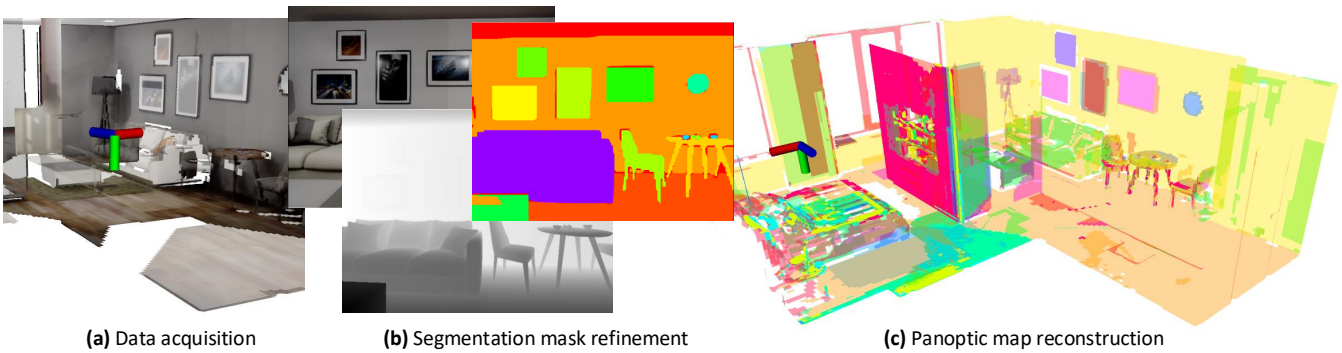
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Volumetric Mapping with Panoptic Refinement using Kernel Density Estimation for Mobile Robots

Khang Nguyen Tuan Dang Manfred Huber



(a) Data acquisition **(b)** Segmentation mask refinement **(c)** Panoptic map reconstruction

Fig. 1: **(a)** Indoor mobile robots operating in an environment with multiple objects **(b)** refines RGB-based segmentation masks using kernel density estimation via depth perception, and **(c)** rebuilds panoptic map with object instances using projective signed distance functions.

***Abstract*— Reconstructing three-dimensional (3D) scenes with semantic understanding is vital in many robotic applications. Robots need to identify which objects, along with their positions and shapes, to manipulate them precisely with given tasks. Mobile robots, especially, usually use lightweight networks to segment objects on RGB images and then localize them via depth maps; however, they often encounter out-of-distribution scenarios where masks over-cover the objects. In this paper, we address the problem of panoptic segmentation quality in 3D scene reconstruction by refining segmentation errors using non-parametric statistical methods. To enhance mask precision, we map the predicted masks into a depth frame to estimate their distribution via kernel densities. The outliers in depth perception are then rejected without the need for additional parameters in an adaptive manner to out-of-distribution scenarios, followed by 3D reconstruction using projective signed distance functions (SDFs). We validate our method on a synthetic dataset, which shows improvements in both quantitative and qualitative results for panoptic mapping. Through real-world testing, the results furthermore show our method's capability to be deployed on a real-robot system. Our source code is available at: https://github.com/mkhangg/refined_panoptic_mapping.**

I. Introduction

Understanding and reconstructing 3D scenes are crucial in robotic perception and manipulation. SDF-based volumetric mapping methods are common in building 3D maps by integrating new observations from RGB-D images. To obtain the semantics of 3D maps, robots need to localize and identify each object in a scene, which is the first and foremost step toward scene understanding, where each voxel in the SDF-based map is assigned a label that can be predicted either from RGB-D images or point clouds. Nevertheless, robotic applications often rely on low-cost computations with information-rich RGB-D images to avoid the expense of computing and storing point clouds.

Traditional methods [1]–[3] segment objects on RGB images using convolutional neural networks (CNNs) and look into depth maps to project 2D segments into 3D segments. Together with this, several methods [4]–[8] have been proposed to refine segments generated by CNNs using depth information. Depth-driven region growing method [4] takes the similarity between objects' depth and connectivity criteria to produce accurate segmentation results, particularly for objects with distinct depth boundaries. Meanwhile, depth-assisted graph cut utilizes graph cut-based segmentation [5]–[8] to produce refined segmentation. However, depending solely on RGB images in the first place induces signification errors, especially on the border of segments if background colors are similar to the objects' colors.

Another approach is to fuse RGB and depth images using deep neural networks at the feature level; for example, the works [9]–[14] design networks to explore the correlation between pixels of the same semantic and their corresponding depth under an assumption that pixels in the same segment should have similar depth and vice versa. Meanwhile, one of earlier works [15]–[17] uses depth cues to support relations between objects to generate better segments. As RGB-D images are encoded in the network, it is difficult to detect minimal errors from the network output, and this largely depends on the nature of the training datasets. In other words, the main problem is that training and segmenting objects with RGB-D fusion does not guarantee the algorithm's performance will be adaptive under diverse real-world factors.

To fill this gap, we propose a novel method to refine the over-covered segmentation masks generated by CNNs via Kernel Density Estimation (KDE) to eliminate the uncertainty in segmentation masks in a statistical manner,

All authors are with the Learning and Adaptive Robotics Laboratory, Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76013, USA. (emails: khang.nguyen8@mavs.uta.edu, tuan.dang@uta.edu, huber@cse.uta.edu)

Features	**KinectFusion** [18]	**Chisel** [19]	**Voxblox** [20]	**Voxblox++** [21]	**Voxfield** [22]	**PanMap** [23]	**Ours**
RGB-D-based panoptic perception	✗	✗	✗	✓	✗	✓	✓
parametric-free segmentation refinement	✗	✗	✗	✗	✗	✗	✓
SDF-based volumetric mapping	✓	✓	✓	✓	✓	✓	✓
on-robot real-time performance	✓	✓	✓	✓	✓	✓	✓

TABLE I: Comparison of features across RGB-D volumetric mapping systems for indoor mobile robots.

which differs from previous research where uncertainty is accumulated. The main advantage is that lightweight CNNs and KDE are suitable for resource-constrained embedded computers on robotic systems rather than point cloud-based segmentation networks. Additionally, with this proposed non-parametric method, we achieve a better result without fine-tuning models and hyperparameters for the pipeline to well-perform under various settings.

II. Related Work

Panoptic Segmentation with RGB-D Perception: Segmentation has recently gained popular attention in the robotics community thanks to its robustness in recognizing objects accurately at the pixel level. Earliest works, such as SemanticFusion [24], Co-Fusion [25], and MaskFusion [26], fuse the semantics of objects onto simultaneous localization and mapping (SLAM) frameworks to build 3D maps; however, this method does not differentiate between objects of the same kind in the environment. To solve this, panoptic segmentation [27] then first brings the concept of semantic segmentation and instance segmentation together, where each image pixel is assigned an object label, benefiting robots to understand indoor entities distinctively. Leveraging panoptic segmentation, an incremental work [28] uses depth cues to segment objects but results in over-cover objects in complex scenes when objects are articulated with each other. Voxblox++ [21] and its subsequent PanMap [23], therefore, bridge the gap between these works by refining the mask and also using depth perception. Nevertheless, this geometric-based segmentation explicitly assumes that cutoff thresholds are provided when assigning 3D segments for each object instance. To ameliorate this, we proposed a depth-based segmentation refinement from the view of a non-parametric statistical approach, which enhances the adaptability of robots in various settings.

Outlier Rejection & Mask Refinement: Outlier removal is a crucial step in segmentation to ensure the robustness and precision of masks, particularly in dealing with learning models that generate noises in segments (*i.e.,* at object boundaries). Classical methods that solely rely on the observation data are sensitive to noises without a good guess of hyperparameters, such as Random Sample Consensus (RANSAC) [29], [30] or distance-based outlier detection, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [31]. Recent approaches leverage the CNNs [7], [8] to encode RGB and depth images into separate encoders and use extracted features to correct the RGB-based segmentation error. Different approaches also fuse latent spaces of both RGB and depth as a single feature to generate correct segmentation masks. However, the primary limitation of these approaches is their dependency on training datasets, and their performance eventually subsides on out-of-distribution images and real-world scenarios. To avoid uncertainty in CNNs, we employ pre-trained 2D models to segment RGB images and refine these segments by applying KDE to each object's mask distribution on depth images.

SDF-Based Volumetric Mapping: Modeling target objects' shapes and textures based on their colored images can be dated back to SDF from Curless and Levoy [32]. This has established the foundations for the recent emergence of robotics and graphics, especially volumetric mapping. Notable works in robotics using SDF can be categorized into object tracking [33]–[35] and volumetric modeling [18], [20]–[23], [36], which are both essential components of mapping in the context of moving cameras. Signed Distance Functions [36] and Voxblox [20] first provide frameworks using SDF to model a 3D obstacle map for autonomous navigation based on prior works, such as KinectFusion [18] and Chisel [19].

Subsequently, Voblox++ [21] introduces functional scene understanding; meanwhile, Voxfield [22] is also developed with a more focus on accurate mapping but consumes lower computation costs. Most recently, PanMap [23] presented a hierarchical semantic submap management. However, PanMap emphasizes the semantics of sub-maps more than object instances for ease of mapping management. Recognition inconsistency, therefore, affects points at the instance level, especially when an object occupies multiple voxels or parts of multiple recognized objects share one common voxel. Moreover, when segmentation masks are not well-refined via depth perception, these perception uncertainties can be increased in out-of-distribution scenes. To address this issue, in this work, we embed per-point semantics into the SDF-based volumetric mapping, where the voxels containing the points within the object of interest are updated incrementally, providing semantic consistency for the scene.

III. Method Overview

We first take an RGB-D image stream as shown in Fig. 1. With the RGB image, the segmentation model segments objects and non-object entities, such as walls, floors, and ceilings. The segmentation blobs are then projected using KDE on depth maps (Sec. IV-B) for depth point-based densities. Thus, the masks are refined based on density lines through depth perception in the previous step without additional hyperparameters (Sec. IV-C). This mechanism effectively matches real-world intuition, in which the robots often encounter out-of-distribution scenes due to the effects of brightness and irregularity in objects' appearances and shapes. Iteratively, the scene is reconstructed via projective SDF, where the 3D points of objects of interest are updated over time (Sec. IV-D) until the robot stops its observation.

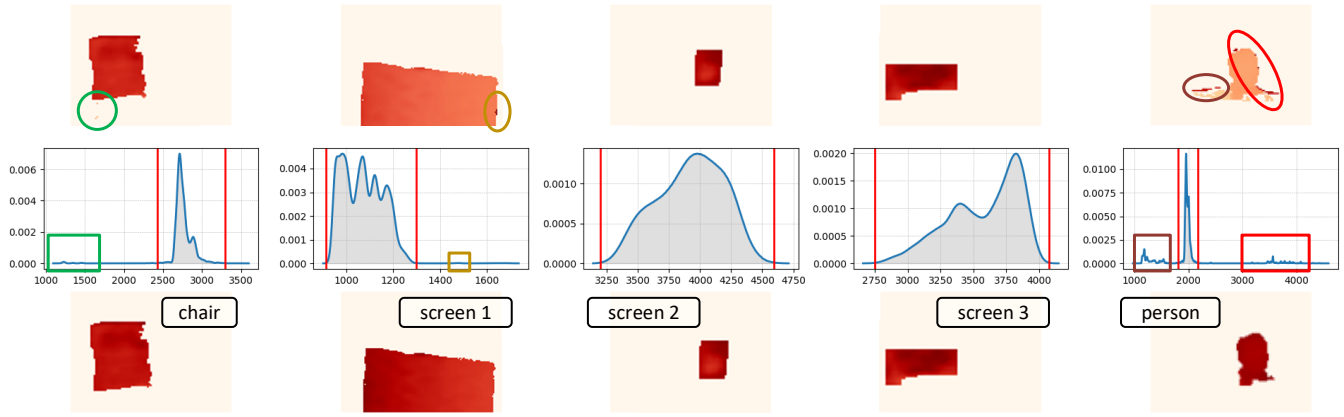


Fig. 2: Depth maps of object instances containing depth outliers (*top row*) due to the imperfection of segmentation models and their density estimations along depth perception (*middle row*), and refined depth maps (*bottom row*). The shaded depth values on the density lines in between vertical red cutoff lines are considered inliers; otherwise, Alg. 1 rejects them as they appear to be outliers. The outliers are encoded by the same colors as Fig. 3 along with corresponding objects presented in the scene.

Algorithm 1: Mask Refinement via Depth Perception

Input : $\mathbf{M} :=$ binary masks of objects of interest
$\mathcal{D} :=$ depth map
Output: $\mathcal{M} :=$ refined masks for objects of interest

function `RefineSegMaskViaDepth` ($\mathbf{M}$, $\mathcal{D}$)
  $\mathcal{M} = [\,]$
  for $\mathbf{M}_i \in \mathbf{M}$ **do**
    $\mathcal{D}[\mathbf{M}_i = 0] = 0$
    `x_kde, y_kde = FFTKDE(`$\mathcal{D}$`.flatten())`
    $peak =$ `max(y_kde)`
    `left_id = find_id(y_kde[:`$peak$`] < 1e-6)`
    `right_id = find_id(y_kde[`$peak$`:] < 1e-6)`
    `low_cutoff = x_kde[left_id]`
    `high_cutoff = x_kde[right_id]`
    $\mathcal{D}\,[d <$ `low_cutoff` **or** $d >$ `high_cutoff`$] = 0$
    $\mathbf{M}_i[\mathcal{D} = 0] = 0$
    $\mathcal{M}$`.append(`$\mathbf{M}_i$`)`
  return $\mathcal{M}$

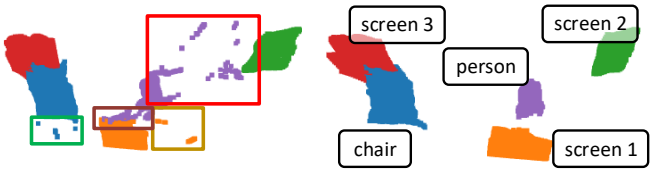
Our approach additionally offers a parametric-free depth outlier rejection for segmentation mask refinement, as depicted in Table I. This improves the quality of SDF-based panoptic mapping and makes robots, particularly mobile indoor robots, adaptive and versatile in out-of-distribution scenarios, which has not been well-addressed in previous works of the same category.

IV. Methodology

Mobile robots usually produce uncertainties in their perception, which are best seen via depth maps, where depth pixels cannot be interpolated from the previous frame, resulting in holes on depth maps. Plus, in 3D semantic perception, RGB-based segmentation models add uncertainty when recognizing objects and mapping their occupancies in the real world. To alleviate this, we consider addressing two uncertainties: (1) depth map hole-filing and (2) refining segmentation masks in 3D space.

A. Holes Filling on Depth Images

To fill possible holes that occurred on depth maps, $\mathcal{D}$, we interpolate the missing values from its $k \times k$ neighboring pixels. Otherwise, the missing value remains as 0 if all of their g^2 neighbors are empty, as follows:



(a) Without outlier removal **(b)** With outlier removal

Fig. 3: The scene of multiple objects with outliers boxed in red (*left*) and the scene without outliers after applying Alg. 1 (*right*).

$$\mathcal{D}(i,j) = \begin{cases} 0, \text{ if } \mathcal{D}(i \pm k, j \pm l) = 0 \\ \sum_{i=u-k}^{u+k} \sum_{j=v-k}^{v+k} \mathcal{D}(i,j) \cdot G(u,v) \end{cases} \tag{1}$$

where $0 \leq k, l \leq g$, (u, v) represents the image coordinates of pixels, $G(u, v)$ is the 2D Gaussian kernel, size of $g \times g$, centered on the (u, v) pixel.

B. Depth Outlier Rejection

Segmentation imperfection also occurs in objects' boundary pixels, leading to inaccurate depth perception when aligning RGB and depth frames. In practice, this is typically ignored by manually truncating depth pixels that exceed a defined threshold. Another approach to overcome this is taking the top-down view, where depth errors are projected on the table surface and spatially compensated within the manipulation process. To solve this, we apply the density function $\hat{f}$ at any given point d to non-parametrically reject depth outliers as follows:

$$\hat{f}(d) = \frac{1}{mH} \sum_{i=1}^{m} \mathcal{G}\left(\frac{d - d_i}{H}\right) \tag{2}$$

where d_i is the depth value from $\mathcal{D}(\cdot, \cdot)$, H represents the optimal bandwidth obtained from the ISJ algorithm for the 1D Gaussian kernel, $\mathcal{G}$, and m indicates the number of depth values on the segmented object.

Re-organizing Eq. 2 in terms of equidistant 1D grid points $\{\mathbf{g}_j\} \in [\mathbf{g}_1, \mathbf{g}_M]$ covering all $\mathbf{p}_i^{\mathcal{K}_t}$, and grid counts $\{c_j\}$ to

represent the number of $\mathbf{p}_i^{\mathcal{K}_t}$'s that are near $\mathbf{g}_j$, for $j = 1, 2, ..., M$ with $M \neq m$, we obtain:

$$\widetilde{f_{d,\mathbf{g}_j}} \coloneqq \tilde{f}(g_j) = \frac{1}{mH} \sum_{i=1}^{M} c_i \cdot \mathcal{G}\left(\frac{\mathbf{g}_j - \mathbf{g}_i}{H}\right) \approx \hat{f}(d) \quad (3)$$

With the FFT's time complexity of $\mathcal{O}(M \log M)$, Eq. 3 is then translated in the form of convolution as:

$$\widetilde{f_{d,\mathbf{g}_j}} = \sum_{i=-(M-1)}^{M-1} c_{j-i}\mathbf{k}_i \text{ with } \mathbf{k}_i = \frac{1}{m} \cdot \mathcal{G}\left(\frac{\mathbf{g}_M - \mathbf{g}_1}{H} \cdot i\right) \quad (4)$$

C. Segmentation Mask Refinement

Alg. 1 illustrates the segmentation mask refinement process via depth perception, given the set of RGB-D binary segmentation masks and the depth map. For each object, its depth values are selected according to its binary mask. Hence, applying Eq. 4, we obtain the density function along with depth values. The maximum peak in the density line is then identified; meanwhile, the cutoff values on its two tails are determined when the density goes to 0 – indicating the disconnection between the object and its fragments in the depth axis, as shown in Fig. 2. The depth map is rectified on regions where the depth pixels sub-ceed the lower cutoff and exceed the upper cutoff, followed by the mask refinement via non-zero regions of the depth map. Iterating through objects of interest, Alg. 1 returns their refined binary masks with outlier rejection via depth knowledge without requiring predefined thresholds.

As an example provided in Fig. 3a, the outliers induced by segmentation masks are visible in the form of point clouds. By applying the refined binary masks returned by Alg. 1, the outliers are also removed in the corresponding point clouds of instances, as shown in Fig. 3b.

D. Integration of Projective SDFs & Semantic Perception

Comprehending the semantics of indoor environments with a mobile robot entails dynamically updating the occupancy in the real world. Embedded with semantic knowledge along with the building process, the robot is also ready for other manipulation tasks besides navigating and exploring such environments. Therefore, to do this in the discretized voxel-like world, our objective is to incrementally update in 3D space via SDF [32] with semantic knowledge in a recursive manner, as follows:

$$D_{t+1}(v, \hat{\mathbf{p}}) = \frac{W_t(v, \hat{\mathbf{p}}) D_t(v, \hat{\mathbf{p}}) + w_{t+1}(v, \hat{\mathbf{p}}) d_{t+1}(v, \hat{\mathbf{p}})}{W_t(v, \hat{\mathbf{p}}) + w_{t+1}(v, \hat{\mathbf{p}})}$$

$$W_t(v, \hat{\mathbf{p}}) = \sum_i w_{t-1}^{(i)}(v, \hat{\mathbf{p}}) + w_t(v, \hat{\mathbf{p}}) \quad (5)$$

$$d_{t+1}(v, \hat{\mathbf{p}}) = \begin{cases} ||\hat{\mathbf{p}} - v||_2, \text{ if } \hat{\mathbf{p}} \in v \\ -||\hat{\mathbf{p}} - v||_2, \text{ if } \hat{\mathbf{p}} \notin v \end{cases}$$

where v is the voxel in 3D, $\hat{\mathbf{p}}$ is the 3D point within in the object of interest, $d_i(\cdot, \cdot)$ represents SDFs with their corresponding weights, $w_i(\cdot, \cdot)$, from refined masked RGB images, and $D_i(\cdot, \cdot)$ indicates the cumulative SDF with $W_i(\cdot, \cdot)$ as the corresponding cumulative weight function.

Using Eq. 5, the voxels containing the points within the object of interest are updated recursively, eventually providing the robot with the object's occupancy voxels.

V. Evaluation on `flat` Dataset

A. Dataset & Evaluation Metrics

We evaluate Alg. 1 on the `flat` dataset to verify our approach's performance in both quantitative and qualitative results against the most recent state-of-the-art method [23]. The `flat` dataset consists of synthetic RGB-D images rendered in Unreal Engine 4 (UE4) and includes ground truth mesh, per-image panoptic annotations, and the camera poses at each RGB-D frame. In terms of quantitative results, we compute the mask intersection over union (IOU) between the ground truth annotations and the predicted masks provided by the segmentation model. Meanwhile, in terms of qualitative results, we compare the results after performing volumetric mapping across the approaches. In addition, we also reconstruct the scene with texture and panoptic masks to better compare the 3D reconstruction qualities.

Approaches	Mask IOU	Changes
(a) PanMap without refinement	16.5150	–
(b) PanMap with refinement	26.2283	**+9.7133** ↑
(c) Our approach without refinement	79.8860	**+53.6577** ↑
(d) Our approach with refinement	90.6077	**+10.7217** ↑

TABLE II: Quantitative results in mask IOU percentage and changes on the `flat` dataset between **(a)** the original approach without mask refinement, **(b)** with mask refinement, **(c)** our approach without mask refinement, and **(d)** with mask refinement.

B. Quantitative Results

The quantitative comparisons between our approach and the panoptic mapping technique [23] are shown in Table II. For the original approach, the mask IOU is at 16.5150; however, after refining their predicted masks, the mask IOU is elevated by approximately 10 percent to 26.2283.

Since the semantic consistency at the point level is not prioritized in the original approach, we retrain the segmentation model [37] with similar objects and produce segmentation masks for the same sequence; however, we do not apply the mask refinement process initially. We thus achieve the mask IOU of 79.8860, which significantly improves from Detectron used in the original approach. Therefore, to see how mask refinement's effectiveness on new sets of segmentation images, we proceed with performing mask refinement for each mask in the set; the result then reaches 90.6077 in terms of mask IOU, again improving by roughly 10 percent from when not applying mask refinement.

Overall, we observe a slight improvement when applying mask refinement to the raw set of segmentation masks. This implies the crucial role of refinement based on depth maps, which strongly correlate to the objects' spatial occupancy and shapes in the 3D world rather than on 2D images.

C. Qualitative Results

We continue to compare the quality of SDF-based volumetric mapping on the same robot moving sequence between

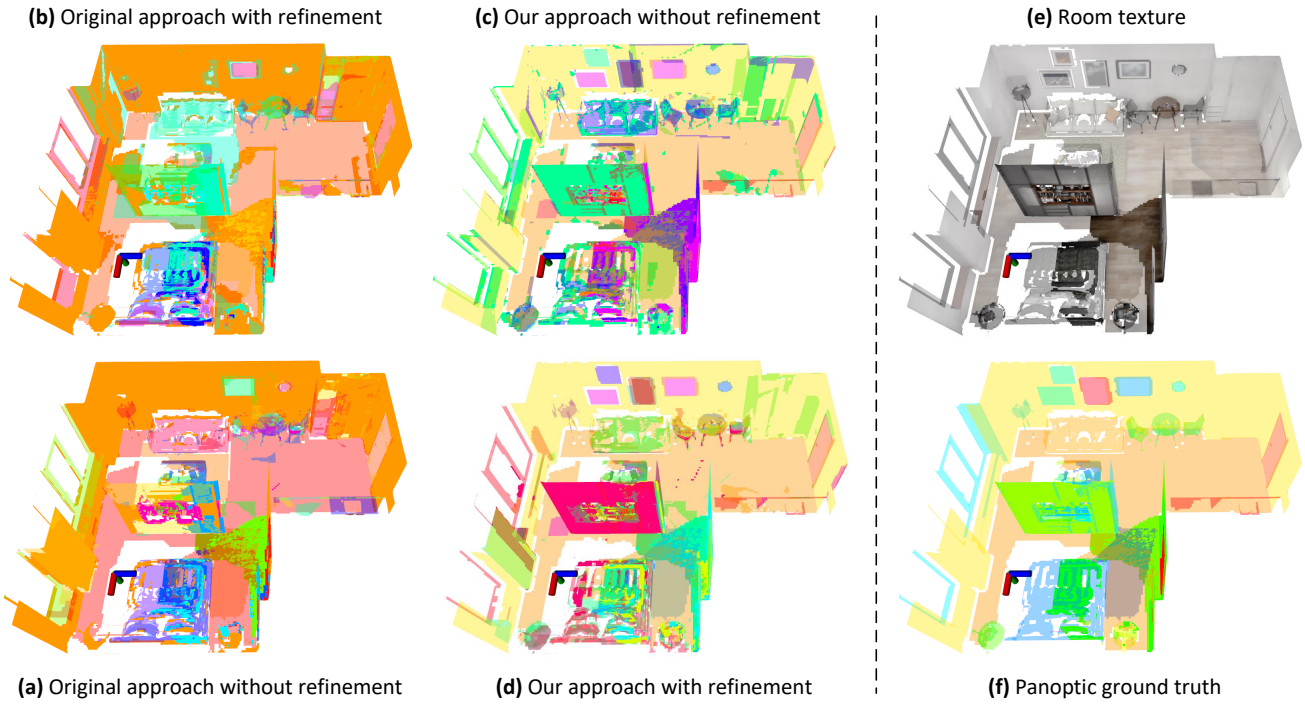


Fig. 4: Qualitative results on the `flat` dataset of **(a)** the original panoptic mapping approach, **(b)** the original approach coupled with mask refinement, **(c)** our approach without mask refinement, and **(d)** our approach with mask refinement. The room texture and its panoptic segmentation ground truth are retrieved based on RGB images and annotation masks provided by the original framework [23]. Note that the robot frame indicating its pose is simplified and represented as the RGB mesh frame in each reconstructed map.

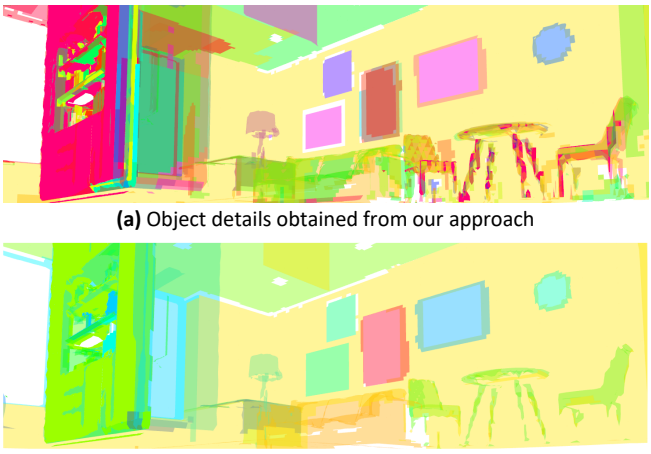


Fig. 5: Comparisons of object detail reconstruction quality between **(a)** our approach with mask refinement and **(b)** from ground truth.

the original and our approaches, as shown in Fig. 4. Also, for each approach, we compare the results with and without the mask refinement step to see how crucial it is qualitatively.

1) Original Approach without Mask Refinement: As shown in Fig. 4a, the raw panoptic mapping approach mainly focuses on guaranteeing the scene's hierarchical map management, where the semantic consistency of submaps is prioritized over the completeness and accuracies of the segmentation. However, this approach remains pitfalls when the recognition inconsistency occurs at the instance level, resulting in point-level inconsistency, especially in circumstances where an object straddles multiple submaps or parts of multiple objects occupy one submap simultaneously. These cases are brittle, especially when segmentation masks are not well-refined via depth perception.

2) Original Approach with Mask Refinement: Therefore, to mitigate this error, we perform the refinement process on each segmentation mask. As shown in Fig. 4b, the other picture on the wall is partially detected and reconstructed, which has been entirely omitted in Fig. 4a. Nevertheless, the qualitative result is far off compared to the ground truth in Fig. 4f. The reason for this is due to the imperfection of naive segmentation inputs, inducing false detections at instance levels.

3) Our Approach without Mask Refinement: To resolve this problem, we use the retrained segmentation model as mentioned in Sec. V-B and perform re-segmentation on RGB images, thus reconstructing the scene with new segmentation masks. The result in Fig. 4c shows that the approach is able to recognize more objects compared to those in Fig. 4b. However, the outlier artifacts are presented due to the lack of segmentation mask refinement.

4) Our Approach with Mask Refinement: Combining the mask refinement process with the approach in Fig. 4, we effectively remove small-sized outliers and achieve a "cleaner" map at the end. Compared to the original result in Fig. 4a, we are able to segment and reconstruct objects of interest in the room, such as tables, chairs, sofas, clocks, wall pictures, etc, with respect to the room texture in Fig. 4e and the panoptic segmentation ground truth in Fig. 4f. In this approach, these object details (Fig. 5a) are also well-constructed compared to the ground truth (Fig. 5b).

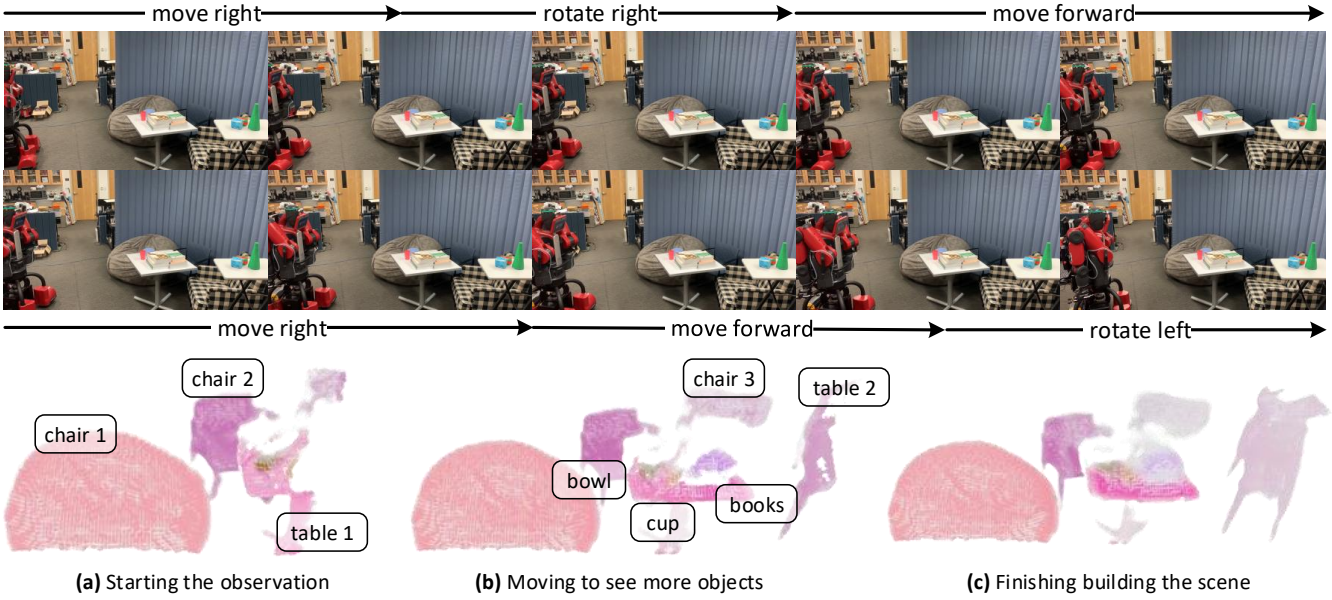


Fig. 6: The refined volumetric mapping process on the Baxter robot: **(a)** starts to observe a part of the scene, **(b)** iteratively updating the scene by moving in the lab's free space, including translations and rotations, and **(c)** finishing building the observed scene.

VI. Real-Robot Experiments

To test the adaptability of the mask refinement procedure in real-world settings, we employ our proposed method on a real-robot system with a setup scene of everyday objects.

A. Software & Hardware Setup

We build our software on distributed computers where all components are synchronized using a Robotic Operating System (ROS). One node in ROS is responsible for acquiring the RGB-D image steam from the depth camera and pre-processing raw data before broadcasting them into the ROS network. The pre-processing step includes segmenting RGB images and refining their segmentation masks using our mentioned method above. Meanwhile, another node performs SDF-based volumetric mapping when receiving refined segments from the data acquisition ROS node. Note that each node runs different computers, and each RGB-D image time-stamp synchronizes raw data.

B. Real-World Performance of Proposed Method

We set up the scene of indoor objects containing chairs and tables of various types with stacks of books and a plastic cup on the table, as shown in the top right of Fig. 6. The robot is operated to walk around the scene arbitrarily with the support of omnidirectional wheels. In Fig. 6a, the robot bootstraps the pipeline and starts observing the setup scene, and the scene is incrementally built as the robot moves. Initially, the bean bag, chair, part of the table, and cup appeared (Fig. 6a). Then, the other chair, the rest of the table, and the stack of books are built (Fig. 6b). Lastly, the robot tries to complete the scene by observing more points of the tables, which are also incrementally updated into the existing reconstructed one (Fig. 6c). The walls and floors are colored in grey for ease of visualizing per-object refinements.

C. Spatial Mask Refinement & Semantic Consistency

Also, as shown in the bottom row of Fig. 6, each reconstructed objects are spatially well-refined using mask refinement. These refinements are visible at corners of objects, such as chair legs and table legs, and in between objects' boundaries. Likewise, per-point semantic consistency is well-preserved via projective SDF-based volumetric mapping. Together with the refinements, these qualitative results underline the adaptability of our proposed method in real-world indoor settings and on a real-robot system.

D. Demonstration

The demonstration video shows the performance of our proposed method on the Baxter mobile robot is available at `https://youtu.be/u214kCms27M`.

VII. Conclusions

In this paper, we present a parametric-free mask refinement process for projective SDF-based volumetric mapping, which improves the quality of 3D scene reconstruction with panoptic understanding. The segmentation-induced outliers appearing when masking onto objects' point clouds are statistically removed by applying KDE along the depth perception, whereas the discontinuity on density lines is rectified automatically without additional hyperparameters. After the segmentation mask refinement process, the point of each object instance is updated via projective SDF, which, together with the mask precision, guarantees point-level semantic consistency when volumetrically building the panoptic map. To verify our proposed method's precision, we conduct evaluations on the synthetic `flat` dataset and achieve better results in both quantitative and qualitative manners. The results of experiments on the Baxter robot with an Intel RealSense D435i RGB-D camera also demonstrate that our method is adaptive and feasible in real-world indoor environments.

REFERENCES

[1] X. Ren, L. Bo, and D. Fox, "Rgb-(d) scene labeling: Features and algorithms," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2759–2766.

[2] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual organization and recognition of indoor scenes from rgb-d images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 564–571.

[3] K. Nguyen, T. Dang, and M. Huber, "Real-time 3d semantic scene perception for egocentric robots with binocular vision," *arXiv preprint arXiv:2402.11872*, 2024.

[4] J. Le Moigne and J. C. Tilton, "Refining image segmentation by integration of edge and region data," *IEEE transactions on geoscience and remote sensing*, vol. 33, no. 3, pp. 605–615, 1995.

[5] D. Freedman and T. Zhang, "Interactive graph cut based segmentation with shape priors," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 755–762.

[6] S. Vicente, V. Kolmogorov, and C. Rother, "Graph cut based image segmentation with connectivity priors," in *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.

[7] Y. Wang, J. Zhang, C. An, M. Cavichini, M. Jhingan, M. J. Amador-Patarroyo, C. P. Long, D.-U. G. Bartsch, W. R. Freeman, and T. Q. Nguyen, "A segmentation based robust deep learning framework for multimodal retinal image registration," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1369–1373.

[8] M. Grcić, J. Šarić, and S. Šegvić, "On advantages of mask-level recognition for outlier-aware segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2936–2946.

[9] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, and L. Lin, "Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 541–557.

[10] J. Wang, Z. Wang, D. Tao, S. See, and G. Wang, "Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 2016, pp. 664–679.

[11] S.-J. Park, K.-S. Hong, and S. Lee, "Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4980–4989.

[12] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13*. Springer, 2017, pp. 213–228.

[13] W. Wang and U. Neumann, "Depth-aware cnn for rgb-d segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 135–150.

[14] Y. Yue, W. Zhou, J. Lei, and L. Yu, "Two-stage cascaded decoder for semantic segmentation of rgb-d images," *IEEE Signal Processing Letters*, vol. 28, pp. 1115–1119, 2021.

[15] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*. Springer, 2012, pp. 746–760.

[16] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.

[17] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*. Springer, 2014, pp. 345–360.

[18] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *2011 10th IEEE international symposium on mixed and augmented reality*. Ieee, 2011, pp. 127–136.

[19] M. Klingensmith, I. Dryanovski, S. S. Srinivasa, and J. Xiao, "Chisel: Real time large scale 3d reconstruction onboard a mobile device using spatially hashed signed distance fields." in *Robotics: science and systems*, vol. 4, no. 1. Citeseer, 2015.

[20] H. Oleynikova, Z. Taylor, M. Fehr, R. Siegwart, and J. Nieto, "Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1366–1373.

[21] M. Grinvald, F. Furrer, T. Novkovic, J. J. Chung, C. Cadena, R. Siegwart, and J. Nieto, "Volumetric instance-aware semantic mapping and 3d object discovery," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 3037–3044, 2019.

[22] Y. Pan, Y. Kompis, L. Bartolomei, R. Mascaro, C. Stachniss, and M. Chli, "Voxfield: Non-projective signed distance fields for online planning and 3d reconstruction," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 5331–5338.

[23] L. Schmid, J. Delmerico, J. L. Schönberger, J. Nieto, M. Pollefeys, R. Siegwart, and C. Cadena, "Panoptic multi-tsdfs: a flexible representation for online multi-resolution volumetric mapping and long-term dynamic scene consistency," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8018–8024.

[24] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "Semanticfusion: Dense 3d semantic mapping with convolutional neural networks," in *2017 IEEE International Conference on Robotics and automation (ICRA)*. IEEE, 2017, pp. 4628–4635.

[25] M. Rünz and L. Agapito, "Co-fusion: Real-time segmentation, tracking and fusion of multiple objects," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4471–4478.

[26] M. Runz, M. Buffier, and L. Agapito, "Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects," in *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2018, pp. 10–20.

[27] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9404–9413.

[28] Y. Nakajima, K. Tateno, F. Tombari, and H. Saito, "Fast and accurate semantic mapping through geometric-based incremental segmentation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 385–392.

[29] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[30] A. Y. Yang, S. R. Rao, and Y. Ma, "Robust statistical estimation and segmentation of multiple subspaces," in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*. IEEE, 2006, pp. 99–99.

[31] H. Bäcklund, A. Hedblom, and N. Neijman, "A density-based spatial clustering of application with noise," *Data Mining TNM033*, vol. 33, pp. 11–30, 2011.

[32] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 303–312.

[33] T. Schmidt, R. A. Newcombe, and D. Fox, "Dart: Dense articulated real-time tracking." in *Robotics: Science and systems*, vol. 2, no. 1. Berkeley, CA, 2014, pp. 1–9.

[34] R. A. Newcombe, D. Fox, and S. M. Seitz, "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 343–352.

[35] A. Walsman, W. Wan, T. Schmidt, and D. Fox, "Dynamic high resolution deformable articulated tracking," in *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017, pp. 38–47.

[36] H. Oleynikova, A. Millane, Z. Taylor, E. Galceran, J. Nieto, and R. Siegwart, "Signed distance fields: A natural representation for both mapping and planning," in *RSS 2016 workshop: geometry and beyond-representations, physics, and scene understanding for robotics*. University of Michigan, 2016.

[37] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics