocol in their publicly released code.

It is important to clarify that the *recall@1* metric stated in [35] differs from the conventional *recall@N* metric used in 2D SG evaluation [12, 36, 38]. In standard *recall@N* evaluation, only the top-N relationship triplets with the highest confidence scores are considered. In contrast, the *recall@1* metric employed in our work and [35] focuses solely on the predicted class labels within a detected triplet. Specifically, for a detected triplet in which both the subject and object match ground truth objects, only the predicted class labels for the subject, object, and predicate with the highest confidence scores are considered. Notably, as our approach does not impose a restriction on the number of detected relationship triplets, the *recall@1* metric in [35] is conceptually more aligned with *recall@*∞ with graph constraints [38]. To mitigate potential confusion, we refer to this metric simply as *recall* throughout our work.

Additionally, the predicate recall metric used in this study does not fully correspond to the conventional predicate classification (**PredCls**) [36] setting, as no ground truth objects are provided.

7. Additional Experimental Results

7.1. Object and Predicate Performance per Class

The per-class performance comparison of FROSS and other baselines is presented in Tables 6 and 7. In addition, FROSS's per-class object and predicate performance on the proposed ReplicaSSG dataset is presented in Table 8.

FROSS excels in detecting object classes that rely heavily on visual information, particularly those with similar geometric structures, such as *bookshelf*, *counter*, *desk*, *picture*, *refrigerator*, *shower curtain*, and *window*. These objects are often box-like or flat. FROSS's ability to capture complex visual features leads to significantly higher performance in both object recall and mean recall.

FROSS's predicate performance is significantly affected by class imbalance, excelling in relationship classes such as *attached to*, *build in*, and *standing on*, while performing poorly on others. Despite retaining only the top seven most frequent relationships, the 3DSSG dataset still exhibits an extreme imbalance, with the top two classes occurring at substantially higher frequencies than the others [35]. While addressing this issue could potentially enhance FROSS's performance, we leave it as future work, as class imbalance is not the primary focus of this research.

Table 6. Per-class performance comparison of 3D SSG generation methods on 3DSSG for object recall (%). The best and second-best results are highlighted in **red**, and blue, respectively.

Method	bath.	bed	bkshf.	cab.	chair	cntr.	curt.	desk	door	floor	ofurn.	pic.	refri.	show.	sink	sofa	table.	toil.	wall	wind.	mean
IMP [36]	0.0	66.7	0.0	38.1	45.3	0.0	47.7	0.0	8.1	95.1	19.9	2.3	0.0	0.0	20.0	47.4	48.5	66.7	77.0	17.9	30.0
VGfM [9]	0.0	66.7	0.0	34.6	49.4	0.0	48.6	4.2	19.8	95.7	14.1	1.1	0.0	0.0	23.3	57.9	56.9	63.0	**78.0**	17.9	31.6
3DSSG [31]	25.0	66.7	0.0	20.0	51.0	25.8	50.5	0.0	47.7	91.4	14.7	3.4	22.2	14.3	25.0	47.4	42.5	25.9	51.9	13.1	31.9
SGFN [34]	75.0	33.3	0.0	50.8	63.6	19.4	40.5	8.3	38.7	**96.9**	23.0	11.4	11.1	0.0	38.3	55.3	62.3	51.9	73.0	13.1	38.3
Wu [35]	75.0	**100.0**	0.0	50.4	**65.6**	19.4	45.9	12.5	34.2	**96.9**	25.1	5.7	0.0	14.3	38.3	57.9	59.9	66.7	76.1	15.5	43.0
FROSS (Ours)	**100.0**	83.3	**28.6**	**56.1**	64.8	**67.7**	**73.0**	**29.2**	**73.3**	91.5	**40.3**	**41.9**	**50.0**	**42.9**	**73.3**	**73.7**	**68.2**	**100.0**	60.9	**57.5**	**63.8**

Table 7. Per-class performance comparison of 3D SSG generation methods on 3DSSG for predicate recall (%). The best and second-best results are highlighted in **red**, and blue, respectively.

Method	attached to	build in	connected to	hanging on	part of	standing on	supported by	mean
IMP [36]	48.4	7.7	21.7	11.9	0.0	1.4	5.5	13.8
VGfM [9]	49.1	2.6	10.9	5.2	0.0	0.5	8.8	11.0
3DSSG [31]	46.6	15.4	10.9	11.9	0.0	1.8	14.3	14.4
SGFN [34]	**58.4**	33.3	32.6	**26.1**	0.0	1.0	**16.5**	24.0
Wu [35]	58.0	33.3	**39.1**	**26.1**	**12.5**	1.5	15.4	**26.6**
FROSS (Ours)	29.4	**43.6**	0.0	1.4	0.0	**47.2**	4.2	18.0

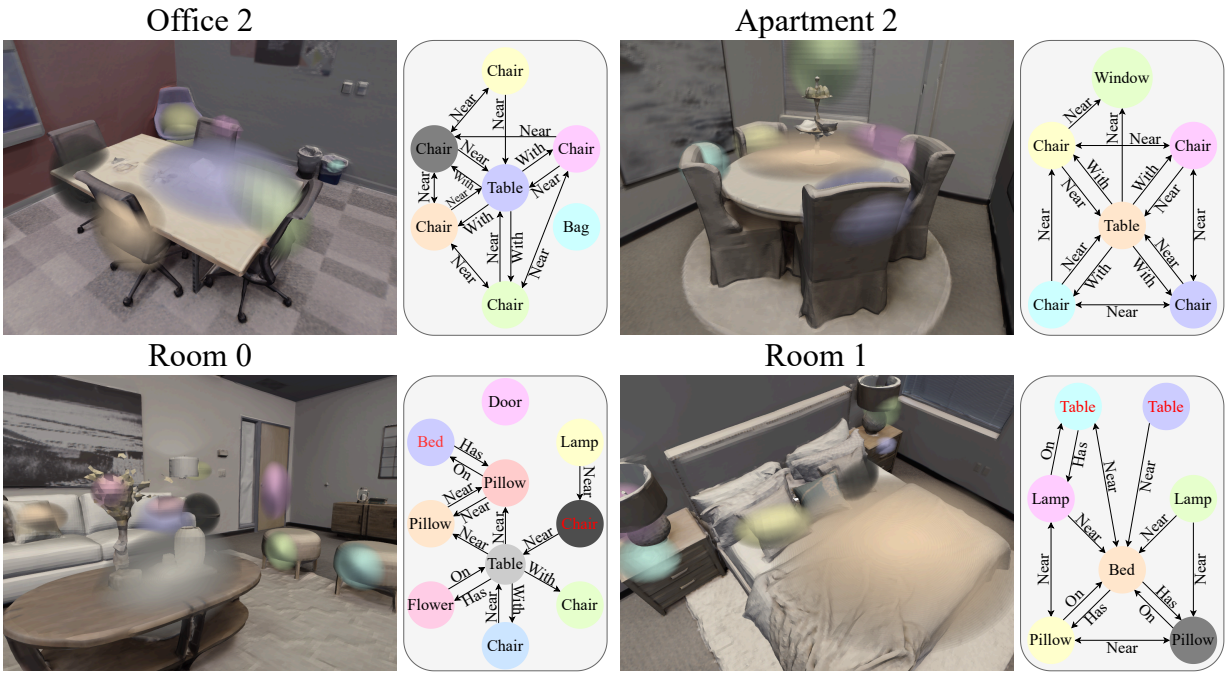

Figure 5. Qualitative results of FROSS on four scenes in the ReplicaSSG dataset. Please note that only representative objects are visualized, with misclassified objects marked in red on the right graph. The node colors in the left visualization correspond to the graphs on the right.

7.2. Additional Qualitative Results

Figure 5 presents 3D SSG generation results from FROSS on the ReplicaSSG dataset. FROSS captures both spatial relationships (e.g., "near" and "on") and semantic relationships (e.g., "has" and "with"). Misclassified objects are likely caused by occlusions from certain viewpoints or unusual viewing angles. These results further demonstrate FROSS's robustness in diverse scene conditions.

7.3. 2D Scene Graph Generation Performance

In this section, we present the evaluation of two models: the original EGTR [12] 2D SG generation model and our modified version employed in FROSS, *RT-DETR+EGTR*. The latter replaces the object detection backbone in the original EGTR with RT-DETR [44] object detector. These models are assessed on three datasets: Visual Genome [18], 3DSSG [31], and the proposed ReplicaSSG, as detailed in Table 9. For these evaluations, the models tested on ReplicaSSG received training on the Visual Genome dataset, whereas the models tested on the other two datasets used their respective training splits. Moreover, both models were optimized and accelerated using TensorRT[3]. The evaluation results demonstrate that *RT-DETR+EGTR* achieves superior performance in object detection (AP@50) and decreases processing latency by more than half. On the other hand, the original EGTR model demonstrates better performance in relationship prediction tasks. The above observations reveal that the integration of RT-DETR as the object detection backbone results in substantial processing efficiency improvements, with only a slight impact on relationship prediction performance for the ReplicaSSG dataset. This trade-off highlights the potential of RT-DETR in enhancing EGTR's practicality for applications that require faster inference speed. Moreover, the per-class object and predicate performance of *RT-DETR+EGTR* are shown in Tables 10 and 11.

[3] https://github.com/NVIDIA/TensorRT

7.4. ORB-SLAM3 Performance on ReplicaSSG

Table 12 presents the root mean square absolute trajectory error (RMS ATE) for ORB-SLAM3 [2] on the proposed ReplicaSSG dataset. The evaluation is conducted using ORB-SLAM3 with its default parameters and RGB-D input. The results are consistent with those reported in the original literature, confirming that ORB-SLAM3 can reliably track trajectories within the ReplicaSSG dataset.

8. Statistics of the ReplicaSSG Dataset

The statistics of the proposed ReplicaSSG Dataset are presented in Figures 6-9. More specifically, Figure 6 and 7 illustrate the occurrence frequency of objects and relationships across all categories in the dataset. In addition, Figures 8 and 9 offer scene-specific statistics that detail the number of objects and relationships in each scene.

Table 8. Per-class performance comparison of FROSS on the ReplicaSSG dataset for object and predicate recall (%).

Object Recall per Class																	
bag	**bskt.**	**bed**	**bench**	**bike**	**book**	**botl.**	**bowl**	**box**	**cab.**	**chair**	**clock**	**cntr.**	**cup**	**curt.**	**desk**	**door**	**mean**
25.0	50.0	0.0	0.0	0.0	1.5	9.1	37.5	4.0	14.3	68.1	66.7	40.0	33.3	9.1	0.0	80.0	
lamp	**pil.**	**plant**	**plate**	**pot**	**rail.**	**scrn.**	**shlf.**	**shoe**	**sink**	**stand**	**table**	**toil.**	**towel**	**umb.**	**vase**	**wind.**	28.8
16.7	41.5	47.4	31.2	7.7	0.0	0.0	11.1	8.3	100.0	0.0	72.2	100.0	0.0	66.7	38.9	0.0	

Predicate Recall per Class							
above	**against**	**attached to**	**in**	**near**	**on**	**under**	**with**	**mean**
22.2	0.0	0.0	33.3	28.8	19.1	10.0	50.0	20.4

Table 9. Evaluation results of two 2D SG generation models across three datasets. 'RT-DETR+EGTR' represents the EGTR model with RT-DETR as its object detector backbone. Latencies are reported in milliseconds. Recall@K (denoted as R@K) provides the class-agnostic average recall, while mean Recall@K (denoted as mR@K) represents the average recall across all relationship categories. All relationship metrics are evaluated with graph constraints as described in [38].

				Relationship					
Dataset	**Method**	**Latency**	**AP@50**	**R@20**	**R@50**	**R@100**	**mR@20**	**mR@50**	**mR@100**
Visual Genome [18]	EGTR	14.6	30.8	23.5	30.2	34.3	5.5	7.9	10.1
	RT-DETR+EGTR	6.82	32.2	15.7	22.0	26.6	3.3	4.9	6.2
ReplicaSSG	EGTR	14.6	21.2	13.2	18.1	22.0	6.9	9.6	11.9
	RT-DETR+EGTR	6.82	23.8	12.4	17.1	21.0	6.5	9.1	11.2
3DSSG [31]	RT-DETR+EGTR	6.82	41.0	43.4	48.2	52.0	23.3	27.3	30.7

Table 10. Per-class object detection performance in 2D SG generation with RT-DETR (AP@50).

Dataset	**Object Detection AP@50 per Class**																			
3DSSG [31]	**bath.**	**bed**	**bkshf.**	**cab.**	**chair**	**cntr.**	**curt.**	**desk**	**door**	**floor**	**ofurn.**	**pic.**	**refri.**	**show.**	**sink**	**sofa**	**table.**	**toil.**	**wall**	**wind.**	**mAP@50**
	100.0	83.3	28.6	56.1	64.8	67.7	73.0	29.2	73.3	91.5	40.3	41.9	50.0	42.9	73.3	73.7	68.2	100.0	60.9	57.5	63.8
ReplicaSSG	**bag**	**bskt.**	**bed**	**bench**	**bike**	**book**	**botl.**	**bowl**	**box**	**cab.**	**chair**	**clock**	**cntr.**	**cup**	**curt.**	**desk**	**door**	**lamp**	**pil.**	**plant**	**mAP@50**
	1.0	13.2	21.6	6.8	52.9	5.1	4.9	15.6	1.4	28.9	50.2	47.7	47.8	6.3	5.2	5.5	39.4	21.5	62.4	45.4	
	plate	**pot**	**rail.**	**scrn.**	**shlf.**	**shoe**	**sink**	**stand**	**table**	**toil.**	**towel**	**umb.**	**vase**	**wind.**							23.8
	14.0	3.4	18.9	34.1	24.5	4.1	43.5	0.7	44.3	61.4	2.3	7.5	30.3	37.0							

Table 11. Per-class relationship extraction performance in 2D SG generation with RT-DETR+EGTR (Recall@K).

Dataset	**Recall@K**	**Relationship Recall@K per Class**									
		attached to	**build in**	**connected to**	**hanging on**	**part of**	**standing on**	**supported by**			**mean**
3DSSG [31]	Recall@20	55.7	32.1	0.3	7.0	5.4	54.1	8.7			23.3
	Recall@50	61.6	37.6	1.4	9.1	11.3	59.5	10.2			27.3
	Recall@100	65.2	42.3	2.1	11.3	18.4	64.2	11.0			30.7
		above	**against**	**attached to**	**has**	**in**	**near**	**on**	**under**	**with**	**mean**
ReplicaSSG	Recall@20	2.1	0.0	0.0	0.0	13.2	11.2	13.1	0.0	19.1	6.5
	Recall@50	4.4	0.0	0.0	0.0	15.3	15.8	17.0	0.0	29.2	9.1
	Recall@100	6.8	0.0	0.0	0.0	16.5	19.9	20.0	0.0	37.4	11.2

Table 12. ORB-SLAM3 RMS ATE (cm) in each ReplicaSSG scene.

Apartment 0	**Apartment 1**	**Apartment 2**	**Office 0**	**Office 1**	**Office 2**	**mean**
4.6	1.9	3.8	0.9	0.5	4.0	
Office 3	**Office 4**	**Room 0**	**Room 1**	**Room 2**	**Hotel 0**	3.6
3.1	1.9	0.9	0.9	1.3	2.9	
FRL Apartment 0	**FRL Apartment 1**	**FRL Apartment 2**	**FRL Apartment 3**	**FRL Apartment 4**	**FRL Apartment 5**	
2.3	5.5	19.9	6.1	2.3	2.5	

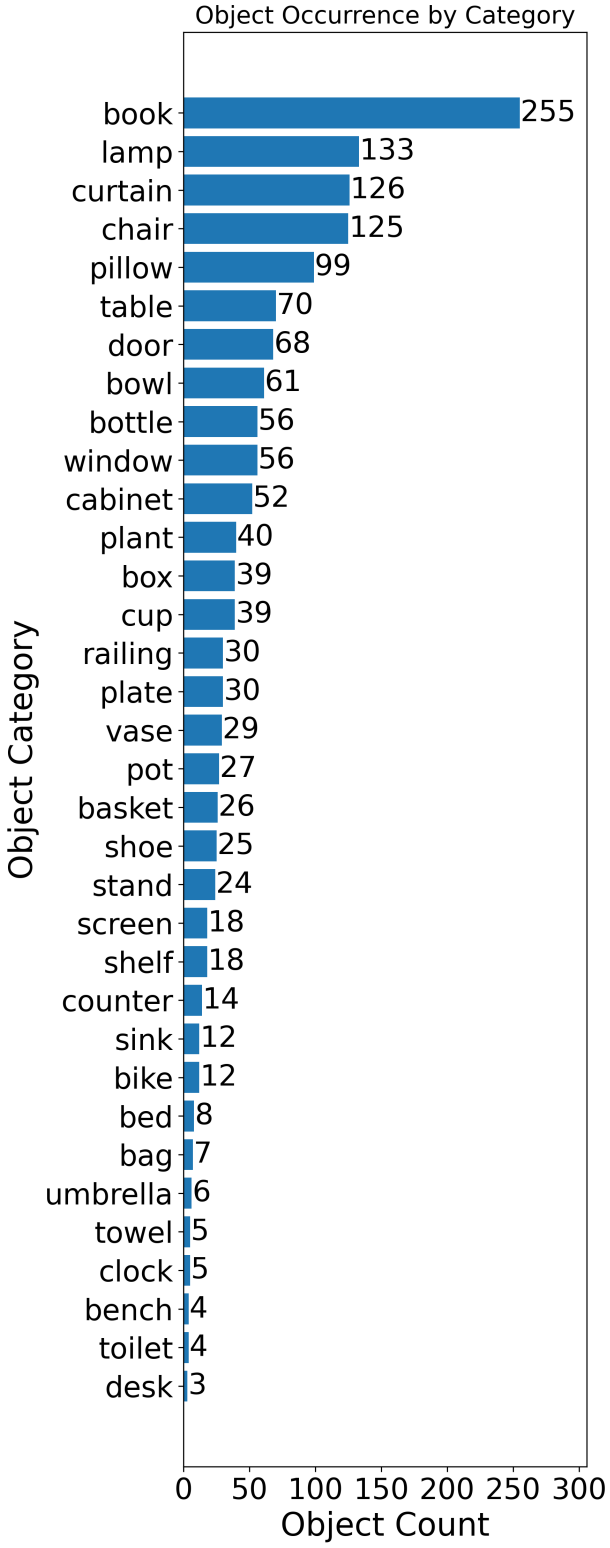

Figure 6. The occurrence frequency of each object category in the ReplicaSSG dataset.

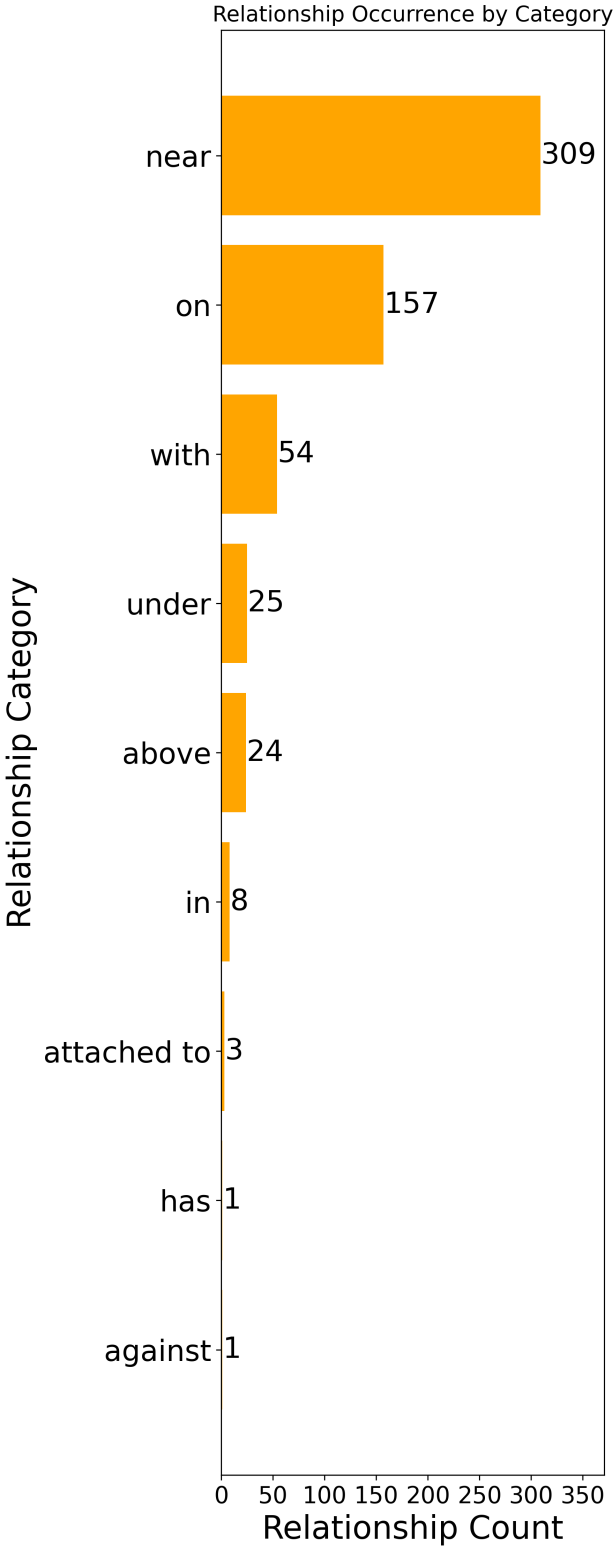

Figure 7. The occurrence frequency of each relationship category in the ReplicaSSG dataset.

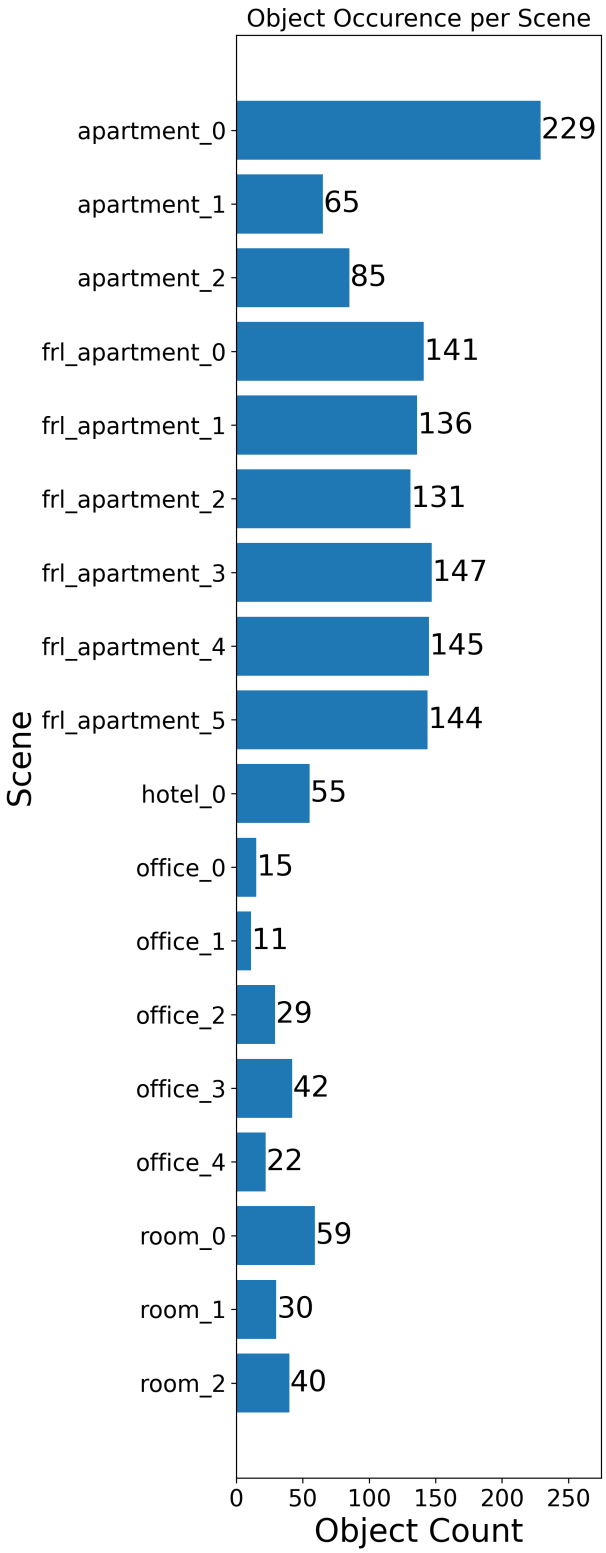

Figure 8. The number of objects present in each scene within the ReplicaSSG dataset.

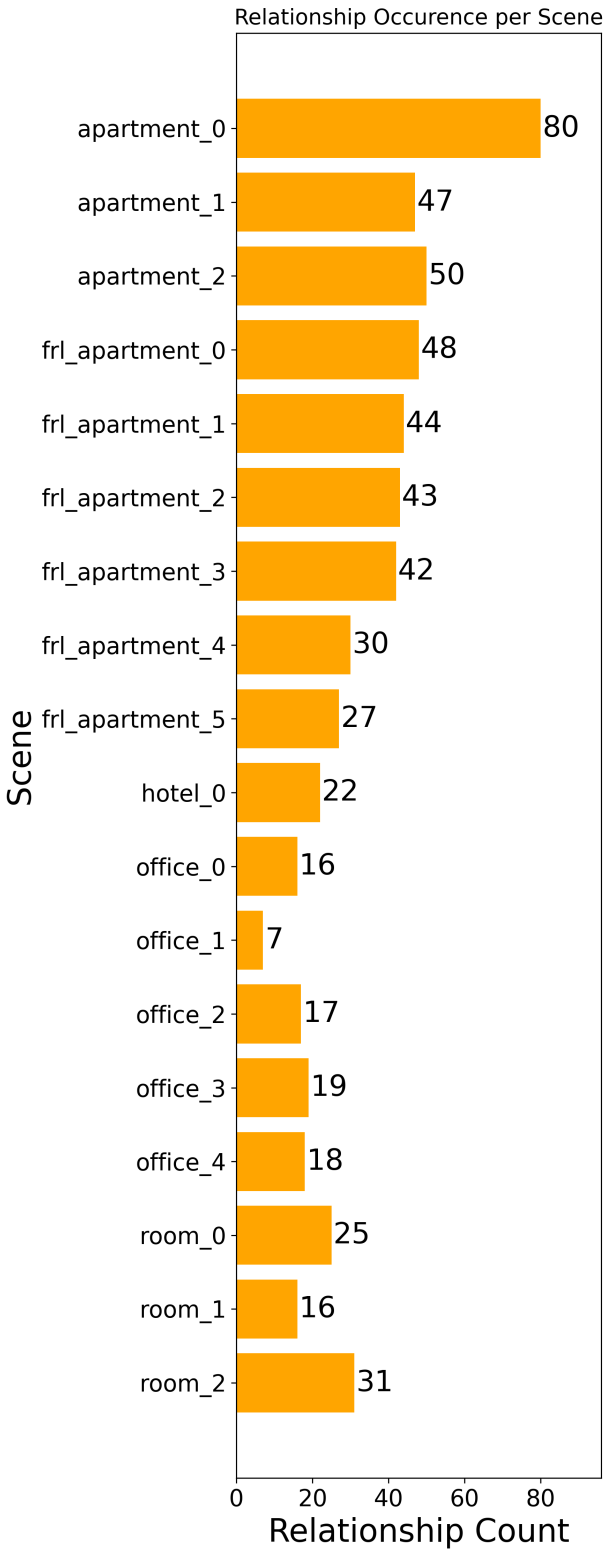

Figure 9. The number of relationships present in each scene within the ReplicaSSG dataset.