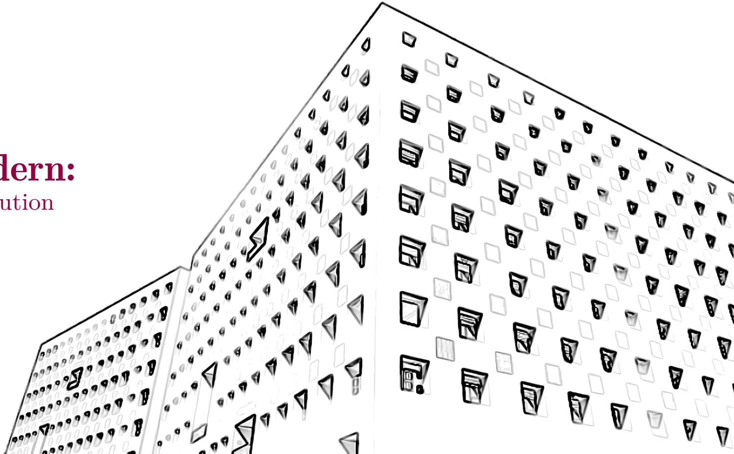
From Traditonal to Modern:

A Comprehensive Review of the Evolution

in 3D Reconstruction

Enyun Xuan

2024 年 1 月 29 日



近期工作

1.18-1.29

在 1.18 组会汇报工作的基础上，进一步细化了文献调研工作。从深度和广度上扩展我的工作。

- 在广度上，从上次调研了十来篇论文，到现在一共浏览了 52 篇相关论文。这些论文是根据三维重建的发展脉络来展开阅读，从传统的方法到 10 年代兴起的基于深度学习的方法，再到今年火热的 NeRF 和最新的 3DGS。
- 在深度上，上次汇报中仅总结了前人的工作内容，且知识体系比较松散，在此基础上，通过进一步的文献调研，逐渐形成了自己对于三维重建中关键技术的理解。

Table of Contents

Traditonal Motheds

1980s-

在传统的三维重建工作中，研究人员通常需要借助各类传感器和无人机等复杂工具的协助，并且需要到实地去勘查和测量。可以根据是否通过传感器主动地向物体照射信号来区分这些重建方法。

目的在于估计三维物体的深度信息。

- **主动式：**
  - structure light[Geng, 2011]
  - Time-of-Flight(ToF)
- **被动式：**
  - stereo vision[Hamzah and Ibrahim, 2016]

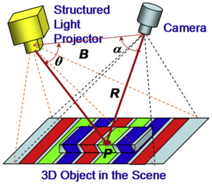


Illustration of structured light.

Figure: structure light

Traditonal Motheds

1980s-

在传统的三维重建工作中，研究人员通常需要借助各类传感器和无人机等复杂工具的协助，并且需要到实地去勘查和测量。可以根据是否通过传感器主动地向物体照射信号来区分这些重建方法。

目的在于估计三维物体的深度信息。

- **主动式：**
  - structure light[Geng, 2011]
  - Time-of-Flight(ToF)
- **被动式：**
  - stereo vision[Hamzah and Ibrahim, 2016]

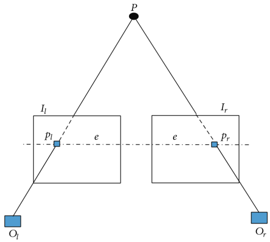


FIGURE 3: Epipolar geometry. The 3D image of target scene at point P.

Figure: stereo vision

The Limitation

传统方法的最显著的局限性在于“不智能”，即在其 pipeline 中需要消耗的人力物力太大。

- **手动操作：** 例如在被动式的方法中，需要人为的对多张图片进行对齐（alignment）
- **精度问题和复杂性：** 实现相对麻烦，精度普遍不高，受到传感器的系统误差的影响，还会受到恶劣天气等影响。
- **计算资源：** 相比于现代的方法，早期技术对计算资源的要求太高。

Table of Contents

2 Deep Learning Based Motheds

Structure From Motion(2012)

2 Deep Learning Based Motheds

SFM[Westoby et al., 2012] 自动求解相机位姿（pose）和生成稀疏点云（sparse point cloud）的低成本高效的方法。

提取三维物体表面的关键点（keypoints）。在 COLMAP[Schonberger and Frahm, 2016] 中的开源代码被许多后续工作相继使用。

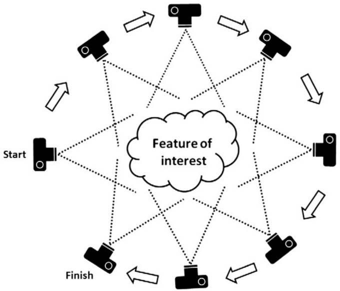


Fig. 1. Structure-from-Motion (SfM). Instead of a single stereo pair, the SfM technique requires multiple, overlapping photographs as input to feature extraction and 3-D reconstruction algorithms.

Multi-view Stereo

2 Deep Learning Based Motheds

MVS[Seitz et al., 2006]

- **目标：** 从不同视角的图片重构物体几何
- **方法：** 使用三角测量原理来手动测量绘制。

目前大部分的方法的输入都是物体的多张图片，然后将图片进行对齐。[Furukawa and Hernández, 2015] 后续的研究有些使用其生成稠密点云（dense point cloud）

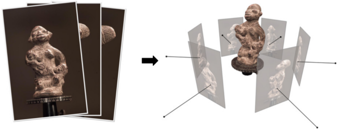


Figure 1.1: Image-based 3D reconstruction. Given a set of photographs (left), the goal of image-based 3D reconstruction algorithms is to estimate the most likely 3D shape that explains those photographs (right).

可以根据输入图片的数量来进行分类。

- single image
- multi images

可以根据对物体的表示方式来分类

- **depth map** 包含物体的深度信息
- voxel grid 占用内存较多
- point cloud 可以作为 mesh 的输入
- mesh 包含了临近信息

Section 2.1

Single Image

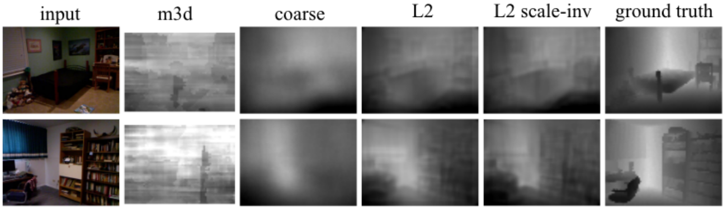
Single Image

2 Deep Learning Based Motheds

必要性：

- 有些时候需要对仅有的一张图片来重建
- 人眼可以通过先验知识，对单张图片识别出深度信息

Eigen[Eigen et al., 2014] 等人首次使用深度学习来预测深度图，该方法通过优化两个网络（coarse and refine）来预测，其中一个网络用于预测全局信息，一个用于局部优化。



Single Image

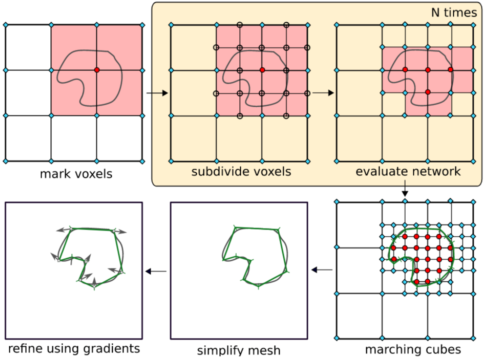
2 Deep Learning Based Motheds

PointOutNet[Fan et al., 2016] 从单张图片预测出一个物体的完整点云，并且强调了该类任务中的不确定性（ambiguity）。
AtlasNet[Groueix et al., 2018]mesh 是通过在 point cloud 的基础上，加入临近信息来得到的。

Implicit Representation

2 Deep Learning Based Motheds

Occupancy Networks[Mescheder et al., 2019] 将三维物体表示成占用率，是一种隐式表达。



贡献：

- 缓解内存问题
- 提高重建质量

Problems

2 Deep Learning Based Motheds

- inherent ambiguity
- self-occlusion in single view input

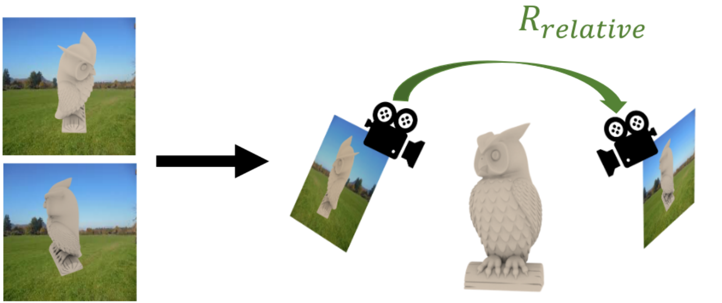
Section 2.2

Multi Images

在多张图片的重建任务中，一个重要问题是：Alignment。

存在的监督方式：

- 3D GT 大多数时候难以获得，且不好优化
- 2D GT
- depth
- course point
- no surpervision [El Banani et al., 2020]
- self surpervision

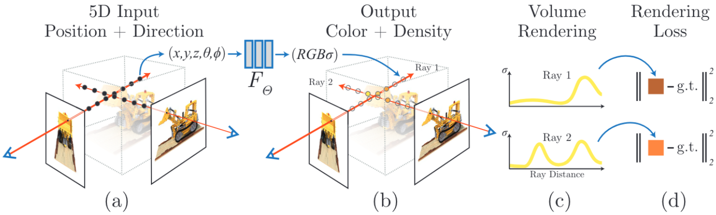


Section 2.3

NeRF(2020)

NeRF[Mildenhall et al., 2020] 将场景表示成神经辐射场（Neural Radiance Fields），是一种隐式表达。

- 隐式表达场景，内存压力小
- 合成质量高
- 密度（density）和颜色（color）
- 使用体渲染（volume rendering）



Improvement

2 Deep Learning Based Motheds

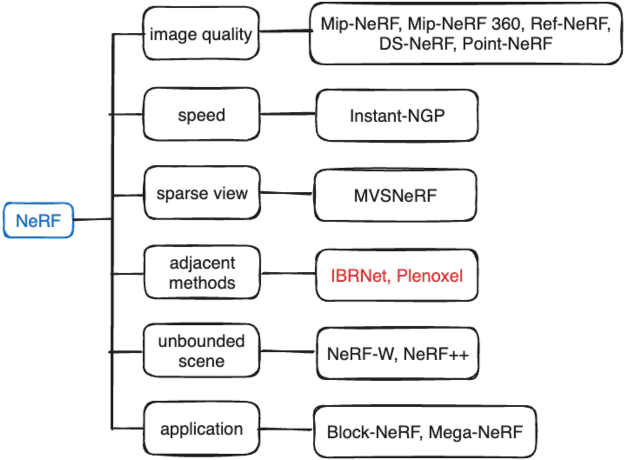


Image Quality

2 Deep Learning Based Motheds

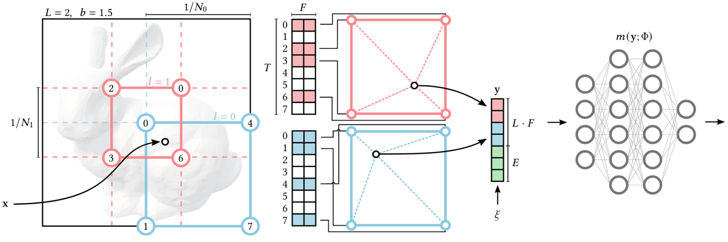
- Mip-NeRF[Barron et al., 2021]，IPE
- Mip-NeRF 360[Barron et al., 2022]，Mip-NeRF + unbounded scene
- DS-NeRF[Deng et al., 2022]，用 SFM 生成点云来监督 depth
- Point-NeRF[Xu et al., 2022]，预提取 depth map 和 features 来建立点云

改进思路：**数学原理、增强监督、特征增强、已存在的技术**

Speed

2 Deep Learning Based Motheds

- Instant-NGP[Müller et al., 2022]　多分辨率编码、哈希函数、简化网络、球谐函数



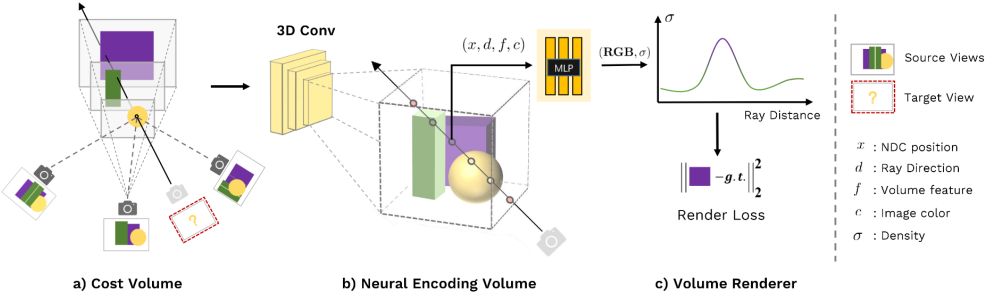
(1) Hashing of voxel vertices　(2) Lookup　(3) Linear interpolation　(4) Concatenation　(5) Neural network

改进思路：**牺牲内存换取速度，插值法**

Sparse View

2 Deep Learning Based Motheds

- MVSNeRF[Chen et al., 2021] 结合了 MVSNet[Yao et al., 2018] 中提出的 cost volume。

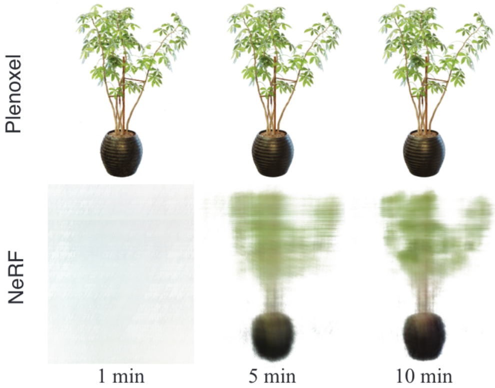


改进思路：**参考前人工作，特征增强**

Adjacent Methods

2 Deep Learning Based Motheds

- IBRNet[Wang et al., 2021]　相邻图片插值生成中间图片
- Plenoxel[Yu et al., 2021]　不使用神经网络，voxel 表示场景，球谐函数表示颜色



Unbounded Scene

2 Deep Learning Based Motheds

这类场景需要分离物体与背景

- NeRF-W[Martin-Brualla et al., 2021] 解决灯光影响场景的问题
- NeRF++[Zhang et al., 2020] 分析并解决 NeRF 中的几何歧义性

(a) Photos (b) Renderings

将 NeRF 应用到实际中

- Block-NeRF[Tancik et al., 2022]　结合了 Mip-NeRF 和 NeRF-W 来对大城市场景重建
- Mega-NeRF[Turki et al., 2022]　结合了 NeRF++ 和 NeRF-W 来对无人机航拍的图片重建

解决思路：**将相应的工作应用到相应的场景中**

Conclusion of NeRF

Contributions：

- Plenoxel 说明了 NeRF 的贡献不在于 MLP，而在于使用颜色和密度来合成图片；
- 体渲染：可微（differentiable）的渲染方式。

Limitation：

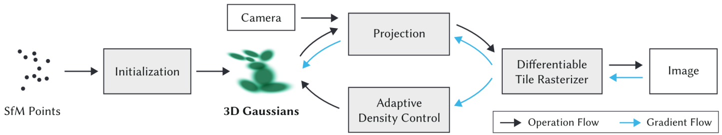
- 速度慢
- 训练网络规模大
- 视觉质量的提升代价大

Table of Contents
3 3DGS(2023)

Contribution of 3DGS

3DGS[Kerbl et al., 2023] 是一种显式的表达方式，在速度和质量上都达到了 SOTA。

- 基于点云（3D Gaussian）
- 球谐函数表示颜色
- 速度快、质量高
- 不需要神经网络
- 光栅化渲染



Application

将 3DGS 应用到实际中，并评测

- SLAM　确定全局地图信息，定位
- Dynamic Scene Modeling　加入时间维度，4D
- Autonomous Driving　使用 3DGS 重建场景

Table of Contents

4 Discussion and Conclusion

一项新的研究灵感来源可以是:

- 前人工作 camP[Park et al., 2023]
- 其他领域的工作
- 数学方法
- 发现本质

本文从时间上的发展脉络，总结了在三维重建领域的关键技术及其发展。从单纯的摄影测量学到与计算机视觉、深度学习等领域相结合，三维重建正在变得越来越“智能化”。

目前 3DGS 作为新视图合成中的 SOTA 方法，会是未来几年内的热门研究，然而在 3 年前兴起的 NeRF 将会有可能被其取代。

Thank You

References

Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., and Srinivasan, P. P. (2021).
Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields Supplemental Material.

Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., and Hedman, P. (2022).
Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields.
In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5460–5469. IEEE.

Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., and Su, H. (2021).
MVSNeRF: Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo.
In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14104–14113. IEEE.

Deng, K., Liu, A., Zhu, J.-Y., and Ramanan, D. (2022).
Depth-supervised NeRF: Fewer Views and Faster Training for Free.
In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12872–12881. IEEE.

Eigen, D., Puhrsch, C., and Fergus, R. (2014).
Depth Map Prediction from a Single Image using a Multi-Scale Deep Network.

El Banani, M., Corso, J. J., and Fouhey, D. F. (2020).
Novel Object Viewpoint Estimation Through Reconstruction Alignment.
In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3110–3119. IEEE.

Fan, H., Su, H., and Guibas, L. (2016).
A Point Set Generation Network for 3D Object Reconstruction from a Single Image.

References

Furukawa, Y. and Hernández, C. (2015).
Multi-View Stereo: A Tutorial.
Number 9,1/2 in Foundation and Trends in Computer Graphics and Vision. Now.

Geng, J. (2011).
Structured-light 3D surface imaging: A tutorial.
3(2):128.

Groueix, T., Fisher, M., Kim, V. G., Russell, B. C., and Aubry, M. (2018).
AtlasNet: A Papier-M\âch\'e Approach to Learning 3D Surface Generation.

Hamzah, R. A. and Ibrahim, H. (2016).
Literature Survey on Stereo Vision Disparity Map Algorithms.
2016:1–23.

Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. (2023).
3D Gaussian Splatting for Real-Time Radiance Field Rendering.

Martin-Brualla, R., Radwan, N., Sajjadi, M. S. M., Barron, J. T., Dosovitskiy, A., and Duckworth, D. (2021).
NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections.
In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7206–7215. IEEE.

References

Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., and Geiger, A. (2019).
Occupancy Networks: Learning 3D Reconstruction in Function Space.
In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4455–4465. IEEE.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2020).
NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis.

Müller, T., Evans, A., Schied, C., and Keller, A. (2022).
Instant Neural Graphics Primitives with a Multiresolution Hash Encoding.
41(4):1–15.

Park, K., Henzler, P., Mildenhall, B., Barron, J. T., and Martin-Brualla, R. (2023).
CamP: Camera Preconditioning for Neural Radiance Fields.

Schonberger, J. L. and Frahm, J.-M. (2016).
Structure-from-Motion Revisited.
In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113. IEEE.

Seitz, S., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. (2006).
A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms.
In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06)*, volume 1, pages 519–528. IEEE.

References

Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B. P., Srinivasan, P., Barron, J. T., and Kretzschmar, H. (2022).
Block-NeRF: Scalable Large Scene Neural View Synthesis.
In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8238–8248. IEEE.

Turki, H., Ramanan, D., and Satyanarayanan, M. (2022).
Mega-NeRF: Scalable Construction of Large-Scale NeRFs for Virtual Fly- Throughs.
In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12912–12921. IEEE.

Wang, Q., Wang, Z., Genova, K., Srinivasan, P., Zhou, H., Barron, J. T., Martin-Brualla, R., Snavely, N., and Funkhouser, T. (2021).
IBRNet: Learning Multi-View Image-Based Rendering.
In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4688–4697. IEEE.

Westoby, M., Brasington, J., Glasser, N., Hambrey, M., and Reynolds, J. (2012).
'Structure-from-Motion' photogrammetry: A low-cost, effective tool for geoscience applications.
179:300–314.

Xu, Q., Xu, Z., Philip, J., Bi, S., Shu, Z., Sunkavalli, K., and Neumann, U. (2022).
Point-NeRF: Point-based Neural Radiance Fields.
In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5428–5438. IEEE.

References

Yao, Y., Luo, Z., Li, S., Fang, T., and Quan, L. (2018).
MVSNet: Depth Inference for Unstructured Multi-view Stereo.
In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision –ECCV 2018*, volume 11212, pages 785–801. Springer International Publishing.

Yu, A., Fridovich-Keil, S., Tancik, M., Chen, Q., Recht, B., and Kanazawa, A. (2021).
Plenoxels: Radiance Fields without Neural Networks.

Zhang, K., Riegler, G., Snavely, N., and Koltun, V. (2020).
NeRF++: Analyzing and Improving Neural Radiance Fields.