Online Scene CAD Recomposition via Autonomous Scanning Supplemental Material

CHANGHAO LI, University of Science and Technology of China, China
JUNFU GUO, University of Science and Technology of China, China
RUIZHEN HU*, Shenzhen University, China
LIGANG LIU, University of Science and Technology of China, China

CCS Concepts: • **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: scene CAD recomposition, autonomous scanning, relation-guided pose optimization, relation-constrained retrieval

ACM Reference Format:
Changhao Li, Junfu Guo, Ruizhen Hu, and Ligang Liu. 2023. Online Scene CAD Recomposition via Autonomous Scanning Supplemental Material. *ACM Trans. Graph.* 42, 6, Article 250 (December 2023), 8 pages. https://doi.org/10.1145/3618339

1 IMPLEMENTATION DETAILS

1.1 Object pose estimation

We use the ShapeNet dataset to construct the training dataset of both our rotation prediction model and complete pose estimation model. For each CAD model C, we first sample 2048 points P_c using farthest point sampling to represent the complete shape of it. Then, we randomly excise 20 to 100 percent of the CAD model C and them sample 2048 points P_p from the remaining part to represent the partial shape of C.

When training the pose estimation model, rather than predicting the pose by using the complete shape of the CAD model, we use the partial shape P_p as input and rotate it randomly in 3-DoF to approximate objects in the real world with arbitrary positions. First, we use the point transformer [Zhao et al. 2020] as our shape encoder *Enc* to estimate the complete shape feature F of the input points with the dimension of 512, and then freeze it as a pre-trained shape encoder for our geometry-aligned retrieval model. Then, the shape feature F is used to decode both the rotation matrix and the pose of complete shape by 3 MLP layers. Besides, we use the FoldingNet to decode the complete shape of input points which also contains 2048 points with the supervision of our weighted chamfer distance loss.

*Corresponding author: Ruizhen Hu (ruizhen.hu@gmail.com)

Authors' addresses: Changhao Li, lch0510@mail.ustc.edu.cn, University of Science and Technology of China, China; Junfu Guo, guojunfu@mail.ustc.edu.cn, University of Science and Technology of China, China; Ruizhen Hu, ruizhen.hu@gmail.com, Shenzhen University, China; Ligang Liu, lgliu@ustc.edu.cn, University of Science and Technology of China, China.

0730-0301/2023/12-ART250 $15.00
https://doi.org/10.1145/3618339

Table 1. Comparison to single-frame baselines, including Mask2CAD [Kuo et al. 2020], ROCA [Gümeli et al. 2021], the combination of Im3D [Zhang et al. 2021] and JointEmbedding [Dahnert et al. 2019] (Im3D+JE).

Method	CA ↑	PA ↑	TE ↓	RE ↓	SE ↓	GE ↓
Mask2CAD	0.332	0.103	0.892	0.196	0.186	0.792
ROCA	0.427	0.266	0.713	0.211	0.184	0.626
Im3D+JE	0.548	0.503	0.146	0.121	0.171	0.028
Ours	**0.552**	**0.617**	**0.104**	**0.097**	**0.164**	**0.013**

1.2 Relation-guided optimization

When all the information of both objects and the layout are gathered, we first build the graph and then encode these information into embedded features. After that, we use three different subnetworks which are both constructed by 5 MLP layers to encode the messages between neighbors of corresponding types of relations, including the relations between objects, between objects and the floor, and between objects and walls. And then, the embedded features of both objects, walls and the floor are updated by passing encoded messages. Besides, we also use a subnetwork with 3 MLP layers to encode the correlations between all neighbors.

2 MORE EXPERIMENTS

2.1 Comparison with single-frame baselines

We additionally compare our method with single-frame CAD recomposition baselines to show the strength of our method on both retrieval accuracy and alignment accuracy even with only single frame as input.

Single-frame baselines. The goal of single-frame methods is to generate the CAD recomposition of the local region contained in the given frame. They take the single RGB image as the input and finally generate the recomposition corresponding to the image with retrieved and aligned CAD models. To the best of our knowledge, *Mask2CAD* [Kuo et al. 2020] and *ROCA* [Gümeli et al. 2021] are two state-of-the-art methods that generate CAD recomposition of objects in the given frame. Besides, we find the work of *Im3D* [Zhang et al. 2021] also estimates the poses of both objects and the layout of the local region in the given frame by considering the possible relations between them, which is similar to our pose estimation method. Thus, we also compare our method with the method of Im3D. However, it only reconstructs the surface of each detected object in the given frame, so we implement a modification (*Im3D+JE*) by combining it with the method of *Joint-Embedding*, which generates the

Fig. 1. Qualitative comparisons with single-frame baselines, including Mask2CAD [Kuo et al. 2020], ROCA [Gümeli et al. 2021], the combination of Im3D [Zhang et al. 2021] and Joint-Embedding [Dahnert et al. 2019] (Im3D+JE) and our method. The top left image of each group is the input for each method, and the top right one is the corresponding GT recomposition result of the input.

Table 2. Comparison between the exploration-oriented scanning strategy of the method Schmid [Schmid et al. 2020] (RelCAD+Explore) and our method on scanning efficiency and recomposition accuracy.

Method	DC ↓	TC ↓	#NBV ↓	SC ↓	CA ↑	PA ↑	PE ↓	GE ↓
RelCAD+Explore	**24.415**	**3.042**	27.872	35.263	0.493	0.588	0.410	0.352
Ours (RelCAD+RelNBV)	37.806	6.045	**23.646**	**5.271**	**0.562**	**0.636**	**0.352**	**0.010**

joint embedding space for both scanned point clouds and CAD models and make them have lower feature distance if they have higher geometry similarity. The pose and the reconstructed surface of each object in the given frame along with the layout are first predicted by Im3D. We then estimate embeddings for objects, choose CAD models with nearest embedding features for them and put CAD models into predicted poses of corresponding objects to get the final CAD recomposition of the given frame. Since our method takes RGBD observations as input, we use the pre-trained *S2R-DepthNet* [Chen et al. 2021] to predict the depth from the given RGB observation. And the RGB observations are randomly generated in the validation set of ScanNet scenes with at least three objects in view.

Comparison results. Table 1 shows the comparisons with single-frame baselines. Both Im3D and our method achieves better alignment accuracy due to the consideration of relations between objects and layout which provides additional constraints of object poses and more accurate understanding of the structure of the local region. On the other hand, both Mask2CAD and ROCA results in much lower alignment accuracy than other methods since they only estimate the pose of each object individually which ignore the structure of the local region. Besides, since we additionally use the shape completion as the additional training supervision for pose estimation of objects, our method achieves the most accurate pose estimation compared with all other methods.

Figure 1 shows the qualitative results between Mask2CAD, ROCA, Im3D+JE and our method. Benefiting from our geometry-aligned retrieval component, we can achieve more precise geometric matching between retrieved CAD models and scanned objects, while other methods ignore the guidance of geometry details on scanned objects and only retrieve CAD models with the encoded features.

2.2 Comparison with exploration-oriented baseline

Although the goal of exploration-oriented scanning strategies have huge differences with ours which only focus on the efficiency of exploration, we also choose a representative method *Schmid* [Schmid et al. 2020] to make a detailed comparison on our task. The method *Schmid* contains an exploration-oriented scanning strategy which considers the efficiency of the exploration first and generates NBVs based on frontier points and an updated *RRT** [Karaman and Frazzoli 2011] algorithm. We replace our NBV generation module with the scanning strategy of Schmid *(OurCAD+Explore)* to compare the performances on our task between strategies designed for different goals.

Comparison results. Table 2 shows the comparison results. When compared with the exploration-oriented scanning strategy, the distance and time consumption of our method are higher since some generated NBVs are used to scan the local regions that overlap with scanned region to improve the accuracy of CAD recomposition instead of just exploring the unknown area. However, the exploration-oriented scanning strategy tends to generate NBVs frequently to make sure the information of the scene is up-to-date, which results in generating more NBVs than our method. Besides,

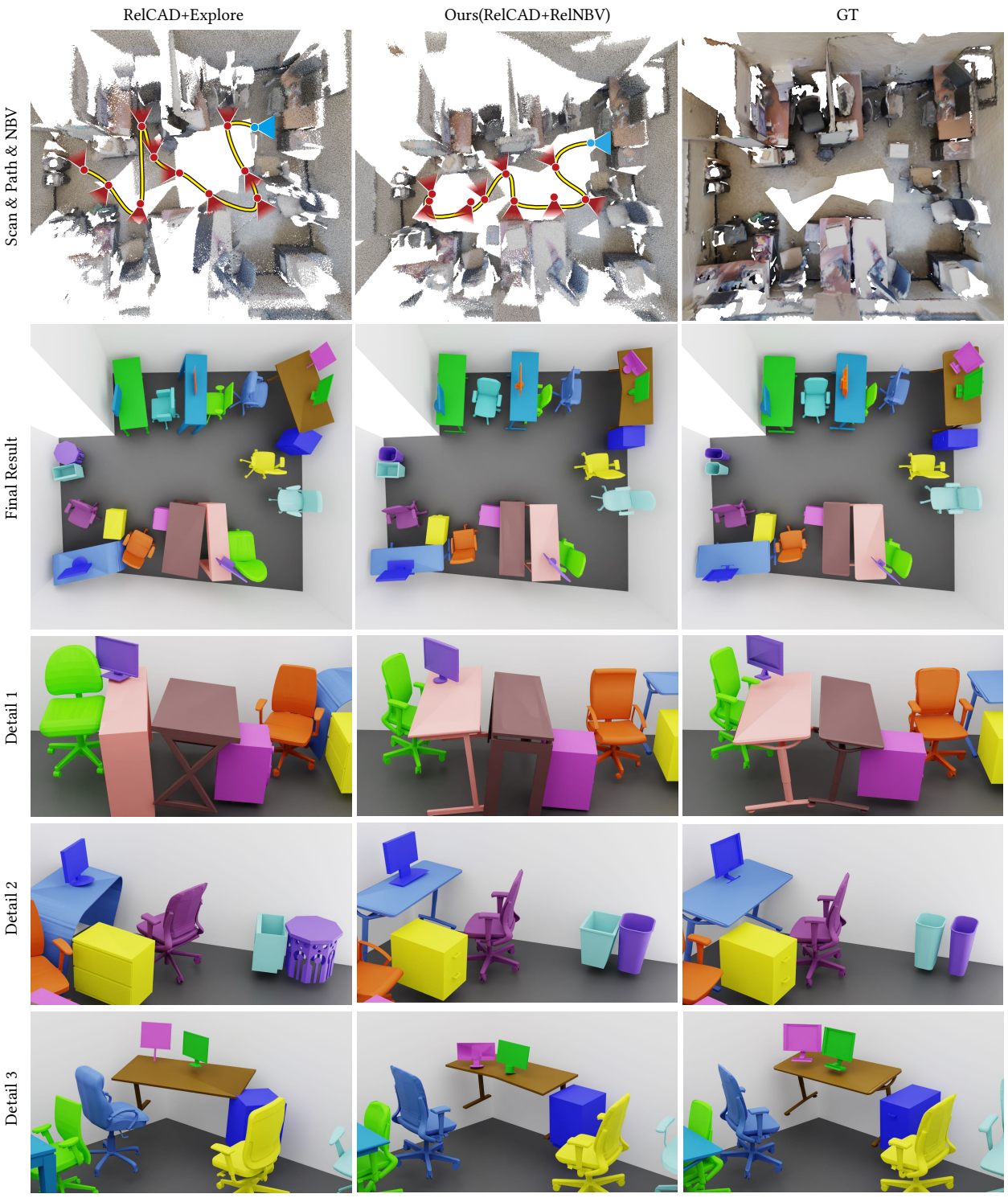

Fig. 2. Qualitative comparison with exploration-oriented scanning strategy Schmid and our method. The observation corresponding to the initial position of the robot is painted blue, and the retrieved CAD models of scanned objects are shown in different colors.

Table 3. Impact of different settings of the retrieval confidence threshold R. An object point will be generated for the scanned object only when the retrieval confidence of the corresponding retrieved CAD model is lower than this threshold.

Method	DC ↓	TC ↓	#NBV ↓	SC ↓	CA ↑	PA ↑	PE ↓	GE ↓
Ours+0.5R	**27.602**	**4.100**	**13.312**	**5.252**	0.496	0.511	0.454	0.024
Ours+0.6R	31.898	4.921	17.372	5.261	0.548	0.632	0.365	0.013
Ours+0.7R	32.210	5.210	19.041	5.263	0.551	0.632	0.363	0.013
Ours+0.8R	32.384	5.267	19.957	5.264	0.553	0.634	0.357	0.011
Ours+0.9R	37.806	6.045	23.646	2.571	**0.562**	**0.636**	**0.352**	**0.010**

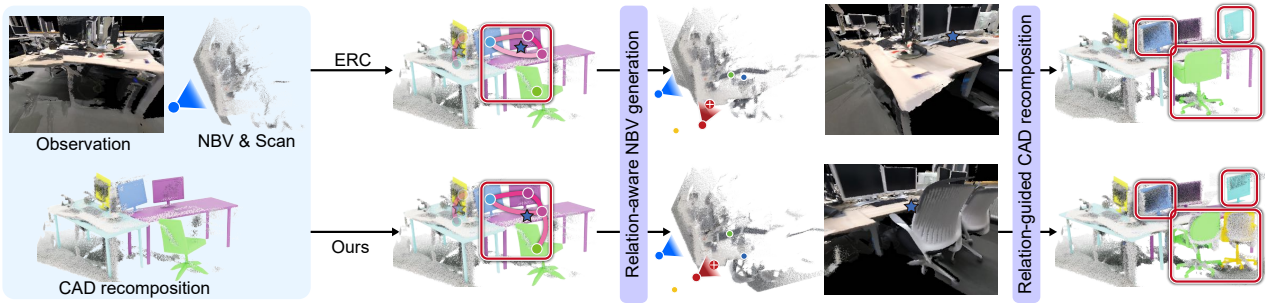

Fig. 3. Detailed comparison between different relation calculation methods, including the explicitly calculated relations (with ER) and the predicted relations by our method, and their impact on newly generated NBV and new CAD recomposition result. The main differences are framed by red boxes.

the lack of collected information of objects brings more uncertainty to the pose estimation of them and makes the recomposition accuracy much lower than our method, which shows that it is difficult to accomplish our task effectively by the scanning strategies which focus on the quick exploration of the scene.

Figure 2 show some visual comparisons. What is remarkable is that the path generated by our scanning strategy is similar to the path of exploration-oriented scanning strategy, which shows that our method can generate more accurate CAD recomposition results by optimizing the generated NBVs and remain the high scanning efficiency at the same time.

2.3 More detailed ablation studies

In addition to the ablation studies on the key components of our method, we also conduct ablation studies on several detailed designs of our method that have a significant impact, including the retrieval confidence threshold and the relation calculation.

Influence of retrieval confidence threshold. We conduct an experiment to figure out the impact of different retrieval confidence threshold and the quantitative results are shown in Table 3. When the threshold becomes smaller, the behavior of the scanning strategy tends to be similar with the exploration-oriented scanning strategies such as [Schmid et al. 2020], while the CAD recomposition result will be more accurate with higher threshold and more NBVs. And finally we choose 0.9 as our default setting since it brings us the highest scene CAD recomposition accuracy with the relatively ideal scanning efficiency.

Influence of relation calculation. In addition to this, we also explore the influence of different relation calculations, including the explicit relation calculation (`with ER`) and the predicted relation by our relation-guided optimization component. Figure 3 shows an detailed example of how do different relations lead to different NBVs and recomposition results. When we generate the CAD recomposition from the first scan, the pose estimation of the single chair is inaccurate due to only few part of it is scanned. Then, the explicitly calculated relation between the chair and the right desk is less than the threshold and ignored when generating the corresponding relation point.

However, our method still remain higher relation between them since our relation-guided optimization component takes both encoded shape features, estimated poses and semantic categories of objects as inputs and learns to predict more accurate relation by fusing all these data. Therefore, the generated relation points which are shown with blue stars of two relations are quite different, which results in the significant difference between generated NBVs and the new CAD recomposition results. From the comparison, the predicted relations are more accurate and stable than explicit relations, and the accurate relations provide the effective guidance for new generated NBVs to gather more information and relations in local regions.

Moreover, as shown in Figure 4, when using explicit relation calculated by the estimated poses of objects, sometimes the inaccurate relations reduce the ability of NBV to capture more information and relations in local regions and thus more NBVs are needed to achieve same accuracy of alignment and retrieval (framed by green boxes).

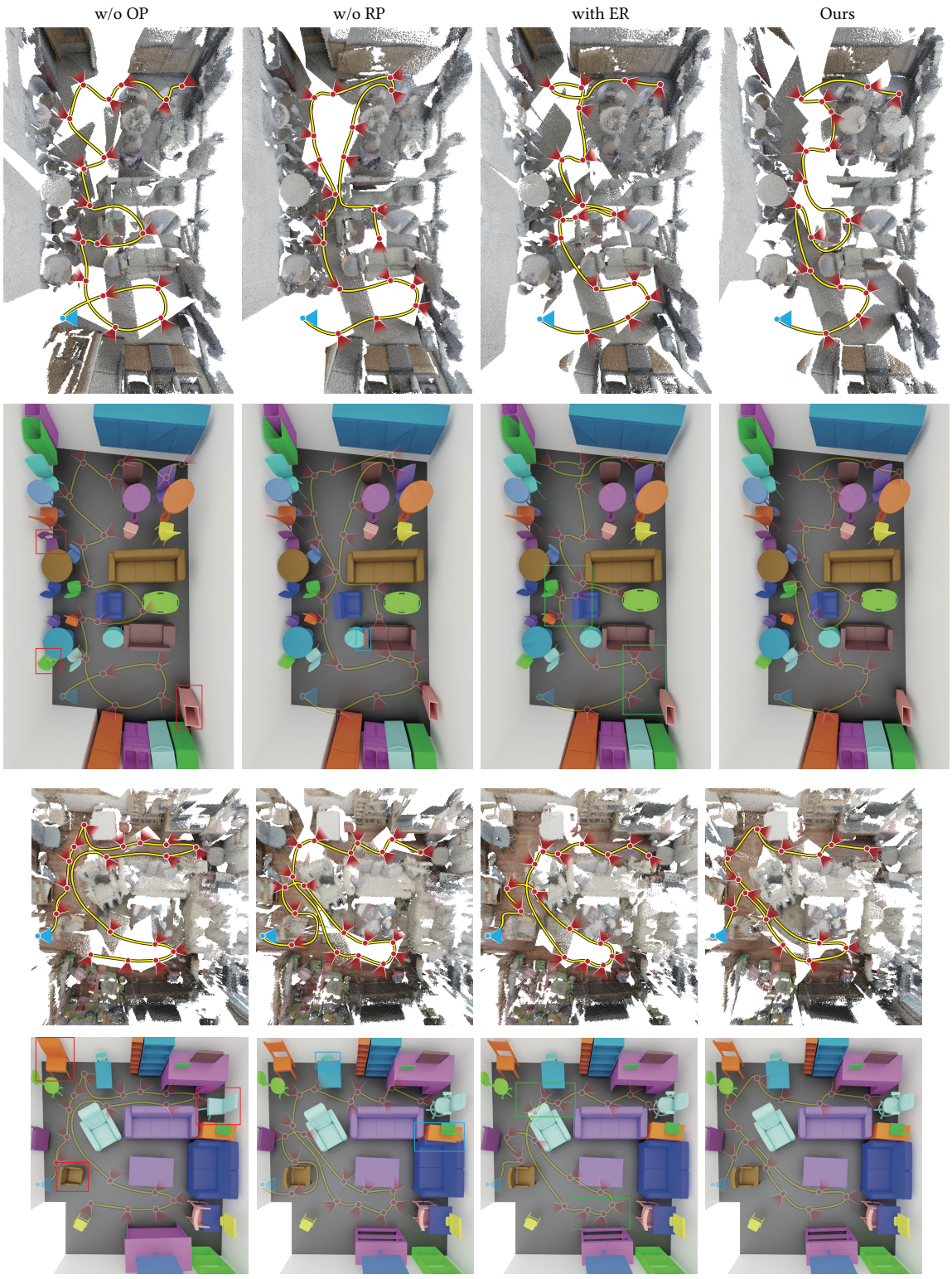

Fig. 4. Qualitative comparisons between different settings of our NBV generation module on scanning paths and CAD recomposition results.

To show the influence of different types of interest points and different calculation of relations, we provide some qualitative results to compare the differences of scanning paths and recomposition results. As shown in Figure 4, when there are no object point (`w/o OP`), some retrieved CAD models (framed by red boxes) are not the CAD models with the highest geometry similarity of scanned objects, which shows that object points mainly focus on local geometry details that differ the most between retrieved CAD models and scanned objects and then instruct the new NBV to pay more attention to these distinctions. Besides, when there are no relation point (`w/o RP`), some poses of retrieved CAD models (framed by blue boxes) overlap with neighbor CAD models since the generated NBVs mainly focus on the exploration and the objects with inaccurate retrieval results and thus ignore some local regions where remain inaccurate relations.

2.4 More comparison to baselines

For the scanning efficiency, our method requires much less storage consumption when comparing to all the baseline methods compared in our paper, as shown in Figure 5. Figure 6 further shows that the scanning efficiency of our method is similar with the fast exploration-oriented scanning strategy while achieving the similar recomposition accuracy by generating only half of NBVs compare with the object-aware scanning method which collect almost the full details of the scene. Besides, we also provide more qualitative results on the offline and online comparisons to further illustrate the rationality and superiority of our method.

As shown in Figure 7, compared with other offline baselines, our method identify objects and recompose the scene with CAD models accurately since we jointly consider the relations between objects and the layout of the scene, which have strong guidance for the estimation of the poses of objects and help us retrieve CAD models with higher geometry similarities.

Moreover, Figure 8 shows more qualitative comparisons between all online baselines and our method. Since ROCA estimates the poses of objects not accurately enough, even scanning the same object repeatedly does not significantly improve the recomposition accuracy. And the exploration-oriented scanning strategy also makes the recomposition result inaccurate because of the lack of gathering information of objects. Besides, our method achieves similar recomposition accuracy with the reconstruction-oriented scanning strategy which collects more information of objects than our method most of times.

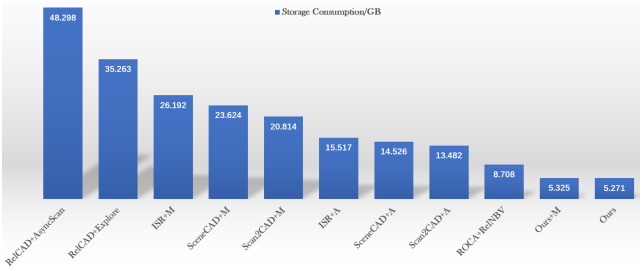

Fig. 5. The storage consumption of the generated intermediate data for the scanned scene between different methods.

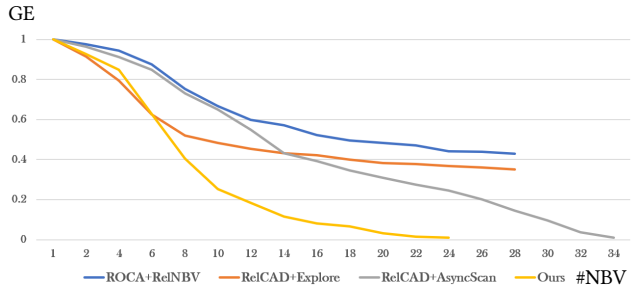

Fig. 6. The relationships between #NBV and GE of online CAD recomposition baselines, where the line ends until the scanning is finished.

REFERENCES

Xiaotian Chen, Yuwang Wang, Xuejin Chen, and Wenjun Zeng. 2021. S2R-DepthNet: Learning a Generalizable Depth-specific Structural Representation. *CoRR* abs/2104.00877 (2021). arXiv:2104.00877 https://arxiv.org/abs/2104.00877

Manuel Dahnert, Angela Dai, Leonidas J. Guibas, and Matthias Nießner. 2019. Joint Embedding of 3D Scan and CAD Objects. *CoRR* abs/1908.06989 (2019). arXiv:1908.06989 http://arxiv.org/abs/1908.06989

Can Gümeli, Angela Dai, and Matthias Nießner. 2021. ROCA: Robust CAD Model Retrieval and Alignment from a Single Image. *CoRR* abs/2112.01988 (2021). arXiv:2112.01988 https://arxiv.org/abs/2112.01988

Sertac Karaman and Emilio Frazzoli. 2011. Sampling-based Algorithms for Optimal Motion Planning. *CoRR* abs/1105.1186 (2011). arXiv:1105.1186 http://arxiv.org/abs/1105.1186

Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. 2020. Mask2CAD: 3D Shape Prediction by Learning to Segment and Retrieve. *CoRR* abs/2007.13034 (2020). arXiv:2007.13034 https://arxiv.org/abs/2007.13034

Lukas Schmid, Michael Pantic, Raghav Khanna, Lionel Ott, Roland Siegwart, and Juan Nieto. 2020. An efficient sampling-based method for online informative path planning in unknown environments. *IEEE Robotics and Automation Letters* 5, 2 (2020), 1500–1507.

Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. 2021. Holistic 3D Scene Understanding from a Single Image with Implicit Representation. *CoRR* abs/2103.06422 (2021). arXiv:2103.06422 https://arxiv.org/abs/2103.06422

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. 2020. Point Transformer. *CoRR* abs/2012.09164 (2020). arXiv:2012.09164 https://arxiv.org/abs/2012.09164

Fig. 7. More qualitative comparisons with offline CAD recomposition baselines including Voxblox++-Scan2CAD (VPP-S2C), Voxblox++-SceneCAD (VPP-SceneCAD), Interactive Scene Reconstruction (ISR) and our method. Scanned data is either the fusion of all RGBD observations obtained in the dataset (`+M`), which is the intermediate result of other methods, or the autonomous scanned data generated by our method (`+A`). The retrieved CAD models of scanned objects are shown in different colors. Note that only VPP-SceneCAD and our method estimates the layout of the scene, we add the same layout estimated by VPP-SceneCAD to the results of VPP-S2C and ISR for a better comparison between previous methods and our method.

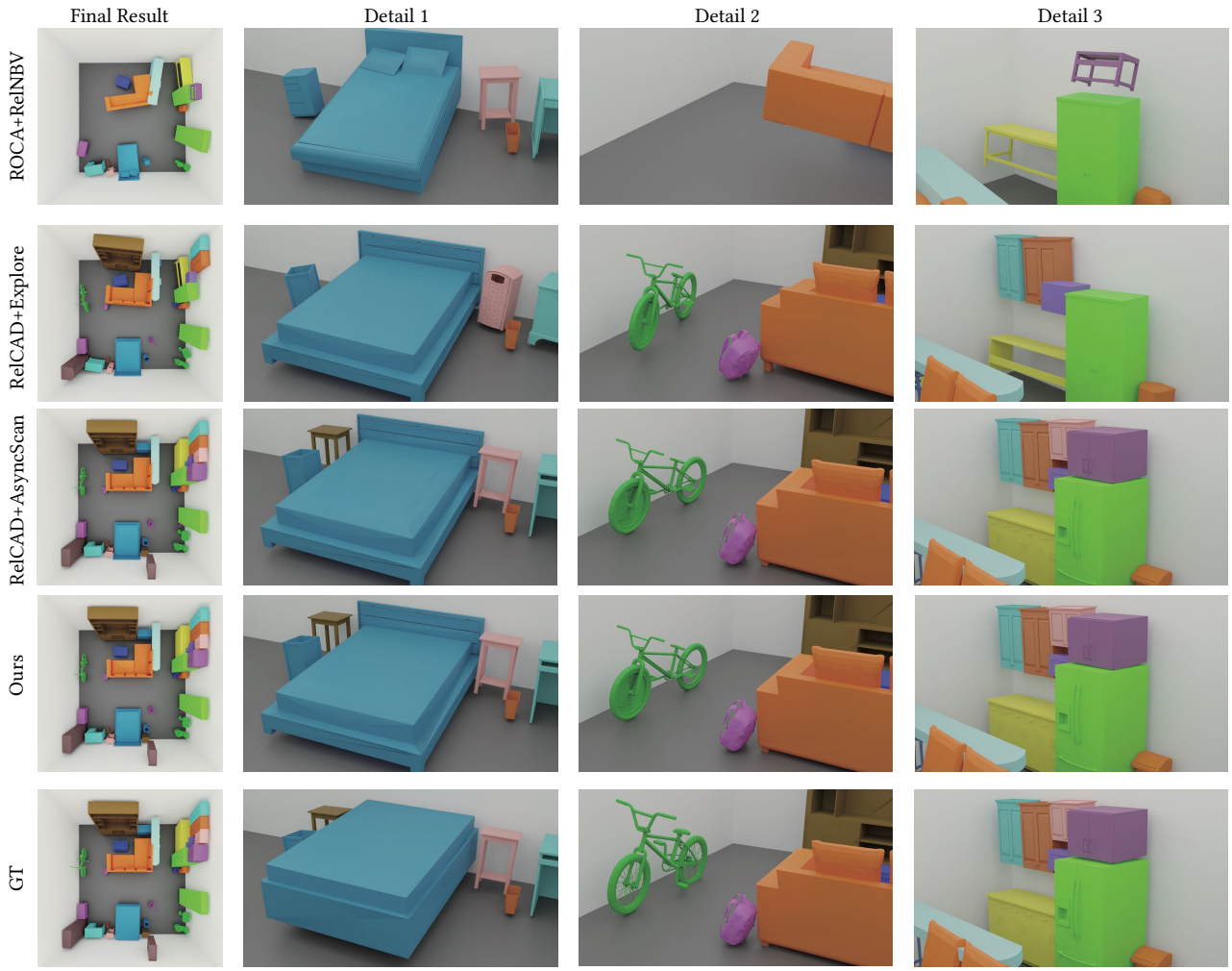

Fig. 8. More qualitative results of online CAD recomposition baselines, including ROCA with our NBV generation module (ROCA+OurNBV), AsyncScan with our CAD recomposition module (OurCAD+AsyScan) and our full method. Scanned data is the fusion of all RGBD observations of generated NBVs by each method. The observation corresponding to the initial position of the robot is painted with blue. The retrieved CAD models of scanned objects are shown in different colors.