3D Scene Reconstruction with the Atlanta-world Assumption

CV Term Project Report

Jiazheng Li, Zixin Huang

Beijing Institute of Technology

December 14, 2023

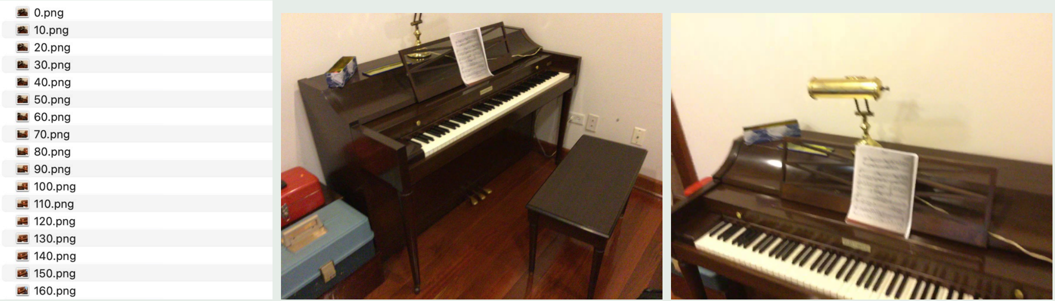
Outline

1. **Introduction**
2. Method
3. Experiments
4. Conclusion
5. Thanks

Introduction

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.



0.png
10.png
20.png
30.png
40.png
50.png
60.png
70.png
80.png
90.png
100.png
110.png
120.png
130.png
140.png
150.png
160.png

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

Introduction

Related Works

Traditional methods typically employ Multi-View Stereo (MVS)[1] 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.

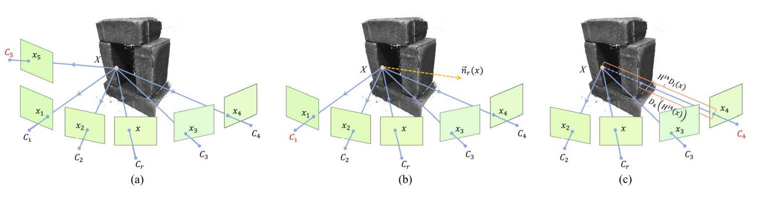


Figure: Multi-View Stere[1].

Introduction

Related Works

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

NeRF[2] 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.

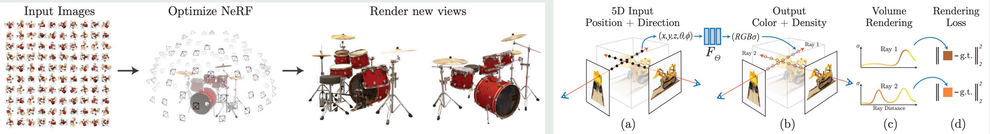


Figure: NeRF[2].

Introduction

Related Works

NeuS[3] and VolSDF[4] 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[5], 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 function has been further improved to achieve better results.

Outline

1. Introduction
2. **Method**
3. Experiments
4. Conclusion
5. Thanks

Method

The Manhattan and Atlanta World Assumption

In artificially defined environments, scenes typically have a structured form (for example, the layout of buildings and many indoor scenes like furniture), which can be represented by a set of parallel and orthogonal planes.
The Manhattan world assumption considers that the world is composed of a single horizontal plane and a single vertical plane, and the normals of these planes can easily describe this world.
In contrast, the Atlanta world assumption posits that artificially defined scenes can be composed of horizontal planes (e.g., the ground) and several vertical planes (e.g., buildings and walls), and then the normals of these planes (referred to as the world frame) can abstractly describe the scene.
In short, the difference between the two world assumptions lies in the number of normal vectors in a certain dimension.

The Manhattan and Atlanta World Assumption

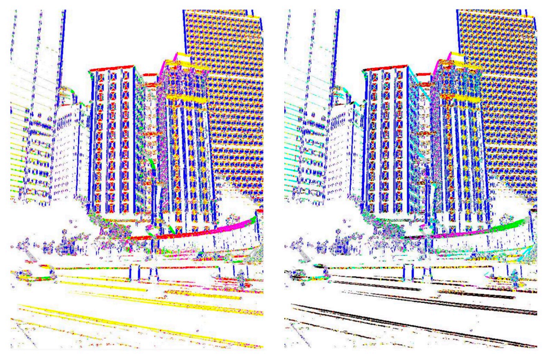


Figure: Left is Manhattan and while right is Atlanta.[6] .

Method

Overview of our method.

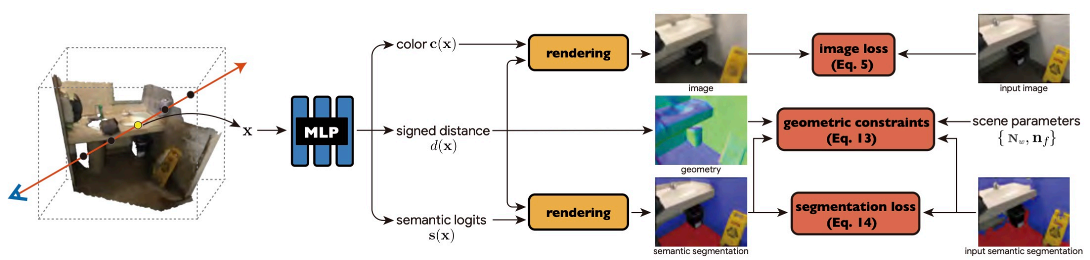

x
MLP
color $\mathbf{c}(\mathbf{x})$
signed distance $d(\mathbf{x})$
semantic logits $\mathbf{s}(\mathbf{x})$
rendering
rendering
image
geometry
semantic segmentation
image loss (Eq. 5)
geometric constraints (Eq. 13)
segmentation loss (Eq. 14)
input image
scene parameters $\{\mathbf{N}_w, \mathbf{n}_f\}$
input semantic segmentation

Figure: Overview of our method. [5]

Method

The implicit neural representation

For an input scene $\mathbf{x}$, three implicit fields are constructed:

color:$\mathbf{c}(\mathbf{x})$

signed distance:$d(\mathbf{x})$

semantic logits:$\mathrm{s}(\mathbf{x})$

We can use MLP to acquire them:

$$(d(\mathbf{x}), \mathbf{z}(\mathbf{x})) = F_d(\mathbf{x})$$
$$\mathbf{c}(\mathbf{x}) = F_{\mathbf{c}}(\mathbf{x}, \mathbf{v}, \mathbf{n}(\mathbf{x}), \mathbf{z}(\mathbf{x}))$$
$$\mathbf{s}(\mathbf{x}) = F_{\mathbf{s}}(\mathbf{x})$$

Here, $\mathbf{z}(\mathbf{x})$ is the geometry feature, $\mathbf{n}(\mathbf{x})$ is computed by the gradient of $d(\mathbf{x})$, and F is implemented as an MLP network.

Method

NeRF Rendering

Using NeRF to render SDF, color, and semantics separately:

$$\sigma(\mathbf{x}) = \begin{cases} \frac{1}{\beta}(1 - \frac{1}{2}\exp(\frac{d(\mathbf{x})}{\beta})) & d(\mathbf{x}) < 0 \\ \frac{1}{2\beta}\exp(-\frac{d(\mathbf{x})}{\beta}) & d(\mathbf{x}) \geq 0 \end{cases}$$

$$\widehat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^{K} T_i(1 - \exp(-\sigma_i\delta_i))\mathbf{c}_i$$

$$\widehat{\mathbf{S}}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))\mathbf{s}_i$$

Here, $\delta_i = \|\mathbf{x}_{i+1} - \mathbf{x}_i\|$ is the distance between adjacent sampling points, $T_i = \exp(-\sum_{j=1}^{i-1}\sigma_j\delta_j)$ is the accumulated transparency, and β is a learnable parameter.

Method

Loss Functions

Based on the preceding preparation and analysis, we have designed the following loss function:
Color loss: $\mathcal{L}_{\text{img}} = \sum_{\mathbf{r}\in\mathcal{R}} \|\widehat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|$.
Eikonal loss: $\mathcal{L}_E = \sum_{\mathbf{y}\in\mathcal{Y}}(\|\nabla_{\mathbf{y}} d(\mathbf{y})\| - 1)^2$.
Depth loss: $\mathcal{L}_d = \sum_{\mathbf{r}\in\mathcal{D}} \|\widehat{\mathbf{D}}(\mathbf{r}) - \mathbf{D}(\mathbf{r})\|$.
Semantic Segmentation loss: $\mathcal{L}_s = -\sum_{\mathbf{r}\in\mathcal{R}}\sum_{k\in\{f,w,b\}} p_k(\mathbf{r})\log\hat{p_k}(\mathbf{r})$.
Geometry loss: $\mathcal{L}_{\text{joint}} = \sum_{\mathbf{r}\in\mathcal{F}}\hat{p_f}(\mathbf{r})\mathcal{L}_f(\mathbf{r}) + \sum_{\mathbf{r}\in\mathcal{W}}\hat{p_w}(\mathbf{r})\mathcal{L}_w(\mathbf{r})$.
Here, $\mathcal{L}_f(\mathbf{r}) = \|1-\mathbf{n}(\mathbf{x}_{\mathbf{r}})\cdot\mathbf{n}_{\mathbf{f}}\|$, $\mathbf{n}_{\mathbf{f}} = (0,0,1)$, $\mathcal{L}_w(\mathbf{r}) = \min_{i\in\{-1,0,1\}} f_{\mathbf{n}_w\in\mathcal{N}_w}\|i - \mathbf{n}(\mathbf{x}_{\mathbf{r}})\cdot\mathbf{n}_{\mathbf{w}}\|$,
$\mathbf{N}_w$ is a set of horizontal baseline vectors of a certain size, and f is $\min/\text{avg}$.
To get a certain f, we will use experiments to compare them. $\mathbf{N}_w$ is composed of a baseline vector rotated by several angles.

Method

Network architecture

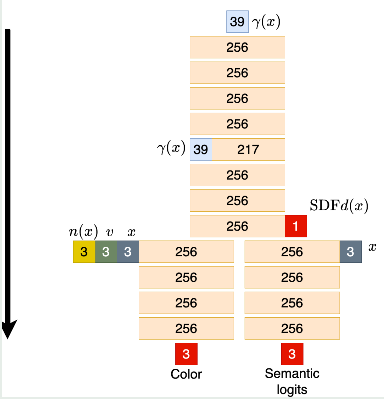


Figure: Network architecture.

Outline

1. Introduction
2. Method
3. **Experiments**
4. Conclusion
5. Thanks

Experiment

Environment and Dataset

The code and supplementary materials are available at `https://github.com/foreverlasting1202/cv-term-pj`.

Experiment

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[7], we consider the **F-score** as the overall metric.

Experiment

Baselines

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

Experiment

Result

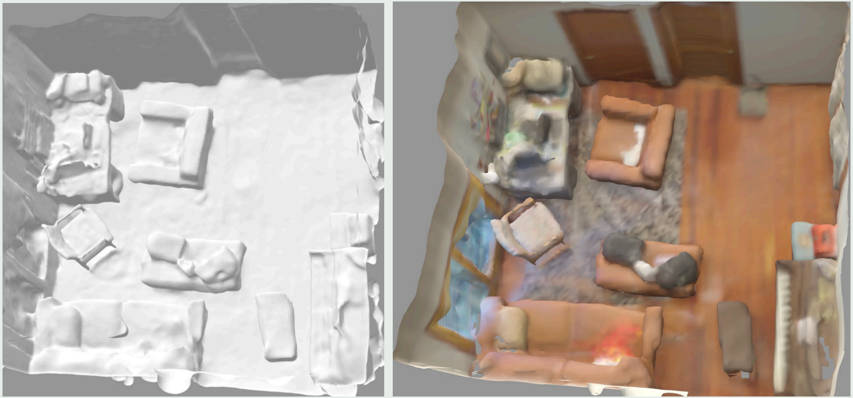


Figure: Result

Experiment

Result

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: Average 3D reconstruction metrics on four datas.

Outline

1. Introduction

2. Method

3. Experiments

4. **Conclusion**

5. Thanks

Conclusion

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.

Conclusion

Future Works

In the Atlanta world representation, there are actually some effective search algorithms[9] 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.

Outline

1. Introduction
2. Method
3. Experiments
4. Conclusion
5. **Thanks**

Thanks

Thank you!

References I

[1] HUANG Z, SHI Y, GONG M. Visibility-aware pixelwise view selection for multi-view stereo matching[A]. 2023. arXiv: 2302.07182.

[2] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[A]. 2020. arXiv: 2003.08934.

[3] WANG P, LIU L, LIU Y, et al. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction[A]. 2023. arXiv: 2106.10689.

[4] YARIV L, GU J, KASTEN Y, et al. Volume rendering of neural implicit surfaces[A]. 2021. arXiv: 2106.12052.

[5] GUO H, PENG S, LIN H, et al. Neural 3d scene reconstruction with the manhattan-world assumption[A]. 2022. arXiv: 2205.02836.

References II

[6] SCHINDLER G, DELLAERT F. Atlanta world: an expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments[C/OL]//Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.: volume 1. 2004: I-I. DOI: 10.1109/CVPR.2004.1315033.

[7] SUN J, XIE Y, CHEN L, et al. Neuralrecon: Real-time coherent 3d reconstruction from monocular video[A]. 2021. arXiv: 2104.00681.

[8] WANG J, ZHONG Y, DAI Y, et al. Deep two-view structure-from-motion revisited[A]. 2021. arXiv: 2104.00556.

[9] LIU Y, CHEN G, KNOLL A. Globally optimal vertical direction estimation in atlanta world [J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020: 1–1. http://dx.doi.org/10.1109/TPAMI.2020.3027047. DOI: 10.1109/tpami.2020.3027047.

[10] FU Q, XU Q, ONG Y S, et al. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction[A]. 2022. arXiv: 2205.15848.

References III

[11] SIDDIQUI Y, ALLIEGRO A, ARTEMOV A, et al. Meshgpt: Generating triangle meshes with decoder-only transformers[A]. 2023. arXiv: 2311.15475.

[12] WANG Z, ZHU X, ZHANG T, et al. 3d face reconstruction with the geometric guidance of facial part segmentation[A]. 2023. arXiv: 2312.00311.