3D Scene Representation Learning

Martin Oswald

Computer Vision Group, University of Amsterdam

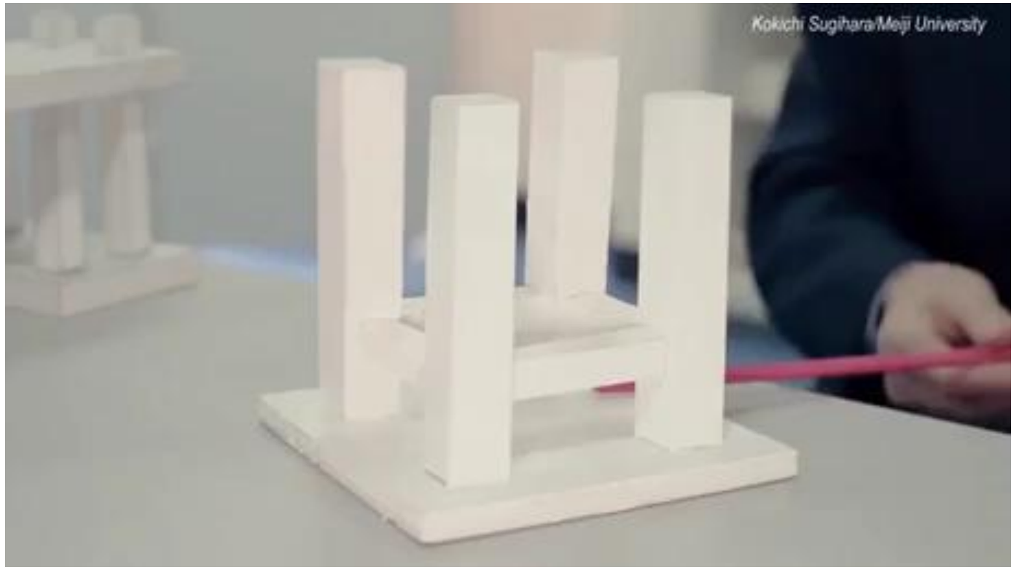
University of Amsterdam

ATLAS LAB.

Motivation: 3D reconstruction is hard!

Motivation: 3D reconstruction is hard!

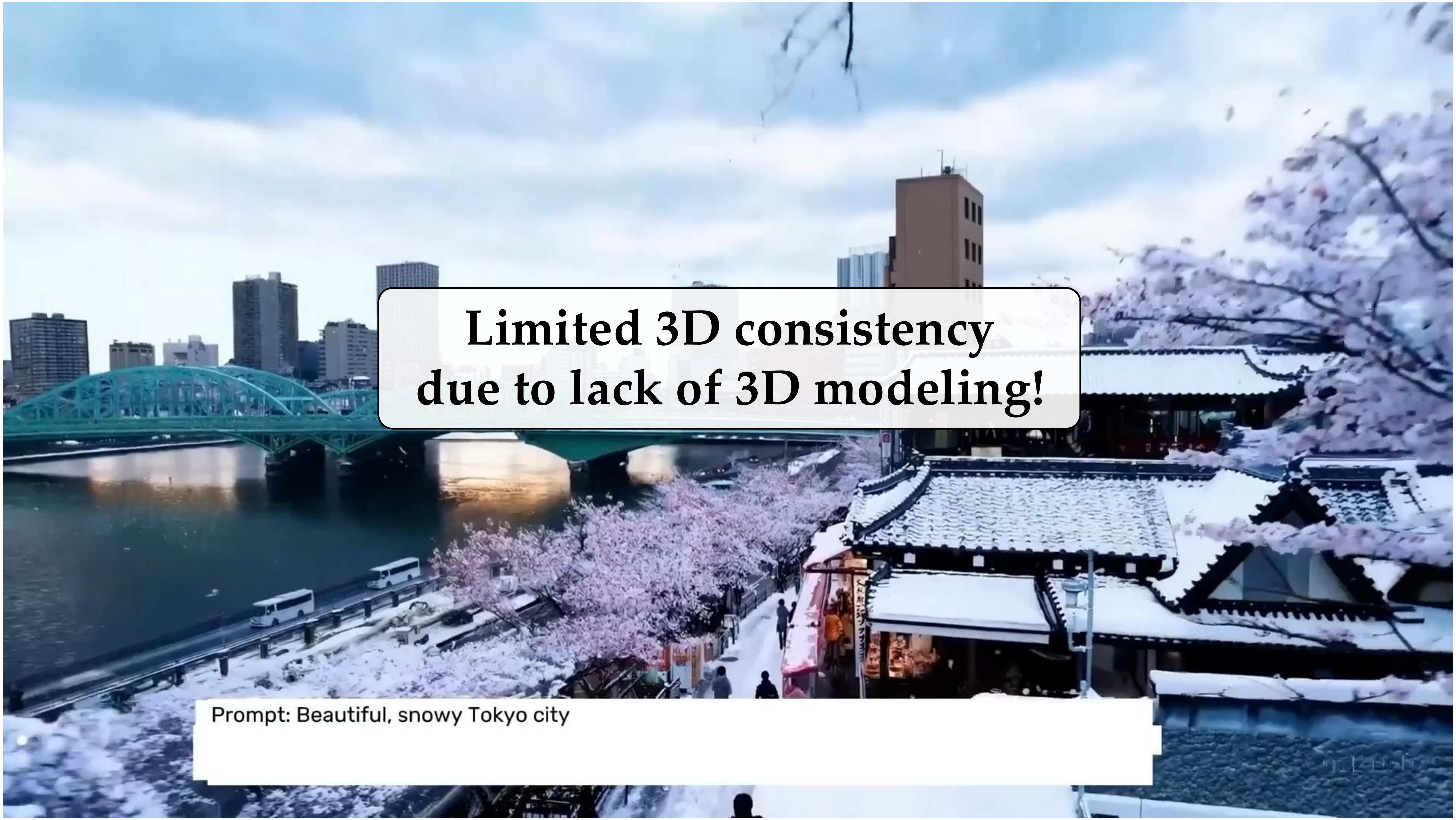
Motivation: 3D reconstruction is hard!



Video Generation: Sora

[OpenAI, 2024]

Video Generation: Sora

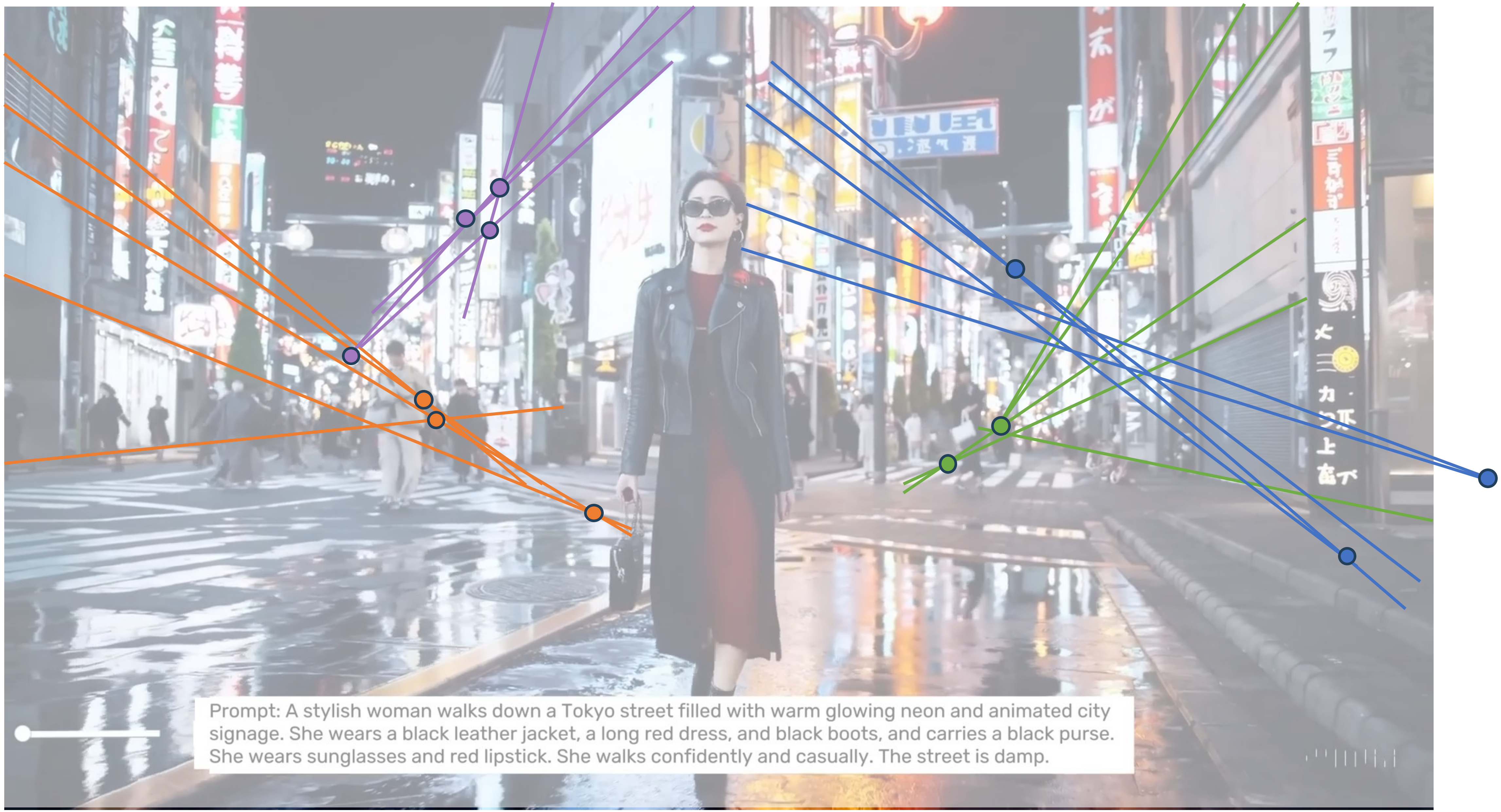


Limited 3D consistency due to lack of 3D modeling!

Vanishing Points!?

Video Generation: Sora

[OpenAI, 2024]



Dall-E3

Man looking through telescope

Woman on a surfboard

[OpenAI, 2024]

“Sora is also a Physics Engine!

“Photorealistic closeup video of two pirate ships battling each other as they sail inside a cup of coffee”

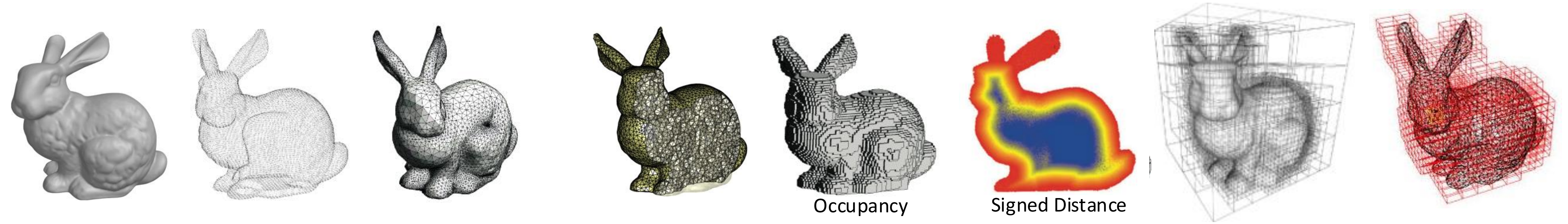
“Sora is also a Physics Engine!

[OpenAI, 2024]

Ski jumping man

Scene Representations

[https://arxiv.org/pdf/1803.03352.pdf]

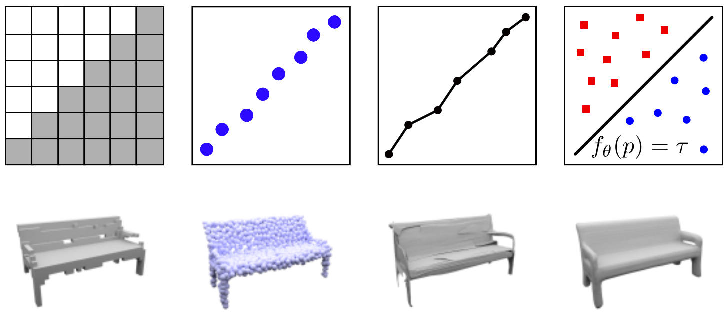


Occupancy | Signed Distance

Spline/NURBS | Point Cloud | Surface Mesh | Tetrahedral Mesh | Voxel Grid | Voxel Octree | Voxel Hashing

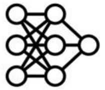
explicit (topology change=>hard) | **implicit (topology change=>simple)**

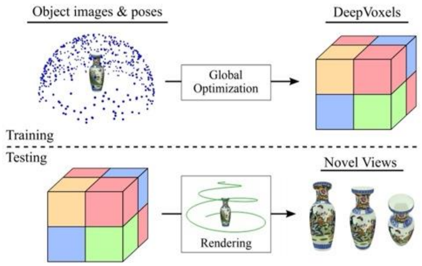
Classical



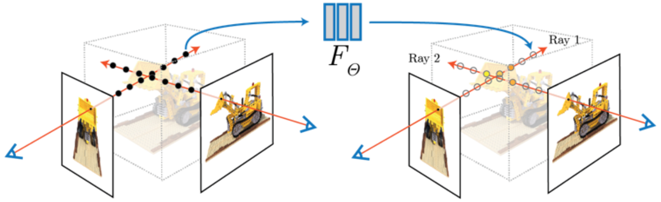
Voxel Grid | Point cloud | Mesh | (Neural) Classifier

Learned / Deep Representations:
OccNet [https://arxiv.org/pdf/1812.03828.pdf]
DeepSDF [https://arxiv.org/pdf/1901.05103.pdf]
IM-Net [https://arxiv.org/pdf/1812.02822.pdf]

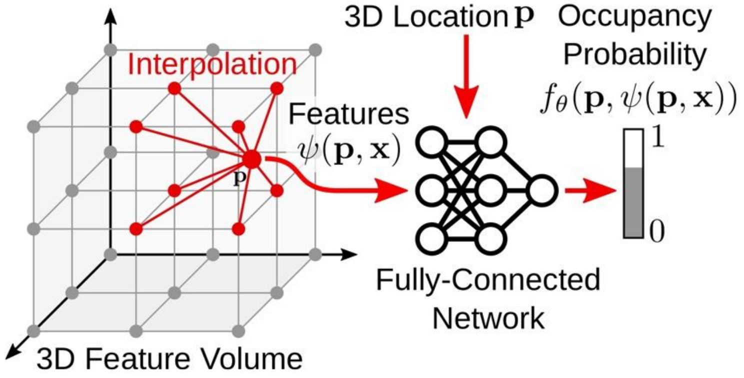




[DeepVoxels CVPR 2019]



[NeRF - Neural Radiance Fields, ECCV 2020]



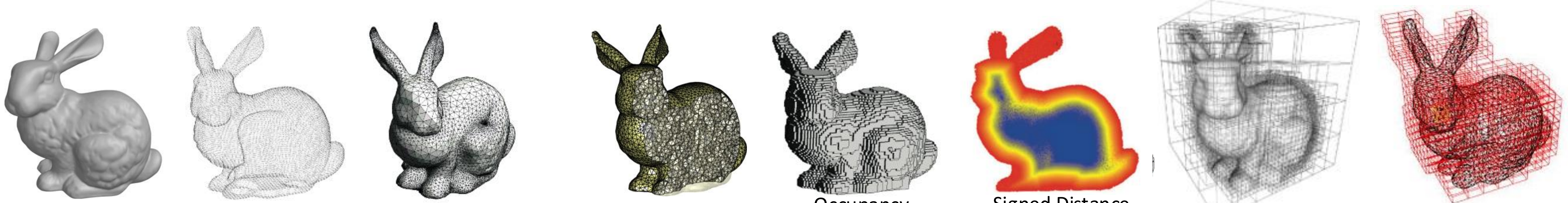
[Peng et al., Convolutional Occupancy Networks, ECCV 2020]

Learned

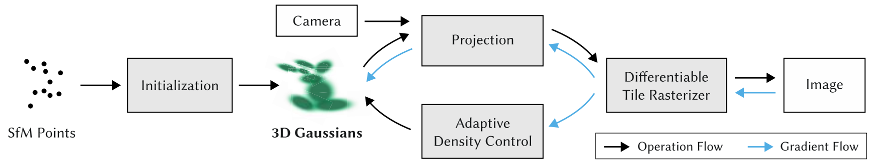
Neural implicit

Scene Representations

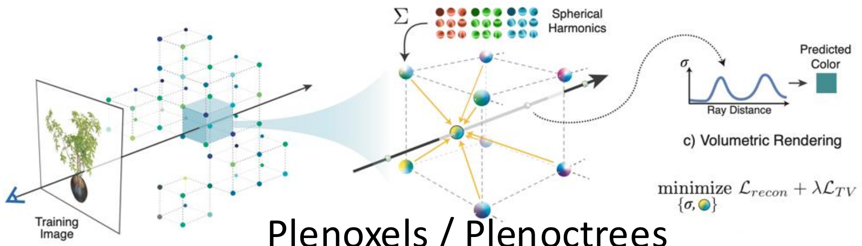
[https://arxiv.org/pdf/1803.03352.pdf]



Spline/NURBS | Point Cloud | Surface Mesh | Tetrahedral Mesh | Voxel Grid | Voxel Octree | Voxel Hashing



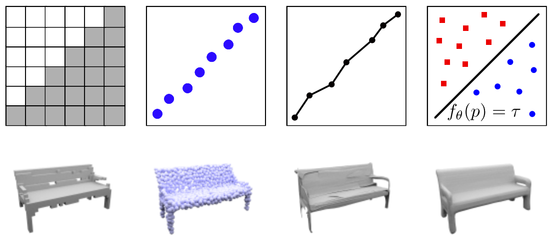
3D Gaussian Splatting



Plenoxels / Plenoctrees

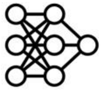
explicit | **implicit**

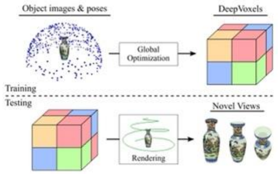
Classical / Non-Neural



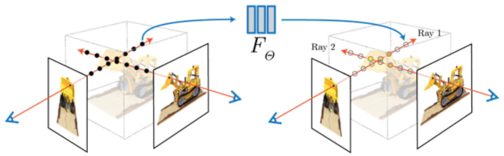
Voxel Grid | Point cloud | Mesh | (Neural) Classifier

Learned / Deep Representations:
OccNet [https://arxiv.org/pdf/1812.03828.pdf]
DeepSDF [https://arxiv.org/pdf/1901.05103.pdf]
IM-Net [https://arxiv.org/pdf/1812.02822.pdf]

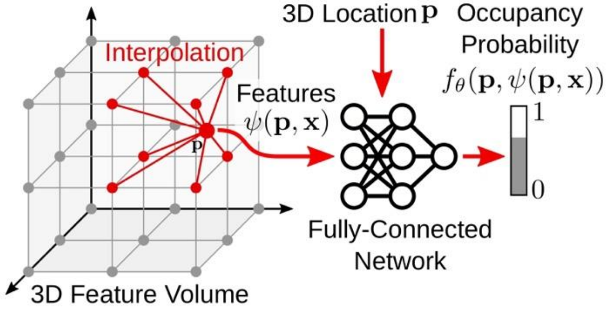




[DeepVoxels CVPR 2019]



[NeRF - Neural Radiance Fields, ECCV 2020]



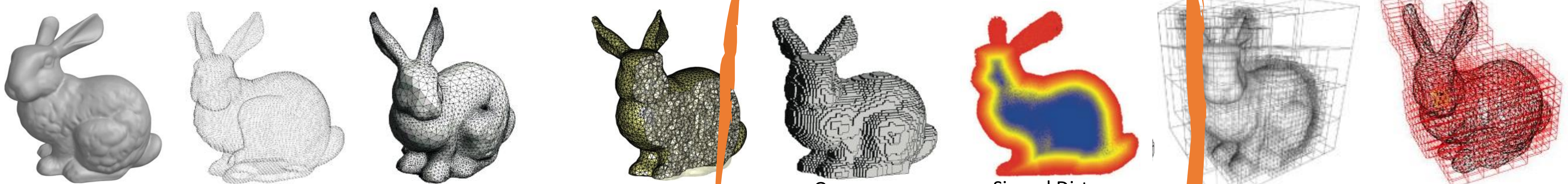
[Peng et al., Convolutional Occupancy Networks, ECCV 2020]

Neural

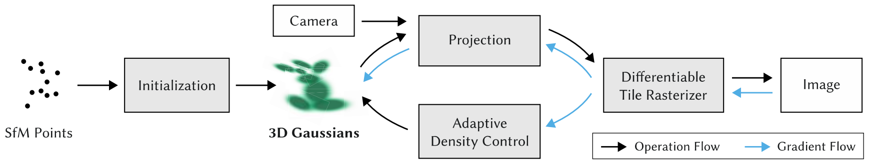
Neural implicit

Scene Representations

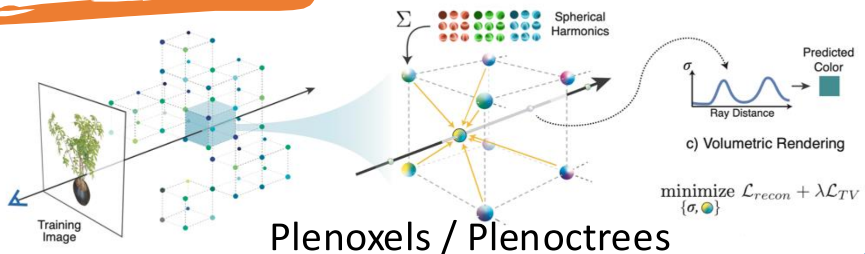
[https://arxiv.org/pdf/1803.03352.pdf]



Spline/NURBS | Point Cloud | Surface Mesh | Tetrahedral Mesh | Voxel Grid | Voxel Octree | Voxel Hashing



3D Gaussian Splatting

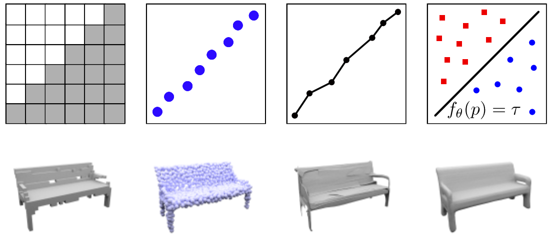


Plenoxels / Plenoctrees

explicit

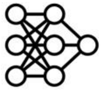
implicit

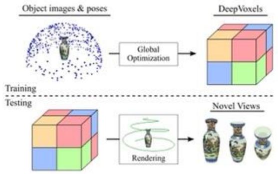
Classical / Non-Neural



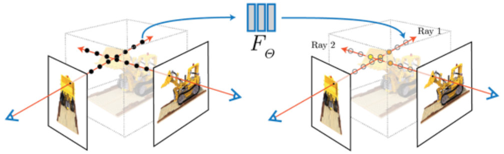
Voxel Grid | Point cloud | Mesh | (Neural) Classifier

Learned / Deep Representations:
OccNet [https://arxiv.org/pdf/1812.03828.pdf]
DeepSDF [https://arxiv.org/pdf/1901.05103.pdf]
IM-Net [https://arxiv.org/pdf/1812.02822.pdf]

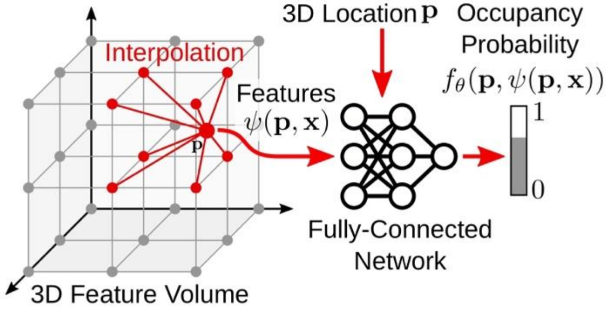




[DeepVoxels CVPR 2019]



[NeRF - Neural Radiance Fields, ECCV 2020]



[Peng et al., Convolutional Occupancy Networks, ECCV 2020]

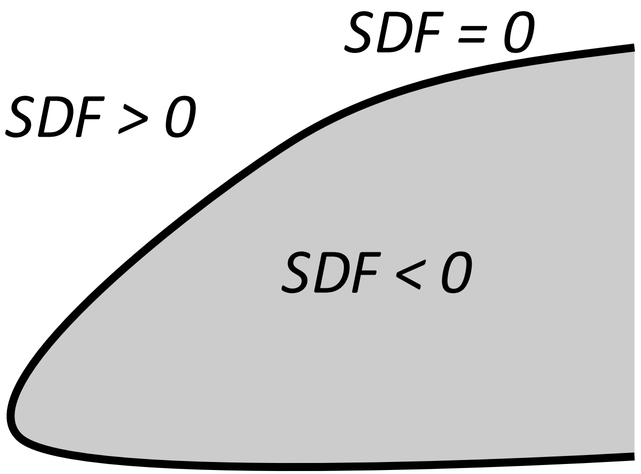
Neural

Neural implicit

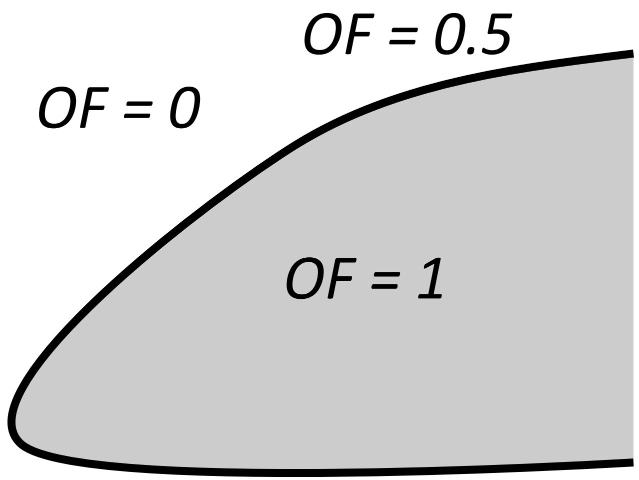
Implicit Volumetric Representation

- *Voxel grid*: sample a volume containing the surface of interest uniformly
- Label each grid point as lying *inside* or *outside* the surface

Signed distance function



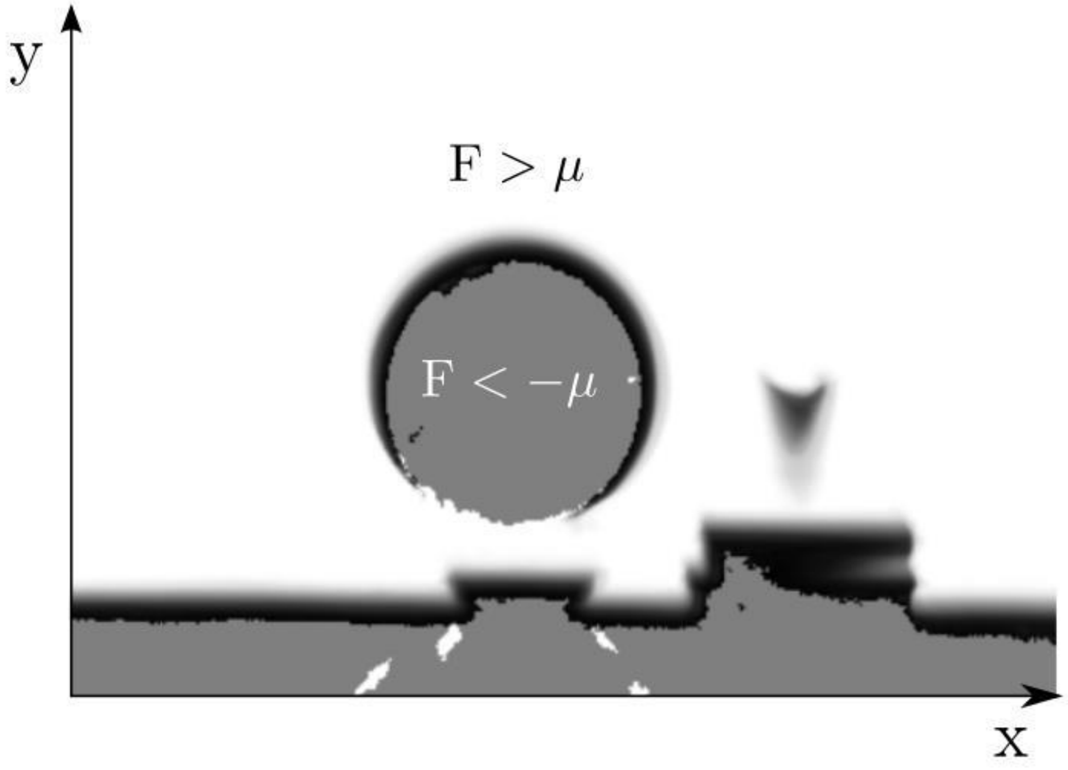
Occupancy function

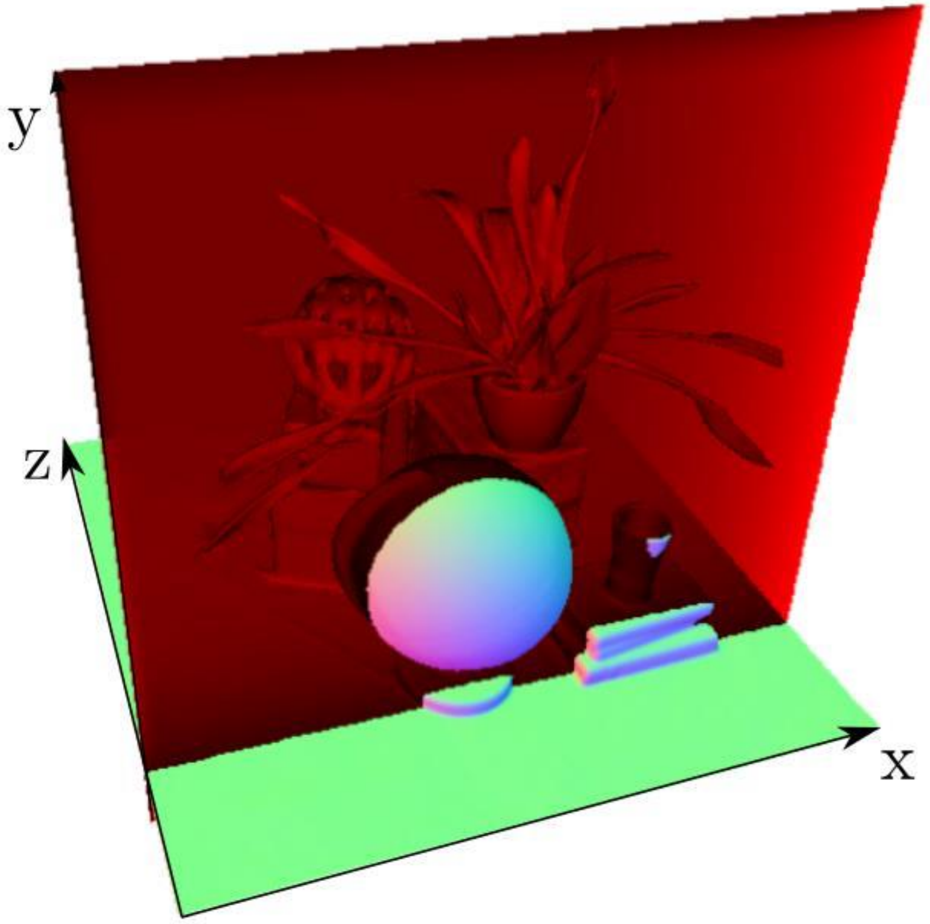


- The modeled surface is represented as an *isosurface* (e.g. SDF = 0 or OF = 0.5) of the labeling (implicit) function

- Advantages: simple handling of topological changes, watertight surfaces, no self-occlusions
  Disadvantages: Large memory requirement, bad scalability to large scenes (cubic growth)

Represent Scenes with TSDFs

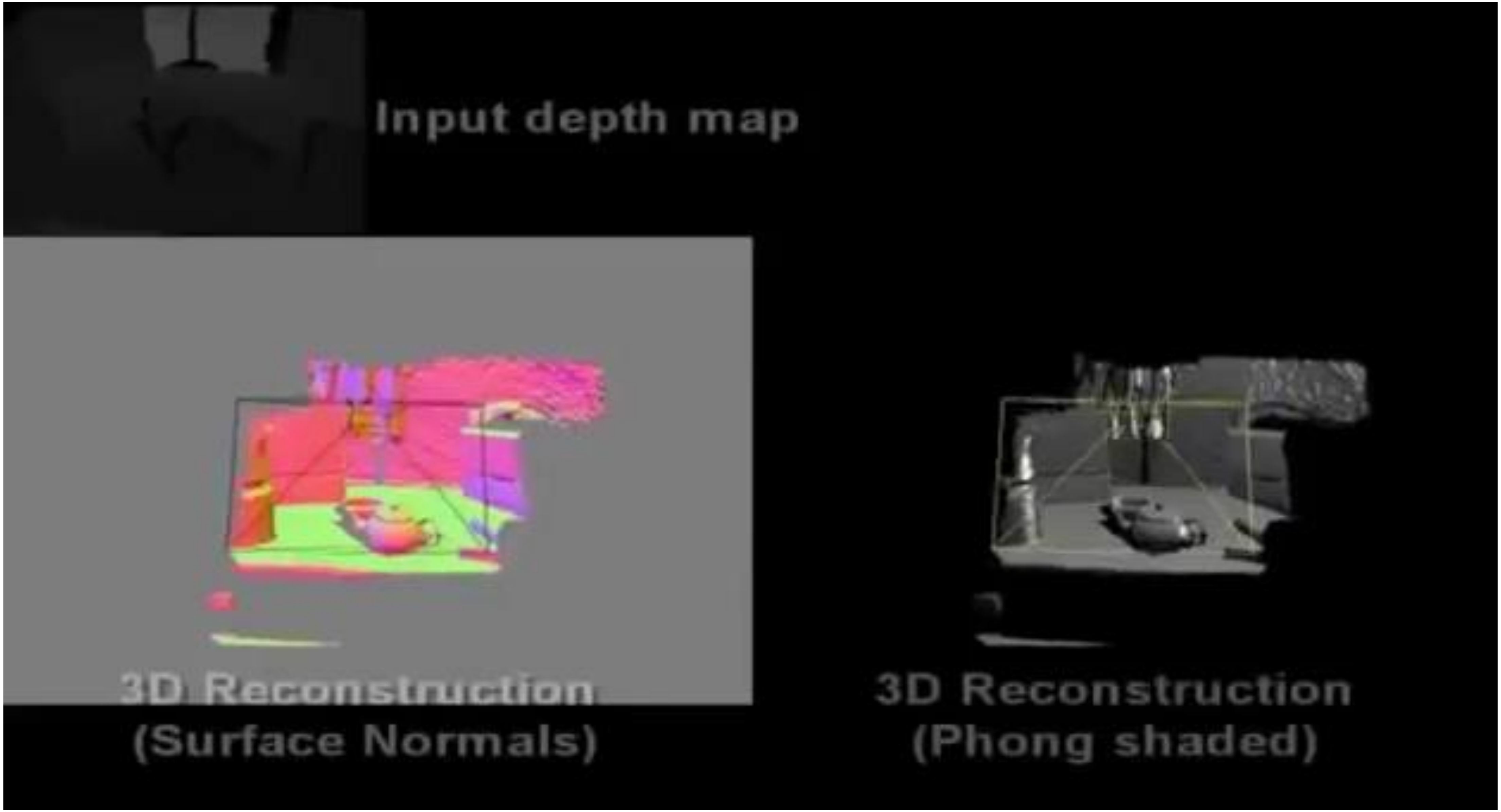




[Newcombe & Lovegrove, Geometric Reconstruction Lecture]

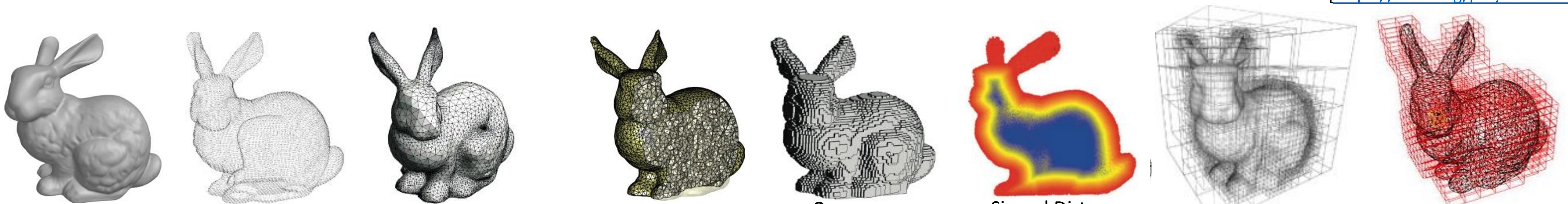
Real-time Mapping - KinectFusion

[Newcombe et al, ISMAR 2011]



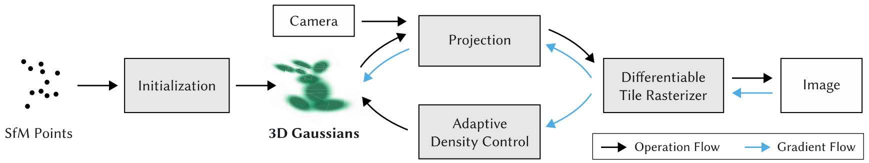
Scene Representations

[https://arxiv.org/pdf/1803.03352.pdf]

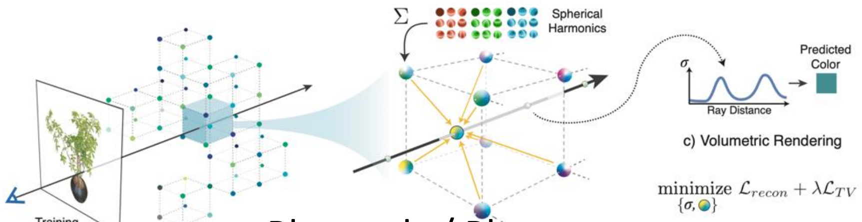


Occupancy Signed Distance

Spline/NURBS | Point Cloud | Surface Mesh | Tetrahedral Mesh | Voxel Grid | Voxel Octree | Voxel Hashing



3D Gaussian Splatting

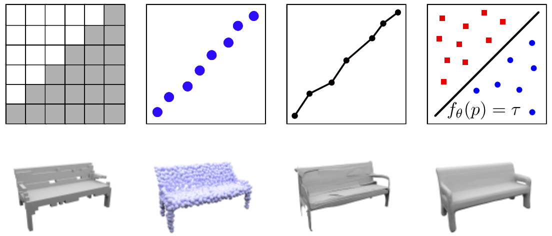


Plenoxels / Plenoctrees

Classical / Non-Neural

explicit

implicit



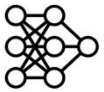
Voxel Grid | Point cloud | Mesh | (Neural) Classifier

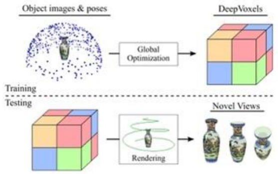
Learned / Deep Representations:

OccNet [https://arxiv.org/pdf/1812.03828.pdf]

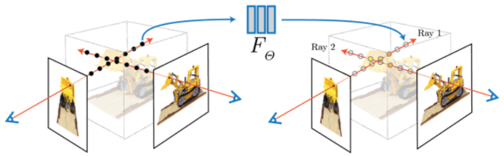
DeepSDF [https://arxiv.org/pdf/1901.05103.pdf]

IM-Net [https://arxiv.org/pdf/1812.02822.pdf]

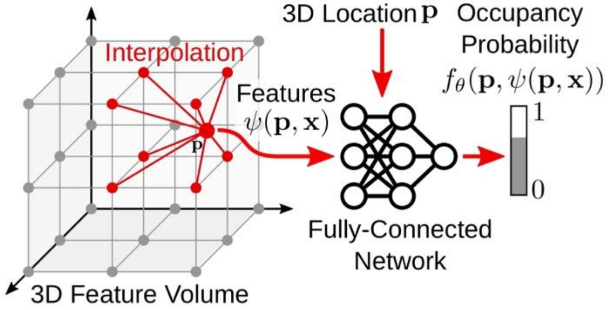




[DeepVoxels CVPR 2019]



[NeRF - Neural Radiance Fields, ECCV 2020]

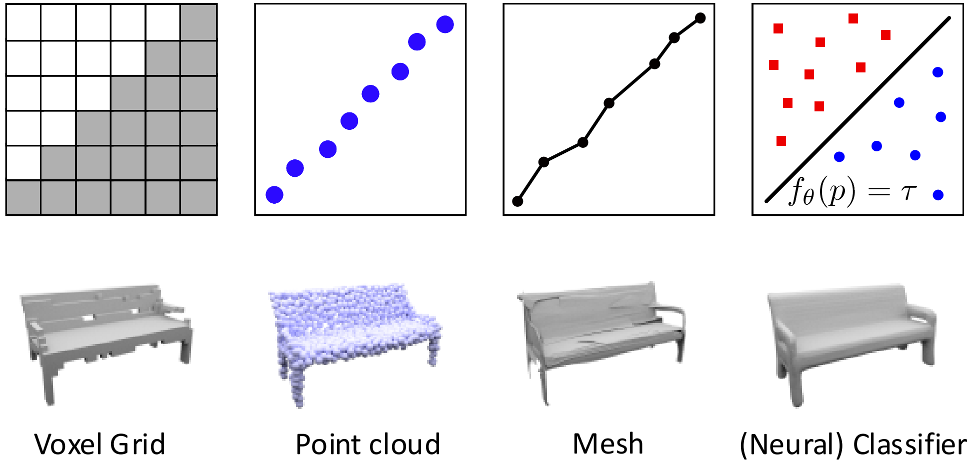


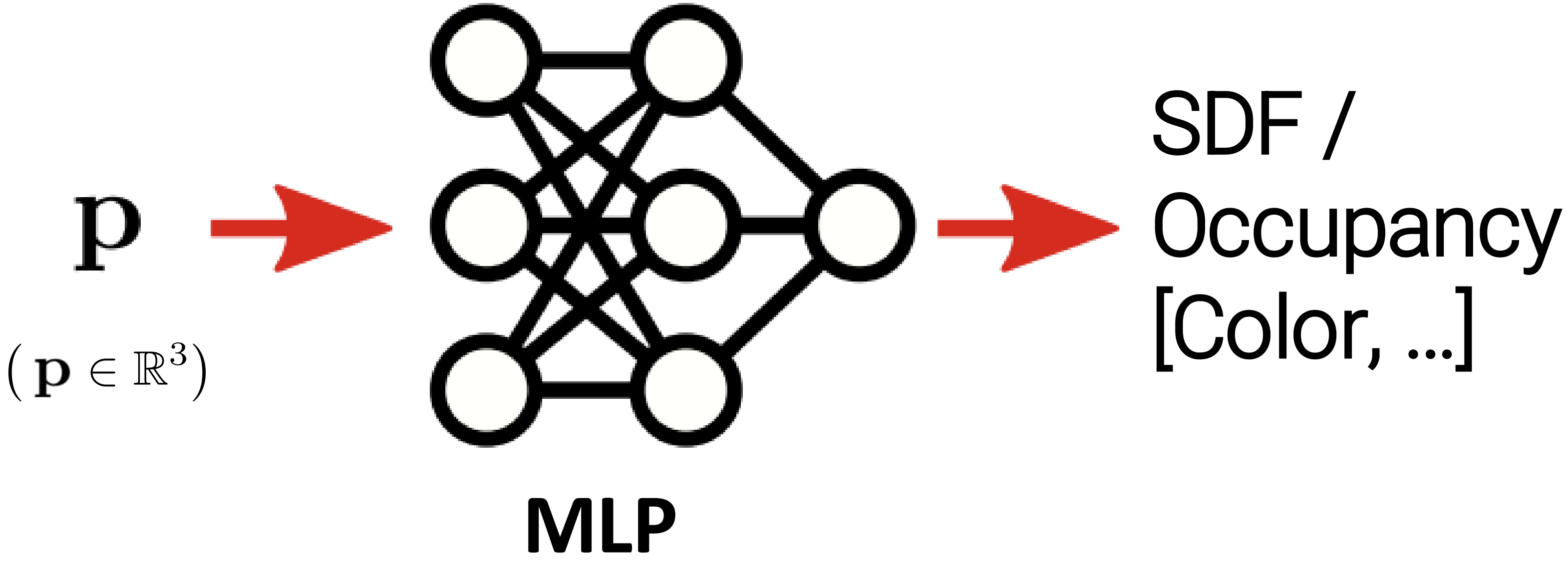
[Peng et al., Convolutional Occupancy Networks, ECCV 2020]

Neural

Neural implicit

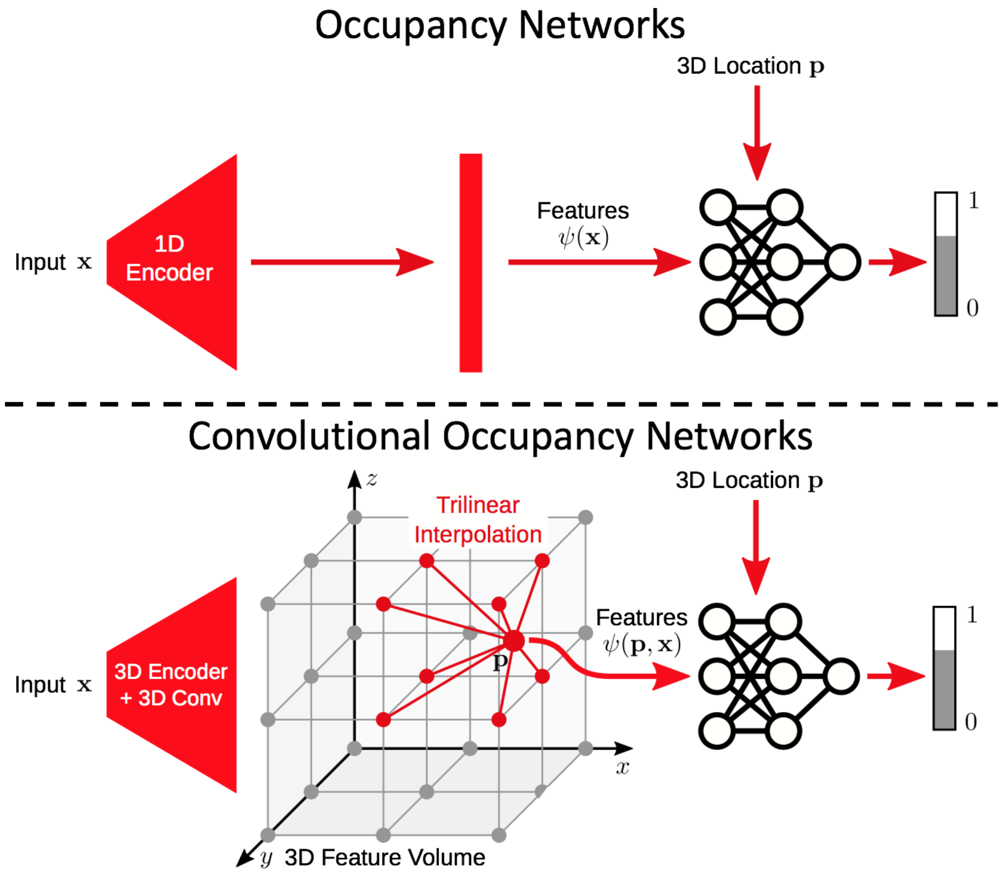
Neural Implicit Scene Representations

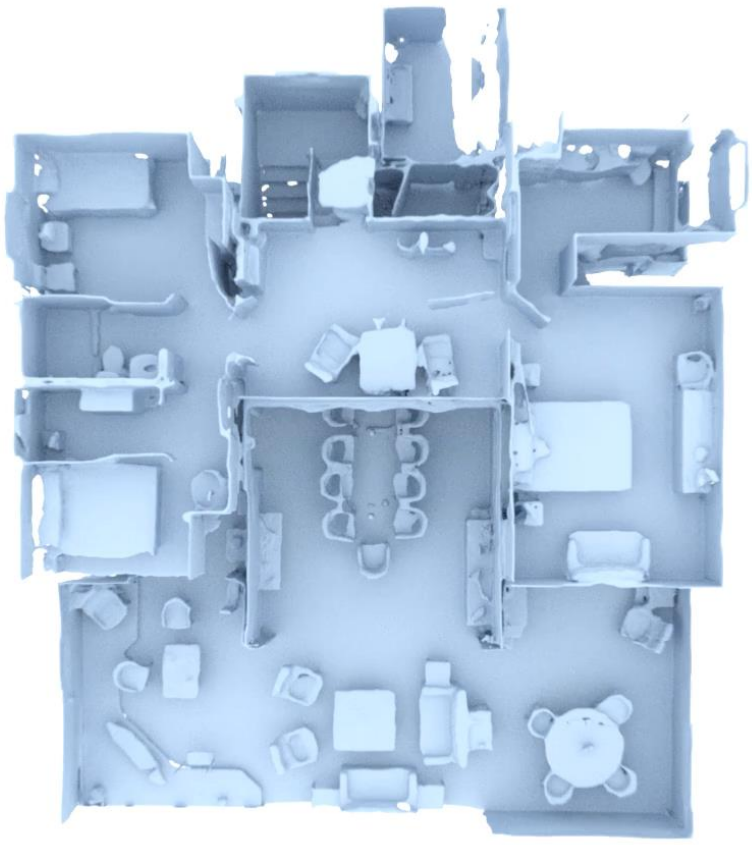




Learned / Deep Representations:
OccNet [https://arxiv.org/pdf/1812.03828.pdf]
DeepSDF [https://arxiv.org/pdf/1901.05103.pdf]
IM-Net [https://arxiv.org/pdf/1812.02822.pdf]

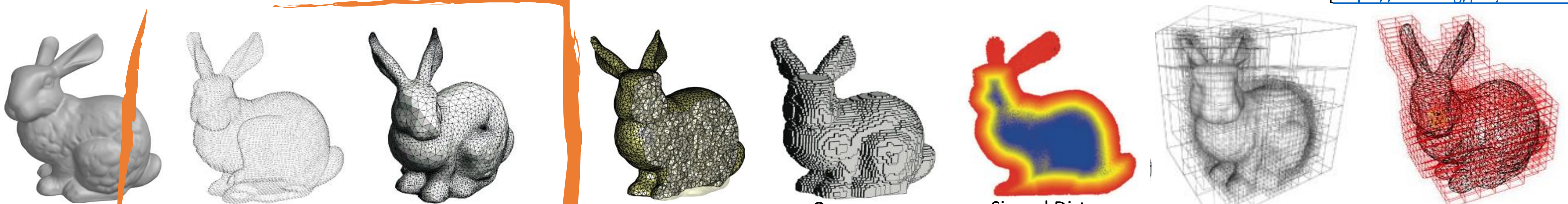
Neural Implicit Representations





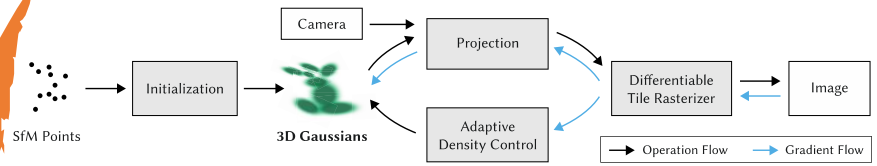
Scene Representations

[https://arxiv.org/pdf/1803.03352.pdf]

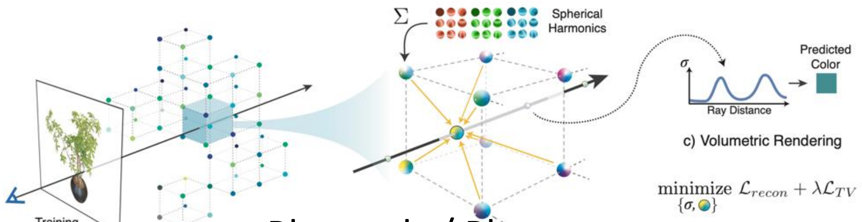


Occupancy | Signed Distance

Spline/NURBS | Point Cloud | Surface Mesh | Tetrahedral Mesh | Voxel Grid | Voxel Octree | Voxel Hashing



3D Gaussian Splatting

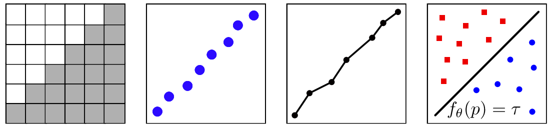


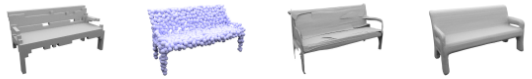
Plenoxels / Plenoctrees

explicit

implicit

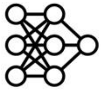
Classical / Non-Neural

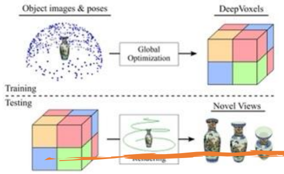




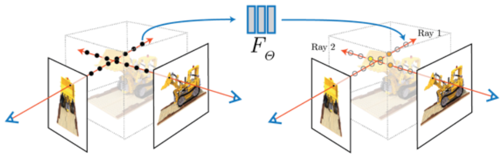
Voxel Grid | Point cloud | Mesh | (Neural) Classifier

Learned / Deep Representations:
OccNet [https://arxiv.org/pdf/1812.03828.pdf]
DeepSDF [https://arxiv.org/pdf/1901.05103.pdf]
IM-Net [https://arxiv.org/pdf/1812.02822.pdf]

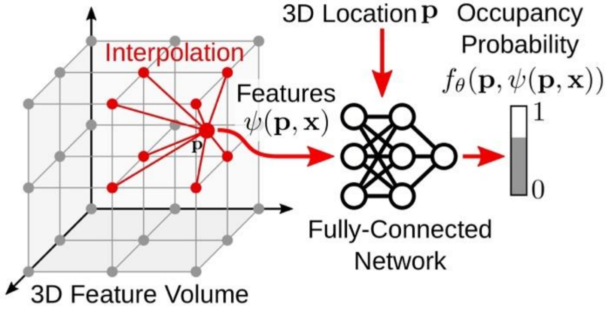




[DeepVoxels CVPR 2019]



[NeRF - Neural Radiance Fields, ECCV 2020]

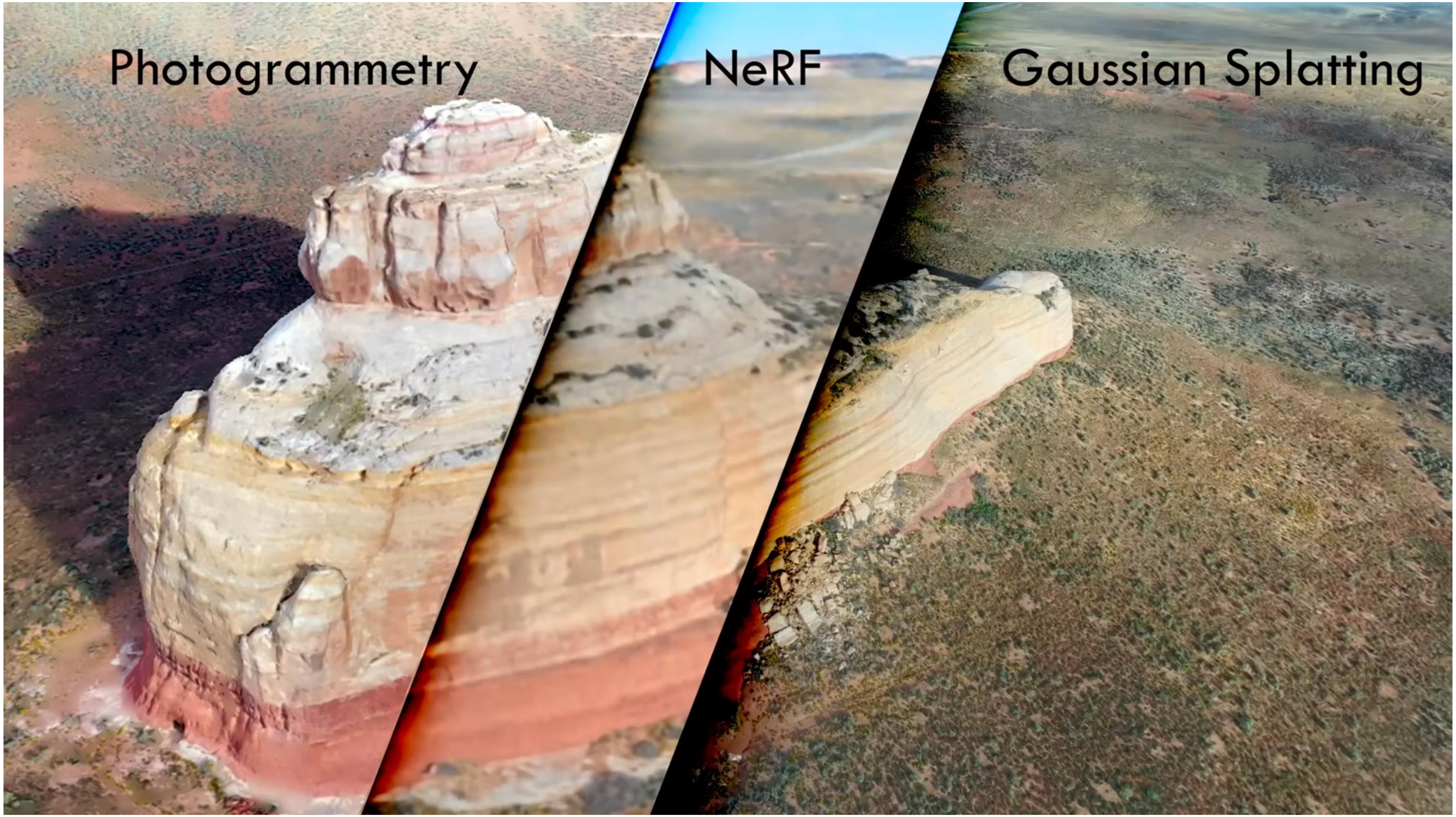


[Peng et al., Convolutional Occupancy Networks, ECCV 2020]

Neural

Neural implicit

Scene Representations for 3D Reconstruction

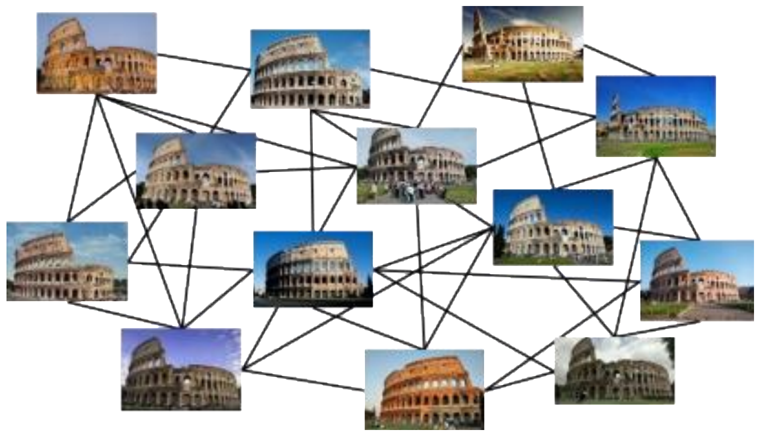


[Metthew Brennan, “Photogrammetry / NeRF / Gaussian Splatting comparison”, YouTube 2023]

Structure-from-Motion

Image Set

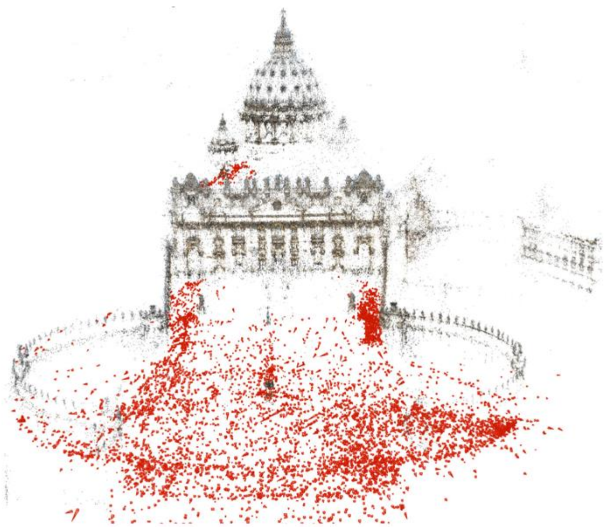
Image Association



Scene Graph

unknown cameras

SfM



Sparse Model

- NeRF
- Gaussian Splatting

MVS

known cameras

(Semi-) Dense Model

[Large-scale 3D Modeling Tutorial, CVPR 2017]

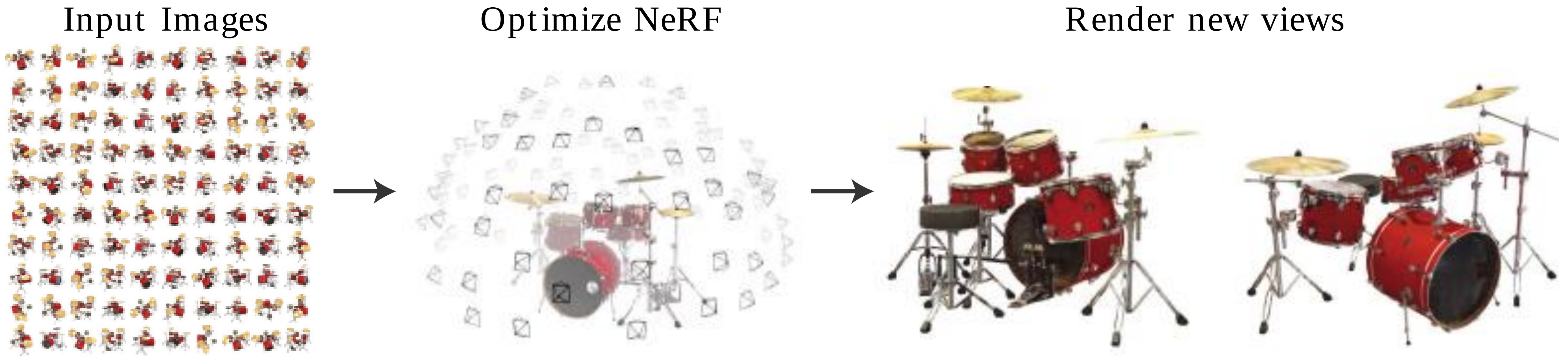
Structure-from-Motion (SfM)

Rome dataset

74,394 images

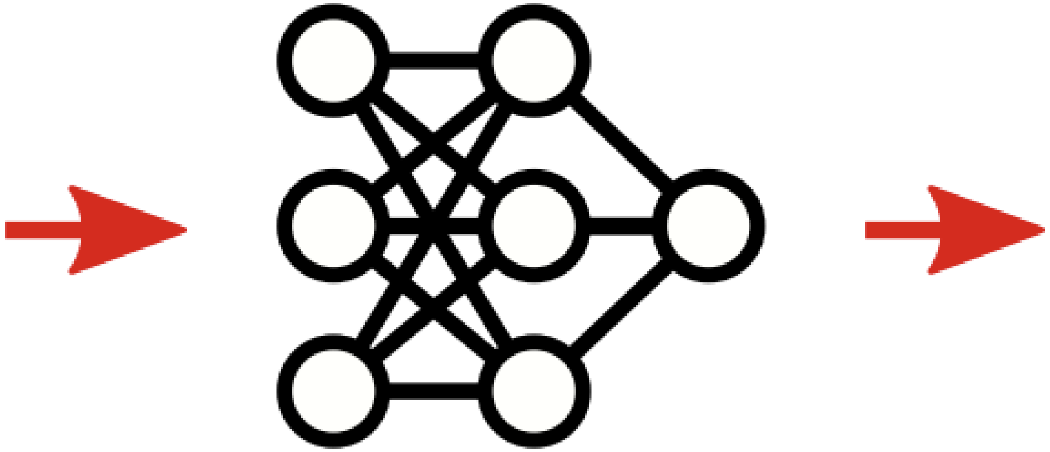
[Johannes L. Schönberger, Jan-Michael Frahm. **Structure-from-Motion Revisited**. CVPR, 2016; COLMAP]

Neural Radiance Fields
(NeRF)



Neural Implicit Representations

[3D point, viewing angle]



[Color, Density]

≈ Occupancy

Why view-dependent colors?

[NeRF: Mildenhall et al., ECCV 2020]

Neural Radiance Fields (NeRFs)

[NeRF: Mildenhall et al., ECCV 2020]

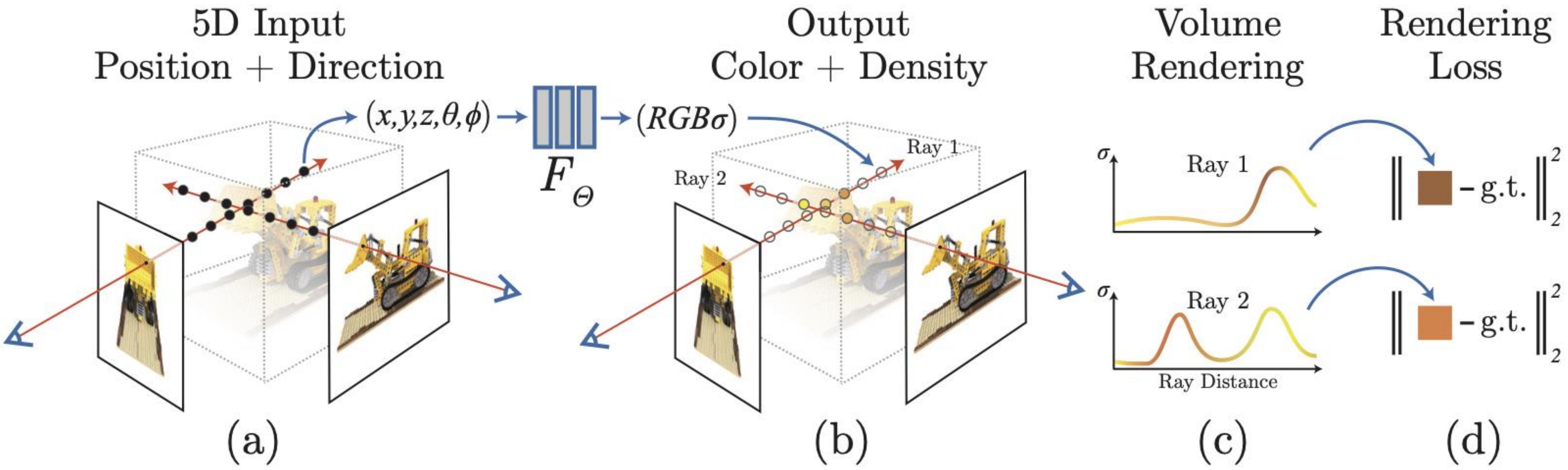
NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis

Ben Mildenhall* UC Berkeley
Pratul P. Srinivasan* UC Berkeley
Matthew Tancik* UC Berkeley
Jonathan T. Barron Google Research
Ravi Ramamoorthi UC San Diego
Ren Ng UC Berkeley

* Denotes Equal Contribution

Neural Radiance Fields (NeRFs)

[NeRF: Mildenhall et al., ECCV 2020]

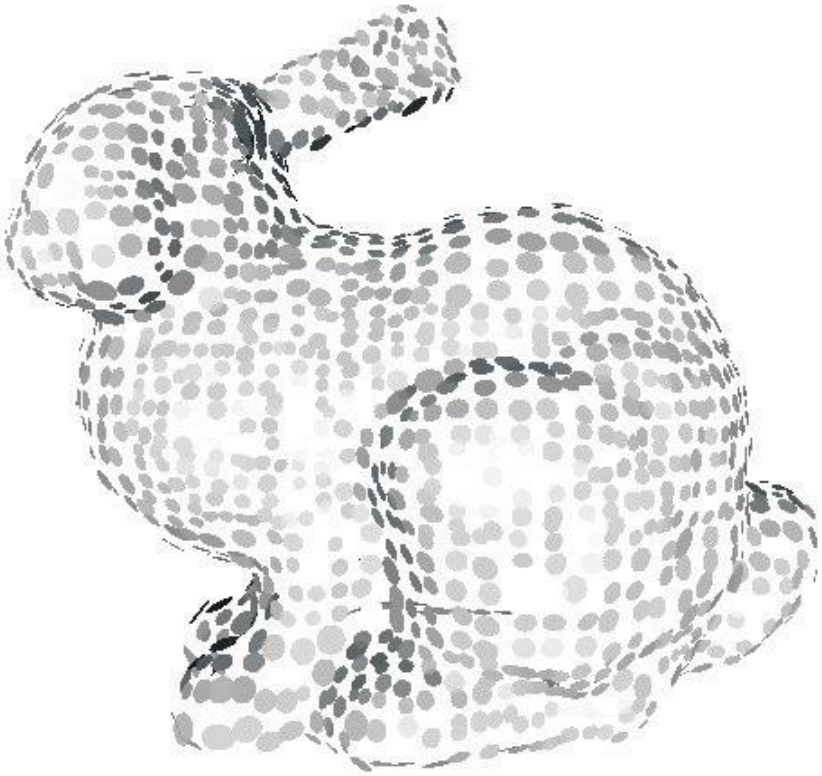


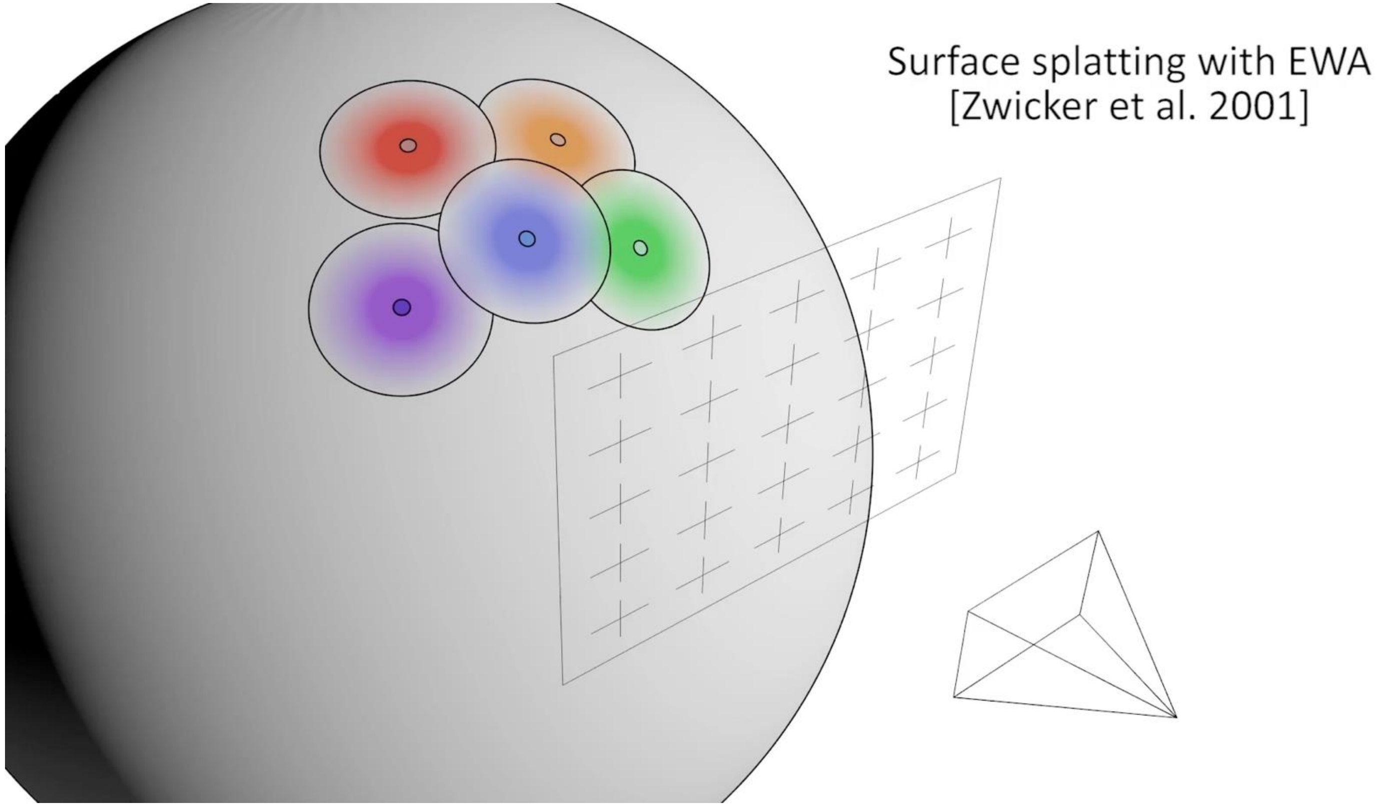
Neural Radiance Fields (NeRFs)

[NeRF: Mildenhall et al., ECCV 2020]

Gaussian Splatting

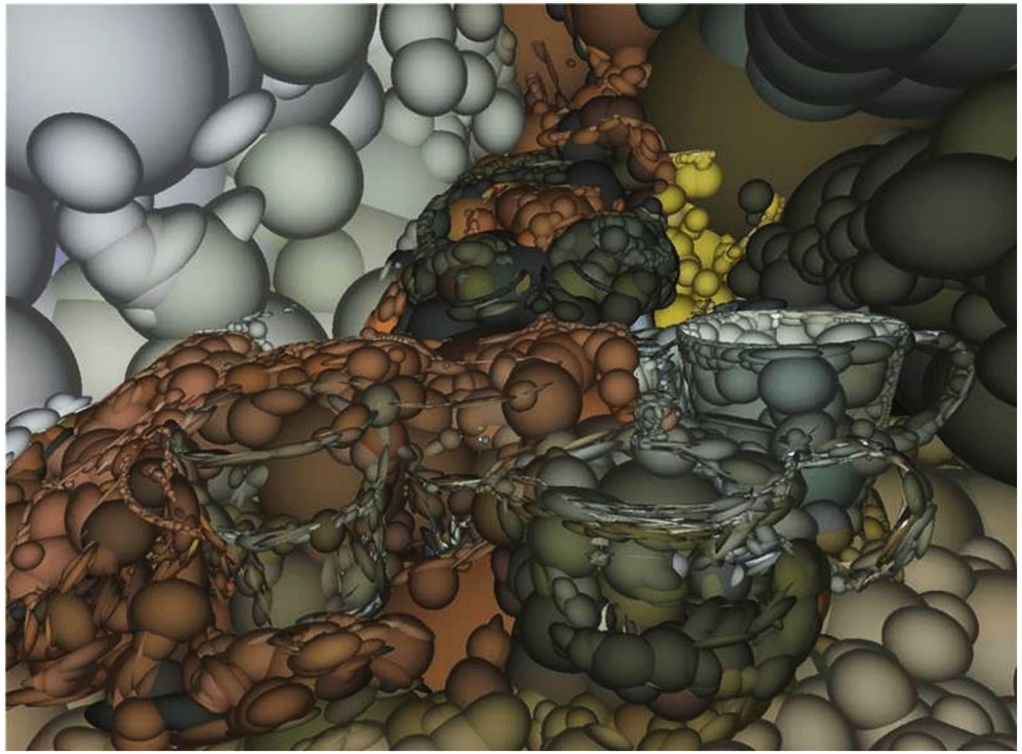
Point Splatting





[Zwicker et al. 2001; Yifan Wang, 2019]

Gaussian Splatting

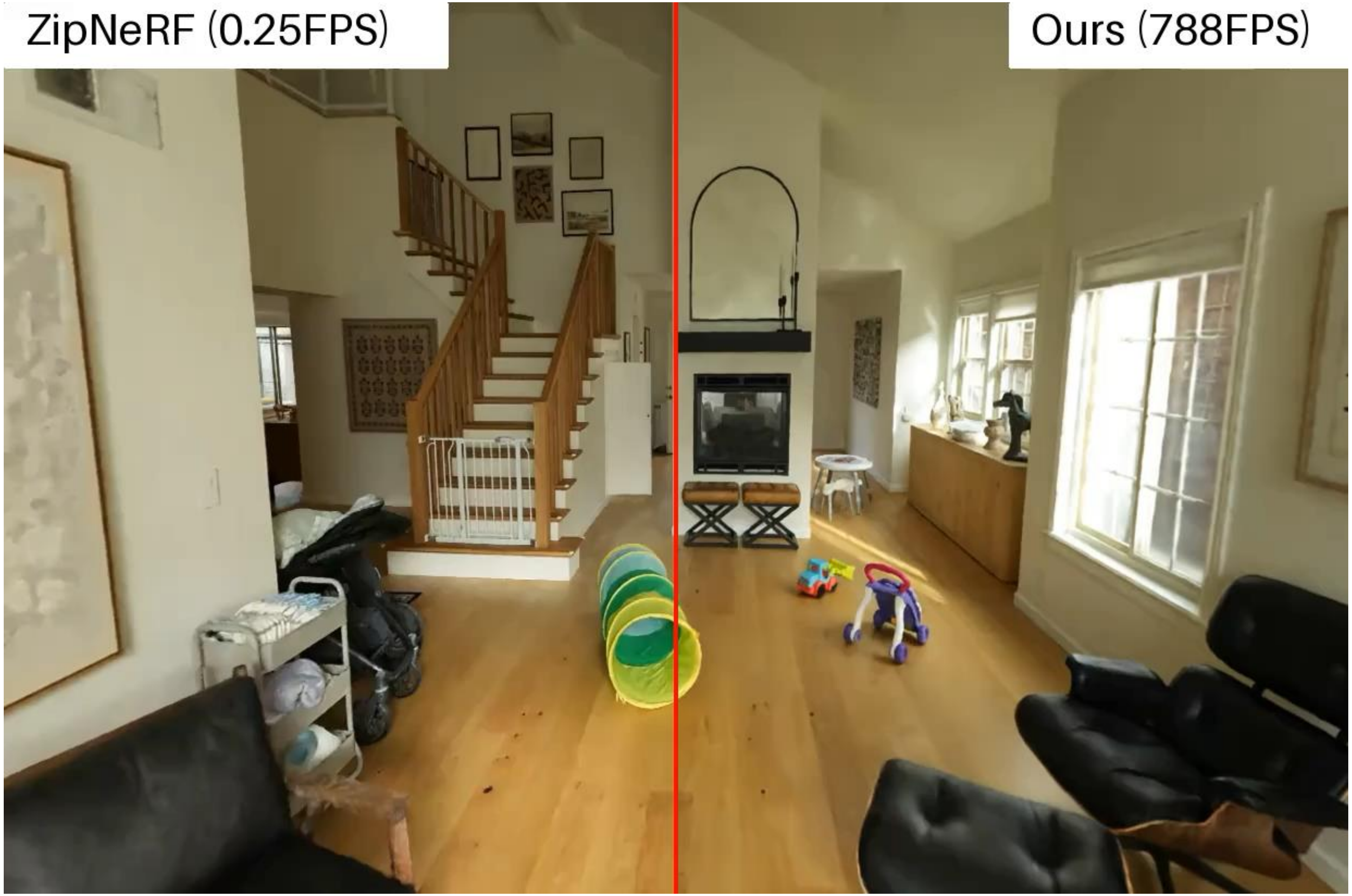


[Matsuki et al., Gaussian Splatting SLAM, CVPR 2024]

Gaussian Splatting

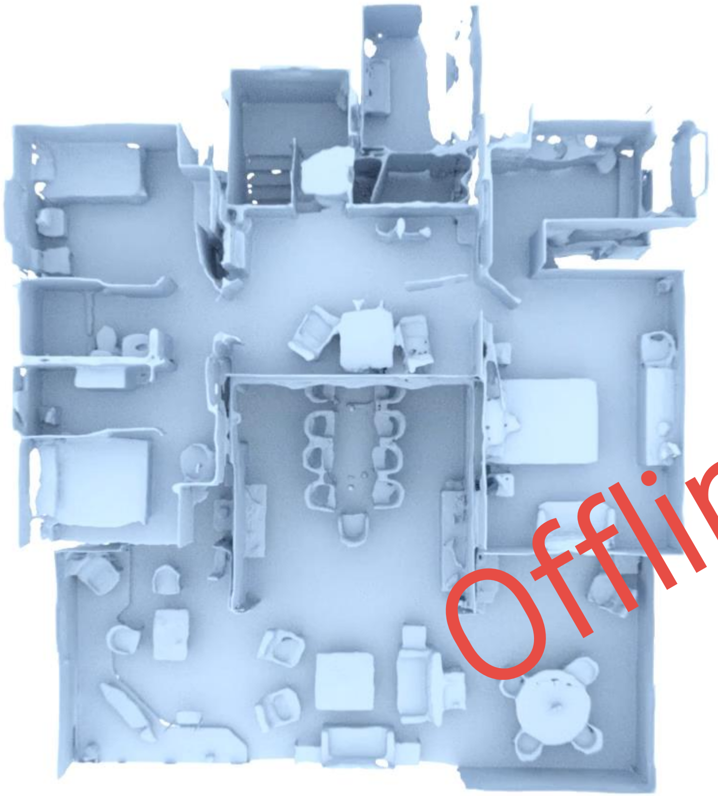
[pinkpocketTV, YouTube 2024]

Hybrid NeRF / GS: RadSplat



[Niemeyer et al., RadSplat: Radiance Field-Informed Gaussian Splatting for Robust Real-Time Rendering with 900+ FPS, Arxiv 2024]

Neural Implicit Representations

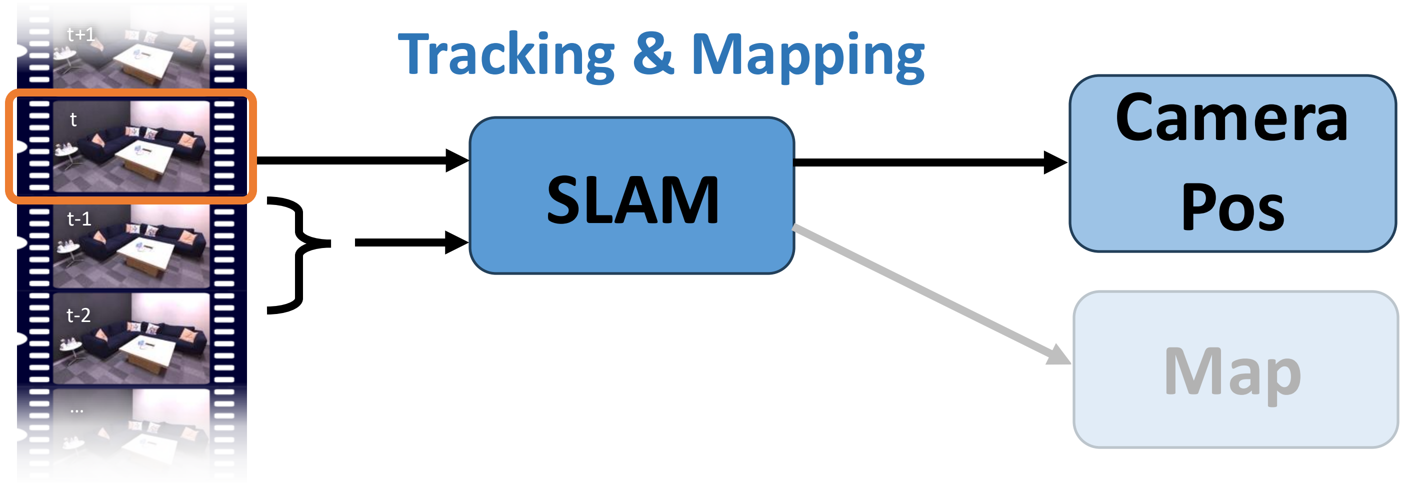


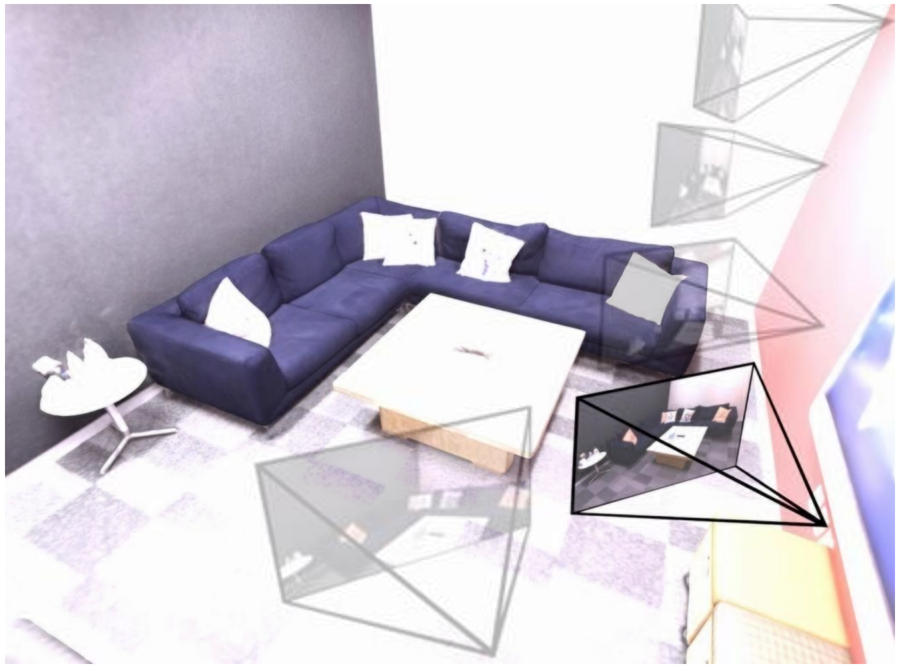
Offline Methods

ConvONet [Peng et al., ECCV'20]

NeRF [Mildenhall et al., ECCV'20]

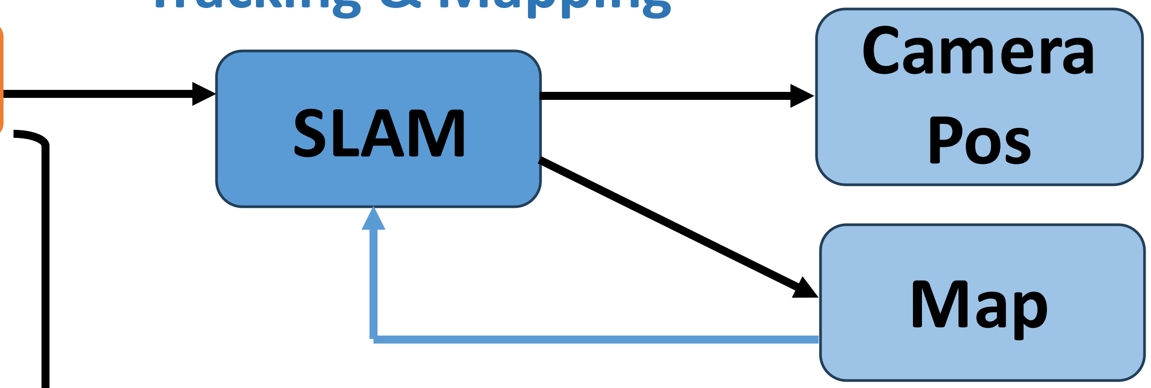
Simultaneous Localization and Mapping





Simultaneous Localization and Mapping

Tracking & Mapping

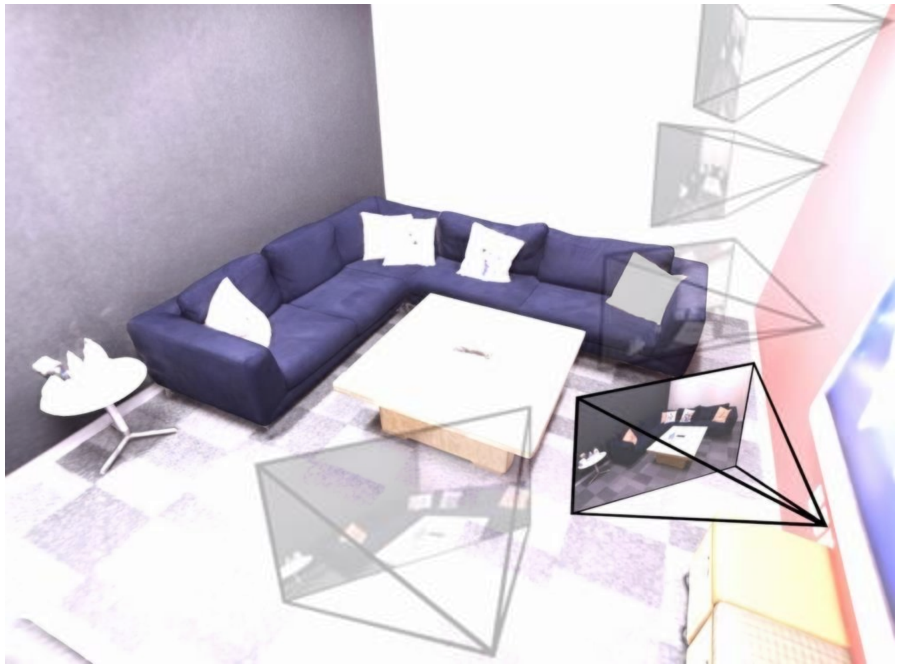


Radiance Field

"Loss-less"

Compression ≈

No storage



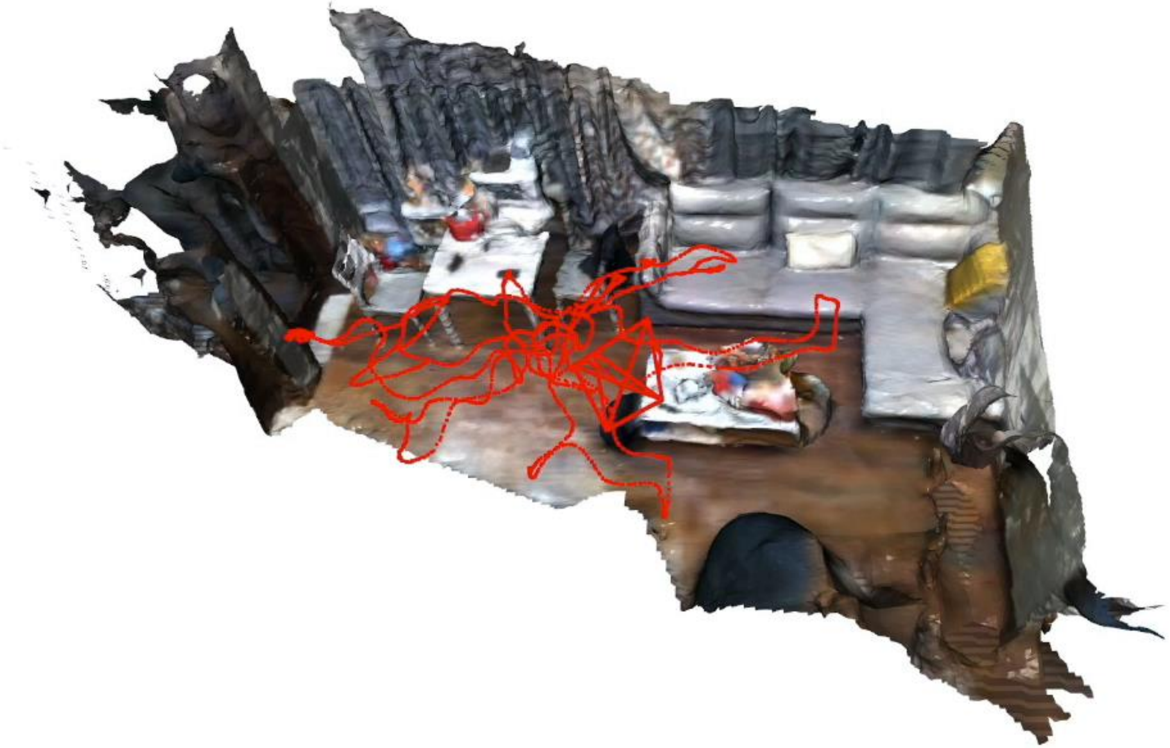
Neural Implicit SLAM: iMAP

[iMAP: Sucar, Liu, Ortiz, Davison, ICCV 2021]

Dense SLAM with a Neural Implicit Scene Represenation

[Zhu, Peng, Larsson, Xu, Bao, Cui, Oswald, Pollefeys, CVPR'22]

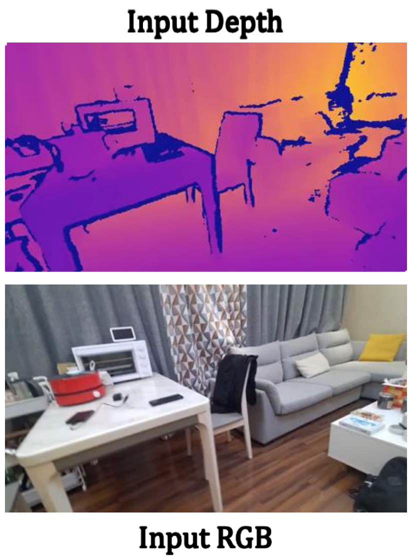
RGB-D Sequences

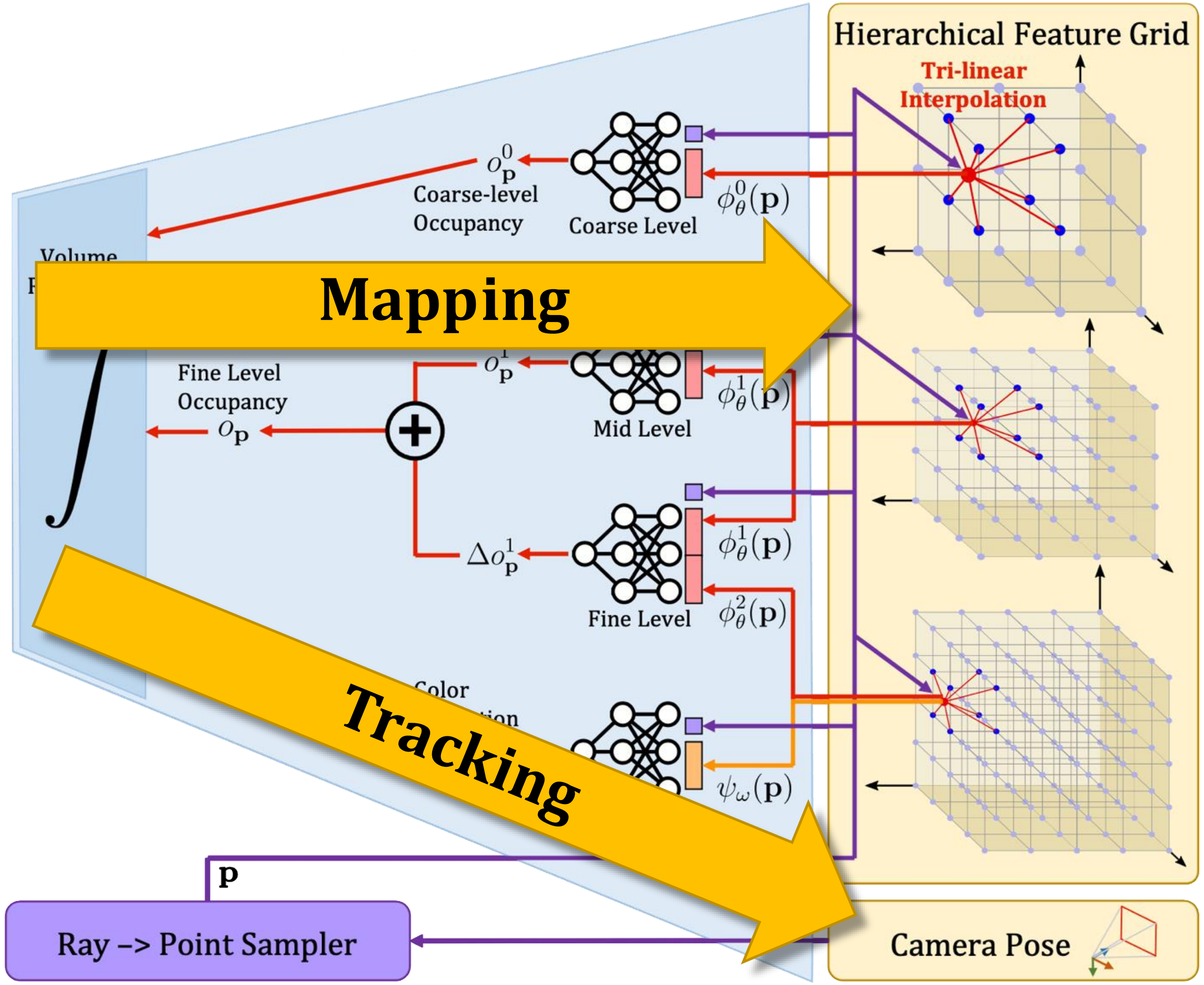


[Zhu et al., NICE-SLAM, CVPR 2022]

Dense SLAM with a Neural Implicit Scene Represenation

[Zhu, Peng, Larsson, Xu, Bao, Cui, Oswald, Pollefeys, CVPR'22]

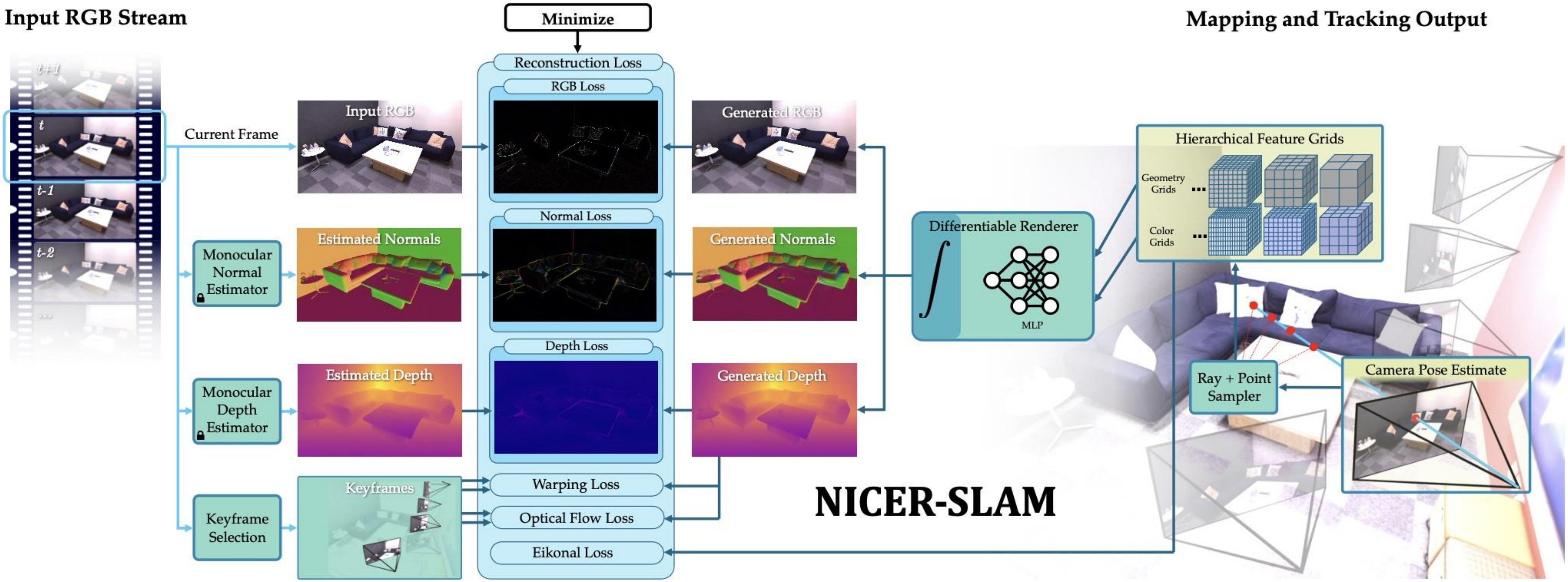




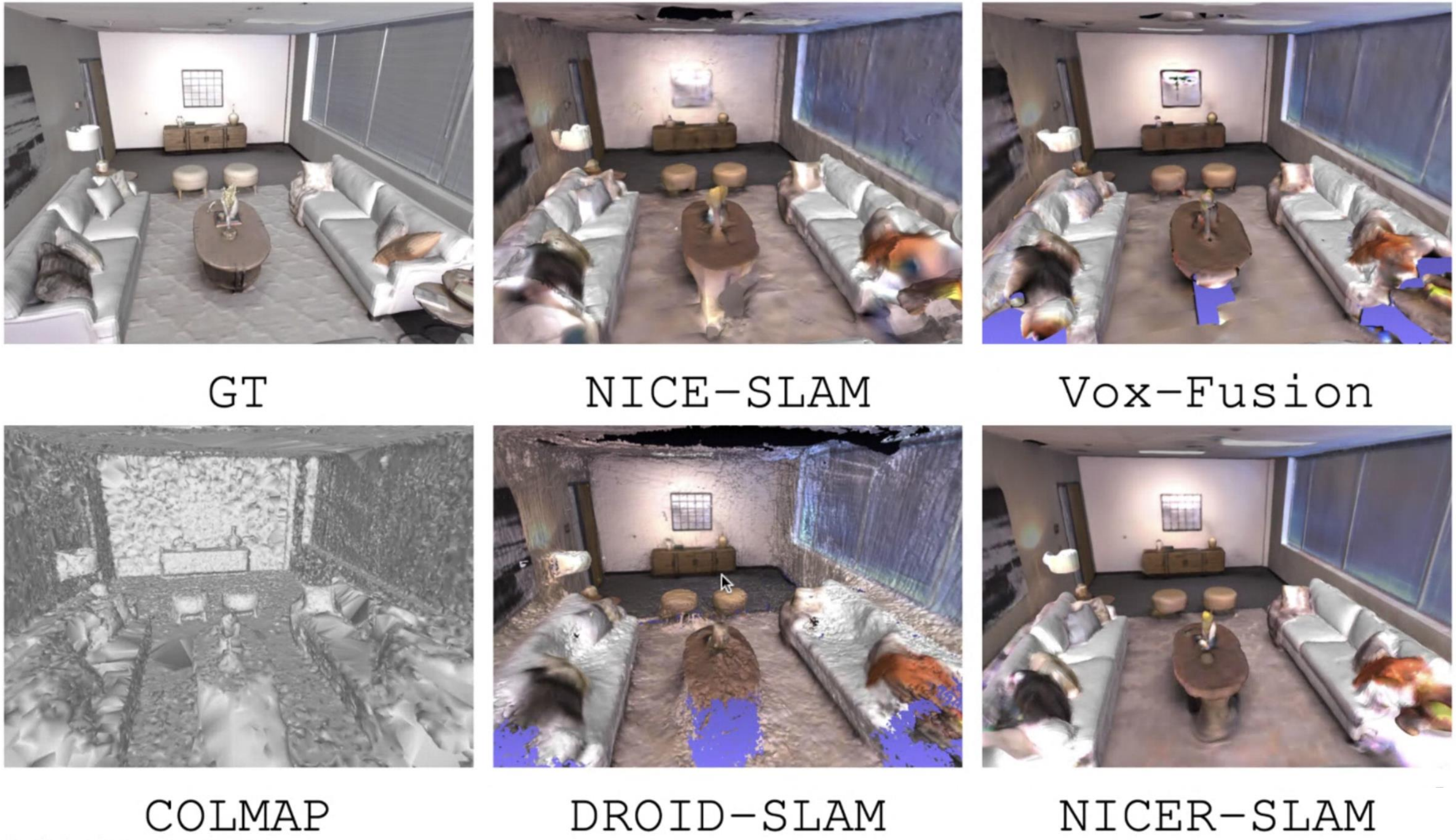
[Zhu et al., NICE-SLAM, CVPR 2022]

NICER-SLAM: RGB-only SLAM

[Zhu, Peng, Larsson, Cui, Oswald, Geiger, Pollefeys, 3DV'24]
Best Paper Honorable Mention Award



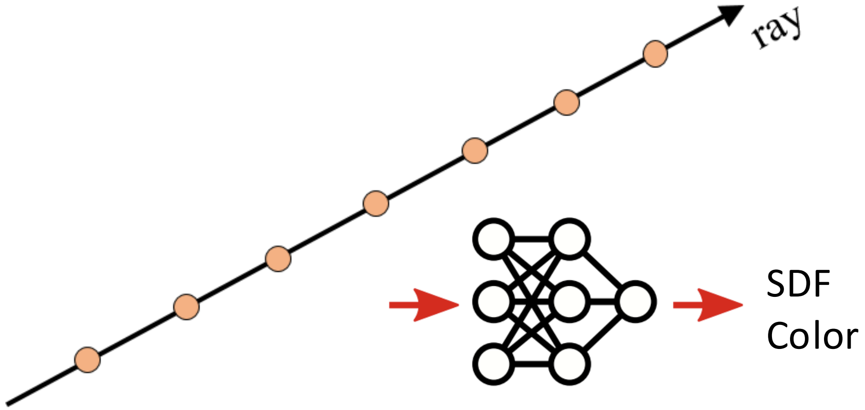
NICER-SLAM: RGB-only SLAM



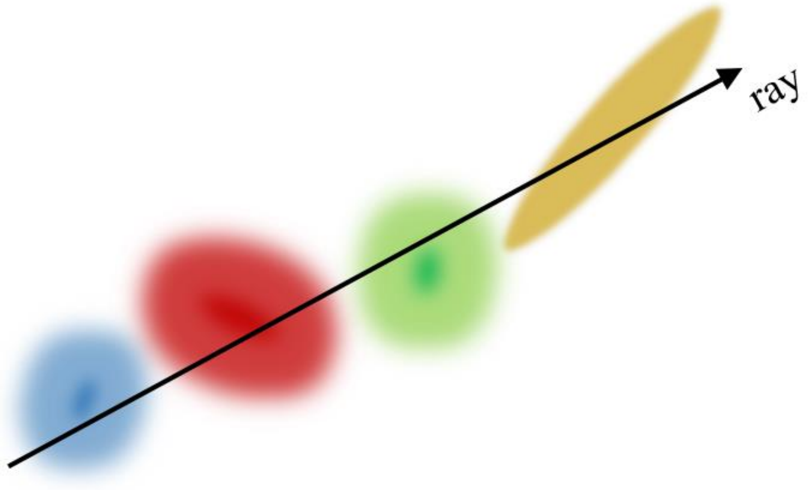
[Zhu, Peng, Larsson, Cui, Oswald, Geiger, Pollefeys, NICER-SLAM , Arxiv 2023]

Gaussian-SLAM: Dense SLAM with Gaussian Splatting

[Yugay, Li, Gevers, Oswald, Arxiv'23]



Radience field sampling & feature aggregation



Set of Gaussians encodes geometry and color

Gaussian-SLAM: Dense SLAM with Gaussian Splatting

[Yugay, Li, Gevers, Oswald, Arxiv’23]

ESLAM

Point-SLAM

Gaussian-SLAM

GT

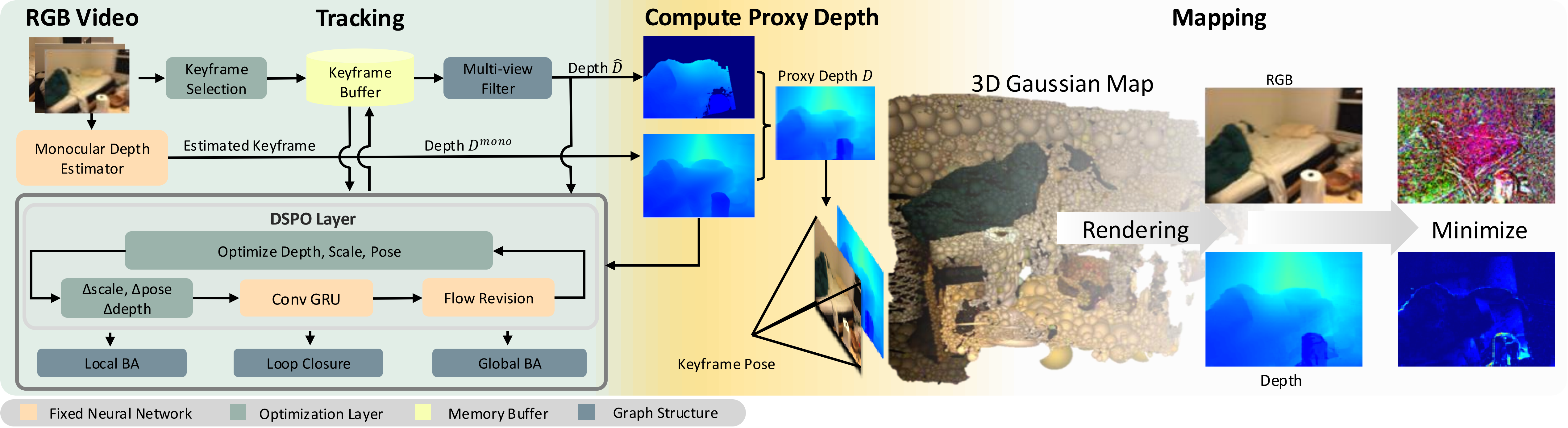
[Yugay, Gevers, Oswald, Arxiv'24]

MAGiC-SLAM: Multi-Agent Gaussian SLAM

Splat-SLAM

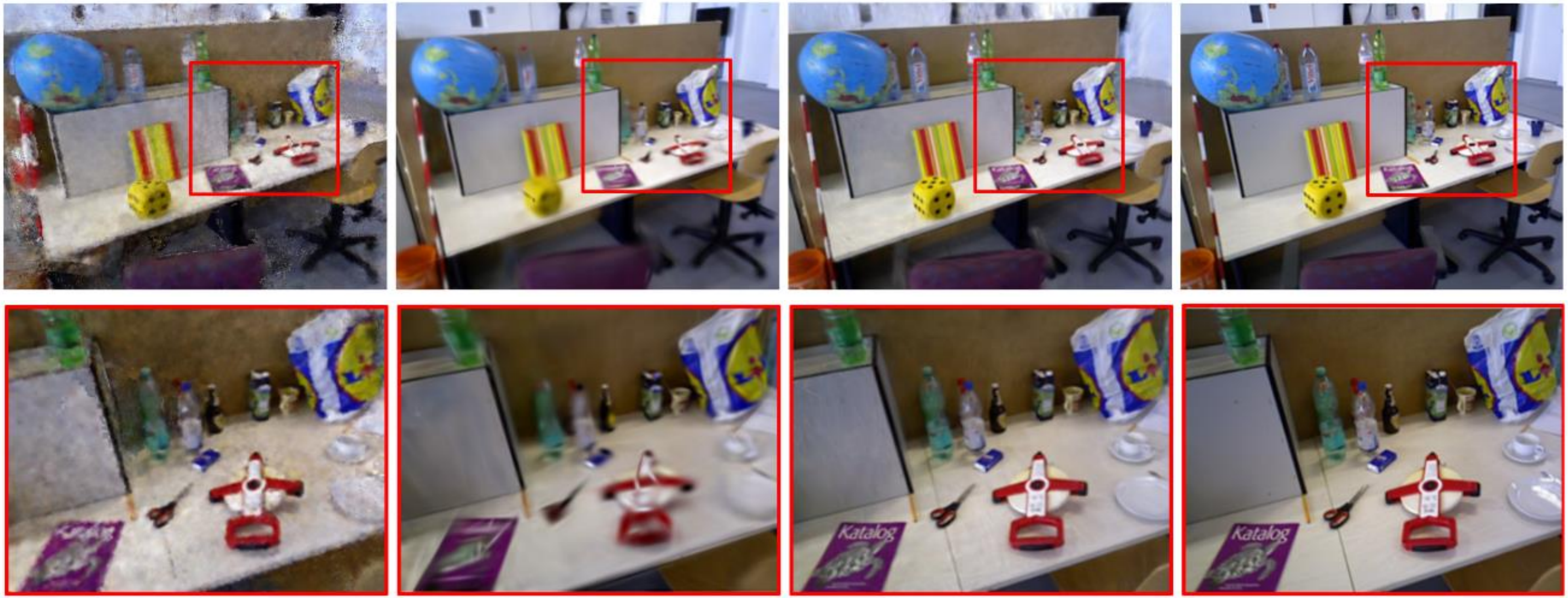
[Sandtröm, Tateno, Oechsle, Niemeyer, Oswald, Tombari, Arxiv'24]

Globally Optimized RGB-only SLAM with 3D Gaussians



Results: Rendering on TUM-RGBD

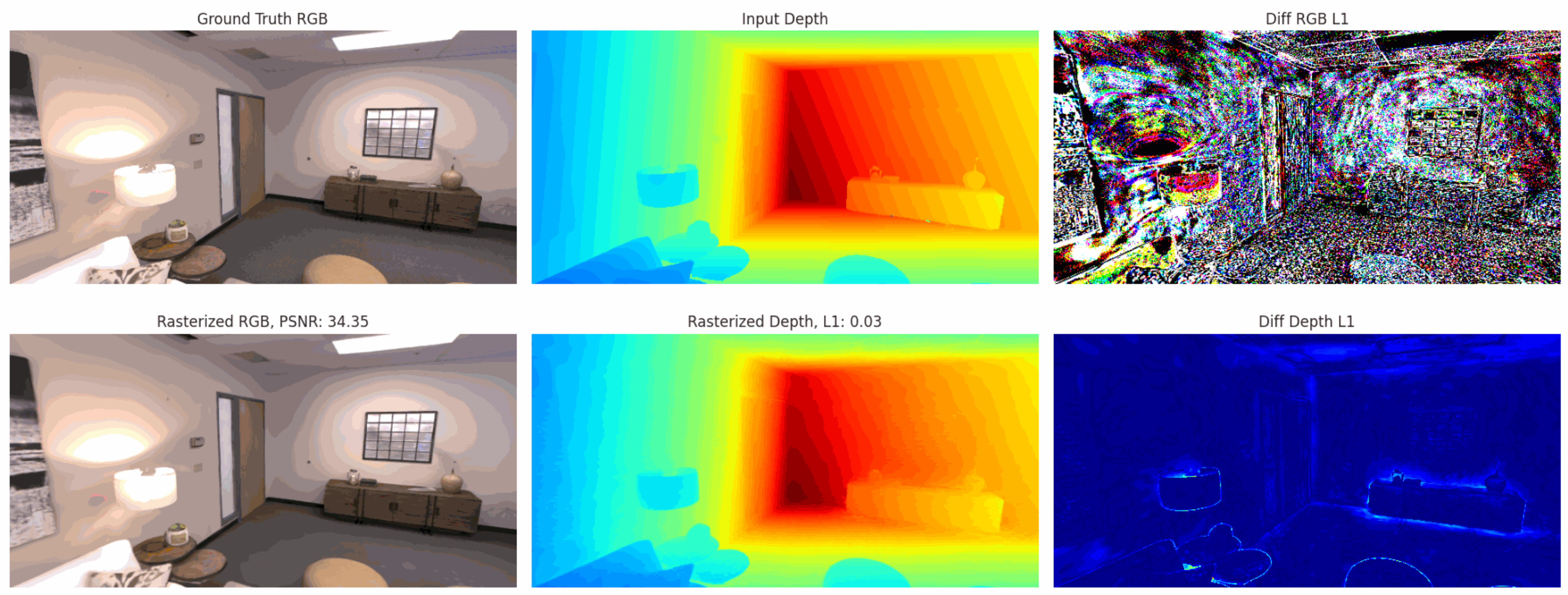
[Sandtröm, Tateno, Oechsle, Niemeyer, Oswald, Tombari, Arxiv'24]



GlORIE-SLAM | MonoGS | Ours | Ground Truth

Results: Color & Depth Rendering on Replica

[Sandtröm, Tateno, Oechsle, Niemeyer, Oswald, Tombari, Arxiv'24]



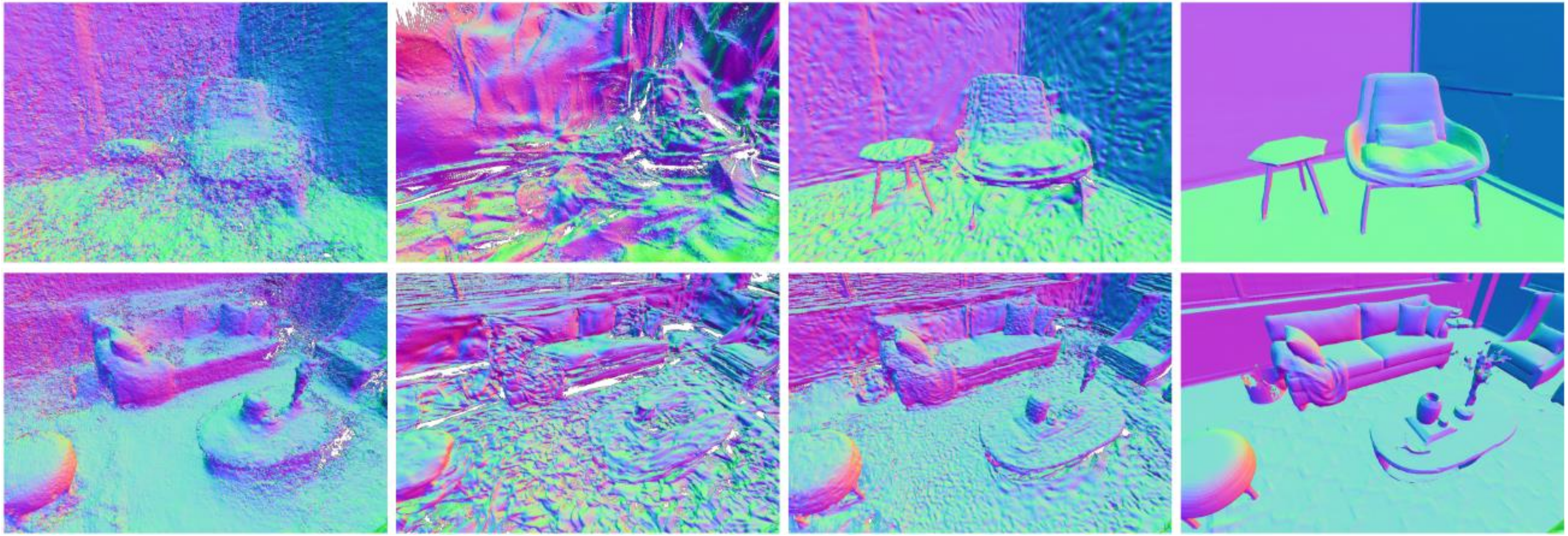
Results: Rendering on ScanNet

[Sandtröm, Tateno, Oechsle, Niemeyer, Oswald, Tombari, Arxiv'24]

Method	Metric	0000	0059	0106	0169	0181	0207	Avg.
RGB-D Input								
SplaTaM [24]	PSNR↑	19.33	19.27	17.73	21.97	16.76	19.80	19.14
	SSIM ↑	0.66	0.79	0.69	0.78	0.68	0.70	0.72
	LPIPS↓	0.44	0.29	0.38	0.28	0.42	0.34	0.36
MonoGS [38]	PSNR↑	18.70	20.91	19.84	22.16	22.01	18.90	20.42
	SSIM ↑	0.71	0.79	0.81	0.78	0.82	0.75	0.78
	LPIPS↓	0.48	0.32	0.32	0.34	0.42	0.41	0.38
Gaussian-SLAM [74]	PSNR↑	28.54	26.21	26.26	28.60	27.79	28.63	27.67
	SSIM ↑	**0.93**	**0.93**	**0.93**	**0.92**	**0.92**	**0.91**	**0.92**
	LPIPS↓	0.27	0.21	0.22	0.23	0.28	0.29	0.25
RGB Input								
GO-SLAM [79]	PSNR↑	15.74	13.15	14.58	14.49	15.72	15.37	14.84
	SSIM ↑	0.42	0.32	0.46	0.42	0.53	0.39	0.42
	LPIPS↓	0.61	0.60	0.59	0.57	0.62	0.60	0.60
MonoGS [38]	PSNR↑	16.91	19.15	18.57	20.21	19.51	18.37	18.79
	SSIM ↑	0.62	0.69	0.74	0.74	0.75	0.70	0.71
	LPIPS↓	0.70	0.51	0.55	0.54	0.63	0.58	0.59
GlORIE-SLAM* [75]	PSNR↑	23.42	20.66	20.41	25.23	21.28	23.68	22.45
	SSIM ↑	0.87	0.87	0.83	0.84	0.91	0.76	0.85
	LPIPS↓	0.26	0.31	0.31	0.21	0.44	0.29	0.30
Splat-SLAM (Ours)	PSNR↑	**28.68**	**27.69**	**27.70**	**31.14**	**31.15**	**30.49**	**29.48**
	SSIM ↑	0.83	0.87	0.86	0.87	0.84	0.84	0.85
	LPIPS ↓	**0.19**	**0.15**	**0.18**	**0.15**	**0.23**	**0.19**	**0.18**

Results: Reconstruction on Replica

[Sandtröm, Tateno, Oechsle, Niemeyer, Oswald, Tombari, Arxiv'24]



Metrics	NeRF-SLAM [62]	DIM-SLAM [28]	GO-SLAM [79]	NICER-SLAM [81]	HI-SLAM [78]	MoD-SLAM* [80]	GlORIE-SLAM* [75]	Mono-GS [38]	Q-SLAM * [46]	**Ours**
Render Depth L1↓	4.49	-	-	-	-	-	-	27.24	2.76	**2.41**
Accuracy ↓	-	4.03	3.81	3.65	3.62	2.48	2.96	30.61	-	**2.43**
Completion ↓	-	4.20	4.79	4.16	4.59	-	3.95	12.19	-	**3.64**
Comp. Rat. ↑	-	79.60	78.00	79.37	80.60	-	83.72	40.53	-	**84.69**

Deblur Gaussian SLAM

Blurry Input

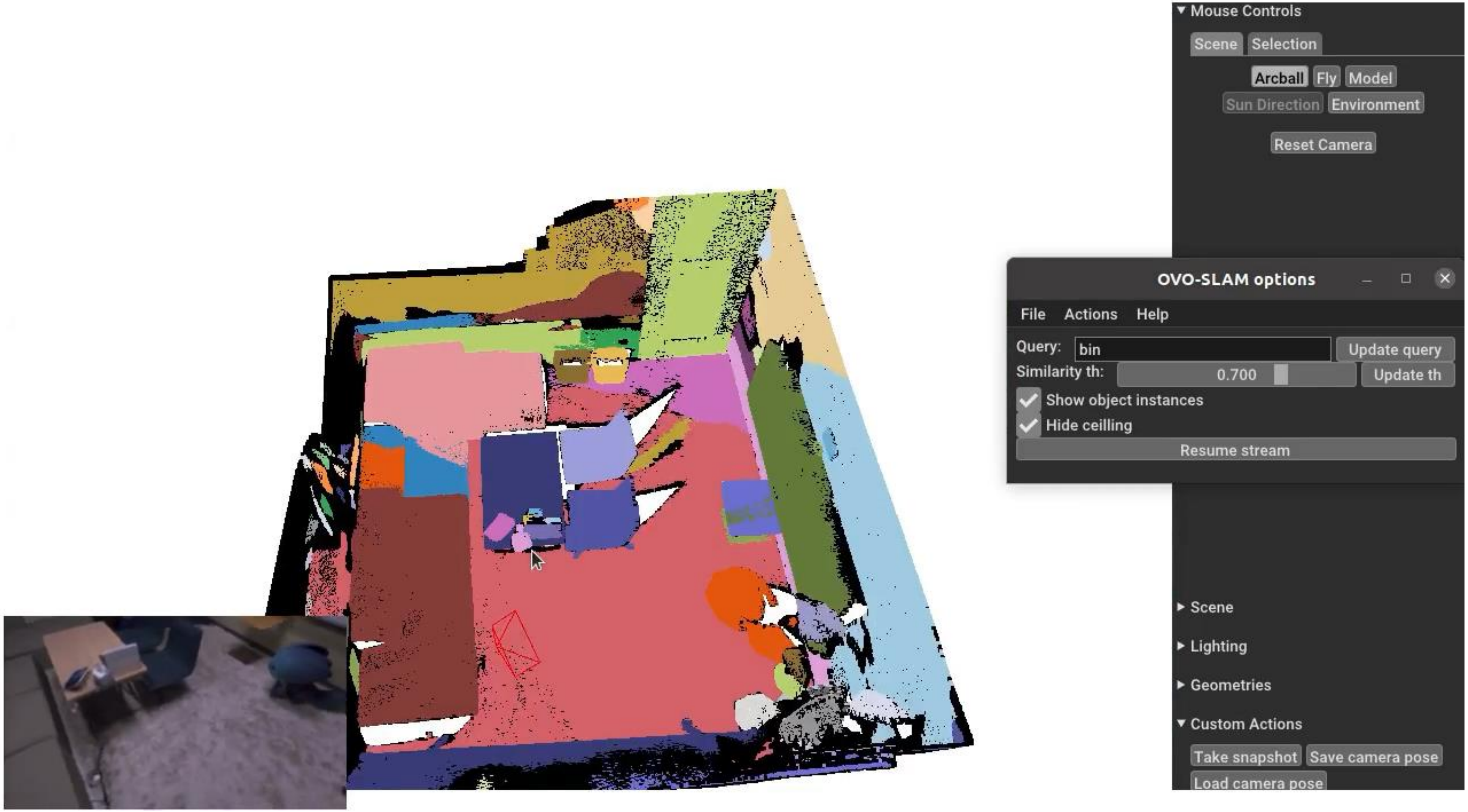
Reconstruction

Deblur Gaussian SLAM

Language and 3D

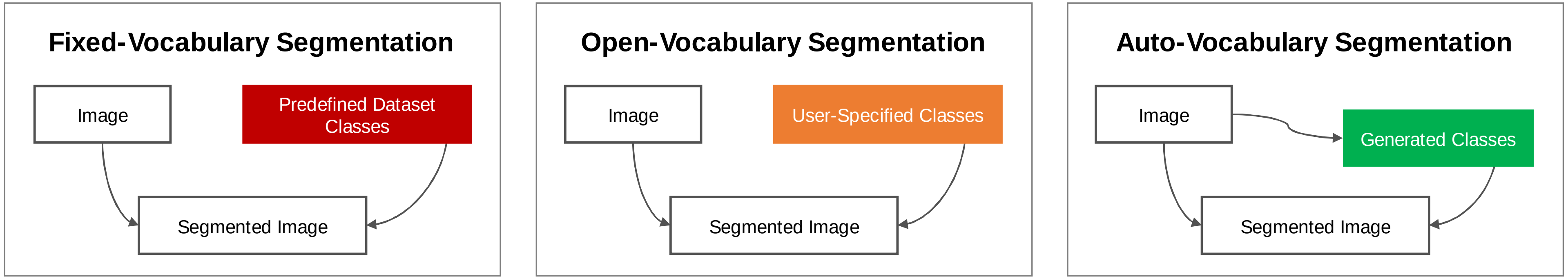
[LERF: Language Embedded Radiance Fields, ICCV 2023]

Open-vocabulary Online SLAM

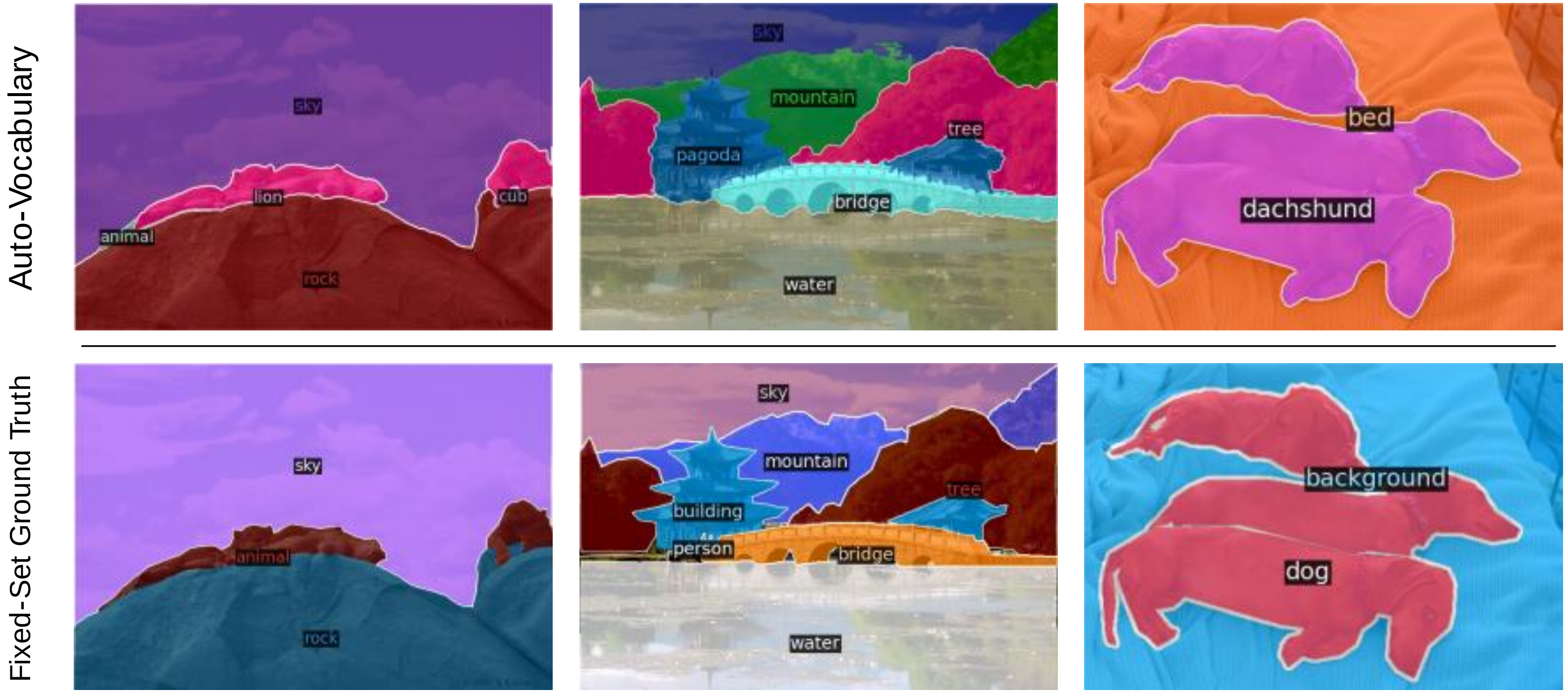


Auto-Vocabulary Segmentation

[Ülger, Kulicki, Asano, Oswald, Arxiv'23]

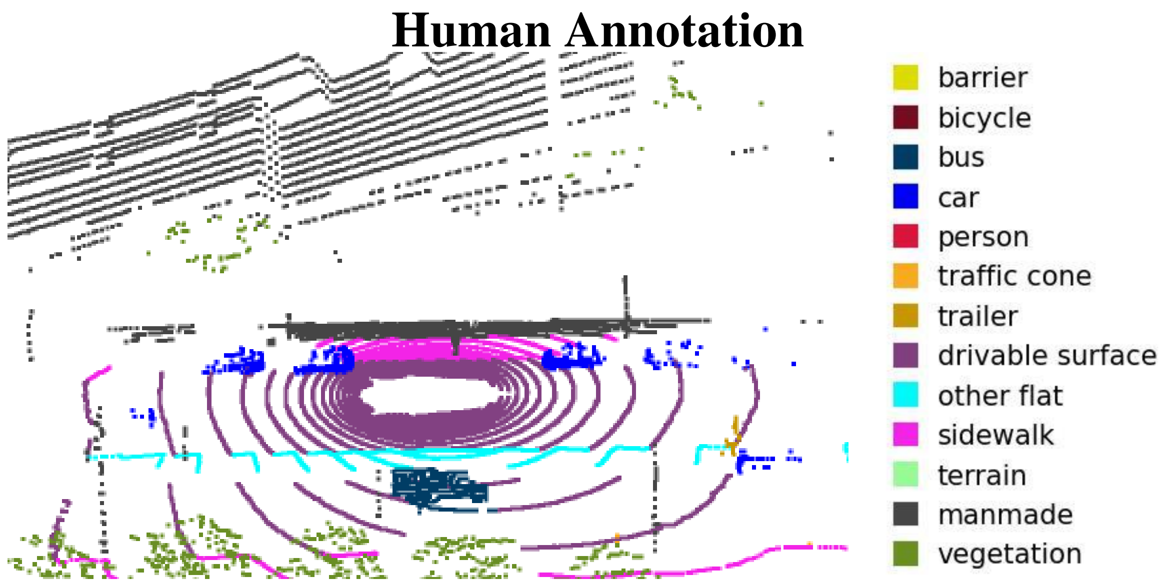


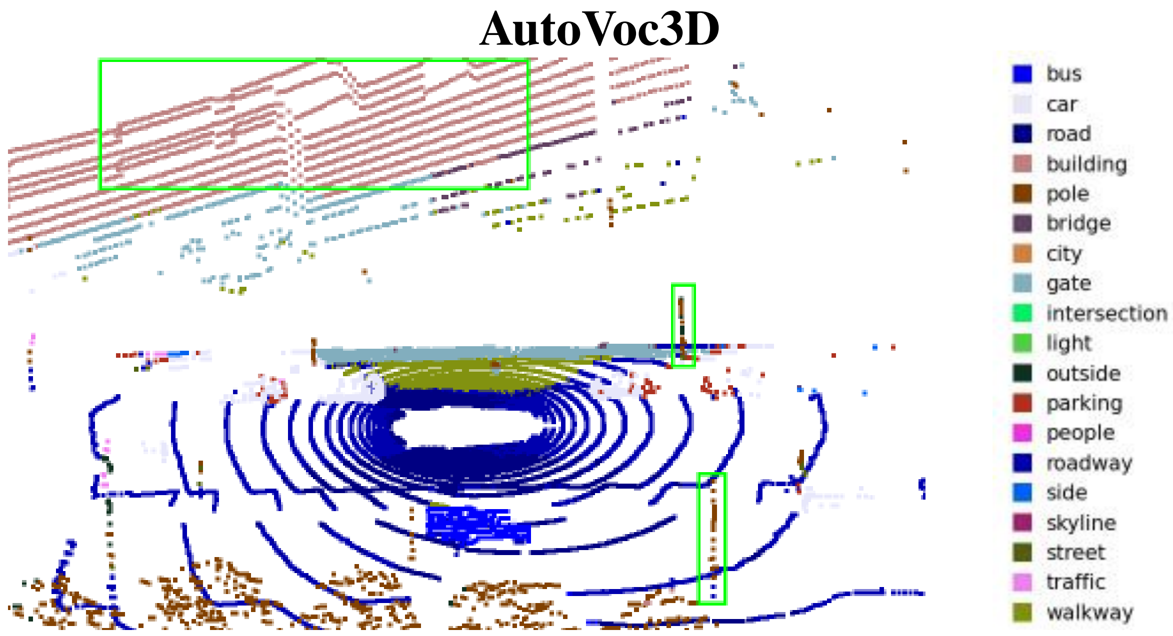
Known & Fixed Vocabulary | **Known & Open Vocabulary** | **Unknown & Open Vocabulary**



3D Auto-Vocabulary Segmentation for LiDAR

[Wie, Ülger, **Karimi** Gevers, Oswald, Arxiv'23]





Conclusion and Take-away

- 3D / 4D computer vision algorithms train faster and require less training data (vs. 2D)
- 3D modeling, but 2D supervision
- Scene understanding requires memory
- Photographic and deformable memory improves accuracy & enables new applications
- Self-supervised learning via re-rendering error minimization
- Scene representation is important (local updates, deformable, catastrophic forgetting)
- SLAM can be a useful stepping stone for continual scene understanding

Future Directions

- Beyond semantics: multi-modal output open-vocabulary & foundation models
- 3D-Language maps and spatial language-based reasoning
- Learning and controlling forgetting (keeping track of task-relevant changes)
- Collaborative / distributed asynchronous learning with multiple agents
- Physics-based scene representations (metric units, weights, gravity, etc.)
- 3D generative multi-modal models
- Dynamic scenes and temporal representations