SCALABLE MAV INDOOR RECONSTRUCTION WITH NEURAL IMPLICIT SURFACES

Haoda Li, Puyuan Yi, Yunhao Liu, Avideh Zakhor

Department of Electrical Engineering and Computer Sciences
University of California, Berkeley

ABSTRACT

In this paper, we propose a fully automated pipeline for reconstructing large and complex indoor scenes with drone-captured RGB images. First, we leverage traditional structure-from-motion methods to obtain camera poses and reconstruct an initial point cloud. Next, we devise a divide-and-conquer strategy to utilize neural surface reconstruction under the Manhattan-world assumption. Our method reduces the point cloud's outliers and significantly improves reconstruction quality on low-texture regions. We simultaneously predict point-wise semantic logits for walls, floors, and ceilings. The semantic segmentation enables category-wise plane fitting and improves reconstruction quality on polygonal geometry. To validate our method, we use a drone to capture videos inside a large-scale, complex indoor scene. Experimental results showed our method can create robust 3D models for indoor environments using RGB images only.

***Index Terms*—** indoor reconstruction, neural implicit representation, drone mapping, multi-view stereo

1. INTRODUCTION

Most current applications of unmanned aerial vehicles in architecture, engineering, and construction are for outdoor scenes where GPS signals can be integrated with captured RGB imagery for 3D reconstruction. In recent years, the availability of smaller micro air vehicles (MAVs) creates an opportunity for indoor 3D reconstruction applications such as recovering floor plans of buildings in a fast, automated way. However indoor environments pose new challenges in terms of the safety of data capture, positioning in the absence of GPS, and sensor payloads a small drone can carry. In this work, we propose an end-to-end system for indoor 3D reconstruction using commercially available MAVs with single monocular cameras only. We leverage both traditional Structure from Motion (SfM) [1] and neural surface reconstruction [2] methods. We recover camera poses and reconstruct an initial point cloud using traditional SfM methods. The point cloud provides per-view depth cues for neural surface methods. Thereafter, we employ a divide-and-conquer strategy, perform neural surface reconstruction while embedding the Manhattan-world assumption [2], and finally merge the block-wise reconstructions through depth refinement. Our pipeline, shown in Figure 1, generates a finer and denser point cloud than the traditional Multi-view Stereo (MVS) approaches. In addition, our method predicts semantic logits of walls, floor, and ceiling regions for each point. The segmentation enables per-category plane-fitting and improves the performance on downstream tasks such as floor plan recovery and polygonal modelling.

We evaluate our system on a complex, multi-floored scene including rooms, staircases, and corridors. The scene is captured from a commercially available drone as a 20-minute video footage. Our system is capable of producing scalable, robust, and efficient 3D indoor representations using only monocular images.

2. RELATED WORK

2.1. Indoor scene reconstruction

3D scene reconstruction usually involves estimating per-image depth maps and fusing them into a 3D model. SfM, a class of traditional MVS approaches [3, 4], utilizes feature matching to find pixel correspondences across images to compute depths and camera poses. However, for indoor scenes, the matching often fails on large low-texture, or repetitive surfaces such as walls and floors, resulting in holes and outliers. Some works address the issue by introducing planar priors [5, 6] but they still perform poorly in large-scale indoor environments.

Learning-based MVS approaches [7–9] have become more popular in recent years. These works first deploy 2D Convolutional Neural Networks (CNN) to extract image features and build cost volumes. They construct 3D CNNs for cost volume regularization and predict depth using soft $argmin$. Skipping cross-image correspondences using the 3D CNNs, the data-driven approaches can hence address the aforementioned low-texture-region challenge, while the edges and corners are often over-smoothed.

2.2. Implicit neural representation

Alternatively, implicit neural representations describe the scene as a 3D field estimated by neural networks (NNs). [10] leverages the volume rendering to learn a implicit radiance

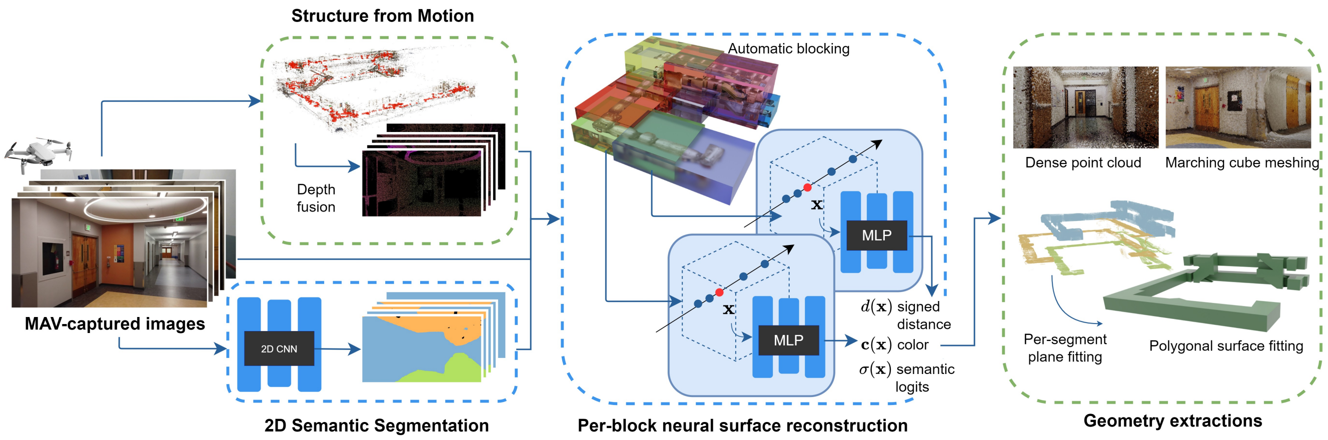

Structure from Motion
Depth fusion
MAV-captured images
2D CNN
2D Semantic Segmentation
Automatic blocking
x
MLP
x
MLP
$d(\mathbf{x})$ signed distance
$\mathbf{c}(\mathbf{x})$ color
$\sigma(\mathbf{x})$ semantic logits
Per-block neural surface reconstruction
Dense point cloud
Marching cube meshing
Per-segment plane fitting
Polygonal surface fitting
Geometry extractions

Fig. 1: Workflow of the proposed system

field from images; [11–13] combine the neural volume rendering with implicit surfaces to enable high-fidelity surface geometry extraction. These methods perform well on single objects with rich textures. However, they tend to result in erroneous or incomplete surfaces in low-texture regions common in indoor scenes.

Recently, novel approaches specifically tackle indoor scene reconstruction by introducing additional priors such as depth [2, 14, 15], geometric consistency [16], and planar region assumptions [2, 17]. These methods perform well in rectangular rooms. However, they fail to demonstrate reasonable reconstructions in more complex indoor scenes. Our method further extends these methods and scale the neural surface representation to work for larger scenes.

3. PROPOSED APPROACH

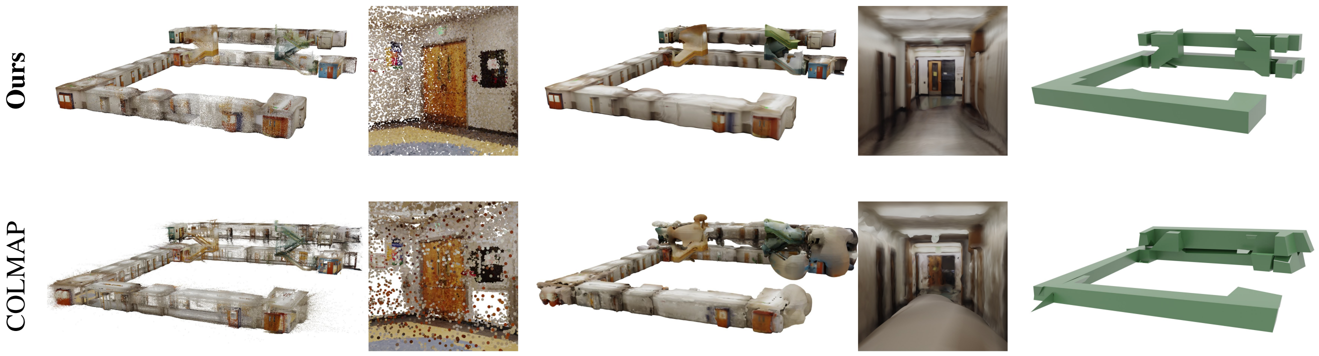
Our system takes an RGB image sequence as inputs and is capable of outputting various 3D representations for indoor modelling tasks. Figure 1 is an overview of our proposed workflow. We first derive camera poses and a point cloud using the traditional SfM and MVS. Then, we generate 2D depth maps by projecting each point back to its corresponding images. In order to apply the neural reconstruction mentioned in Section 2.2 to a larger scale, a divide-and-conquer strategy is leveraged by dividing the camera views into blocks. Simultaneously, we segment out walls, floors and ceilings in the input RGB images with a 2D CNN. Each block is then reconstructed with a modified ManhattanSDF [2] method with 2D depth maps and segmentation supervision. Thereafter, the blocks are aligned by a depth-based refinement step. At this stage, the scene is represented as block-wise signed distance and appearance fields. Finally, various geometry representations can be extracted for different downstream tasks, including a dense color point cloud from fusing viewing ray depth, a high-fidelity textured mesh using marching cubes [18], and an accurate polygonal model using our improved plane fitting.

3.1. Structure from Motion

Given an RGB image sequence of uncalibrated images, the SfM phase in the pipeline aims to reconstruct 3D geometries. We incorporate the traditional SfM and MVS [1, 3] pipeline to retrieve both camera poses and the fused point cloud. During SfM, we put scale constraints, such as the starting height of drone and widths of some corridors, so that the model is scaled to real-world units. Since geometric consistency is enforced by the fusion step [1], most points accurately depict surfaces. Due to the insufficiency of the SfM on indoor scenes mentioned in Section 2.1, the point cloud is expected to be sparse in low-texture regions with outliers across the scene. Surface reconstruction directly on such point clouds may perform poorly, necessitating refinement with neural surface approaches to be described in the next section. To utilize the point cloud as a depth prior, we project each point to frame coordinates of the corresponding images and acquire per-image sparse depth maps.

3.2. Neural Surface Reconstruction

We adopt an implicit neural surface approach [11] to fill the incomplete regions and refine the noisy surface. Specifically, each scene is represented by a signed distance field (SDF) F_d, a radiance field $F_{\mathbf{c}}$, and an additional semantic logits field F_σ. All the fields are parameterized using an MLP, mapping each 3D location $\mathbf{x}$ to a signed distance $d_\Omega(\mathbf{x})$, a color $\mathbf{c}(\mathbf{x})$, and a semantic logits $\sigma(\mathbf{x})$. These fields can be learned from images by accumulating points along viewing rays and supervising the rendered color by the corresponding image pixel. Compared to single-room scenes, large-scale scenes are more challenging due to more sparse view-ports, and larger low-texture areas. Among prior works on neural scene reconstruction [15–17], we observe that embedding the Manhattan world assumption is likely the most robust for our task [2]. Thus, we apply [2] as our backbone. To adapt it for larger scenes, we also make some modifications as follows:



(a) Dense point cloud (b) Mesh reconstruction (c) PolyFit [19] model

Fig. 2: Qualitative results of our reconstruction results comparing with COLMAP.

We acquire per-view sparse depth maps by projecting the reconstructed point cloud in Section3.1. The depth maps are of higher accuracy and better consistency than photometric estimated depth maps. In addition, since indoor-scene images are taken inside-out, we filter out obvious outliers by a maximum depth, and use them as the depth loss:

$$\mathcal{L}_D = \sum_{\mathbf{r}\in\mathcal{D}} |D(\mathbf{r}) - \hat{D}(\mathbf{r})| \tag{1}$$

where $\mathcal{D}$ denotes rays going through pixels with depth values.

Manhattan-world assumption [20], leveraged by ManhattanSDF, puts a geometric constraint by assuming floors, ceilings, and walls are mutually orthogonal. Guo *et al.* [2] observes that the low-texture regions obey the Manhattan-world assumption. Extending ManhattanSDF to our scenario, we train a 2D semantic segmentation network [21] to mask walls, floors, and ceilings. The masks regularize the planar region surface as follows:

$$\mathcal{L}_f = \sum_{\mathbf{r}\in\mathcal{F}} |1 - \mathbf{n}(\mathbf{x}_\mathbf{r}) \cdot \mathbf{n}_f| \tag{2}$$

$$\mathcal{L}_c = \sum_{\mathbf{r}\in\mathcal{C}} |1 + \mathbf{n}(\mathbf{x}_\mathbf{r}) \cdot \mathbf{n}_f| \tag{3}$$

$$\mathcal{L}_w = \sum_{\mathbf{r}\in\mathcal{W}} |\mathbf{n}(\mathbf{x}_\mathbf{r}) \cdot \mathbf{n}_f| + \alpha \min_{i\in\{-1,0,1\}} |i - \mathbf{n}(\mathbf{x}_\mathbf{r}) \cdot \mathbf{n}_w| \tag{4}$$

$$\mathcal{L}_{\text{geo}} = \mathcal{L}_f + \mathcal{L}_c + \mathcal{L}_w \tag{5}$$

where $\mathcal{F}, \mathcal{C}$, and $\mathcal{W}$ denote rays going through pixels masked with floor, ceiling, and wall, respectively. The floor normal $\mathbf{n}_f$ and wall normal $\mathbf{n}_w$ are specified as:

$$\mathbf{n}_f = [\sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi), \cos(\theta)] \tag{6}$$

$$\mathbf{n}_w = [\sin(\theta + \frac{\pi}{2})\cos(\phi), \sin(\theta + \frac{\pi}{2})\sin(\phi), \cos(\theta + \frac{\pi}{2})] \tag{7}$$

where θ, ϕ are initialized as 0 and are optimized through training. We also add another hyperparameter $\alpha \in [0, 1]$ to soften the Manhattan world constraint, because in practice walls are not always orthogonal or parallel to $\mathbf{n}_w$. We also include the semantic field to estimate semantic logits at each 3D location, the field is trained with the joint optimization loss described in [2].

3.3. Scaling reconstruction with blocks

For large scenes, neural surface reconstruction forsakes details and eventually becomes intractable. Therefore, we use a divide-and-conquer strategy to make our approach scalable. We partition camera poses along the trajectory into several segments and ensure enough overlapping poses for adjacent segments. We partition the poses by distance so that each segment is bounded within a $30m \times 30m \times 5m$ box and the overlapping volume between each pair of segment is at least $10m \times 10m \times 5m$. Subsequently, we initialize neural surface reconstruction on each block. Despite the sparse depth supervision, the surfaces show some mismatches in the overlapping regions. To overcome this, we propose a depth-based boundary refinement step to align each pair of implicit surfaces. For each camera view in the overlapping area, we replace the sparse depth with the estimated depth from each of the neighboring blocks as:

$$\mathcal{L}_{D,i} = \sum_{\mathbf{r}} |\hat{D}_i(\mathbf{r}) - \hat{D}_j(\mathbf{r})| \tag{8}$$

where $\hat{D}_i(\mathbf{r})$ and $\hat{D}_j(\mathbf{r})$ are the depth estimations from blocks i and j, respectively.

3.4. Geometry Extraction and Polygonal Surface Fitting

We initialize a voxel grid volume to bound our scene to extract geometries from the block-wise neural implicit fields. For each voxel, we query all the implicit signed distance fields that bound its location and assign the minimum distance to the coordinate. With the signed distance grid, we can either generate a dense point cloud by fusing sampled ray depth or extract a mesh using the marching cubes [18] algorithm.

Polygonal surface reconstruction generates simplified models from point cloud. The method is suitable for many downstream tasks in man-made environments by eliminating unnecessary surface details while preserving sharp features. In general, the reconstruction starts with random sample consensus (RANSAC) planar primitive fitting and reassembles the planes to form a geometry. However, the non-planar objects introduce noise for RANSAC plane fitting, leading to erroneous planes in indoor scenes. Our per-point semantic logits described in Section 3.2 conveniently solve this problem as they enable a categorical plane fitting. We execute RANSAC exclusively for walls, floors or ceilings by removing those categorized as none of the above three by the semantic logits.

4. EXPERIMENTS

4.1. Data capture

We used a DJI Mini 2 drone, which only weighs 249g, to capture videos. The drone traversed a trajectory whereby the optical axis of its camera was varying over time from being perpendicular to parallel to the main axis of the hallway. All videos are captured in 2720 × 1530 at 24 FPS. Our example reconstruction used two clips captured in two flights for a total of 20 minutes. Frames are extracted at 2 FPS. We further filtered out overly blurred images by a threshold over variance of image Laplacian and removed redundant views by a threshold on distances among estimated poses. 1971 frames were extracted and resized to 1360 × 765.

4.2. Implementation

We used Pix4D Mapper for camera pose recovery, then use COLMAP's dense reconstruction [3] to generate a point cloud. For 2D semantic segmentation, we trained a DeepLab [21] model on ADE20K dataset [22]. For neural surface reconstruction, we partitioned the space into 9 blocks, each containing 300-400 images. The training was performed on an NVIDIA TITAN RTX GPU. We first trained each block with batches of 2048 rays for 10k iterations and then perform boundary refinement for 2k iterations. Each block took approximately 2 hours to train. For polygonal surface modeling, our system fitted planes to each of wall, floor, and ceiling segments as is mentioned in Section 3.4. PolyFi [19] is applied to find plane boundaries and assemble planes into a polygonal model. We compared our proposed 3D reconstruction framework with a traditional robust pipeline COLMAP [3] with its built-in Poisson meshing [23].

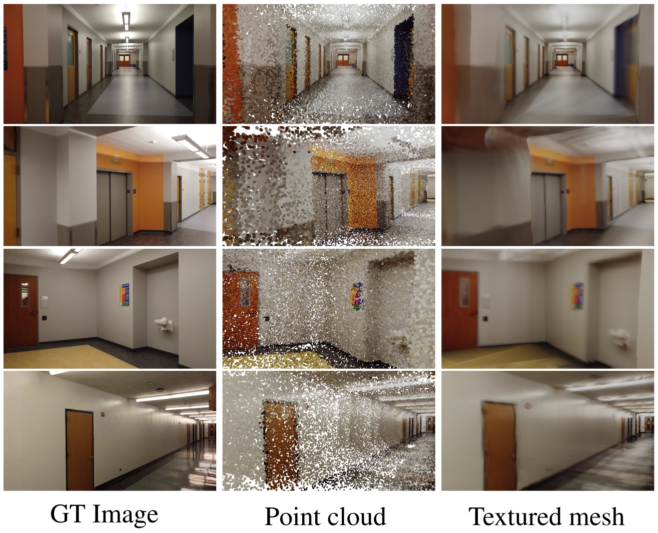


Fig. 3: Close-up observation of our reconstructions.

4.3. Results

As shown in Figure 2, we compared our method with COLMAP in terms of dense point cloud quality, meshing, and polygonal surface fitting. In Figure 2(a), the point cloud from our method is dense and accurate even though the selected indoor scene contains a large number of low-texture white walls, while the point cloud from COLMAP contains a large number of holes and outliers. In Figure 2(b), our mesh accurately models the surface, while the one from COLMAP is erroneous due to point cloud noises. Finally, as shown in Figure 2(c), our system successfully generates a simple polygonal model by filtering out non-planar noise.

Figure 3 contrasts our reconstructed geometry against the RGB pictures at approximately the same viewpoint. Our reconstruction accurately predicts the surface even for low-texture regions. However, the surface details in our reconstruction are somewhat smooth, especially in corners associated with short wall segments. We speculate it to be caused by Manhattan-world assumption being too strong to recover some details. More results and a complete walk-through are shown in this video: https://youtu.be/l23-fPahw38

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have developed a completely end-to-end system to build 3D models for indoor scenes from RGB images using drones. We utilize both the traditional methods and the neural rendering approaches with a block-by-block strategy. The Manhattan-world assumption is used to improve point cloud quality, followed by plane fitting to create polygonal 3D models. Our system is proven robust enough for large-scale indoor modeling. For our future work, we will improve the recovery of geometric details and textures.

6. REFERENCES

[1] Johannes Lutz Schönberger and Jan-Michael Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou, "Neural 3d scene reconstruction with the manhattan-world assumption," in *CVPR*, 2022.

[3] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[4] Yasutaka Furukawa and Jean Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376, 2010.

[5] Michal Jancosek and Tomas Pajdla, "Multi-view reconstruction preserving weakly-supported surfaces," in *CVPR 2011*, 2011, pp. 3121–3128.

[6] Andrea Romanoni and Matteo Matteucci, "Tapa-mvs: Textureless-aware patchmatch multi-view stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10413–10422.

[7] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," *European Conference on Computer Vision (ECCV)*, 2018.

[8] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan, "Recurrent mvsnet for high-resolution multi-view stereo depth inference," *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] Gu Xiaodong, Fan Zhiwen, Zhu Siyu, Dai Zuozhuo, Tan Feitong, and Tan Ping, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[10] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[11] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman, "Volume rendering of neural implicit surfaces," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[12] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *NeurIPS*, 2021.

[13] Michael Oechsle, Songyou Peng, and Andreas Geiger, "Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction," in *International Conference on Computer Vision (ICCV)*, 2021.

[14] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou, "Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo," in *ICCV*, 2021.

[15] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger, "Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction," *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[16] Qiancheng Fu, Qingshan Xu, Yew-Soon Ong, and Wenbing Tao, "Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction," *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[17] Yusen Wang, Zongcheng Li, Yu Jiang, Kaixuan Zhou, Tuo Cao, Yanping Fu, and Chunxia Xiao, "Neuralroom: Geometry-constrained neural implicit surfaces for indoor scene reconstruction," *ACM Trans. Graph.*, vol. 41, no. 6, nov 2022.

[18] William E. Lorensen and Harvey E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *SIGGRAPH Comput. Graph.*, vol. 21, no. 4, pp. 163–169, aug 1987.

[19] Liangliang Nan and Peter Wonka, "Polyfit: Polygonal surface reconstruction from point clouds," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2353–2361.

[20] James M Coughlan and Alan L Yuille, "Manhattan world: Compass direction from a single image by bayesian inference," in *Proceedings of the seventh IEEE international conference on computer vision*. IEEE, 1999, vol. 2, pp. 941–947.

[21] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[22] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 302–321, 2019.

[23] Michael Kazhdan and Hugues Hoppe, "Screened poisson surface reconstruction," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 3, pp. 1–13, 2013.