Scalable MAV Indoor Reconstruction with Neural Implicit Surfaces

Anonymous ICCV submission

Paper ID ****

Abstract

Many previous works achieved impressive reconstruction results on room-scale indoor scenes from multi-view RGB images, but capturing and reconstructing multistory, complex indoor scenes is still a challenging problem. 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-textured 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 achieved better PSNR in view synthesis tasks and higher floor plan IOU than traditional reconstruction solutions such as COLMAP.

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. Existing mobile reconstruction methods use either a robot on wheels or a human operator to capture data; the former is limited in capturing detailed images from all angles and all locations especially near the ceiling; the latter is limited in SLAM reconstruction due to unavoidable human body movements such as pitch and roll. As such, most high fidelity systems use tripod in a stop-and-go capture which is laborious and time-consuming. In contrast, drones enable rapid and effortless capture from a variety of perspectives and positions, including confined areas and result in stable camera trajectory not achievable with robots or humans.

However, indoor environment poses 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 a single sensor, fast data capture methodology and processing pipeline to overcome these challenges. Specifically, we developed a drone based data capture strategy by choosing the direction of motion to be parallel to optical axis of the camera resulting in stable pose recovery, while swaying to right and left in a periodic fashion to capture high resolution images of the walls. We leverage both traditional Structure from Motion (SfM) [19] and neural surface reconstruction [9] 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 [9], 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.

Compared to recent advancements in indoor scene reconstruction with multi-view images, our method focuses on large-scale indoor scenes beyond single, rectangular rooms and could include staircases. Existing methods such as MonoSDF [32], NeuRIS [22], and NeuralRoom [24] are limited to a single room, while our system can deal with scenes that are at least one order of magnitude larger with much faster capture time. 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 drone captured images.

2. Related Work

2.1. MAV-based reconstruction

In recent years, drones have proven to be a cost-efficient and reliable solution for large objects reconstruction. The aerial perspective enables accurate reconstruction without additional sensory data and its maneuverability provides rapid and interactive capture experiences. For example, there are existing methods using drone imagery for 3D facade reconstruction. Daftry *et al.* [3] presented an incremental reconstruction system that provides users with online feedback in real time. Wudunn *et al.* [26] allows accurate building footprint reconstructions with only monocular image data.

Recent advancements have made MAVs smaller and safer, leading to new applications in indoor environments. Some works have used drone to obtain fast camera motion to estimate human poses within scenes [34, 8]. Other works combine MAVs with other devices. For example, Gao *et al.* [7] uses aerial maps created from MAVs to guide and localize ground robots for reconstruction. However, there is no existing work producing high-fidelity reconstructions with the MAV alone.

2.2. 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 [20, 6], utilize 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 [10, 18] but they still perform poorly in large-scale indoor environments.

Learning-based MVS approaches [29, 30, 27] 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.

2.3. Implicit neural representation

Alternatively, implicit neural representations describe the scene as a 3D field estimated by neural networks (NNs). [14] leverages the volume rendering to learn a implicit radiance field from images; [31, 23, 17] 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 [25, 32, 9], geometric consistency [5], and planar region assumptions [9, 24]. 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 multistory indoor scenes.

3. Proposed Approach

Our system takes an RGB image sequence captured with a drone as input 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 2.3 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 [9] 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 [13], 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 [20, 19] 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 [19], most points accurately depict surfaces. Due to the insufficiency of the SfM on indoor scenes mentioned in Section 2.2, the point cloud is sparse in low-texture regions with outliers across the scene. Surface reconstruction directly on such point clouds performs poorly, necessitating refinement with neural surface approaches to be described in the next section. To utilize the point cloud as a depth prior, we

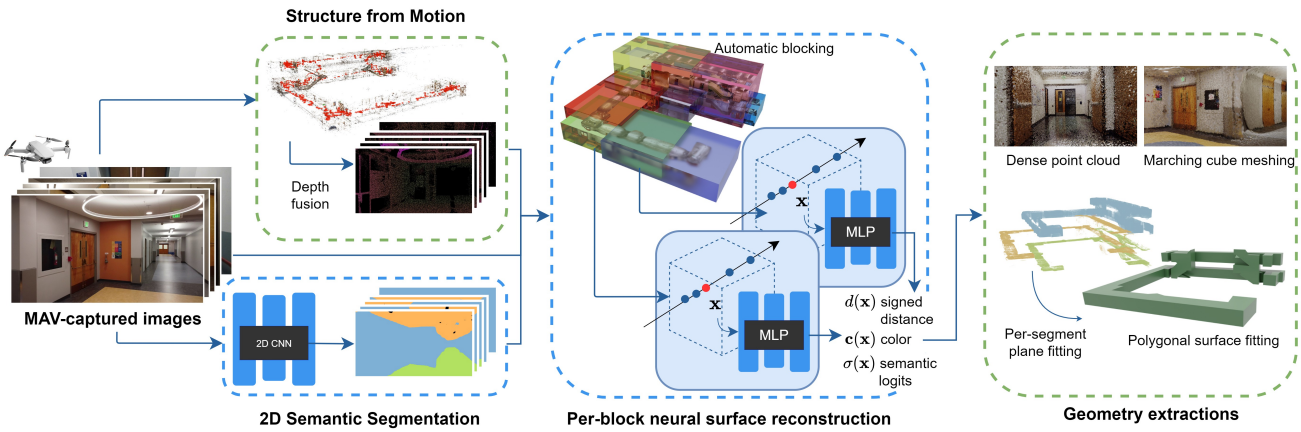


Figure 1: Overview of the proposed system. Our system scales uses drone captured images to reconstruct multistory scenes. We scale the implicit neural surface reconstruction with a divide-and-conquer strategy. Our system can output various geometry representations for different downstream tasks.

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 [31] to fill the incomplete regions and refine the noisy surface. Specifically, each scene is represented by a geometry field F_{d_Ω} and radiance field $F_\mathbf{c}$. The fields are parameterized using MLPs. Specifically, at each 3D position $\mathbf{x}$ and viewing direction $\mathbf{v}$, the signed distance field F_d is defined as

$$(d_\Omega, \mathbf{z}) = F_{d_\Omega}(\mathbf{x}) \tag{1}$$

where d_Ω is the signed distance and $\mathbf{z}$ is the geometric features. The radiance field $F_\mathbf{c}$ is defined as

$$\mathbf{c} = F_\mathbf{c}(\mathbf{x}, \mathbf{d}, \mathbf{n}(\mathbf{x}), \mathbf{z}(\mathbf{x})) \tag{2}$$

where $\mathbf{c}$ is the color and $\mathbf{n}$ is the surface normals obtained by the gradient of signed distance $d_\Omega(\mathbf{x})$. These fields can be learned from images in a differentiable volume rendering manner [14, 31]. For each pixel, we shoot ray $\mathbf{r} = (\mathbf{o}, \mathbf{v})$ as ray origin $\mathbf{o}$ and viewing direction $\mathbf{v}$, and sample N points $\mathbf{x}_i = \mathbf{o} + t_i\mathbf{v}$ for $i = 1, 2, ..., N$. At each point $\mathbf{x}_i$, we obtain the signed distance $d_{\Omega,i}$, color $\mathbf{c}_i$ and semantic logits $\mathbf{s}_i$. The signed distances $d_{\Omega,i}$ are converted into volume density σ_i via a transformation with learnable parameter β as

$$\sigma_i = \begin{cases} \frac{1}{\beta}(1 - \frac{1}{2}\exp(\frac{d_{\Omega,i}}{\beta})) & d_{\Omega,i} < 0 \\ \frac{1}{2\beta}\exp(-\frac{d_{\Omega,i}}{\beta}) & d_{\Omega,i} \geq 0 \end{cases} \tag{3}$$

Then we accumulate colors along the ray as

$$C(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i(t_{i+1} - t_i)))\mathbf{c}_i \tag{4}$$

where $T_i = \sum_{j=1}^{i-1} \sigma_j(t_{j+1} - t_j)$ is the accumulated transmittance. Therefore, the fields can be optimized by the RGB loss

$$\mathcal{L}_C = \sum_{\mathbf{r}\in\mathcal{R}} |C(\mathbf{r}) - C(\mathbf{r})| \tag{5}$$

where $\mathcal{R}$ is the set of rays in multi-view images and $\hat{C}(\mathbf{r})$ is the corresponding ground-truth pixel color.

3.3. Embedding Manhattan-world Assumptions

Compared to existing indoor scanning datasets [4, 21], the drone-captured images have more reliable poses and consistent gravitational direction. However, the larger scenes are more challenging due to more sparse viewpoints and larger low-texture areas. Many prior works used monocular estimated depths and normals as supervision [24, 32, 5], while they were unreliable in large complex scenes, and often led to incorrect reconstruction. Instead, we observed that embedding the Manhattan world assumption is the most robust for our task. Thus, we apply [9] as our backbone. To adapt it for larger scenes, we also make some modifications as follows:

We acquire per-view sparse depth maps by projecting the reconstructed point cloud in Section 3.1. 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 depth range, and use the valid depths as the ground-truth $\hat{D}$ to supervise a sparse depth loss

$$L_D = \sum_{\mathbf{r}\in\mathcal{D}} |D(\mathbf{r}) - \hat{D}(\mathbf{r})| \tag{6}$$

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

Manhattan-world assumption [2], leveraged by ManhattanSDF, puts a geometric constraint by assuming floors, ceilings, and walls are mutually orthogonal. Guo *et al.* observes that the low-texture regions obey the Manhattan-world assumption. Extending ManhattanSDF to our scenario, we train a 2D semantic segmentation network [1] to mask walls, floors, and ceilings. The masks regularize the planar region surface as follows:

$$\mathcal{L}_f = \sum_{\mathbf{r}\in\mathcal{F}} |1 - \mathbf{n}(\mathbf{r}) \cdot \mathbf{n}_f| \tag{7}$$

$$\mathcal{L}_c = \sum_{\mathbf{r}\in\mathcal{C}} |1 + \mathbf{n}(\mathbf{r}) \cdot \mathbf{n}_f| \tag{8}$$

$$\mathcal{L}_w = \sum_{\mathbf{r}\in\mathcal{W}} |\mathbf{n}(\mathbf{r}) \cdot \mathbf{n}_f| + \alpha \min_{i\in\{-1,0,1\}} |i - \mathbf{n}(\mathbf{r}) \cdot \mathbf{n}_w| \tag{9}$$

where $\mathbf{n}(\mathbf{r})$ is the estimated normal and $\mathcal{F}, \mathcal{C}$, and $\mathcal{W}$ denote rays going through pixels masked with floor, ceiling, and wall, respectively. Since the drone has consistent gravitational direction, the floor normal and wall normal is simply

$$\mathbf{n}_f = (0, 0, 1), \mathbf{n}_w = (\sin(\theta), \cos(\theta), 0) \tag{10}$$

where θ is a learnable parameter. 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 $\mathbf{n}_w$. To obtain more reliable estimation of walls, floors, and ceiling in 3D, we also include the semantic field $F_\mathbf{s}(\mathbf{x})$ to estimate semantic logits at each 3D location. The semantic logits are accumulated similar to the color accumulation, and is optimized against the estimated semantic segmentation map using the joint optimization loss described in [9].

3.4. 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 iteratively optimize the depth using the estimated depth from each of the neighboring blocks as

$$L_{D,i} = \sum_{\mathbf{r}} |D_i(\mathbf{r}) - D_j(\mathbf{r})| \tag{11}$$

where $D_i(\mathbf{r})$ and $D_j(\mathbf{r})$ are the depth estimations from blocks i and j, respectively.

3.5. 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 [13] 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.

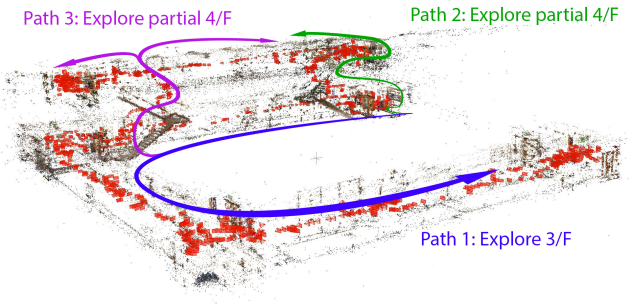
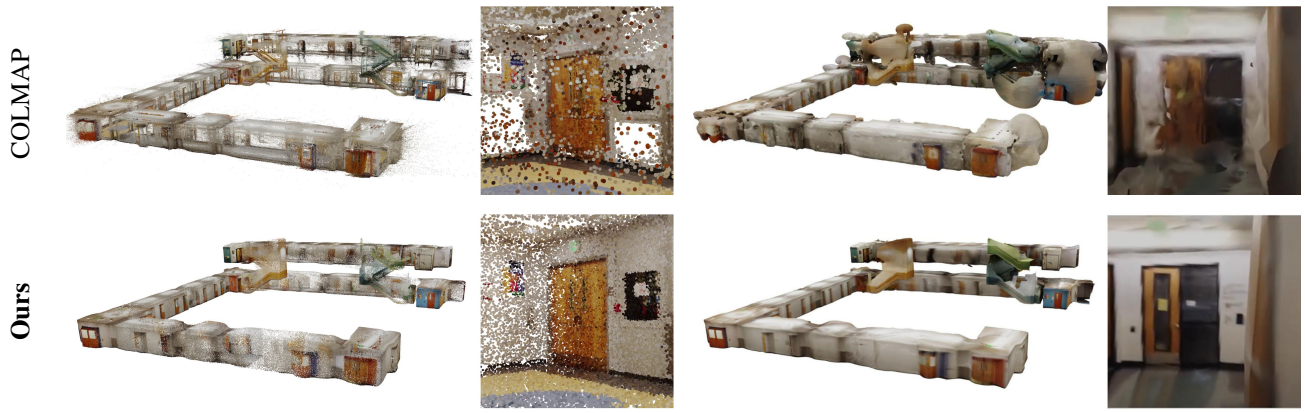


Figure 2: Camera trajectory of our captured data. We shown the three flights along with the sparse point cloud from SfM.

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 clips captured in three flights for a total of 20 minutes, the flight trajectories are shown in



(a) Point cloud (b) Textured mesh

Figure 3: Qualitative results of our reconstruction results. (a) Our dense point cloud is dense on low-textured areas and is far less noisy than COLMAP. (b) Our method reconstructs more accurate textured mesh and is more robust than COLMAP on reflective and low-textured surface.

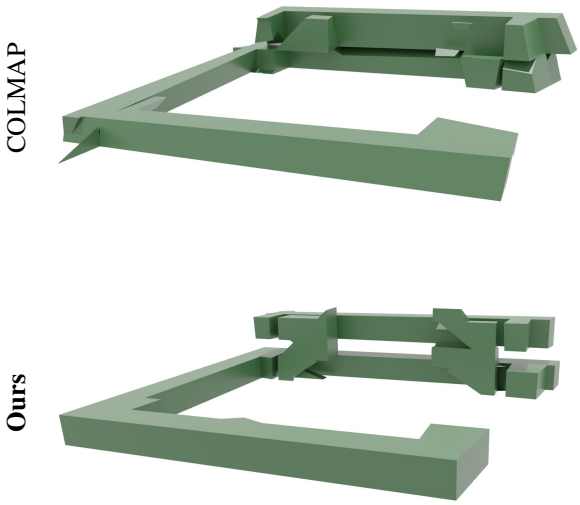


Figure 4: Polygonal surface reconstruction results.

Figure 2. 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. Compared to existing indoor reconstruction datasets, our data has more sparse viewpoints and includes long corridors, staircases, and large open areas.

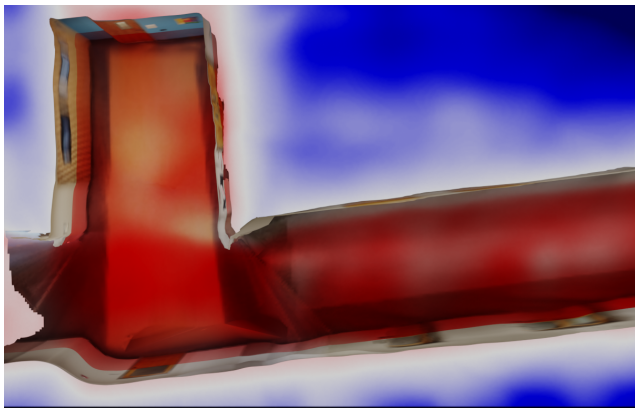
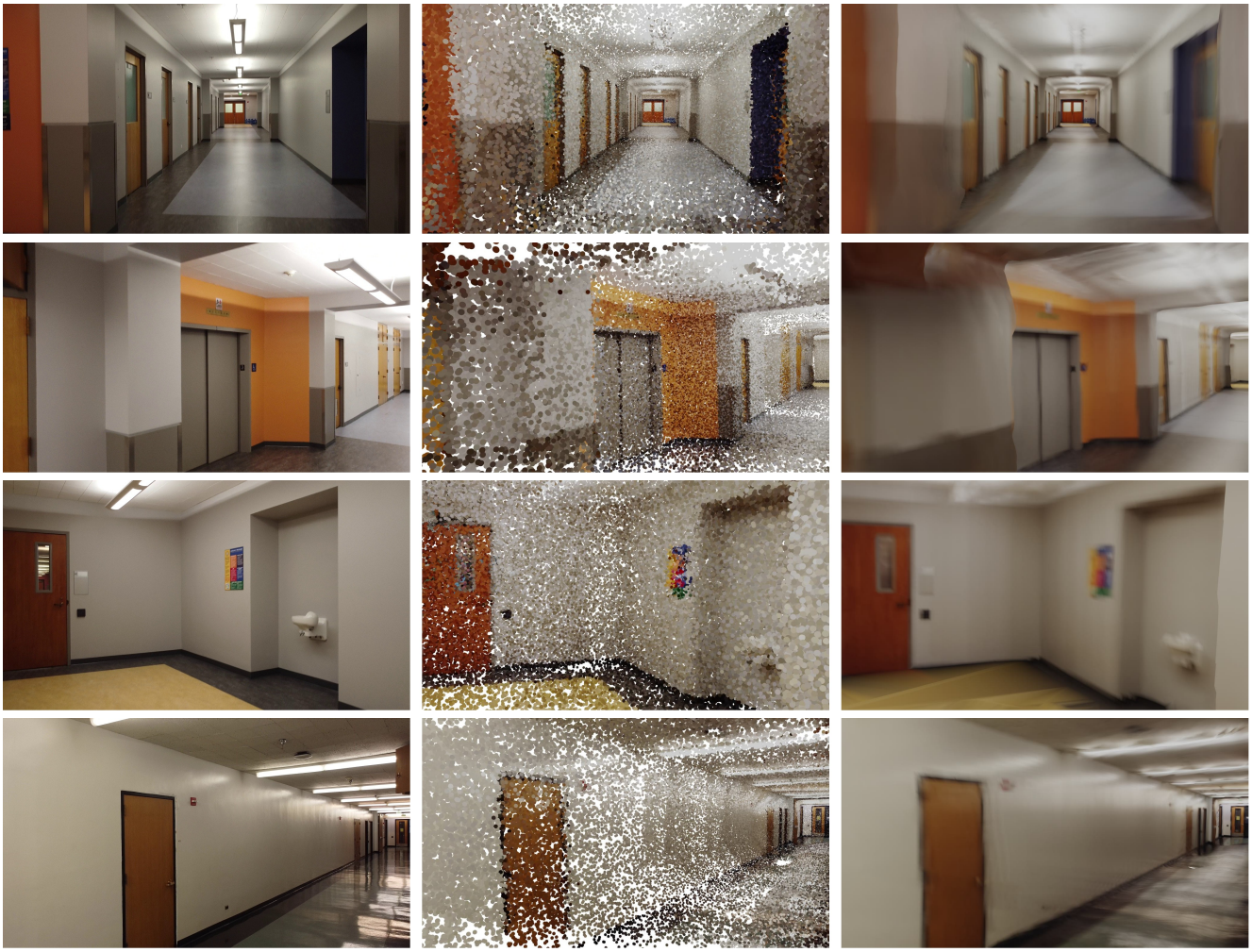


Figure 5: A cross-section of signed distance field. the boundary regions of the surface are shown in white color.

4.2. Implementation

We used Pix4D Mapper for camera pose recovery, then use COLMAP's dense reconstruction [20] to generate a point cloud. For 2D semantic segmentation, we trained a DeepLab-V3+ [1] model on ADE20K dataset [33] with remapped labels. For neural surface reconstruction, we partitioned the space into 9 blocks, each containing 300-400 images and approximately covers $30m \times 30m$ area. 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 performed boundary refinement for 2k iterations. Each block took approximately 1.5 hours to train. For polygonal surface modeling, our system fitted planes to each of wall, floor, and ceiling segments as mentioned in Section 3.5. PolyFit [16] is applied to find plane



GT Image | Point cloud | Textured mesh

Figure 6: Close-up observation of our reconstructions.

boundaries and assemble planes into a polygonal model. We compared our proposed 3D reconstruction framework with a traditional robust pipeline COLMAP [20] with its built-in Screened Poisson meshing [11].

4.3. Results

We compared our method with COLMAP in terms of dense point cloud quality, meshing, and polygonal surface fitting. In Figure 3(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 3(b), our mesh accurately models the surface, while the one from COLMAP is erroneous due to point cloud noises. Moreover, our method significantly outperforms COLMAP on recovering the reflective surfaces. As shown in the closeup image, while our method successfully models the reflective floor, where COLMAP had catastrophic failures. In Figure 4, our system successfully generates a simple polygonal model by filtering out non-planar noise.

We present the quantitative results in Table 1. We render novel view images using the reconstructed geometry from our system and COLMAP. Our system has $4dB$ higher average PSNR and 0.13 higher SSIM than COLAMP, and preserves more local features. In addition, We extracted the floor plan by computing the zero-crossing of the SDF on a 2D cross section, as demonstrated in Figure 5. We then computed the IOU value against the ground-truth floor plan, and our method achieves 14% higher IOU values than COLMAP.

Figure 6 contrasts our reconstructed geometry against

Figure 7: Our reconstruction method over-smooths local details.

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. which is shown in Figure 7. 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 the supplementary material.

Table 1: Quantitative comparisons between our method and COLMAP.

	PSNR(dB)	SSIM	IOU
Ours	20.76	0.74	0.83
COLMAP	16.55	0.61	0.73

5. Conclusions and Future Works

Albeit the wide utilization of drones in modeling the terrain and building facades, indoor mapping with MAVs to generate 3D geometry remains unsolved and challenging. In this paper, we proposed a fully automated pipeline for reconstructing large and complex indoor scenes with RGB images collected from drones. First, we leverage traditional structure-from-motion methods to obtain camera poses and reconstruct the initial point cloud. Finally, we devise a divide-and-conquer strategy to utilize neural surface reconstruction under the Manhattan-world assumption. Our method reduces point clouds' outliers and improve reconstruction quality on low-texture regions. Our method demonstrates the great scalability, efficiency, and accuracy of drone-based indoor mapping, and out-performs COLMAP with 4dB in average PSNR.

For future work, we suggest optimizing the pipeline to achieve real-time positioning and modeling. Currently, the implicit neural representation model in our pipeline is a canonical MLP. There are many other novel differentiable rendering representations wielding better model structures to alleviate the computational complexity problem, such as hash-encoding [15], neural point cloud [28, 12]. With a more rapid model reconstruction, we can replace our structure-with-motion module with a simultaneous localization and mapping (SLAM) module and achieve real-time modeling.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 833–851, Cham, 2018. Springer International Publishing. 4, 5

[2] 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*, volume 2, pages 941–947. IEEE, 1999. 4

[3] Shreyansh Daftry, Christof Hoppe, and Horst Bischof. Building with drones: Accurate 3d facade reconstruction using mavs. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3487–3494, 2015. 2

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE, CVPR*, pages 5828–5839, 2017. 3

[5] 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. 2, 3

[6] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010. 2

[7] Xiang Gao, Lingjie Zhu, Hainan Cui, Zhanyi Hu, Hongmin Liu, and Shuhan Shen. Complete and accurate indoor scene capturing and reconstruction using a drone and a robot. *IEEE Sensors Journal*, 21(10):11858–11869, 2021. 2

[8] Stuart Golodetz, Madhu Vankadari, Aluna Everitt, Sangyun Shin, Andrew Markham, and Niki Trigoni. Real-time hybrid mapping of populated indoor scenes using a low-cost

monocular uav. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 325–332, 2022. 2

[9] 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. 1, 2, 3, 4

[10] Michal Jancosek and Tomas Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *CVPR 2011*, pages 3121–3128, 2011. 2

[11] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 6

[12] Ruofan Liang, Jiahao Zhang, Haoda Li, Chen Yang, and Nandita Vijaykumar. Spidr: Sdf-based neural point fields for illumination and deformation. *arXiv preprint arXiv:2210.08398*, 2022. 7

[13] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH Comput. Graph.*, 21(4):163–169, aug 1987. 2, 4

[14] 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. 2, 3

[15] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 7

[16] Liangliang Nan and Peter Wonka. Polyfit: Polygonal surface reconstruction from point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2353–2361, 2017. 5

[17] 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. 2

[18] Andrea Romanoni and Matteo Matteucci. Tapa-mvs: Textureless-aware patchmatch multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10413–10422, 2019. 2

[19] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[20] 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. 2, 5, 6

[21] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3

[22] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, page 139–155, Berlin, Heidelberg, 2022. Springer-Verlag. 1

[23] 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. 2

[24] 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.*, 41(6), nov 2022. 1, 2, 3

[25] 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. 2

[26] Marc Wudunn, Avideh Zakhor, Samir Touzani, and Jessica Granderson. Aerial 3d building reconstruction from rgb drone imagery. In *Defense + Commercial Sensing*, 2020. 2

[27] 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. 2

[28] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 7

[29] 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. 2

[30] 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. 2

[31] 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. 2, 3

[32] 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. 1, 2, 3

[33] 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*, 127(3):302–321, 2019. 5

[34] Xiaowei Zhou, Sikang Liu, Georgios Pavlakos, Vijay Kumar, and Kostas Daniilidis. Human motion capture using a drone. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2027–2033, 2018. 2