Computer Vision Term Project Report

Zixin Huang[1]
[1] Beijing Institute of Technology

Abstract

This paper presents a novel approach for 3D indoor scene reconstruction using the Manhattan-world assumption. It leverages Multi-Layer Perceptrons (MLPs) to represent scene geometry through signed distance functions, integrating planar constraints in floor and wall regions based on semantic segmentation. The method optimizes both scene geometry and semantics in 3D space, significantly enhancing reconstruction quality, especially in low-textured planar regions. The approach outperforms previous methods in terms of 3D reconstruction quality on ScanNet. The code and supplementary materials are available at https://github.com/foreverlasting1202/cv-term-pj.

1. Introduction

3D scene reconstruction generally refers to the input of a series of image sequences captured in indoor scenes, with the aim of obtaining their 3D models.

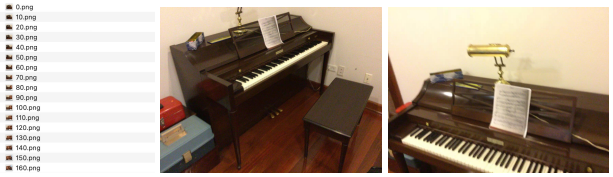

Figure 1. There are a total of 464 pictures. Here are some of them.

Deep convolutional networks are the primary tools for image-based methods, but they experience rapid growth in spatial and temporal complexity when directly extended to the third spatial dimension. More classical and compact surface representation methods, such as triangle or quadrilateral meshes, face challenges during training due to the potential need to handle an unknown number of vertices and arbitrary topological structures. These challenges limit the quality, flexibility, and fidelity of deep learning methods when processing input 3D data or generating 3D inferences for object segmentation and reconstruction [9].

Since the advent of deep learning, generative models have made significant strides in producing realistic high-resolution images, yet this success has not been replicated in the 3D domain. Existing representations are broadly categorized into three types: voxel-based, point-based, and mesh-based representations. Contrasting this, a paper proposes considering the continuous decision boundary of a classifier [7], such as a deep neural network, as a 3D surface, enabling the extraction of 3D meshes at any resolution.

Traditional methods typically employ Multi-View Stereo (MVS) [5] for scene reconstruction, initially estimating depth maps for each viewpoint based on multi-view matching, and then fusing the depths of each viewpoint in three-dimensional space. The biggest issue with these methods is the difficulty in handling weakly textured areas and non-Lambertian surfaces, as these regions are challenging to match, leading to incomplete reconstructions.

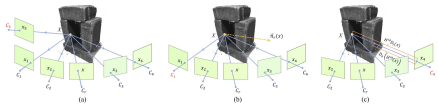

Figure 2. Multi-View Stere [5].

Recently, methods have been proposed for 3D reconstruction based on implicit neural representations.

NeRF [8] learns an implicit radiance field from images through differentiable volumetric rendering. NeRF can achieve realistic viewpoint synthesis, but its geometric reconstruction results are severely noisy, mainly due to the lack of surface constraints.

NeuS [14] and VolSDF [15] model the scene's geometry using SDF (signed distance fields) and achieve SDF-based volumetric rendering, resulting in smoother geometric reconstruction compared to NeRF. However, as these methods are based on the principle of photometric consistency, they struggle with weakly textured areas and perform poorly in indoor scene reconstruction.

Given these reasons, some papers have proposed methods based on the Manhattan world assumption [3], but the scenarios where the Manhattan world assumption is applicable are still too limited.

In this paper, building upon this foundation, the Manhattan world is advanced to an Atlanta world, and the loss

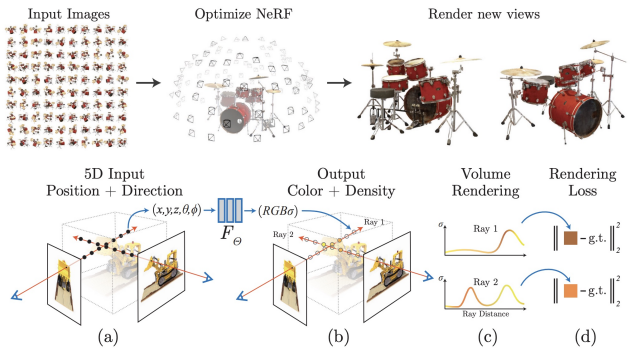

Figure 3. NeRF [8].

function has been further improved to achieve better results.

In summary, our contributions are as follows:

- A novel loss function that optimizes semantic labels along with scene geometry.
- A novel scene reconstruction approach that integrates the Atlanta-world constraint into the optimization of implicit neural representations.

2. Related work

Implicit Neural Representation. Implicit neural representations have gained significant traction in the field of computer vision and 3D modeling, presenting an innovative approach to representing complex geometries and functions. These representations employ neural networks to implicitly encode data, such as images, shapes, or even sounds, in a continuous manner. NeRF (Neural Radiance Fields), Introduced by Mildenhall et al. in their seminal work "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis" [8], NeRF has set a new standard in synthesizing novel views of complex 3D scenes with unprecedented detail and photorealism. This method uses a fully connected neural network to map 5D coordinates (spatial location and direction) to color and density, creating highly realistic renderings from sparse viewpoints.

MVS. Multi-View Stereo (MVS) techniques have seen significant advancements in recent years, driven by the growing demand for high-quality 3D reconstructions in various applications, from virtual reality to autonomous navigation. Below is an overview of some recent notable works in the field of MVS, along with relevant references. Deep Learning-based MVS, the integration of deep learning with traditional MVS techniques has led to substantial improvements in accuracy and efficiency. A prominent example is "DeepMVS: Learning Multi-View Stereopsis" by Huang et al [4]., which introduced a neural network-based approach to aggregate information across multiple views, enhancing the depth estimation process.

Semantic segmentation. Semantic segmentation, a key task in computer vision, has seen significant advancements recently, primarily fueled by deep learning technologies. High-Resolution Segmentation: Efforts to maintain high-resolution details in segmentation maps, as seen in "HRNet: High-Resolution Network for Semantic Segmentation" [12], which focuses on maintaining high-resolution representations through the network. Addressing the challenge of adapting models trained on one domain to another, as explored in "Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training" [16], which proposes a self-training method to refine model predictions on a new domain.

SDF. Recent research in Signed Distance Functions (SDF) has seen notable progress across various domains in computer vision and 3D reconstruction. The integration of neural SDFs with volume rendering, as demonstrated by Wang et al. in "NeuS," [14] enhances differentiable 3D reconstruction. Meanwhile, Chen et al. have extended the use of SDFs to dynamic scene reconstruction [1], addressing the challenges of temporal changes. Furthermore, combining traditional voxel-based methods with SDF, as explored by Shen et al., has significantly improved resolution and detail in 3D reconstructions [10], demonstrating the growing versatility and effectiveness of SDFs in these fields.

3. Implementation details

We implement our method with PyTorch and adopt DeepLabV3+ from Detectron2 for implementing 2D semantic segmentation network. The network training is performed on one NVIDIA 3090 GPU.

4. Experiments

Dataset. We perform the experiments on ScanNet (V2) [2]. ScanNet is an RGB-D video dataset that contains 1613 indoor scenes with 2.5 million views. It is annotated with ground-truth camera poses, surface reconstructions, and instance-level semantic segmentations.

Metrics. For 3D reconstruction, we use RGB-D fusion results as ground truth and evaluate our method using 5 standard metrics: accuracy, completeness, precision, recall, and **F-score**. Following the recommendation of [11], we consider the **F-score** as the overall metric.

Baselines. We use five baselines:

1. Classical MVS method: COLMAP [13].
2. NeRF [8].
3. NeuS [14].
4. VolSDF [15].
5. Manhattan [3].

Ablation study.

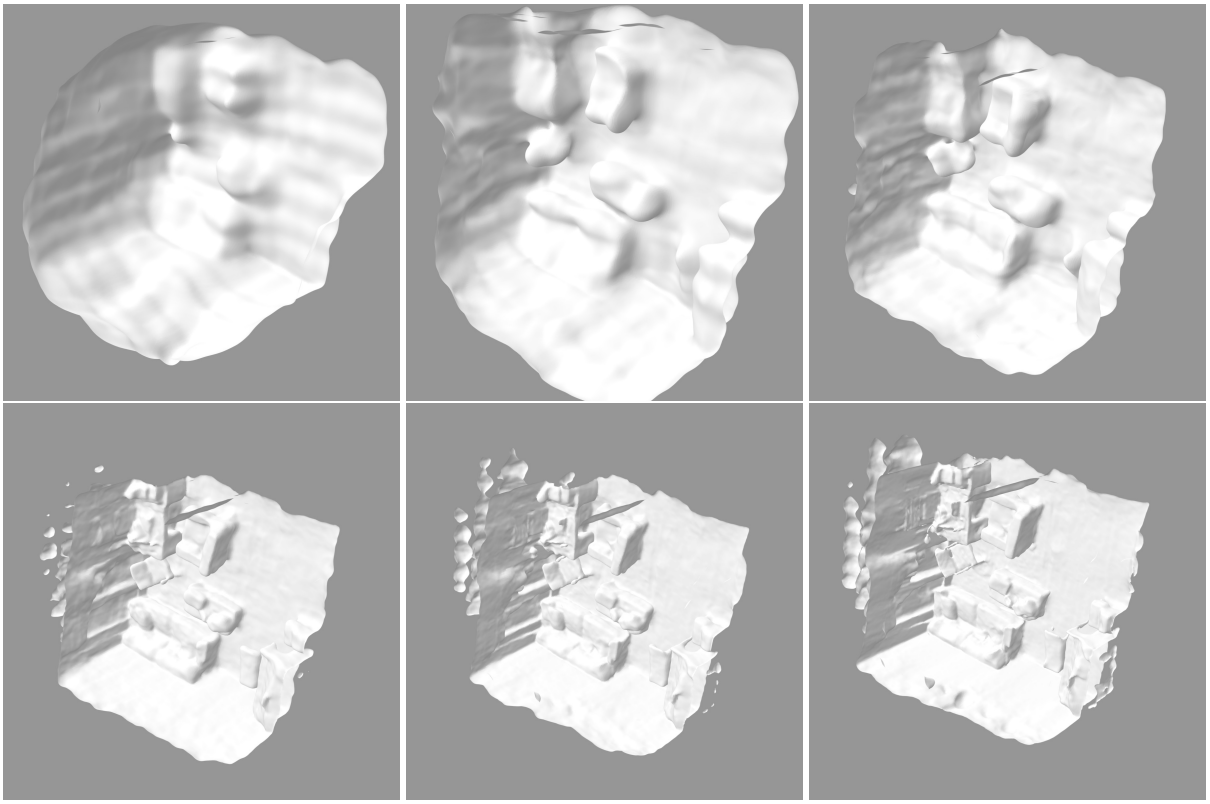

Figure 4. 3D Reconstruction changes on scannet_0050(epoch 0, epoch 1, epoch 2, epoch 10, epoch 20, epoch 49).

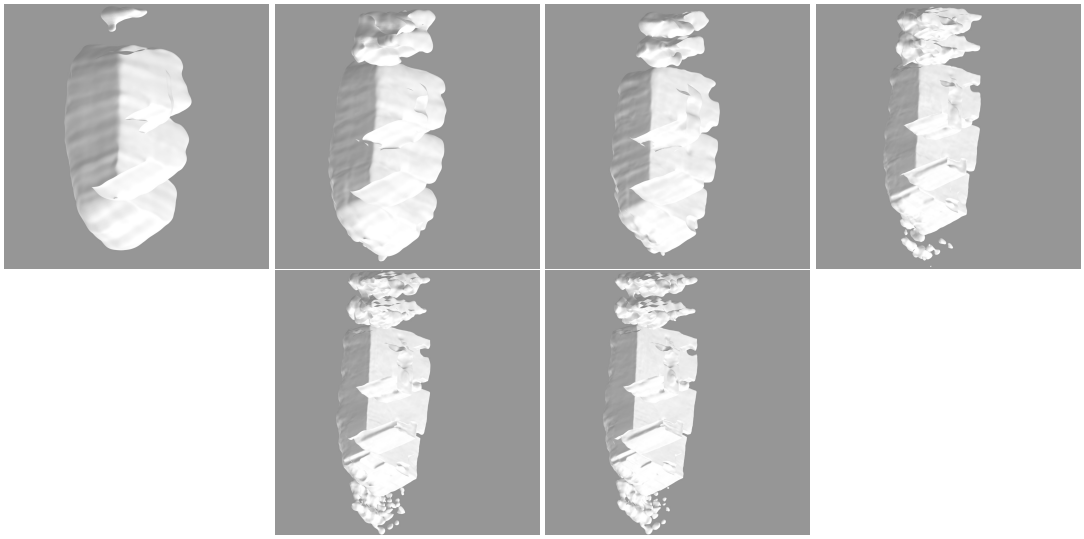

Figure 5. 3D Reconstruction changes on scannet_0084(epoch 0, epoch 1, epoch 2, epoch 10, epoch 20, epoch 49).

As can be seen from Table 1, our results perform better than the baseline in terms of Comp and F-score, but are not as effective as MVS in other aspects.

The COLMAP algorithm based on MVS considers filtering points with inconsistent forms across multiple views during the fusion stage to achieve higher precision, but this leads to a decrease in recall. NeRF performs extremely poorly, mainly because its volume density lacks sufficient constraint on geometry, resulting in poor reconstruction of complex geometrical structures like toilet bowls. NeuS and

Method	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑
COLMAP	**0.047**	0.235	**0.711**	0.441	0.537
NeRF	0.735	0.177	0.131	0.290	0.176
NeuS	0.179	0.208	0.313	0.275	0.291
VolSDF	0.414	0.120	0.321	0.394	0.346
Manhattan	0.072	0.068	0.621	**0.586**	0.602
Ours(min Method)	0.069	0.049	0.631	0.581	0.579
Ours(avg Method)	0.075	**0.046**	0.642	0.583	**0.611**

Table 1. Average 3D reconstruction metrics on four datas.

Figure 6. On the left is the framework, and on the right is the same with added color. It can be observed that the framework is reasonably processed, but the added color is not quite appropriate. A better method is needed to add color.

VolSDF perform slightly better than NeRF due to SDF's excellent constraint on structural surfaces, but they still struggle with complex structures like toilet bowls. Reconstructions based on world assumptions yield better results for such structures, though there is still room for improvement in terms of precision.

5. Conclusion

In this paper, we adopted a 3D reconstruction method based on the Atlanta world assumption. We expressed a vector group by rotating the baseline vector of the horizontal plane multiple times and used its average value as the geometric loss. From the results, some metrics are better than the baseline, but since the dataset is still not very large, we cannot make a final judgment and need to spend time running more datasets.

Future Works. In the Atlanta world representation, there are actually some effective search algorithms [6] that can be used to update the vectors' conditions. In fact, we believe that a more promising direction for 3D reconstruction is based on GPT. Representing the geometric world entirely with triangles and using GPT to predict the next triangle for reconstructing the model is quite efficient. This approach may become mainstream.

6. Code

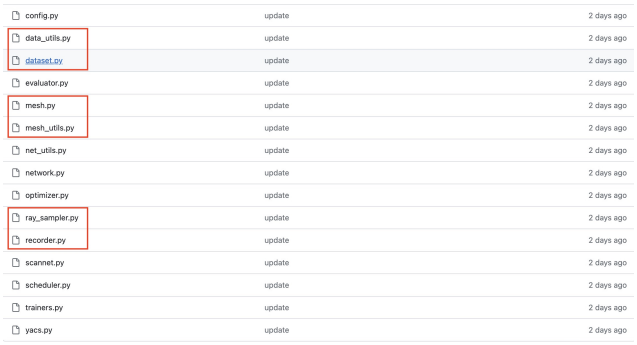

Figure 7. The code has been made publicly available on GitHub. The parts marked with a red circle are the sections I am responsible for.

References

[1] Decai Chen, Haofei Lu, Ingo Feldmann, Oliver Schreer, and Peter Eisert. Dynamic multi-view scene reconstruction using neural implicit surface, 2023. 2

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017. 2

[3] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption, 2022. 1, 2

[4] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis, 2018. 2

[5] Zhentao Huang, Yukun Shi, and Minglun Gong. Visibility-aware pixelwise view selection for multi-view stereo matching, 2023. 1

[6] Yinlong Liu, Guang Chen, and Alois Knoll. Globally optimal vertical direction estimation in atlanta world. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2020. 4

[7] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space, 2019. 1

[8] 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, 2020. 1, 2

[9] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation, 2019. 1

[10] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis, 2021. 2

[11] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video, 2021. 2

[12] Jian Wang, Xiang Long, Guowei Chen, Zewu Wu, Zeyu Chen, and Errui Ding. U-hrnet: Delving into improving semantic representation of high resolution network for dense prediction, 2022. 2

[13] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Stan Birchfield, Kaihao Zhang, Nikolai Smolyanskiy, and Hongdong Li. Deep two-view structure-from-motion revisited, 2021. 2

[14] 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, 2023. 1, 2

[15] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces, 2021. 1, 2

[16] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Domain adaptation for semantic segmentation via class-balanced self-training, 2018. 2