Online Scene CAD Recomposition via Autonomous Scanning

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

Fig. 1. With the rough reconstruction (left) based on spare viewpoints selected via autonomous scanning (middle), our scene CAD recomposition result (right) can faithfully reflect the object geometry and arrangement in the given scene. Note that the corresponding CAD models are retrieved and their relative poses are optimized online with the few automatically optimized scanning viewpoint, even though the scene is only partially reconstructed.

Autonomous surface reconstruction of 3D scenes has been intensely studied in recent years, however, it is still difficult to accurately reconstruct all the surface details of complex scenes with complicated object relations and severe occlusions, which makes the reconstruction results not suitable for direct use in applications such as gaming and virtual reality. Therefore, instead of reconstructing the detailed surfaces, we aim to recompose the scene with CAD models retrieved from a given dataset to faithfully reflect the object geometry and arrangement in the given scene. Moreover, unlike most of the previous works on scene CAD recomposition requiring an offline reconstructed scene or captured video as input, which leads to significant data redundancy, we propose a novel online scene CAD recomposition method with autonomous scanning, which efficiently recomposes the scene with the guidance of automatically optimized Next-Best-View (NBV) in a single online scanning pass. Based on the key observation that spatial relation in the scene can not only constrain the object pose and layout optimization but also guide the NBV generation, our system consists of two key modules: relation-guided CAD recomposition module that uses relation-constrained global optimization to get accurate object pose and layout estimation, and relation-aware NBV generation module that makes the exploration during the autonomous scanning tailored for our composition task. Extensive experiments have been conducted to show the superiority of our method over previous methods in scanning efficiency and retrieval accuracy as well as the importance of each key component of our method.

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. *ACM Trans. Graph.* 42, 6, Article 250 (December 2023), 16 pages. https://doi.org/10.1145/3618339

1 INTRODUCTION

With the increasing demand for applications such as augmented and virtual reality, gaming, and robotics, there are intensive studies focusing on the surface reconstruction of indoor scenes with RGB-D sensors [Charrow et al. 2015; Huang et al. 2020; Wu et al. 2014; Xu et al. 2015]. However, significant noise and errors will be introduced into the reconstructed results due to the increase of the number of objects, the influence of different materials of objects, and the effect of lighting that changes over time in the real world. Since the realistic scene with high-quality meshes is desperately needed by applications such as gaming and virtual reality, instead of reconstructing the detailed surfaces, we aim to recompose the scene with high-quality CAD models retrieved from a given dataset to faithfully reflect the object geometry and arrangement in the given

*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

scene, which is referred to as *scene CAD recomposition* [Avetisyan et al. 2018a; Izadinia and Seitz 2018].

Most of the previous works on scene CAD recomposition require an offline reconstruction of the scene as input [Avetisyan et al. 2018a, 2019, 2020], where the recomposition task is usually split into two sub-processes: the representation of the entire scene such as the mesh or the distance field is first constructed, and then the corresponding CAD models are retrieved and aligned with the objects in the reconstructed results. As a result, additional time consumption and computational cost are introduced due to the redundant data capture for full scene reconstruction. There are also works trying to directly recompose the scene with CAD models from the given sequence of scanning observations [Han et al. 2021; Maninis et al. 2020], without explicitly reconstructing the scene. However, the scanning sequence is still captured offline instead of tailored for the recomposition task, it is not sure how to obtain a video that is most appropriate for the task, and the data redundancy problem also remains.

Compared with surface reconstruction which tries to fully restore the rich details of the scene, CAD recomposition is a more abstract expression of the scene aiming to reflect the arrangement of objects and their relationships and thus requires much less scanning data, so it drives us to optimize the viewpoint selection strategy for scene CAD recomposition. Moreover, inspired by recent works on autonomous reconstruction that reconstruct the scene with higher scanning efficiency and less data redundancy by autonomous scanning using robots with optimized viewpoints [Guo et al. 2022; Liu et al. 2018, 2021; Schmid et al. 2020; Xu et al. 2017], our work aims to solve CAD recomposition problem of unknown indoor scenes *online* with one-pass autonomous scanning.

The key challenges of the online CAD recomposition problem are two-fold: accurate CAD recomposition with partial scans and appropriate viewpoint selection for recomposition. Different from the scanned objects with complete geometric shapes in the reconstructed results, the objects during the autonomous scanning process are often incomplete, and thus it is more difficult to retrieve and further align corresponding CAD models for partially-scanned objects. Moreover, as the goal of CAD recomposition is different from that of surface reconstruction, with our task focusing more on object approximation and arrangement recovery other than surface details, the guideline for the Next-Best-View (NBV) generation during autonomous scanning would be quite different. To address the above challenges, our key observation is that spatial relations in the scene can be used to not only get more accurate CAD recomposition but also guide the NBV generation. The relations between different objects as well as the relations between objects and layout (i.e., the floor and walls of the scene) can help us determine whether the categories of objects are accurate and reasonable, guide the optimization of object poses as well as the following retrieval, and discover unscanned objects and important interaction regions for NBV generation.

Based on the above key observation, we propose an online scene CAD recomposition method with autonomous scanning consisting of two key components: relation-guided CAD recomposition and relation-aware NBV generation. For relation-guided CAD recomposition, we make the retrieval benefit from the global object pose optimization with relation guidance. Different from most existing works of scene recomposition [Avetisyan et al. 2018a, 2019, 2020; Izadinia and Seitz 2018], which first retrieves CAD models for objects and then optimize their poses with relation guidance, we perform a global optimization of object poses and scene layout with relation guidance first and then use the optimized more accurate object poses to retrieve CAD models with better geometry alignment for partial objects. For relation-aware NBV generation, other than the traditional frontier points for exploration of unknown regions, we add object points for retrieval adjustment and relation points for spatial relation refinement. All those three types of points will be considered comprehensively and fused according to their different importance to determine the region of interest (ROI) in the current scene to guide the generation of NBV pointing to the point with the highest interest. Hence, the NBV generated by our method endows the robot with different perception capabilities, which enables the robot to simultaneously explore the unknown area, optimize relations between both objects and layout and retrieve corresponding CAD models for objects in a more efficient way.

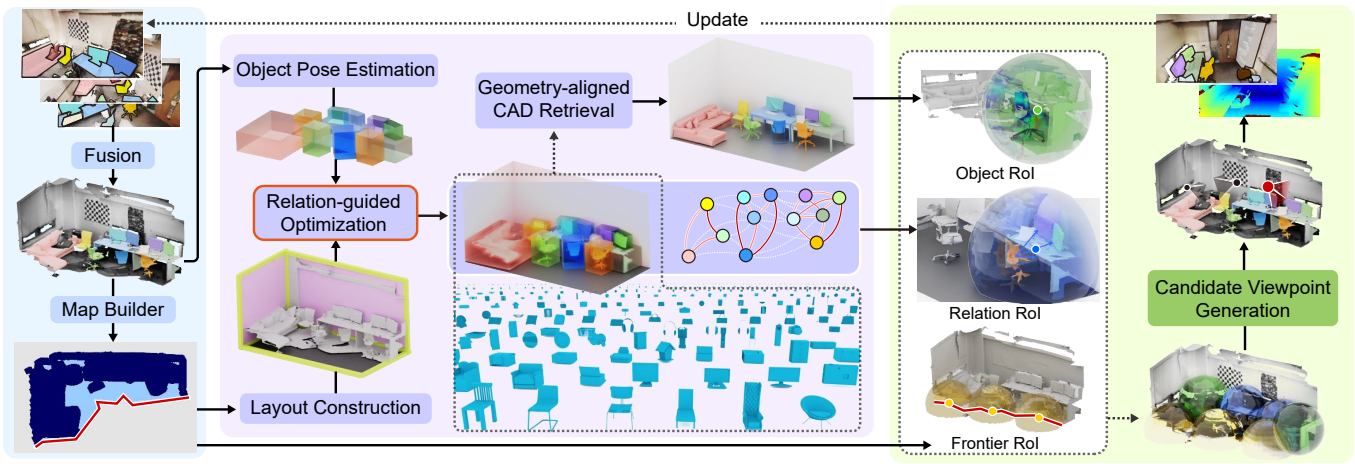
To summarize, we propose a novel online scene CAD recomposition method with autonomous scanning which guides the robot to efficiently scan and analyze the unknown scene, and eventually recompose the scene with retrieved high-quality CAD models and estimated room layout that faithfully resemble the object geometry and arrangement. We show that compared to existing offline recomposition methods and other baselines, our online method obtains higher retrieval accuracy as well as scanning efficiency. Ablation studies are also conducted to validate the importance of each key component of our method. Moreover, we also test our method on a real robot to show the applicability of our method in the real world. Our technical contributions include:

- An online scene CAD recomposition system that can guide the robot to autonomously scan and recompose unknown 3D scenes with high efficiency.
- A relation-guided CAD recomposition method that uses relation-constrained global optimization to get accurate object pose and layout estimation for more accurate object retrieval.
- A relation-aware NBV generation method that makes the exploration during the autonomous scanning tailored for our composition task.

2 RELATED WORK

2.1 Scene CAD recomposition

As the noise contained in the results of surface reconstruction is unacceptable to be served for applications such as gaming and virtual reality, some works focus on the offline scene CAD recomposition to estimate the arrangement from the given scanned data such as meshes and point clouds of the single object [Kim et al. 2013] or the scene [Avetisyan et al. 2018a; Ishimtsev et al. 2020; Izadinia and Seitz 2018; Li et al. 2015; Salas-Moreno et al. 2013]. The work of Avetisyan et al. [2018a] first proposes the concept of *Scan2CAD*, which aims to retrieve corresponding CAD models and align CAD models with objects from the given reconstructed result, but they predict the keypoints in the scanned data and then estimate the matching score between objects corresponding to each keypoint



(a) Info update (b) Relation-guided scene CAD recomposition (c) Relation-aware NBV generation

Fig. 2. Overview of our online scene CAD recomposition method. Given the current information fused by RGBD images captured in previous iterations (a), a relation-guided scene CAD recomposition module is applied to retrieve CAD models with poses optimized for relation constraints and local geometry alignment (b), and then the Next-Best-View (NBV) is generated to guide the autonomous scanning and refine the recomposition results by taking different kinds of ROIs into consideration.

and all CAD models, which is time-consuming. More recent works [Avetisyan et al. 2019, 2020] try to retrieve CAD models with the embedded feature and then align CAD models with objects with high efficiency. Some other works do not require reconstructed results as input, but instead recompose the objects directly from a single image [Gümeli et al. 2021; Kuo et al. 2020, 2021; Uy et al. 2021] or a sequence of RGB images [Han et al. 2021; Maninis et al. 2020; Shao et al. 2012]. Moreover, the work of [Avetisyan et al. 2020] also tries to optimize the poses of objects globally after corresponding CAD models are retrieved.

Compared with those previous works that either rely on an intermediate reconstruction result or a pre-shot video, generating the scene recomposition through one-pass scanning can not only save time and computational consumption but also enhance the retrieval accuracy via optimizing the viewpoints based on the current results on the way. Therefore, our work that aims to achieve an online method by generating the recomposition of the scene from scanned RGB-D observations in one pass has a quite different setting, which also results in new challenges. Different from the complete shapes of objects in the reconstructed result, the objects are often partially scanned during the scanning process which makes it difficult to get accurate retrieval and alignment results for objects. To overcome this obstacle, inspired by works [Wang et al. 2020; Zhang et al. 2021a] that use object relation to improve the 2D object detection accuracy or 3D object reconstruction quality, we proposed a relation-guided scene CAD recomposition, which utilizes relationships between objects and the layout to get more accurate object poses first and then further utilize the input geometry together with the optimized poses to guide the retrieval.

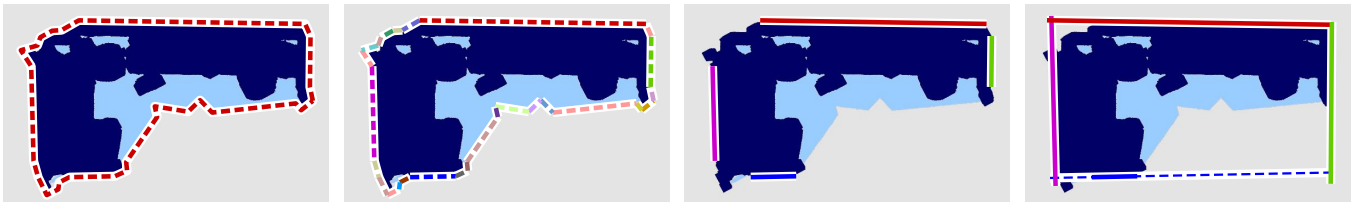
2.2 Autonomous scene reconstruction

Robots assisting people in conducting various tasks is a growing trend these days, and some researchers have started to focus on autonomous reconstruction to make a 3D digital representation of the real world. Unlike manual scanning, autonomous scanning with the robot is goal-driven, which reduces data redundancy and time consumption greatly. In the beginning, many works are dedicated to the scanning of single object [Krainin et al. 2011; Vasquez-Gomez et al. 2014; Wu et al. 2014], and then the goal is extended to the entire scene [Charrow et al. 2015; Ramanagopal and Le Ny 2016; Schmid et al. 2020; Xu et al. 2015, 2016]. There are also works trying to enhance the reconstruction quality simultaneously with a fast covering of the unknown scene by paying more attention to objects of the scene [Guo et al. 2022; Heng et al. 2015; Liu et al. 2018, 2021; Roberts et al. 2017; Xu et al. 2017].

Since autonomous scanning greatly improves the efficiency and quality of surface reconstruction, we also try to generate the scene CAD recomposition for the unknown indoor scene efficiently by autonomous scanning. However, different from previous works that are reconstruction-oriented, when aiming for the scene CAD recomposition task, the generated viewpoints should be able to help retrieve more accurate CAD models and obtain more reliable relationships between both objects and layout, instead of more surface details of the objects. Therefore, we propose a relation-aware NBV generation module, which uses the estimated relations to guide the selection of viewpoints that help both retrieving and pose estimating at most. Note that the method of [Liu et al. 2018] also retrieves CAD models for partially scanned objects during the process to guide the selection of viewpoints, however, the purpose of this method is still to fully reconstruct all surfaces in the scene and this leads to a fundamental difference in the choice of viewpoints.

3 OVERVIEW

Given a random initial pose of the robot in an unknown indoor scene, our goal is to automatically output a sequence of optimized viewpoints to guide the online scanning of the scene for CAD recomposition, which consists of the room layout and a set of retrieved



(a) (b) (c) (d)

Fig. 3. Room layout construction. Given the 2D occupancy map, we first extract the boundary edges using the minimum polygon fitting method [Douglas and Peucker 1973], which results in many short edges (a), and thus we further cluster the short edges based on parallel relationships (b) to obtain several longer and dominant boundary lines around the obstacles instead of frontiers (c), which are then eventually extended to get the final enclosed floor region (d).

CAD object models with both geometry and arrangement resembling the 3D scene. Figure 2 shows an overview of one iteration of our CAD recomposition and NBV generation method.

In each iteration, our method first fuses all the historical RGBD scans captured in the past as well as their instance segmentation results obtained using Mask R-CNN [He et al. 2017] to get the partial scan of each object, where object instance is maintained by the KNN method [Cover and Hart 1967] based on the bounding boxes of objects. Then the fused point cloud is further projected to the ground to get the updated 2D occupancy grid, as shown in Figure 2 (a). The updated information will be passed to the relation-guided scene CAD recomposition module to get the optimized layout and object pose together with the corresponding retrieved CAD models, as shown in Figure 2 (b). Then, the updated occupancy grid of the scene, the retrieval results of objects, and the relation groups will be collected and fused by the NBV generator to obtain a new viewpoint to enter the next iteration, as shown in Figure 2 (c). Iteration ends if no more new viewpoints can be generated.

Relation-guided CAD recomposition. Given the partial scan of each object, a global shape feature is extracted together with its 9DoF pose represented by the oriented bounding box (OBB) through the object pose estimation module. In the meanwhile, the room layout is also constructed based on the 2D occupancy grid. Then they are passed to the relation-guided optimization module to get more accurate estimations with the relation constraints. Finally, for each object, its partial scan and its refined pose will be used to retrieve the most similar CAD model from a given dataset that also aligns well with the partial input geometry. Note that during the relation-guided optimization, we also output the relation confidence between different objects, which is later used to help guide the following NBV generation. More details about the relation-guided CAD recomposition module are provided in Section 4.

Relation-aware NBV generation. To find the viewpoint that serves our task best, three different kinds of interest points are considered to guide the NBV generation, including frontier points, object points, and relation points. The frontier points are generated from the occupancy grid to guide the robot to scan unknown areas. The object points are generated from the comparison between each object with its retrieved CAD model to obtain the region to be observed that can improve the retrieval confidence of the object. The relation points tell the robot the relations of which areas need to be confirmed and updated in the scene. We determine an ROI region around each interest point and fuse all the ROI regions together to build an interest map of the current scene so that it can be used to identify several locations with high interest, then we generate the viewpoint that points to the selected location with the minimal transfer effort of the robot given the current pose. More details about the interest point construction and NBV generation can be found in Section 5.

4 RELATION-GUIDED CAD RECOMPOSITION

In each iteration, we get the fused partial scan of each object, and our goal is to retrieve a similar CAD model for each object with an accurate pose. Our key observation is that as we only get a partial observation of each object during the scanning process, directly using the global latent code of the partial input to simultaneously retrieve the object and estimate the pose as in previous works [Avetisyan et al. 2018a, 2019, 2020; Ishimtsev et al. 2020; Izadinia and Seitz 2018] will lead to inaccurate results, and the retrieval will benefit with a more accurate pose estimation. Therefore, we first get an initial object pose estimation when considering each object separately, and then later use the relation between the objects as well as the relation between the object and layout to perform a global refinement of all the object poses and the layout. The refined object poses together with the partial point cloud are then used to get a better retrieval result with higher similarity and more accurate alignment.

4.1 Room layout construction

Once the new point cloud captured by the new observation is fused with the existing point cloud of the scene, we construct the room layout consisting of the floor and walls based on the corresponding 2D occupancy grid.

The 2D occupancy grid is generated with a side length of 5cm for each pixel and a dilation radius of 10cm for obstacles, and each pixel is assigned one of the three states: unknown, free, or obstacle. Based on the assumption that the floor must contain both free and obstacle areas in the occupancy grid, we first approximate such areas with a polygon using the minimum polygon fitting method [Douglas and Peucker 1973]. However, the results obtained from the polygon fitting method usually contain too many short edges, while in reality, the floor boundary usually consists of long and perpendicular edges. Therefore, we further use a heuristic method to group those short edges into longer edges. In detail, edges with angles less than 2 degrees are considered to be parallel. We first merge all short neighboring edges if they are parallel, and then find

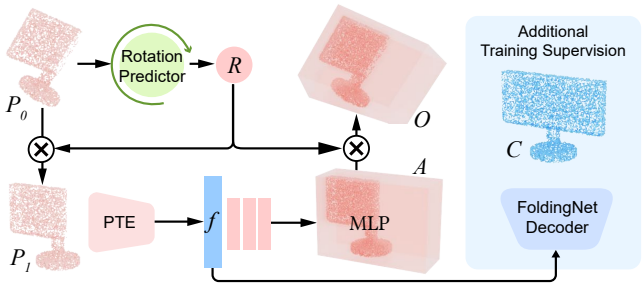


Fig. 4. The network structure for object pose estimation. We first predict the rotation matrix R for the normalized partial scanned point cloud of the object P_0 to get the point cloud P_1 in canonical space. Then the shape feature f is extracted and the axis-aligned bounding box A corresponding to the complete shape of the object is predicted. An additional shape completion supervision is activated only when training to help with the bounding box prediction. The oriented bounding box O of the input point cloud P_0 is finally generated by rotating A with the inverse rotation matrix $R^\top$.

and remove the local bands that appear in the middle of two parallel edges to get all long edges. Finally, we use the extension lines of the top few long edges to subdivide the space and enclose the floor region, which results in a boundary polygon as shown in Figure 3.

To further construct the walls, we generate a plane with a predefined height passing through the edge and perpendicular to the floor if there are sufficient 3D points projecting on such edge. As a result of the layout construction step, we will get a floor boundary and several walls.

4.2 Object pose estimation

For the partial scan of each object, we consider the pose estimation as a variant of the amodal detection problem [Deng and Latecki 2017; Qi et al. 2019; Yu et al. 2022], where the oriented bounding box (OBB) matching the *complete* shape of the object is predicted. However, as we find that it is not robust to predict the OBB directly, we adopt a two-step prediction method, where we first rotate the input into a canonical space to make it axis-aligned and then predict the axis-aligned bounding box (AABB) in such space. The predicted AABB can then be transformed back to the original world coordinate to get the desired OBB. Note that as our input is a *partial* point cloud while our goal is to predict the bounding box of the corresponding *complete* shape, we add the shape completion in the canonical space as an auxiliary task to help improve the accuracy of the second AABB prediction step.

Figure 4 shows the network structure of the object pose estimation. Given the partial point cloud of the object, we first sample n=2048 points using farthest point sampling (*FPS*) and represent the point cloud using the local coordinate to get the input of our network P_0. Then we predict a 6D rotation vector R as in [Zhou et al. 2018] with pointnet [Qi et al. 2016] as the encoder to transform P_0 into the canonical space. The updated point cloud P_1 is then passed to a point transformer encoder [Zhao et al. 2020] to get a global feature f. Then this global feature f is used to present the AABB A and the completed shape C, where the decoder for AABB prediction consists of 3 MLP layers while the decoder for shape completion uses FoldingNet [Yang et al. 2017] as in the SOTA shape completion method [Yu et al. 2021]. Finally, the predicted AABB can be transformed back with the inverse rotation matrix $R^\top$ to get the desired oriented bounding box O.

The loss function of the pose estimation network is then defined as:

$$L_{\text{obj}} = \omega_{\text{r}} L_{\text{r}} + \omega_{\text{b}} L_{\text{b}} + \omega_{\text{c}} L_{\text{c}} \tag{1}$$

where L_{r}, L_{b}, and L_{c} are the rotation, AABB, and completion losses against the ground-truth. In our experiments, we set the weights ω_{r}, ω_{b} and ω_{c} as 1, 1 and 10, respectively. The rotation loss L_{r} is simply L_2 loss. For the AABB prediction loss L_{b}, we use a modified *3D Focal EIoU Loss* based on the method [Zhang et al. 2021b], making it more sensitive to the difference between two axis-aligned bounding boxes and leading to more accurate prediction result. For the completion loss L_{c}, we propose a weighted chamfer distance loss by adding a combination weight on the second term of standard chamfer distance function to take the distance of a completed point to the input partial points into consideration. Intuitively, when completing a shape from a partial input, it is usually more difficult to generate missing points that are far away from the input, so we set a weight to give more priority to those points. The completion loss L_c is defined as:

$$L_c(P_1, C, \hat{C}) = \sum_{p \in C} d^2(p, \hat{C}) + \frac{1}{D(P_1, \hat{C})} \sum_{q \in \hat{C}} d(q, P_1) d^2(q, C) \tag{2}$$

where P_1 is the partial input transformed in the canonical space, C and $\hat{C}$ are the predicted and ground-truth completed shapes, respectively. The point-to-set and set-to-set distances are defined as $d(p, Q) = \min_{q \in Q} ||p - q||_2$ and $D(P, Q) = \max_{q \in Q} d(q, P)$.

4.3 Relation-guided optimization

With the extracted layout and estimated pose for each object, we further utilize their relations to perform a global optimization for refinement. To make full use of the relationships between objects and the layout, the floor and walls are considered separately, since the floor mainly provides direct or indirect supporting relationships to all objects while the walls may provide possible parallel or perpendicular relations to adjacent objects. For all the element pairs, including object-object, object-wall, and object-floor, we would like to further ensure that there is no collision. To consider and constrain all kinds of relations mentioned above, we construct a scene graph and design a *relation-guided graph convolutional network (RGCN)* for optimization.

Scene graph construction. The scene graph we constructed is a fully connected graph connecting all the elements in the scene, including objects, walls, and the floor. For each object node, we denote it as $O = (f_O, B_O)$, where f_O is the encoded shape feature during pose estimation, and B_O is the corner set of the predicted OBB in Section 4.2. Note that as the OBB of each object is derived from the predicted AABB in the canonical space, we can get a consistent corner for all the object boxes. For each wall node, we denote it as $W = (n_W, c_W, B_W)$, where n_W is the unit normal vector, c_W is the center position, and B_W is the corner set of the wall boundary. For the floor node, we denote it as $F = (n_F, c_F, z_F, B_F)$, where n_F is the

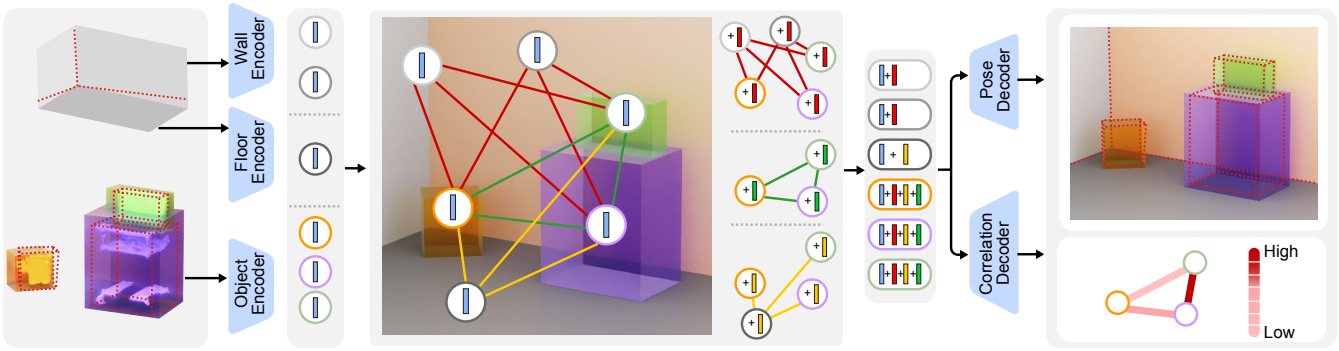


Fig. 5. The network structure of our *relation-guided graph convolutional network* (*RGCN*). With the constructed scene graph, we first encode the node features using three different encoders based on the node type, and then three different subgraphs obtained by grouping different types of edges (indicated using different colors) are used to compute the updated messages separately, which are then added to the initial node features to get all features updated. The optimized pose for each object and the correlation confidence between each pair of objects are then predicted from the updated features, respectively. The ground truth layout and object bounding boxes are drawn using the dashed lines, while the predicted ones before and after the optimization are drawn using transparent boxes for comparisons. Note how the object poses have been refined through our relation-guided optimization.

unit normal vector, c_F is the center position, z_F is the z value of the floor in the world coordinate system where we assume that the floor is perpendicular to the z-axis, and B_F also denotes the corner set of the floor boundary. The full set of graph nodes can be represented as $V = \{O_1, O_2, \ldots, W_1, W_2, \ldots, F\}$. For each edge E_{ij} connecting two nodes V_i and V_j, we denote it as $E_{ij} = (d_{ij}, r_{ij})$, where d_{ij} is the distance between two nodes and r_{ij} is the ratio between the overlapped region and the union volume (IoU) of the bounding boxes of those two nodes. Note that there are three different types of edges considering the types of connecting nodes, including *OO* for object-object, *OW* for object-wall, and *OF* for object-floor. d_{ij} is equal to the distance of centers between two objects for *OO* edges and the minimal distance between the center of the object and the floor or the wall otherwise. Moreover, for the floor and walls that were originally just sliced, we added a small thickness to get the bounding box to the computation of overlap.

Support distance calculation. To make sure that object poses are physically valid with sufficient support, we extract the support relations between objects and calculate the support distance for each object to indicate how much it is floating or overlaps with other objects. First, the oriented bounding boxes and the centroids of all scanned objects are projected onto the floor along the z-axis, and the objects are clustered into different groups based on whether the projected centroid of an object is in the projected oriented bounding box of another object. Then, for each group, we construct the support relations by comparing the z values of centroids to get a sorted object list $O = \{obj_0, obj_1, obj_2, ..., obj_k\}$ with increasing z values and obj_0 referring to the floor. After that, we calculate the distance between each pair of adjacent objects in O, and define the support distance for each object obj_i as follows:

$$d_s(obj_i) = \sum_{j=0}^{i-1} SD(B_j, B_{j+1}) \tag{3}$$

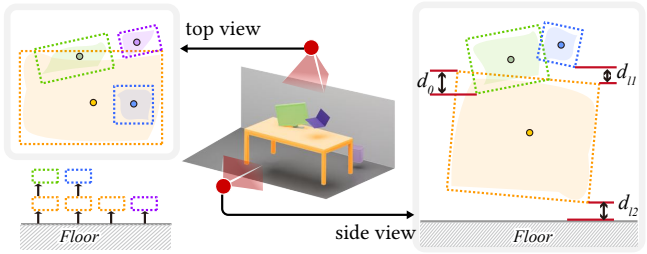


Fig. 6. Support relation identification (left) and support distance term computation (right). We project both the bounding box and the centroid of the point cloud of each object from the top to the floor to determine the support relation. For two vertically adjacent objects, the one on the top will be supported by the one on the bottom if their projecting boxes are overlapped and the top centroid is inside the bottom projecting box as well. For the support distance, we compute both overlap distances, e.g., d_o and levitation distances, e.g., d_{l1} and d_{l2}, and accumulate all the distances between the object and the floor following the support relationship shown in the left to get the final support distance term for each object.

where B_j is the oriented bounding box of obj_j in O, and $SD(B_j, B_{j+1})$ is the distance between the top surface of B_j and the bottom surface of B_{j+1}, which is defined as the absolute value of the minimal signed distance between points on those two surfaces. Figure 6 gives an example of the extraction of support relations and the computation of support distance.

Relation-guided graph convolutional network (RGCN). We use three different node encoders to encode three different types of graph nodes into features with the same dimension and use the same edge encoder for all the edges. Then for the message passing in RGCN, as we want the network to automatically learn the different impacts of different types of relations on the final optimization, we use different subnetworks to compute the message pass for different subgraphs grouped based on edge types. Once the information gets accumulated, all the node features will be updated

and used for the object OBB and layout refinement through a pose decoder. Moreover, we also use the updated node features to output the *correlation confidence* between each pair of objects, which will later be used for NBV generation to check the objects with high correlation in more detail.

Figure 5 illustrates the main structure of the RGCN and the process of the relation-guided optimization. More network details can be found in the supplementary material. When training the RGCN, we use the following loss function:

$$L_{scene} = \omega_{\mathrm{v}} L_v + \omega_{\mathrm{f}} L_f + \omega_{\mathrm{w}} L_w + \omega_{\mathrm{c}} L_c \tag{4}$$

where L_v is the overlap loss measured for each pair of graph nodes, L_f is the support loss measured between the floor and each object node, and L_w is the orientation loss measured between each wall and its adjacent objects, whose distance to the wall is less than 20cm. L_c is the correlation loss that measures the probability of two objects being correlated, i.e., close-by or aligned. In our experiments, we set $\omega_{\mathrm{v}} = 0.2$, $\omega_{\mathrm{f}} = 0.5$, $\omega_{\mathrm{w}} = 0.1$, and $\omega_{\mathrm{c}} = 0.2$.

All these loss terms are defined in more detail as follows:

$$L_v = \sum_{V_i \neq V_j \in V} IoU(V_i, V_j) \tag{5}$$

$$L_f = \sum_{O_i \in V} d_f(O_i) \tag{6}$$

$$L_w = \sum_{W_j \in V, O_i \in \mathrm{Ad}(W_j)} d_w(O_i, W_j) \tag{7}$$

$$L_c = CE(\{C(e)\}_{e \in OO}, \{C_{gt}(e)\}_{e \in OO}) \tag{8}$$

where IoU is calculated using the volumetric representation with the dimension of 10^3, d_f is the accumulated support distance of the object, d_w is computed as the minimum angle to make either x-axis or y-axis of the object OBB parallel or perpendicular to the wall. CE is the cross-entropy loss function, and $C(e)$ and $C_{gt}(e)$ are the predicted correlation and the ground truth correlation extracted using the GT boxes of the edge e. Moreover, the ground truth correlation C_{gt} between each pair of objects with poses $P = (t_p, r_p)$ and $Q = (t_q, r_q)$ connected by the edge $\boldsymbol{e}$ is defined as:

$$C_{gt}(e) = \exp(w_S SD(P, Q) + w_T TE(t_p, t_q) + w_R RE(r_p, r_q)) \tag{9}$$

where SD is the distance between two oriented bounding boxes previously used in Equation 3, and TE and RE are the translation error and rotation error defined as follows:

$$TE(t_p, t_q) = ||t_p - t_q||_2 \tag{10}$$

$$RE(r_p, r_q) = \frac{\arccos(1 - ||(r_p - r_q) \cdot \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^\top ||_2^2 / 2)}{100} \tag{11}$$

Note that if there does not exist a support relation between P and Q, we set the value $SD(P, Q)$ as 1. The combination weights w_S, w_T, and w_R are set as -0.6, -0.2, and -1 in our experiment.

4.4 Geometry-aligned retrieval

Unlike previous works, which usually use an implicit latent feature only for CAD model retrieval and lead to unstable results when the input is only partially scanned, we split the retrieval task into two steps, where we first use the implicit feature to retrieve top-k CAD models and select the one aligned best with the input partial point cloud given the estimated OBB.

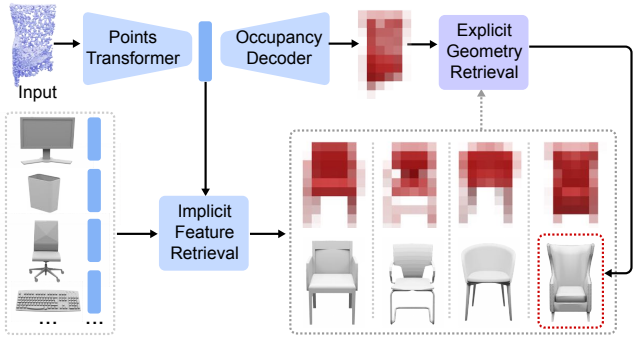


Fig. 7. The network structure of for geometry-aligned retrieval. Given the partial point cloud of the object as input, we first retrieve top-k candidates CAD model from the dataset by matching the implicit feature. Then the occupancy of the input point cloud is decoded and compared with the occupancies of all candidate models to select the one with the best geometry-matching as our final output.

To compute the geometry alignment efficiently, we adopt an encoder-decoder network structure to predict the complete occupancy of the input partial scan inside the estimated OBB, where the resolution is set to be 10^3, as shown in Figure 7. At the same time, we also ensure that the embedded feature is similar to the one obtained by the corresponding ground-truth CAD model during the training. During inference, we first use the embedded feature to retrieve top-k CAD models from the given dataset, and then for each CAD model, we align its bounding box with the estimated OBB of the partial input and select the one with the lowest matching error, where the matching error between the occupancy of the CAD model M and the predicted occupancy of the partial scan P is defined as:

$$E_{\mathrm{occ}}(M, P) = \omega_P d_{\mathrm{occ}}(P, M) + \omega_M d_{\mathrm{occ}}(M, P) \tag{12}$$

where

$$d_{\mathrm{occ}}(A, B) = ||\max(\mathrm{OCC}(A) - \mathrm{OCC}(B), 0)||_2 \tag{13}$$

with $OCC(*)$ indicating the occupancy of the corresponding input scan. We set $\omega_P = 0.8$ and $\omega_M = 0.2$ to give more penalty for parts that are only located on the partial scans but not occupied by CAD models.

Once the CAD model is selected and its initial pose is determined by the bounding box fitting, we perform 10 iterations of the ICP method [Besl and McKay 1992] to fine-tune the alignment between the CAD model and the partial points to get the final CAD recomposition.

5 RELATION-AWARE NBV GENERATION

Once the CAD recomposition is generated, our method automatically selects the next viewpoint to explore the unknown scene as well as further improve the recomposition quality via autonomous scanning. As the CAD recomposition can benefit from more object and relation information, other than the frontier points commonly used to guide the exploration, we further define a set of interest

points based on current CAD recomposition result to localize the region of interest (ROI) and then generate NBV pointing to the most interesting region with less moving efforts.

5.1 Interest point construction

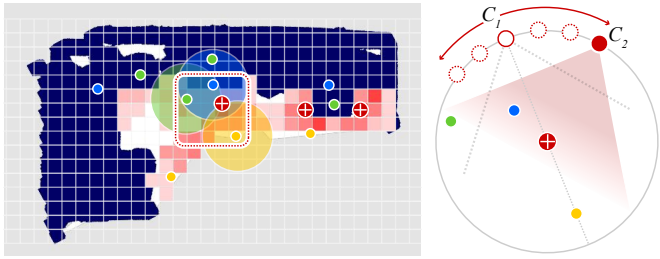
Frontier point. The frontiers of the scanned scene are generated from the updated occupancy grid via the Canny algorithm [Canny 1986]. All pixels lying on the frontiers are clustered by the connectivity. For each pixel cluster of frontiers, we regard the diagonal length L_{diag} of its minimal bounding box as the approximate length of the frontier corresponding to the cluster. The frontier points are extracted by the farthest point sampling method with the number of $N_{exp} = L_{diag}/\delta_{exp}$, where $\delta_{exp} = 0.5m$, which results in at least one frontier point for each cluster.

Object point. For each object, we compare the occupancy grid of the input scan and the retrieved CAD model. If more than half of the corresponding grids have a difference larger than 0.5, then we generate an object point. We first find the point p_m on the CAD model that is the farthest to the input scan, and then find the point p_s on the input scan that is closest to p_m. To make the interest point can give more information to guide better retrieval but at the same time provide a new scan that has sufficient overlap with the current scan, the new object point is defined as $p = 0.75p_m + 0.25p_s$.

Relation point. For the relation point, we would like to discover object pairs that are supposed to have strong relationships but are not satisfied in the current CAD recomposition, where the deviation weights are assigned to the corresponding edges to form interest groups and one relation point is generated for each group.

In more detail, we mainly care about two types of relationships: *support* and *correlation*. For support, we can first identify object pairs that have a support relation based on the current object poses, denoted as the *support confidence*, as well as computing their current vertical distances to serve as the *support deviation* weight on the corresponding edges. For correlation, we can also identify object pairs that are supposed to have correlation based on the correlation probability obtained together with the refined object poses during the global optimization in Section 4.3, and then compute the *correlation deviation* weight based on the current object poses to add to the corresponding edges. Then, both support deviation and correlation deviation weights are combined to get the final *deviation* weights. Moreover, we also accumulate the likelihood of two objects having an important relationship, referred to as *relation confidences*, which is a combination of support confidence and correlation confidence mentioned above.

In order to find the regions where the poses of objects need further refinement, we first select objects having higher *deviation* with others and then cluster them into groups based on their *relation confidence*. In detail, the edges with *deviation* larger than 0.5, named *relation edges*, are selected, sorted, and grouped by the *relation confidence* from high to low. To split relation edges into several groups located in local regions, we iteratively select all the groups. Each time, we first select the edges with the highest relation confidence. We then select the edges from existing relation edges, making sure no more than 4 nodes remain in the current group. Once a group is selected, its nodes, along with the connected edges cannot be selected again. When the iteration ends, there are some groups of objects generated. In order to make the viewpoint cover as many objects as possible in the group, a relation point is generated at the center of the midpoints of all edges in each group. The robot can use this relation point to capture more accurate relations of the poses of objects in the group.



(a) (b)

Fig. 8. Candidate viewpoint generation. (a) Once the voxelized heatmap is constructed, we select several candidate ROI points (red points) with high heat values. (b) For each ROI point, taking the ROI point inside the dashed box for an example, we first generate an initial viewpoint C_1 pointing to both the candidate point and the frontier point (yellow point), and then sample new viewpoints on both sides to find the one that has all interest points in view, including all yellow, green, and blue points assigned to this candidate ROI point, to get the corresponding viewpoint C_2.

5.2 NBV generation

Our goal is to select the next viewpoint that can not only explore the unknown area as much as possible but also help refine the CAD recomposition results in the aspect of both object pose and CAD retrieval. As exploration still plays the dominant role during autonomous scanning, we first generate a set of candidate viewpoints, which observe as much of the unscanned area as possible and at the same time ensure that there are at least one object point and one relation point included in the observation, and then select the one with minimal traveling cost considering the current robot pose as the NBV. Note that if the candidate viewpoint is unreachable, the corresponding traveling cost will be infinite and such viewpoint will not be selected.

More specifically, once all points of interest are obtained, we generate spherical regions with a radius of 1m centering each point as the point-wise ROI. A voxelized heatmap is then constructed with a side length of 20cm per voxel, and the heat value on each voxel v at the region of free areas is defined as:

$$h(v) = \omega_{exp}\mathbb{1}_{exp}(v) + \omega_{obj}\mathbb{1}_{obj}(v) + \omega_{rel}\mathbb{1}_{rel}(v) \qquad (14)$$

where $\mathbb{1}_{exp}$, $\mathbb{1}_{obj}$ and $\mathbb{1}_{rel}$ are indicator functions that indicate whether the voxel v is contained in the any of the ROIs determined by the set of frontier points, object points, and relation points. We set $\omega_{exp} = 0.7$, $\omega_{obj} = 0.2$, and $\omega_{rel} = 0.1$ to reflect importance of exploration.

Candidate viewpoints are then generated based on the constructed voxelized heatmap, as shown in Figure 8. We first select several discrete voxels with the highest heat values as the candidate points,

ensuring the distances between each two candidate points are no less than 1m, shown as the red points in Figure 8 (a). Note that during the selection, we further ensure that the candidate points are near the frontier points by requiring the heart value to be larger than 0.7, and the scanning process terminates if no more points satisfy the condition. For each candidate point, we generate an initial viewpoint C_1 observing the most unknown area (pointing to the yellow frontier point) and then sample new viewpoints on both sides of C_1 on a 1m-radius circle centering at the candidate point until all nearby interest points are in view to get the corresponding viewpoint candidate C_2; see Figure 8 (b). Finally, we choose the viewpoint candidate that takes minimal effort for the robot to transfer from the current position as our NBV.

6 RESULTS AND EVALUATION

6.1 Experiment setup

Environment. We use *Habitat* [Savva et al. 2019] as our simulation environment for rapid experiment iterations, which supports the standard control of the robot and simulates the interactions with the real world on GPU. To better simulate the camera noise in reality, we add the synthetic noise proposed in the work of [Handa et al. 2014] to the depth observation. A desktop PC processes all data with AMD Ryzen 9 5900X CPU (3.7GHz×12), 48GB RAM, and an Nvidia GeForce RTX3080 GPU.

When conducting experiments in the real world, we use `Fetch` [Wise et al. 2016] with a `RealSense Depth Camera D435i` held in hand for scanning and a `SICK 2D sensor` for tracking and navigating.

Dataset. We conduct our experiments on the *Scan2CAD* dataset [Avetisyan et al. 2018b], which consists of 1506 realistic scenes from the *ScanNet* dataset [Dai et al. 2017] with 14225 aligned CAD models from the *ShapeNet* dataset [Chang et al. 2015] containing 55 common object categories. We train our networks on 1204 scenes (80%) and test on the remaining 302 scenes (20%) following the split in *Scan2CAD* [Avetisyan et al. 2018b].

6.2 Evaluation metrics

As our goal is to recompose the unknown indoor scene with CAD models via autonomous scanning, we evaluate our method in two aspects: *scanning efficiency* and *recomposition accuracy*.

Scanning efficiency. To evaluate the efficiency of autonomous scanning, we mainly focus on the resource and time consumption during the scanning process.

We first use the same metrics **Distance Consumption (DC)** and **Time Consumption (TC)** as in the work of [Guo et al. 2022], which are measured by the average distance of the total scanning path in meters and the average scanning time of the robot in minutes. Then, we use **NBV Count (#NBV)** to measure the average number of the generated NBVs during the scanning process, and **Storage Consumption (SC)** to measure the average of the maximum RAM and GPU memory consumption for storing intermediate data during the scanning process in GBs.

Recomposition accuracy. To evaluate the accuracy of CAD recomposition, we mainly focus on the accuracy of semantic categories of retrieved CAD models, denoted as **Class Accuracy (CA)**, and their alignment with GT CAD models.

Table 1. Quantitative comparison with offline CAD recomposition baselines, including `Scan2CAD` [Avetisyan et al. 2018a], `SceneCAD` [Avetisyan et al. 2020] and `Interactive Scene Reconstruction (ISR)` [Han et al. 2021], with either manual scans (`Manual`) provided in the `ScanNet` [Dai et al. 2017] dataset or autonomous scans (`Auto`) generated by our method as input.

#Frame	Method	TC ↓	SC ↓	CA ↑	PA ↑	PE ↓	GE ↓
Manual (94)	Scan2CAD	10.373	20.814	0.484	0.307	0.445	0.049
	SceneCAD	11.763	23.624	0.491	0.524	0.442	0.042
	ISR	3.118	26.192	0.412	0.496	0.440	0.024
	Ours	**2.940**	**5.325**	**0.552**	**0.628**	**0.357**	**0.011**
Auto (24)	Scan2CAD	9.854	13.482	0.467	0.284	0.509	0.054
	SceneCAD	11.072	14.526	0.483	0.492	0.515	0.046
	ISR	0.868	15.517	0.391	0.473	0.520	0.026
	Ours	**0.791**	**5.271**	**0.562**	**0.636**	**0.352**	**0.010**

For alignment accuracy considering the object poses only, we follow the metric used in related works [Avetisyan et al. 2018a, 2019, 2020; Gümeli et al. 2021; Maninis et al. 2020], which considers the translation error (**TE**), rotation error (**RE**) and scaling error (**SE**) between the predicted pose $P_p = (t_p, r_p, s_p)$ and the ground truth pose $P_{gt} = (t_{gt}, r_{gt}, s_{gt})$ of the scanned object, where TE and RE are defined in Equation 10 and 11, and SE is defined as:

$$SE(s_p, s_{gt}) = \frac{||s_p - s_{gt}||_2}{||s_{gt}||_2} \tag{15}$$

where t_* is the center position of the object, r_* is the 3×3 rotation matrix of the object, and s_* are the XYZ-ordered scaling ratios of the oriented bounding box of the object.

We then use **Pose Accuracy (PA)** to measure the proportion of retrieved CAD models with $TE \leq 0.2$, $RE \leq 0.2$, and $SE \leq 0.2$ as in previous works, and use **Pose Error (PE)** to further measure average alignment error:

$$PE(P_p, P_{gt}) = TE(t_p, t_{gt}) + RE(r_p, r_{gt}) + SE(s_p, s_{gt}) \tag{16}$$

For alignment accuracy further considering the object geometry, we define a new metric **Geometry Error (GE)** to calculate the chamfer distance between the GT CAD model and the corresponding retrieved CAD model. Let us denote the set of GT CAD models as O and the retrieved CAD models as C, then the **Geometry Error (GE)** is defined as follows:

$$GE(O, C) = \frac{1}{|O|} \sum_{\boldsymbol{o} \in O} TCD(\boldsymbol{o}, C) \tag{17}$$

where TCD is the trunked chamfer distance function defined as:

$$TCD(\boldsymbol{o}, C) = \begin{cases} min(CD(\boldsymbol{o}, \boldsymbol{c_o}), 1) & \exists \boldsymbol{c_o} \in C \\ 1 & else \end{cases} \tag{18}$$

where CD is the chamfer distance and $\boldsymbol{c_o}$ is the corresponding retrieved CAD model for GT CAD model $\boldsymbol{o}$. We truncate the chamfer distance with 1 to ignore the intolerable value brought by the terrible retrieval result, and at the same time add a penalty for the object that has annotated CAD model but without retrieved CAD models.

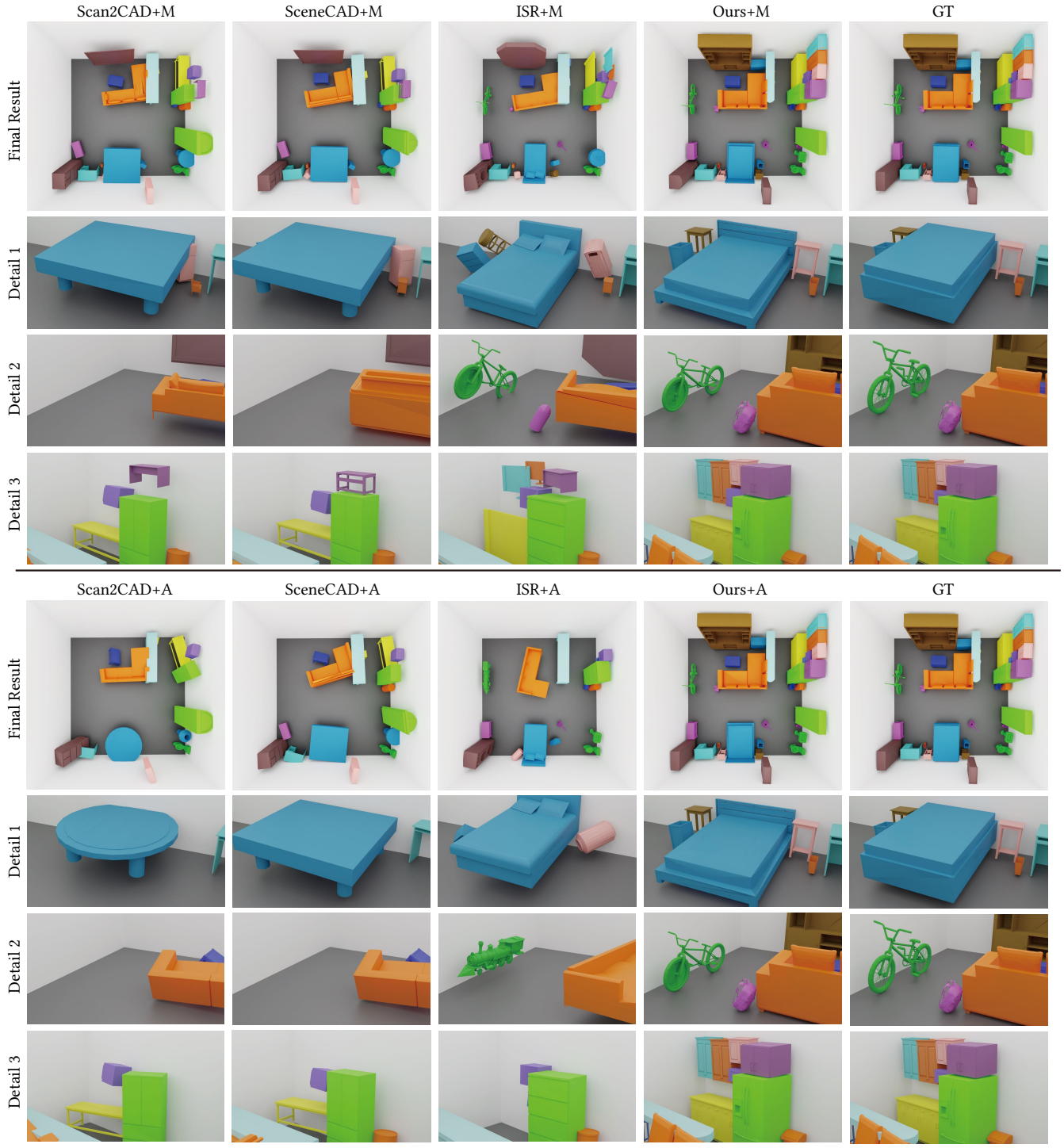


Fig. 9. Qualitative comparison with offline CAD recomposition baselines, including Scan2CAD [Avetisyan et al. 2018a], SceneCAD [Avetisyan et al. 2020] and Interactive Scene Reconstruction (ISR) [Han et al. 2021], with either manual scans (Manual) provided in the ScanNet [Dai et al. 2017] dataset or autonomous scans (Auto) generated by our method as input.

6.3 Comparison with offline CAD recomposition baselines

We first compare our method to existing CAD recomposition methods, including Scan2CAD [Avetisyan et al. 2018a], SceneCAD [Avetisyan et al. 2020] and Interactive Scene Reconstruction (ISR) [Han et al. 2021]. Note that all those existing CAD recomposition methods are offline methods, requiring either the reconstructed scene or a pre-captured RGBD scanning sequence of the scene as input. Thus, we conduct comparisons with either manual scans (Manual) provided in the ScanNet [Dai et al. 2017] dataset or autonomous scans (Auto) generated by our method as input. The extra scene reconstructions required by Scan2CAD and SceneCAD are obtained using the method of *Voxblox++* [Grinvald et al. 2019] .

The quantitive comparisons in terms of scanning efficiency and recomposition accuracy are shown in Table 1. We can see that our method gets consistently better performance when compared to those baseline methods with either Manual and Auto input.

For *scanning efficiency*, the storage consumption of our method is much lower than all other methods since other methods need to save the detailed dense point cloud of the entire scene to make sure the reconstructed result is accurate. For *recomposition accuracy*, both Scan2CAD and SceneCAD take the volumetric representation of the entire reconstructed scene as input and such rough input representation leads to inaccurate object pose detection. ISR gets slightly better results than those two baselines as it retrieves CAD models during the process of reconstruction, where the instance segmentation of scanned objects is estimated once a new framework is given to guide a more accurate retrieval and achieve a lower GE. For our method, other than the instance segmentation, we further predict the OBB of the corresponding complete shape with the provided partial point cloud, This complete OBB prediction together with our relation-guided optimization and geometry-aligned retrieval results in significant improvement in recomposition accuracy.

When comparing the results between Manual and Auto inputs, we can see a general performance degradation of all other methods, as the generated frames are quite sparse and it is difficult to estimate accurate object poses and further retrieve corresponding CAD models given the partially scanned objects. However, the performance of our method is relatively stable as the scans are selected by our relation-aware NBV generation module tailored for the CAD recomposition task.

Some qualitative examples are shown in Figure 9. Note that since only SceneCAD and our method estimate the layout of the scene in addition to the retrieved objects, we add the layout generated by our method for other methods with a consistent rendering style. We can see that our method is able to detect more objects compared to Scan2CAD and SceneCAD, and further retrieve corresponding CAD models with the guidance of relations between objects compared to ISR, leading to the overall best results. Besides, our method maintains similar CAD recomposition capabilities on the autonomous scanned data compared with the manual scanned data, while all other methods tend to recognize only a few objects in the scene and, in the meanwhile, the predicted poses of recognized objects are inaccurate under the sparse observations.

6.4 Comparison with online CAD recomposition baselines

As far as we know, we propose the first method for online CAD recomposition problem and there are no existing works that can be directly compared to. As our method consists of two important modules: relation-guided CAD recomposition (RelCAD) described in Section 4 and relation-aware NBV generation (RelNBV) described in Section 5, we replace either of those two modules with the state-of-the-art method to derive two baselines for comparison:

- ROCA+RelNBV, where we replace our RelCAD module with ROCA [Gümeli et al. 2021]. ROCA is the state-of-the-art method to retrieve corresponding CAD models for objects in a single RGB image. To incorporate ROCA into our online framework, for each new frame, we can first obtain a set of newly retrieved CAD models using ROCA, and then merge them with the existing retrieval results by calculating IoU and comparing semantic categories.
- RelCAD+AsyncScan, where we replace our RelNBV module with AsyncScan [Guo et al. 2022] . AsyncScan is the state-of-the-art method for reconstruction-oriented autonomous scanning, where the scanning strategy is designed to increase the coverage of the scene as well as the surface completeness of scanned objects.

Table 2 shows the quantitative comparison of our method to those two baselines. As ROCA retrieves the CAD model for each scanned object separately without considering their relationship, while our method uses object relations specifically to guide the object pose optimization, our method gets much higher recomposition accuracy than the ROCA+RelNBV baseline. For the RelCAD+AsyncScan baseline, AsyncScan is a reconstruction-oriented method that leads to more complete scene reconstruction and thus more accurate CAD recomposition results, however, it has a much higher storage consumption (SC) as it needs to reconstruct the scene during the whole scanning process. As a comparison, our method gets similar CAD recomposition results with much less storage consumption and scanning efforts thanks to our relation-aware NBV generation module. Some qualitative comparisons are shown in Figure 10. Note that all the results reported in Table 2 only reflect the differences between final CAD recomposition results, and although the final performances between AsyncScan and our method are quite similar in terms of recomposition accuracy, the intermediate results obtained with different viewpoints generated using different scanning strategies are actually quite different. Different scanning paths of different methods are shown in the first row in Fig. 10. We can see that the path generated by our scanning strategy (RelNBV) is much shorter than the reconstruction-oriented scanning strategy (AsyncScan), while the results are similar when equipped with our relation-aware CAD recomposition module (RelCAD). The final CAD recomposition results obtained by our method are quite similar to those of AsyncScan with high accuracy, which also proves that our relation-aware NBV generation strategy helps us find the most effective viewpoint for improving CAD recomposition results.

6.5 Ablation studies

In this section, we conduct several ablation studies to validate our design in both CAD recomposition and NBV generation modules. Qualitative examples can be found in the supplementary materials.

Table 2. Quantitative comparison with online CAD recomposition baselines, with either of the key modules of our method (RelCAD + RelNBV) replaced by the state-of-the-art method. ROCA+RelNBV refers to the baseline that replace RelCAD with ROCA [Gümeli et al. 2021], and RelCAD+AsyncScan refers to the baseline that replace RelNBV with AsyncScan [Guo et al. 2022].

Method	DC ↓	TC ↓	#NBV ↓	SC ↓	CA ↑	PA ↑	PE ↓	GE ↓
ROCA+RelNBV	45.533	7.116	27.675	8.708	0.482	0.342	0.705	0.387
RelCAD+AsyncScan	47.725	9.973	32.284	48.298	0.558	**0.636**	**0.352**	**0.010**
Ours (RelCAD+RelNBV)	**37.806**	**6.045**	**23.646**	**5.271**	**0.562**	**0.636**	**0.352**	**0.010**

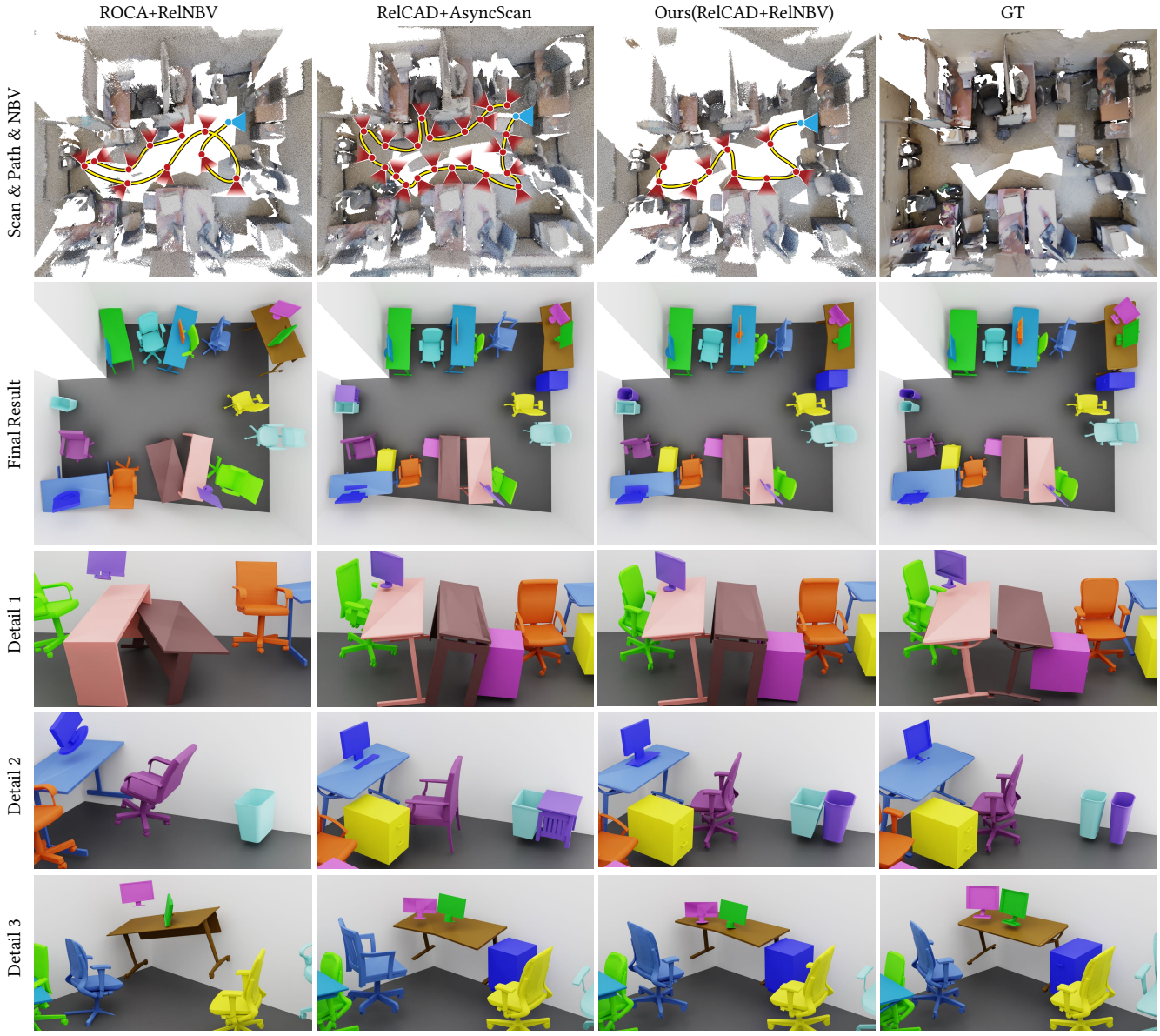


Fig. 10. Qualitative comparison with online CAD recomposition baselines ROCA+RelNBV and RelCAD+AsyncScan. In the first row, we show the fusion of all captured RGBD scans with the sequence of NBVs generated by each method, where the initial view is colored in blue.

Table 3. Ablation studies on key components of our CAD recomposition module.

Method	CA ↑	PA ↑	TE ↓	RE ↓	SE ↓	GE ↓
w/o Layout	0.513	0.548	0.122	0.106	0.173	0.028
w/o Pose	0.469	0.209	0.421	0.170	0.607	0.061
w/o Opt	0.476	0.311	0.167	0.128	0.352	0.048
w/o Geo	**0.562**	0.497	0.110	0.100	0.211	0.035
Ours	**0.562**	**0.636**	**0.101**	**0.089**	**0.162**	**0.010**

Ablation studies on our CAD recomposition module. Our relation-guided scene CAD recomposition module consists of four key steps, including room layout construction (`Layout`), object pose estimation (`Pose`), relation-guided optimization (`Opt`), and geometry-aligned CAD retrieval (`Geo`). To justify the design choices made in our CAD recomposition module, we compare the performance in terms of recomposition accuracy in the following settings:

- `w/o Layout`, where the layout construction module is removed and thus will affect the following relation-guided optimization.
- `w/o Pose`, where the new weighted chamfer distance loss defined in Equation 2 is replaced by the original chamfer distance loss.
- `w/o Opt`, where the relation-guided optimization module is removed and the CAD model for each object is retrieved separately once the corresponding pose is estimated. Note that without the relation-guided optimization module, relation will not be considered in the NBV generation step, thus no relation point will be generated for selection.
- `w/o Geo`, where the CAD models are retrieved directly based on their implicit features without considering the geometric alignment.

Table 3 shows the results of all those settings compared to our full pipeline. We can see that our full pipeline gets the consistently better performance, especially on the GE metric, which shows that our designs can indeed improve the accuracy of CAD recomposition. In more detail, when compared to the (`w/o Layout`) setting, the introduction of layout in relation-guided optimization helps improve the object pose estimation, especially the ones that have a close relation to the layout, and thus improve the overall results. For object pose estimation, we notice that it's quite important to get a relatively accurate initialization for later optimization, and without our weighted chamfer distance loss (`w/o Pose`), the initial object pose estimation is inaccurate and thus the final performance drops significantly. When lacking the relation-guided optimization component in the (`w/o Opt`) setting, some objects will overlap with other objects or the layout, resulting in large errors and undesired results. The retrieved CAD models with only implicit features of objects obtained in the (`w/o Geo`) setting can have correct semantic categories but less accurate matching in geometry and pose, leading to worse performance in alignment accuracy.

Ablation studies on our NBV generation module. Our relation-aware NBV generation module consists of three different types of interest points, including frontier points, object points, and relation points. As frontier points are always needed during autonomous scanning to ensure the exploration of unknown scenes, we make comparisons to two settings, including (`w/o OP`) by removing the object points and (`w/o RP`) by removing the relation points. Moreover, as the relations used to generate the relation points are predicted from the relation-guided optimization module, to show its robustness, we also compare with the setting where the relation scores are calculated explicitly using Equation 9, denoted as (`with ER`) .

Table 4. Ablation studies on key designs of our NBV generation module.

Method	DC ↓	TC ↓	#NBV ↓	CA ↑	PA ↑	PE ↓	GE ↓
w/o OP	39.866	6.427	26.589	0.497	0.524	0.450	0.026
w/o RP	39.971	7.389	28.977	0.511	0.501	0.474	0.030
with ER	38.423	6.145	26.417	0.537	0.630	0.371	0.016
Ours	**37.806**	**6.045**	**23.646**	**0.562**	**0.636**	**0.352**	**0.010**

Table 4 shows the comparison results. Without either object points or relation points, the NBV generation module tends to explore the unknown area, and thus fewer viewpoints are generated to gather the information about objects or their relations in the scene, which results in lower recomposition accuracy.Moreover, the explicitly calculated relations can be inaccurate for the relation between partially scanned objects due to the unstable object pose estimation, and thus the corresponding generated NBVs cannot really reflect the current recomposition state and improve the results. We found that the relation predicted by our optimization module is more robust to the object pose noises and leads to more efficient and stable NBV generation. Some example results can be found in the supplementary materials.

6.6 Qualitative results

Figure 11 shows some results we obtained in our simulation environment, where the virtual scenes are shown on the left and the corresponding CAD recomposition results are shown on the right. We can see that our method can work well on indoor scenes with different layouts. Note how those irregular room boundaries are correctly constructed by our method. Moreover, most of the objects in those given scenes are successfully detected and accurately aligned with the retrieved CAD models, even those the input scene itself is somewhat incomplete.

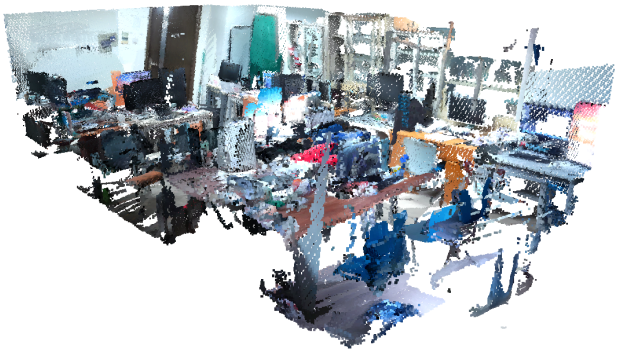
Other than the simulated environment, we also tested our method in the real world by scanning two unknown indoor scenes with different scales, including one office room and one meeting room. Figure 12 shows one example result of our method on real-world scenes, and the other result can be found in the supplementary material. Although the arm of `Fetch` introduces some noise for camera pose and the position deviation for localization when moving, we are still able to recompose the scene with accurately retrieved and aligned CAD models that have the same spatial structure and distribution as the real scene.

7 CONCLUSIONS

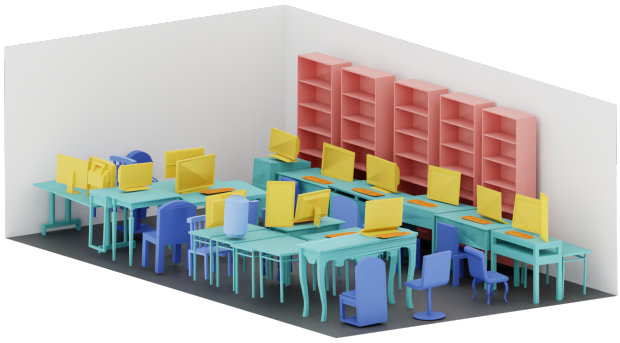
We present an online scene CAD recomposition method with one-pass autonomous scanning to retrieve corresponding CAD models for objects and estimate the layout of the unknown indoor scene. A novel relation-guided CAD recomposition module is designed to use relation-constrained global optimization to get accurate object pose

Virtual scene Our result

Fig. 11. Example results generated by our online scene CAD recomposition method, with the corresponding virtual scene shown on the left.



(a) Raw scan of real scene



(b) Our result

Fig. 12. Our CAD recomposition result for a real scene. (a) The fused scanned data from all the viewpoints generated by our method. (b) The final scene CAD recomposition, where objects with the same category are rendered with the same color.

and layout estimation for more accurate object retrieval. Besides, a novel relation-aware NBV generation module is proposed to make the exploration during the autonomous scanning tailored for our composition task by considering the retrieval and relation accuracy. Extensive experiments and comparisons are adopted to validate the feasibility and effectiveness of our algorithm.

Limitations and future work. Our current method for auto-scanning and retrieving still has several limitations. As a complex system with the incorporation of several components and dependence on some off-the-shelf methods, the final results may be affected in different aspects. First, there are times when the instance segmentation does not recognize the object or recognition error leading to the wrong retrieval results. Second, the localization error of the camera sometimes leads to failure of pose estimation, which will further influence the final retrieval results. Third, to enforce the support relationship between objects, the poses of objects sometimes become a little bit far away from their GT poses, which makes the object points and relation points inaccurate and further decreases the scanning efficiency.

There are several directions to improve our method in the future. First, the generated interest points can be fused with learning-based methods. Not all interest points are equally important since they are denoted as different perspectives of the scene, and different decisions are required for different scanning states. Therefore, it is possible to train an agent to make better decisions via reinforcement learning or other methods. Second, it will be an interesting future direction to generate the scene recomposition with RGB observations only. Recently some works have focused on 3D object pose estimation from a single RGB image. These works make it possible to retrieve CAD models and optimize the relationships directly from the RGB image to skip saving the explicit data representations for objects. Third, new types of sub-tasks can be introduced to our system. For example, various interactive behaviors can be created for different kinds of CAD models. Thus, plenty of difficult tasks can be defined for the scene and learned in the future. Finally, there may exist some strategies with higher efficiency by defining and learning a new relation hierarchy or other higher-level semantic structural information like functionality, which will bring the combination of virtual and reality to a higher level.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable comments. This work is supported by the National Key R&D Program of China (2022YFB3303400), National Natural Science Foundation of China (62025207, 62322207), Shenzhen Science and Technology Program (RCYX20210609103121030), and GD Natural Science Foundation (2021B1515020085).

REFERENCES

Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X. Chang, and Matthias Nießner. 2018a. Scan2CAD: Learning CAD Model Alignment in RGB-D Scans. *CoRR* abs/1811.11187 (2018). arXiv:1811.11187 http://arxiv.org/abs/1811.11187

Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X. Chang, and Matthias Nießner. 2018b. Scan2CAD: Learning CAD Model Alignment in RGB-D Scans. *CoRR* abs/1811.11187 (2018). arXiv:1811.11187 http://arxiv.org/abs/1811.11187

Armen Avetisyan, Angela Dai, and Matthias Nießner. 2019. End-to-End CAD Model Retrieval and 9DoF Alignment in 3D Scans. *CoRR* abs/1906.04201 (2019). arXiv:1906.04201 http://arxiv.org/abs/1906.04201

Armen Avetisyan, Tatiana Khanova, Christopher B. Choy, Denver Dash, Angela Dai, and Matthias Nießner. 2020. SceneCAD: Predicting Object Alignments and Layouts in RGB-D Scans. *CoRR* abs/2003.12622 (2020). arXiv:2003.12622 https://arxiv.org/abs/2003.12622

P.J. Besl and Neil D. McKay. 1992. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14, 2 (1992), 239–256. https://doi.org/10.1109/34.121791

John Canny. 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8, 6 (1986), 679–698. https://doi.org/10.1109/TPAMI.1986.4767851

Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. 2015. ShapeNet: An Information-Rich 3D Model Repository. *CoRR* abs/1512.03012 (2015). arXiv:1512.03012 http://arxiv.org/abs/1512.03012

Benjamin Charrow, Gregory Kahn, Sachin Patil, Sikang Liu, Ken Goldberg, Pieter Abbeel, Nathan Michael, and Vijay Kumar. 2015. Information-Theoretic Planning with Trajectory Optimization for Dense 3D Mapping.. In *Robotics: Science and Systems*, Vol. 11. Rome, 3–12.

T. Cover and P. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13, 1 (1967), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. *CoRR* abs/1702.04405 (2017). arXiv:1702.04405 http://arxiv.org/abs/1702.04405

Zhuo Deng and Longin Jan Latecki. 2017. Amodal Detection of 3D Objects: Inferring 3D Bounding Boxes from 2D Ones in RGB-Depth Images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 398–406. https://doi.org/10.1109/CVPR.2017.50

David H. Douglas and Thomas K. Peucker. 1973. ALGORITHMS FOR THE REDUCTION OF THE NUMBER OF POINTS REQUIRED TO REPRESENT A DIGITIZED LINE OR ITS CARICATURE. *Cartographica: The International Journal for Geographic Information and Geovisualization* 10 (1973), 112–122.

Margarita Grinvald, Fadri Furrer, Tonci Novkovic, Jen Jen Chung, Cesar Cadena, Roland Siegwart, and Juan I. Nieto. 2019. Volumetric Instance-Aware Semantic Mapping and 3D Object Discovery. *CoRR* abs/1903.00268 (2019). arXiv:1903.00268 http://arxiv.org/abs/1903.00268

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

Junfu Guo, Changhao Li, Xi Xia, Ruizhen Hu, and Ligang Liu. 2022. Asynchronous Collaborative Autoscanning with Mode Switching for Multi-Robot Scene Reconstruction. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–13.

Muzhi Han, Zeyu Zhang, Ziyuan Jiao, Xu Xie, Yixin Zhu, Song-Chun Zhu, and Hangxin Liu. 2021. Reconstructing Interactive 3D Scenes by Panoptic Mapping and CAD Model Alignments. *CoRR* abs/2103.16095 (2021). arXiv:2103.16095 https://arxiv.org/abs/2103.16095

Ankur Handa, Thomas Whelan, John McDonald, and Andrew J. Davison. 2014. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*. 1524–1531. https://doi.org/10.1109/ICRA.2014.6907054

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. *CoRR* abs/1703.06870 (2017). arXiv:1703.06870 http://arxiv.org/abs/1703.06870

Lionel Heng, Alkis Gotovos, Andreas Krause, and Marc Pollefeys. 2015. Efficient visual exploration and coverage with a micro aerial vehicle in unknown environments. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1071–1078.

Pengdi Huang, Liqiang Lin, Kai Xu, and Hui Huang. 2020. Autonomous Outdoor Scanning via Online Topological and Geometric Path Optimization. *IEEE Transactions on Intelligent Transportation Systems* (2020).

Vladislav Ishimtsev, Alexey Bokhovkin, Alexey Artemov, Savva Ignatyev, Matthias Nießner, Denis Zorin, and Evgeny Burnaev. 2020. CAD-Deform: Deformable Fitting of CAD Models to 3D Scans. *CoRR* abs/2007.11965 (2020). arXiv:2007.11965 https://arxiv.org/abs/2007.11965

Hamid Izadinia and Steven M. Seitz. 2018. Scene Recomposition by Learning-based ICP. *CoRR* abs/1812.05583 (2018). arXiv:1812.05583 http://arxiv.org/abs/1812.05583

Young Min Kim, Niloy J Mitra, Qixing Huang, and Leonidas Guibas. 2013. Guided real-time scanning of indoor objects. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 177–186.

Michael Krainin, Brian Curless, and Dieter Fox. 2011. Autonomous generation of complete 3D object models using next best view manipulation planning. In *2011 IEEE International Conference on Robotics and Automation*. IEEE, 5031–5037.

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

Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. 2021. Patch2CAD: Patchwise Embedding Learning for In-the-Wild Shape Retrieval from a Single Image. *CoRR* abs/2108.09368 (2021). arXiv:2108.09368 https://arxiv.org/abs/2108.09368

Yangyan Li, Angela Dai, Leonidas Guibas, and Matthias Nießner. 2015. Database-assisted object retrieval for real-time 3d reconstruction. In *Computer graphics forum*, Vol. 34. Wiley Online Library, 435–446.

Ligang Liu, Xi Xia, Han Sun, Qi Shen, Juzhan Xu, Bin Chen, Hui Huang, and Kai Xu. 2018. Object-aware guidance for autonomous scene reconstruction. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–12.

Yilin Liu, Ruiqi Cui, Ke Xie, Minglun Gong, and Hui Huang. 2021. Aerial Path Planning for Online Real-Time Exploration and Offline High-Quality Reconstruction of Large-Scale Urban Scenes. *ACM Transactions on Graphics (Proceedings of SIGGRAPH ASIA)* 40, 6 (2021), 226:1–226:16.

Kevis-Kokitsi Maninis, Stefan Popov, Matthias Nießner, and Vittorio Ferrari. 2020. Vid2CAD: CAD Model Alignment using Multi-View Constraints from Videos. *CoRR* abs/2012.04641 (2020). arXiv:2012.04641 https://arxiv.org/abs/2012.04641

Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. 2016. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. *CoRR* abs/1612.00593 (2016). arXiv:1612.00593 http://arxiv.org/abs/1612.00593

Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. 2019. Amodal Instance Segmentation With KINS Dataset. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3009–3018. https://doi.org/10.1109/CVPR.2019.00313

Manikandasriram Srinivasan Ramanagopal and Jerome Le Ny. 2016. Motion planning strategies for autonomously mapping 3d structures. *arXiv preprint arXiv:1602.06667* (2016).

Mike Roberts, Debadeepta Dey, Anh Truong, Sudipta Sinha, Shital Shah, Ashish Kapoor, Pat Hanrahan, and Neel Joshi. 2017. Submodular trajectory optimization for aerial 3d scanning. In *Proceedings of the IEEE International Conference on Computer Vision*. 5324–5333.

Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. 2013. Slam++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1352–1359.

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. 2019. Habitat: A Platform for Embodied AI Research. *CoRR* abs/1904.01201 (2019). arXiv:1904.01201 http://arxiv.org/abs/1904.01201

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.

Tianjia Shao, Weiwei Xu, Kun Zhou, Jingdong Wang, Dongping Li, and Baining Guo. 2012. An Interactive Approach to Semantic Modeling of Indoor Scenes with an RGBD Camera. *ACM Trans. Graph.* 31, 6, Article 136 (Nov. 2012), 11 pages. https://doi.org/10.1145/2366145.2366155

Mikaela Angelina Uy, Vladimir G. Kim, Minhyuk Sung, Noam Aigerman, Siddhartha Chaudhuri, and Leonidas J. Guibas. 2021. Joint Learning of 3D Shape Retrieval and Deformation. *CoRR* abs/2101.07889 (2021). arXiv:2101.07889 https://arxiv.org/abs/2101.07889

J Irving Vasquez-Gomez, L Enrique Sucar, Rafael Murrieta-Cid, and Efrain Lopez-Damian. 2014. Volumetric next-best-view planning for 3D object reconstruction with positioning error. *International Journal of Advanced Robotic Systems* 11, 10 (2014), 159.

Chu Wang, Babak Samari, Vladimir G Kim, Siddhartha Chaudhuri, and Kaleem Siddiqi. 2020. Affinity graph supervision for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8247–8255.

Melonee Wise, Michael Ferguson, Derek King, Eric Diehr, and David Dymesich. 2016. Fetch and freight: Standard platforms for service robot applications. In *Workshop on autonomous mobile service robots*. 1–6.

Shihao Wu, Wei Sun, Pinxin Long, Hui Huang, Daniel Cohen-Or, Minglun Gong, Oliver Deussen, and Baoquan Chen. 2014. Quality-driven poisson-guided autoscanning. *ACM Transactions on Graphics* 33, 6 (2014).

Kai Xu, Hui Huang, Yifei Shi, Hao Li, Pinxin Long, Jianong Caichen, Wei Sun, and Baoquan Chen. 2015. Autoscanning for coupled scene reconstruction and proactive object analysis. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–14.

Kai Xu, Yifei Shi, Lintao Zheng, Junyu Zhang, Min Liu, Hui Huang, Hao Su, Daniel Cohen-Or, and Baoquan Chen. 2016. 3D attention-driven depth acquisition for

object identification. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 1–14.

Kai Xu, Lintao Zheng, Zihao Yan, Guohang Yan, Eugene Zhang, Matthias Niessner, Oliver Deussen, Daniel Cohen-Or, and Hui Huang. 2017. Autonomous reconstruction of unknown indoor scenes guided by time-varying tensor fields. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 1–15.

Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. 2017. FoldingNet: Interpretable Unsupervised Learning on 3D Point Clouds. *CoRR* abs/1712.07262 (2017). arXiv:1712.07262 http://arxiv.org/abs/1712.07262

Hong-Xing Yu, Jiajun Wu, and Li Yi. 2022. Rotationally Equivariant 3D Object Detection. https://doi.org/10.48550/ARXIV.2204.13630

Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. 2021. PoinTr: Diverse Point Cloud Completion with Geometry-Aware Transformers. *CoRR* abs/2108.08839 (2021). arXiv:2108.08839 https://arxiv.org/abs/2108.08839

Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. 2021a. 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

Yi-Fan Zhang, Weiqiang Ren, Zhang Zhang, Zhen Jia, Liang Wang, and Tieniu Tan. 2021b. Focal and Efficient IOU Loss for Accurate Bounding Box Regression. *CoRR* abs/2101.08158 (2021). arXiv:2101.08158 https://arxiv.org/abs/2101.08158

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

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2018. On the Continuity of Rotation Representations in Neural Networks. *CoRR* abs/1812.07035 (2018). arXiv:1812.07035 http://arxiv.org/abs/1812.07035