SCONE:

Surface Coverage Optimization in uNknown Environments by Volumetric Integration

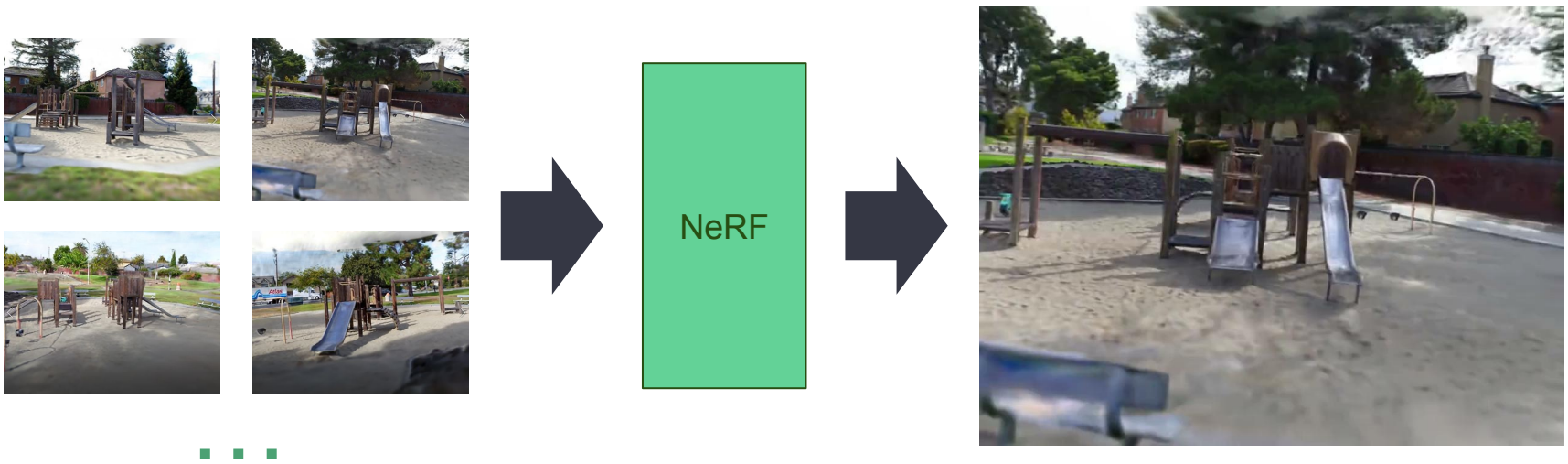
Antoine Guédon, Pascal Monasse, Vincent Lepetit

LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, France

Introduction

3D Scene Rendering from images

Yu *et al.*, Plenoxels: Radiance Fields without Neural Networks (CVPR 2022)



3D Reconstruction: NeRF-related models, SfM and MVS (COLMAP), etc.

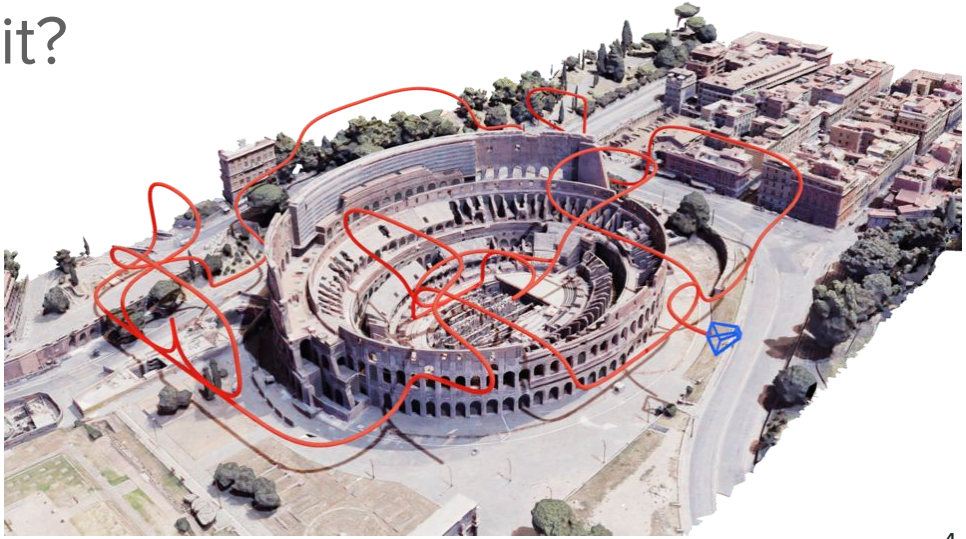
3D Scene Rendering from images

Questions:

1. How do we acquire these images?
2. Can we learn how to do it?

➔ Path Planning

➔ Next Best View (NBV)



Next Best View (NBV)

- **Long standing problem** in robotics
- **Depth sensors** are commonly used (video stream, so even with RGB we consider we have access to partial point clouds)
- **Classic approaches**: Usually, discretization into known and unknown voxels, then ray-casting entropy metrics. Very slow, not suited to real-time exploration.
- **Surface Coverage** is a common metric to evaluate NBV

Next Best View (NBV)

What about learning-based approaches?

- → **Volumetric approaches**: Rely on voxelization. Suited to path planning in scenes, but not scalable and far less precision.
- → **Surface approaches**: Work directly on the dense point cloud. Still rely on a single global encoding to regress NBV => Only suited to floating, small scale, single objects inside a sphere.

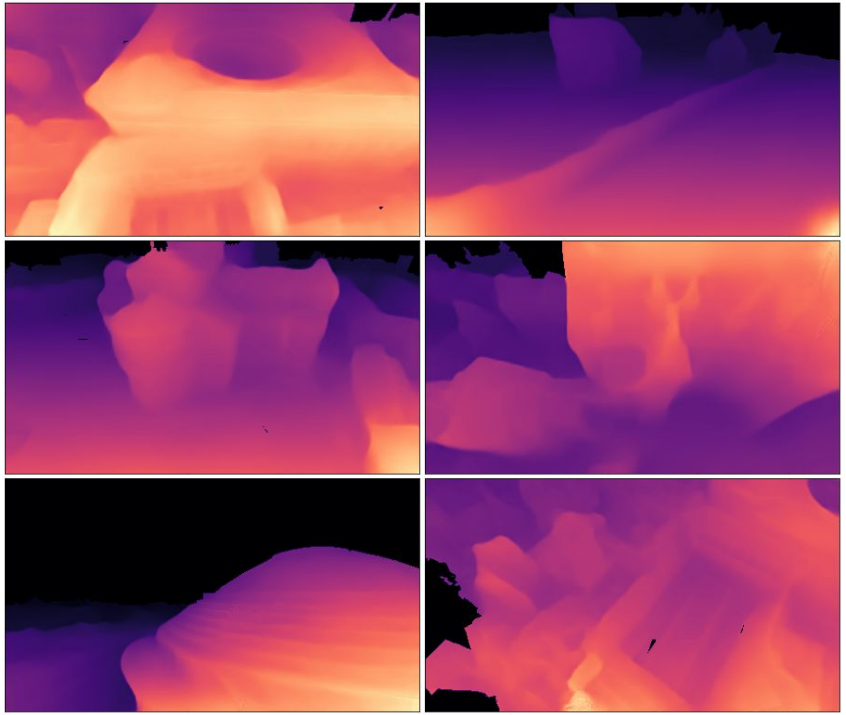
Our approach: SCONE

Input

1. Partial point clouds gathered by a LiDAR-class sensor or reconstructed from an RGB video stream (see DM)

2. Camera poses

3. Bounding Box (if needed, to delimit exploration area)



Strategy

SCONE: Surface Coverage Optimization in Unknown Environments by Volumetric Integration (NeurIPS 2022)

1. Make a **prediction about unseen geometry** in the scene
2. **Predict visibility of surface points** you can't see right now

NBV = Camera with the most new visible surface points
(highest coverage gain)

Predict unseen geometry

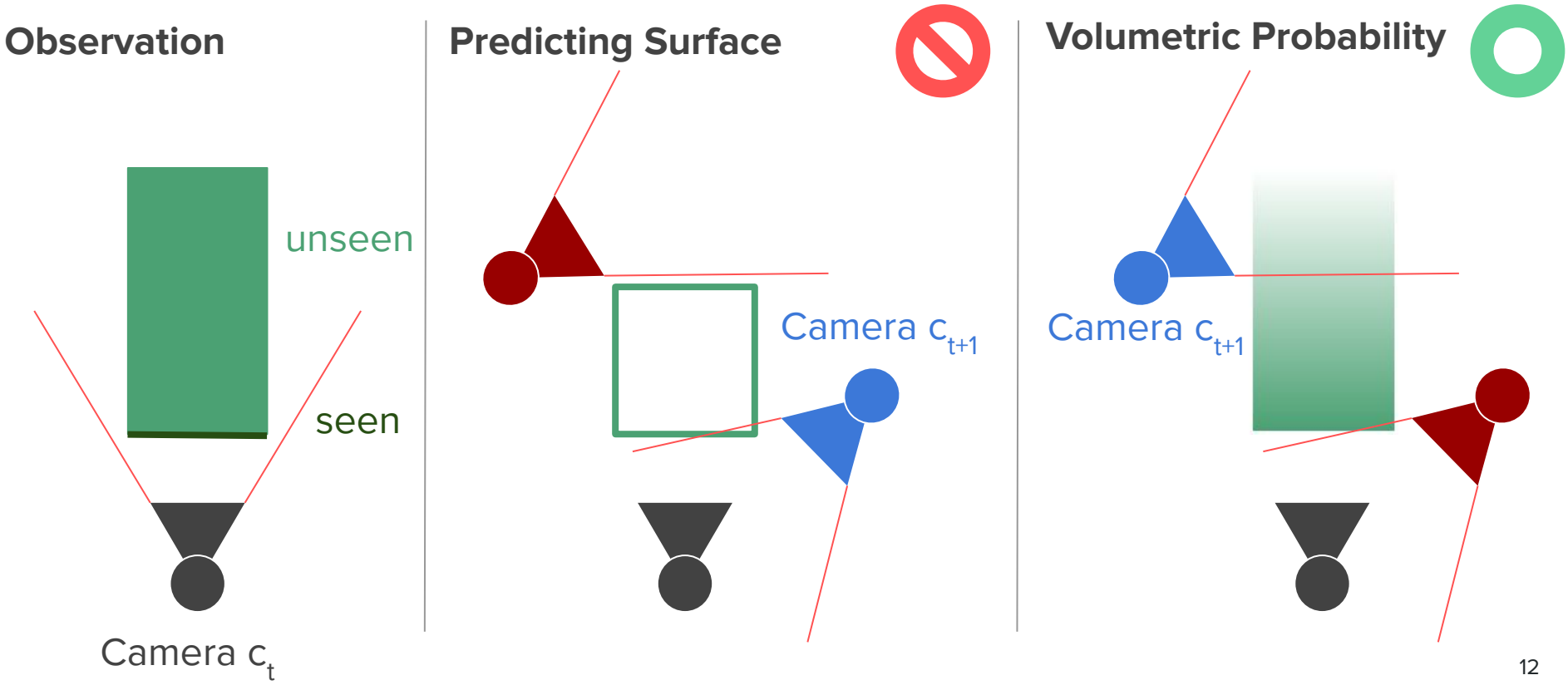
Naive approach

1. Predict unseen surface
2. Compute surface coverage for any camera from this surface.

In practice, does not work in unknown environments.

Predicting a surface requires high confidence in prediction.

The need for a volumetric, probabilistic mapping



Predict an Occupancy Probability Field

Our needs:

1. **Virtually infinite resolution**

Deep Implicit Function (very trendy in CV)

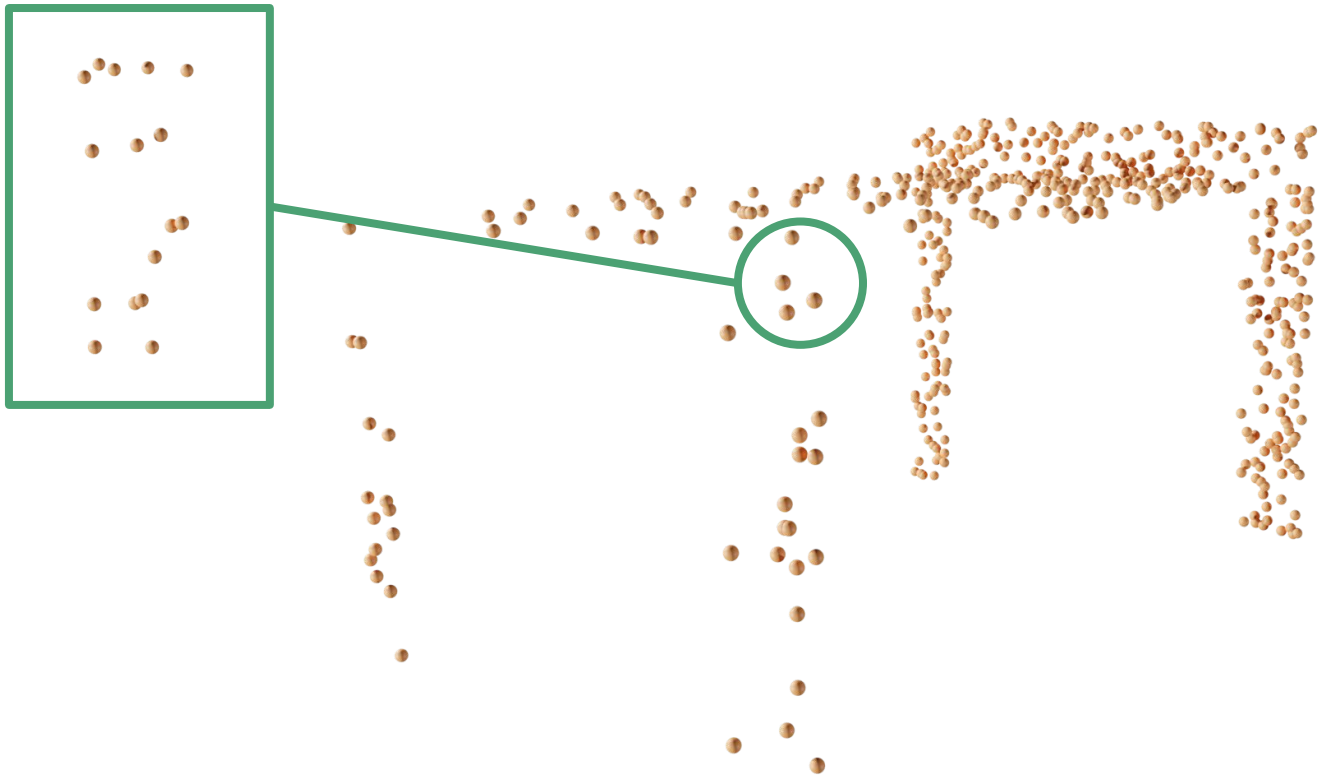
2. **Generalization and Scalability**

Focus on **local features** rather than a single global encoding

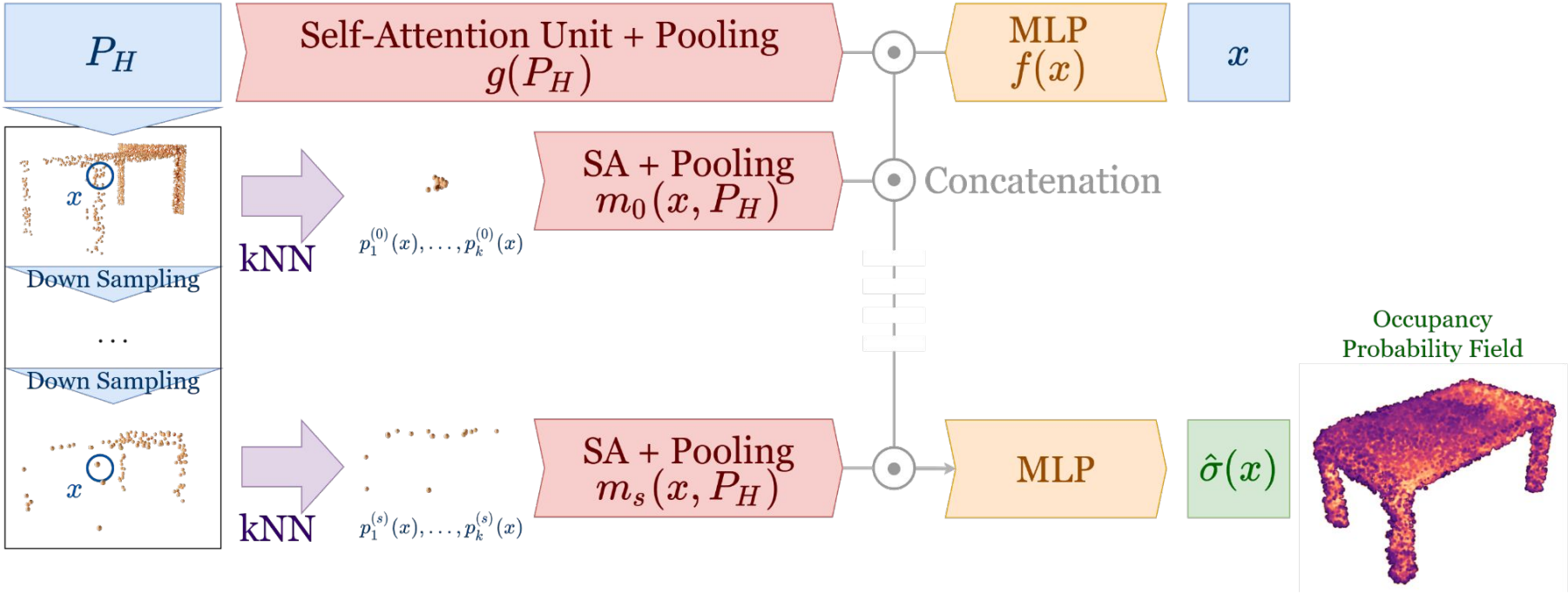
Predict an Occupancy Probability Field

Predict an Occupancy Probability Field

Predict an Occupancy Probability Field



Predict an Occupancy Probability Field



Supervision: MSE between predicted occupancy probabilities and GT occupancy map

Estimate surface coverage gain

Maximizing Surface Coverage with Volumetric Integration

Surface Coverage is a surface metric. Yet, we want to integrate on a volumetric occupancy probability field.

➔ “Relax” the definition of visibility to a volumetric visibility field

➔ Asymptotically, maximizing the volumetric integral is equivalent to maximizing the surface integral

Theorem 1. *Under the previous regularity assumptions on the volume χ of the scene and its surface $\partial\chi$, there exist $\mu_0 > 0$ and $M > 0$ such that for all $\mu < \mu_0$, and any camera $c \in \mathcal{C}$:*

$$\left| \frac{1}{|\chi|_V} \int_\chi g_c^H(\mu; x)dx - \mu \frac{|\partial\chi|_S}{|\chi|_V} G_H(c) \right| \leq M\mu^2 , \quad (6)$$

where $|\chi|_V$ is the volume of χ.

More details in the paper.

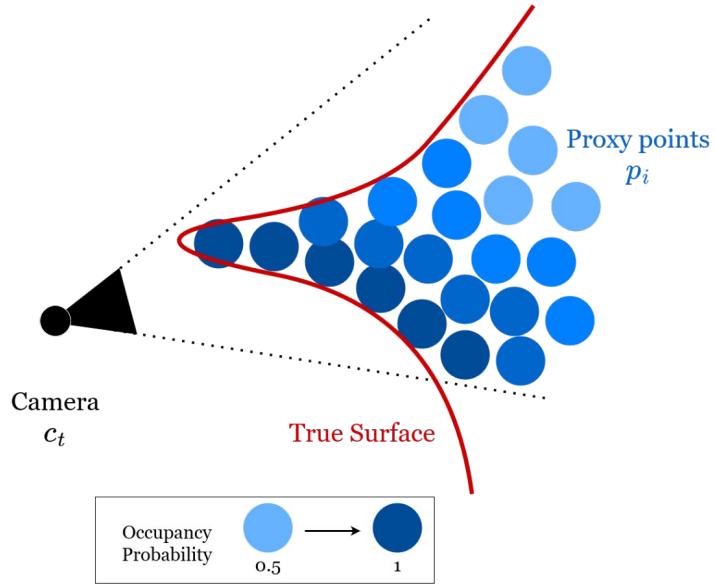
How?

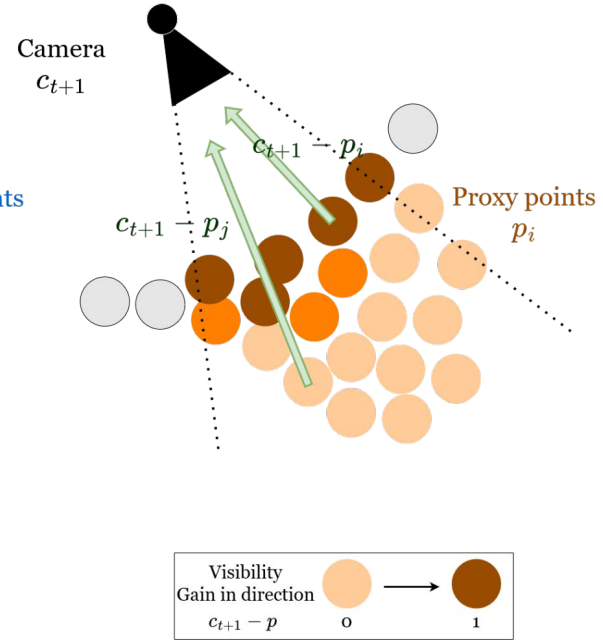
We want to avoid ray-casting operations.

Instead, we use directional, attention-based features between proxy points!

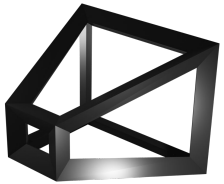
Predicting Visibility Gain

1. **Sample *Proxy Points* in the scene**, based on their occupancy probability
2. For any new camera pose *c*, we **predict a Visibility Gain Score** for each *probabilistic* proxy point

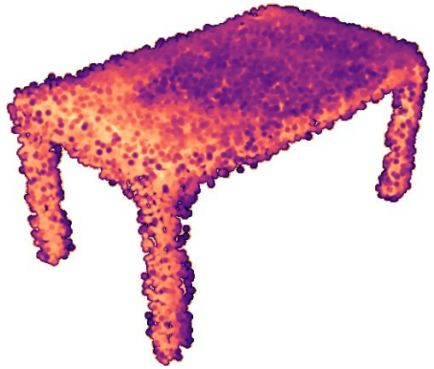




Predicting Visibility Gain: Input

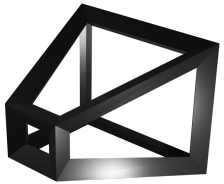


Camera c_t

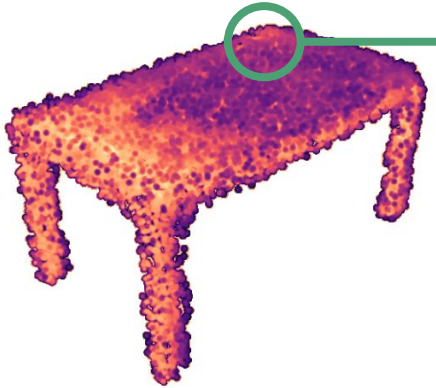


Proxy Points with
Occupancy Probability

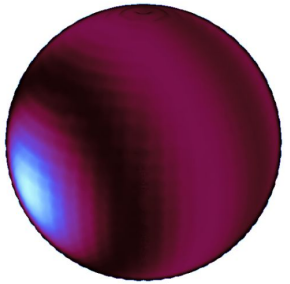
Predicting Visibility Gain: Input



Camera c_t

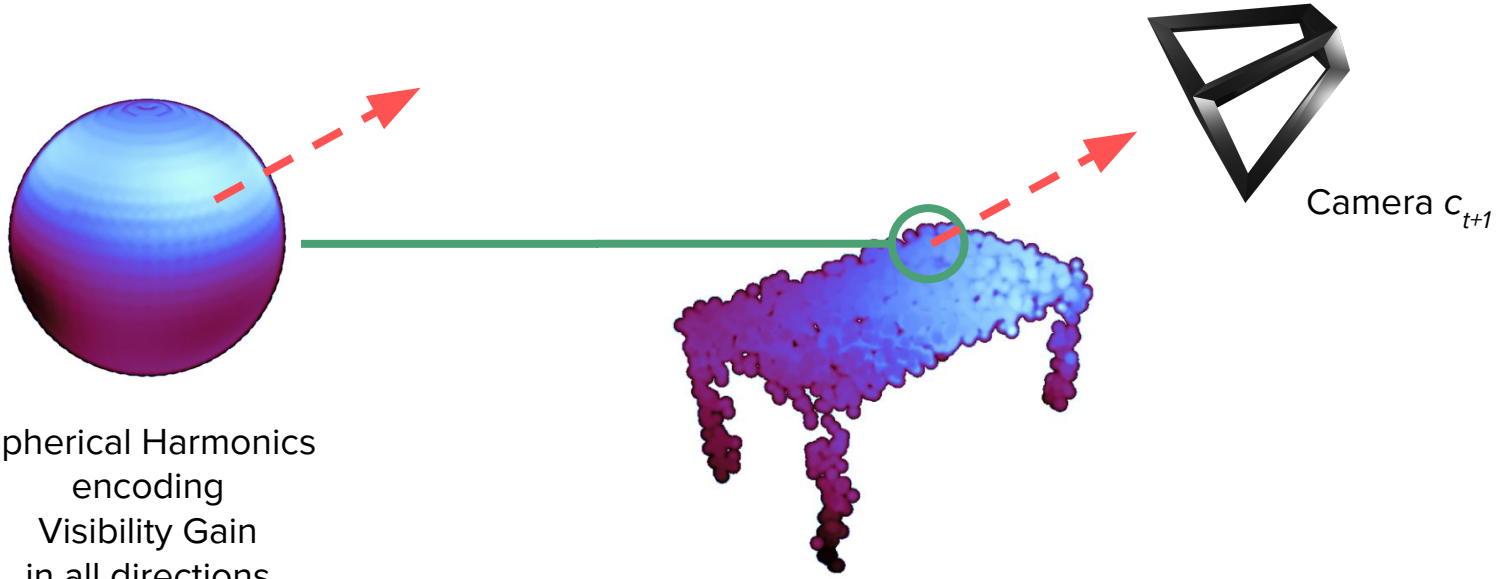


Proxy Points with
Occupancy Probability



Spherical Harmonics
encoding
Camera History

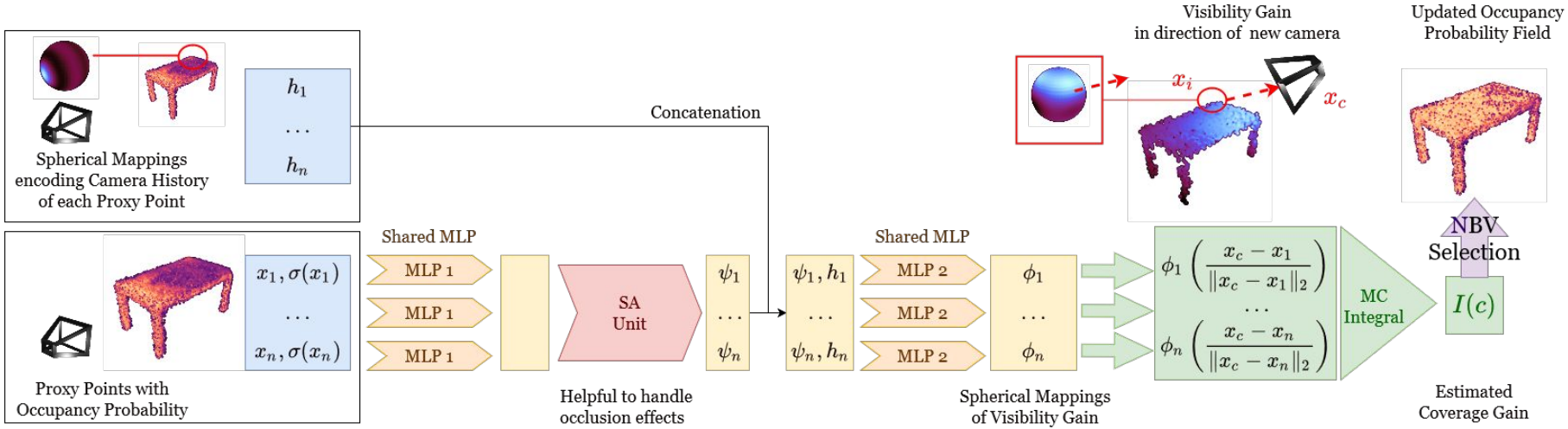
Predicting Visibility Gain: Output



Spherical Harmonics
encoding
Visibility Gain
in all directions

Proxy Points with Visibility Gain
in direction of new camera

Predicting Coverage Gain



Supervision: Softmax then KL Divergence between predicted volumetric coverage gains and GT surface coverage gains, over a distribution of cameras

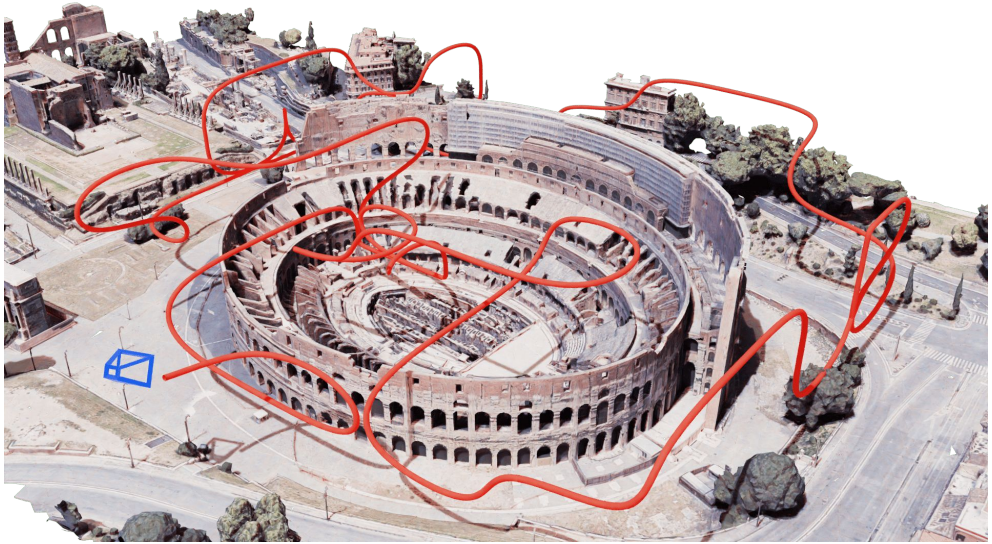
Results

Single Object Reconstruction: Quantitative Results

	Categories seen during training								
Method	Airplane	Cabinet	Car	Chair	Lamp	Sofa	Table	Vessel	Mean
Random	0.745	0.545	0.542	0.724	0.770	0.589	0.710	0.674	0.662
Proximity Count [8]	0.800	0.596	0.591	0.772	0.803	0.629	0.753	0.706	0.706
Area Factor [25]	0.797	0.585	0.587	0.751	0.801	0.627	0.725	0.714	0.698
NBV-Net [16]	0.778	0.576	0.596	0.743	0.791	0.599	0.693	0.667	0.680
PC-NBV [36]	0.799	0.612	**0.612**	**0.782**	0.800	0.640	0.760	0.719	0.716
SCONE (Ours)	**0.827**	**0.625**	0.591	**0.782**	**0.819**	**0.662**	**0.792**	**0.734**	**0.729**

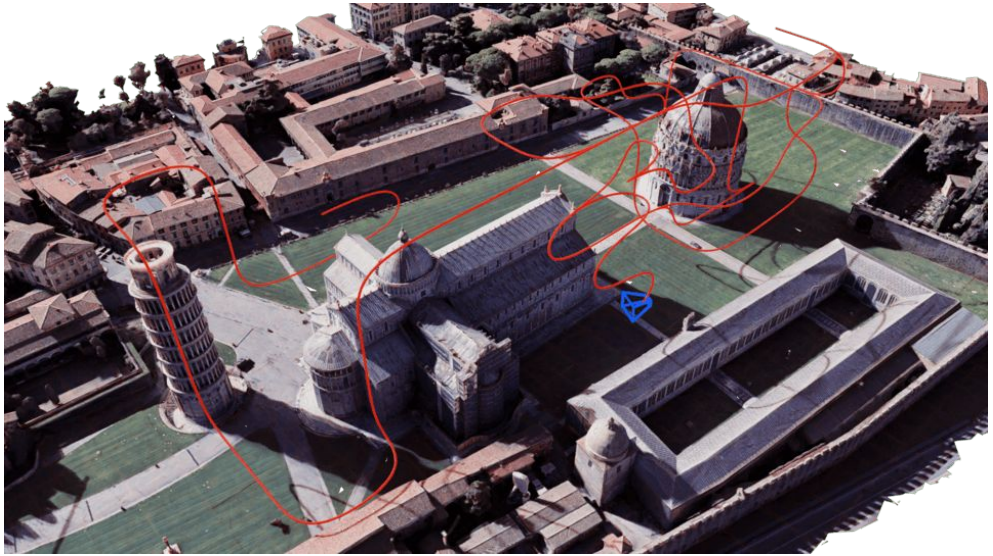
	Categories not seen during training								
Method	Bus	Bed	Bookshelf	Bench	Guitar	Motorbike	Skateboard	Pistol	Mean
Random	0.609	0.619	0.695	0.795	0.795	0.672	0.768	0.614	0.694
Proximity Count	0.646	0.645	**0.749**	0.829	0.854	0.705	0.828	0.660	0.740
Area Factor	0.629	0.631	0.742	0.827	0.852	0.718	0.799	0.660	0.732
NBV-Net	0.654	0.628	0.729	0.824	0.834	0.710	0.825	0.645	0.731
PC-NBV	0.667	0.662	0.740	**0.845**	0.849	**0.728**	0.840	0.672	0.750
SCONE (Ours)	**0.694**	**0.689**	0.746	0.832	**0.860**	**0.728**	**0.845**	**0.717**	**0.764**

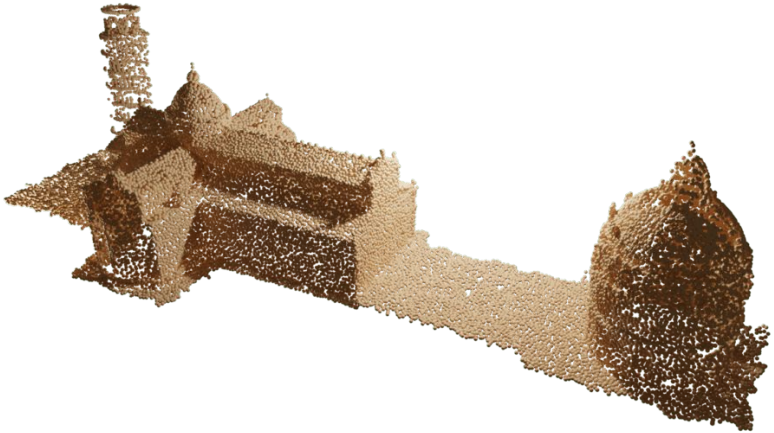
Active View Planning in 3D scenes: Qualitative Results



Colosseum

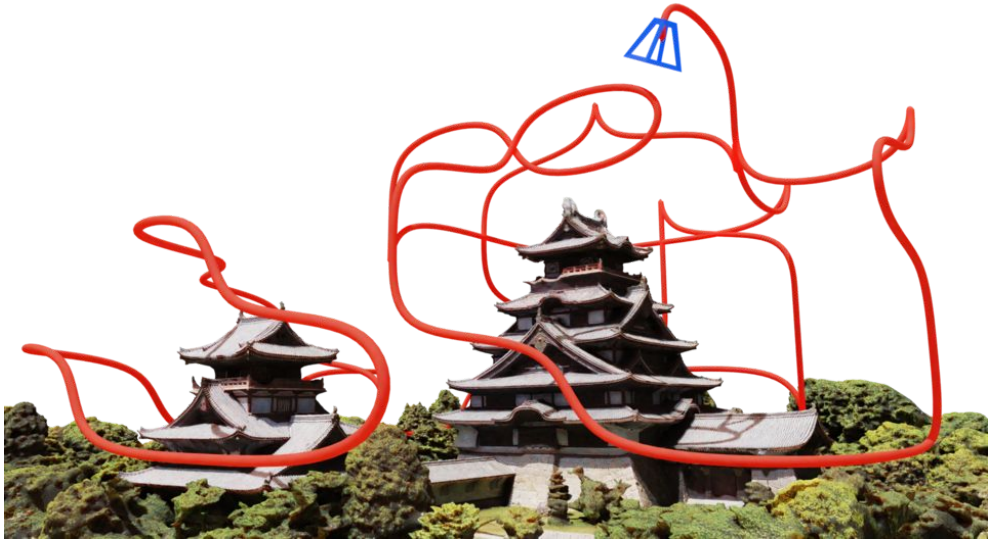
Active View Planning in 3D scenes: Qualitative Results





Pisa Cathedral

Active View Planning in 3D scenes: Qualitative Results

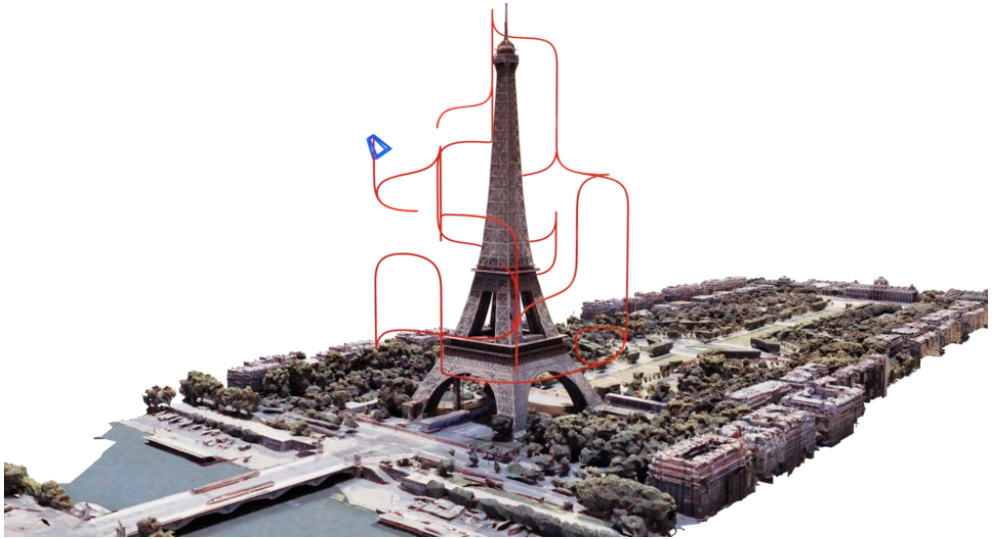


Fushimi Castle

Active View Planning in 3D scenes: Qualitative Results

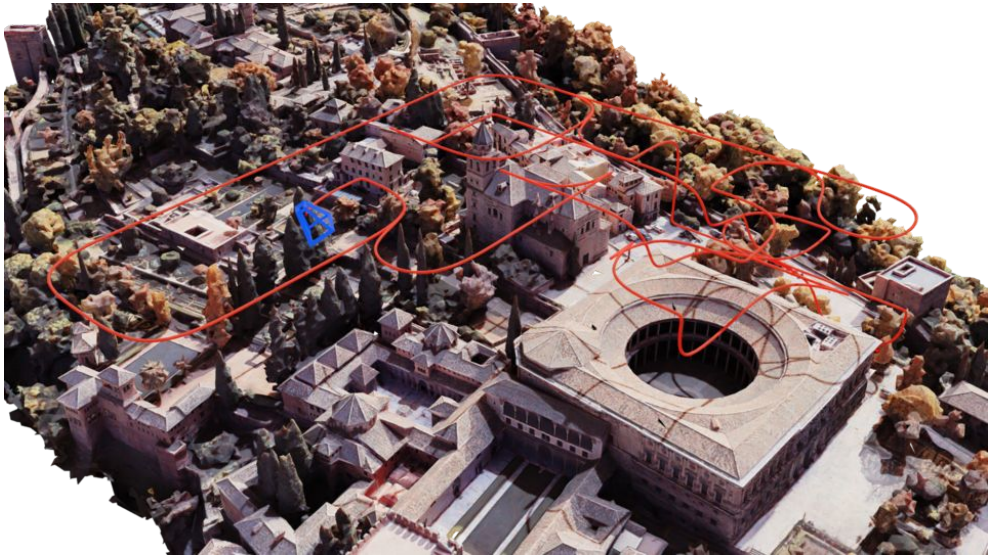
Christ the Redeemer

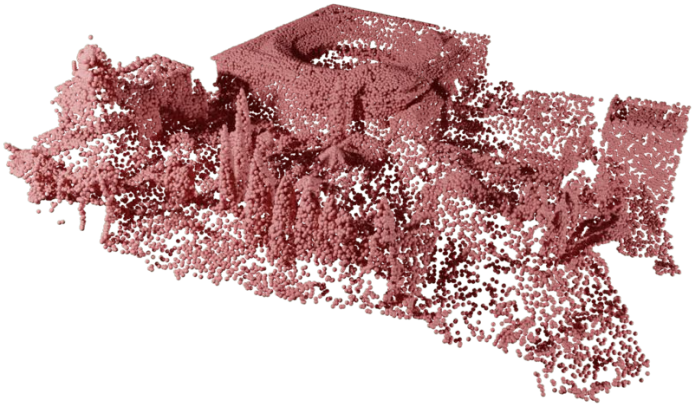
Active View Planning in 3D scenes: Qualitative Results



Eiffel Tower

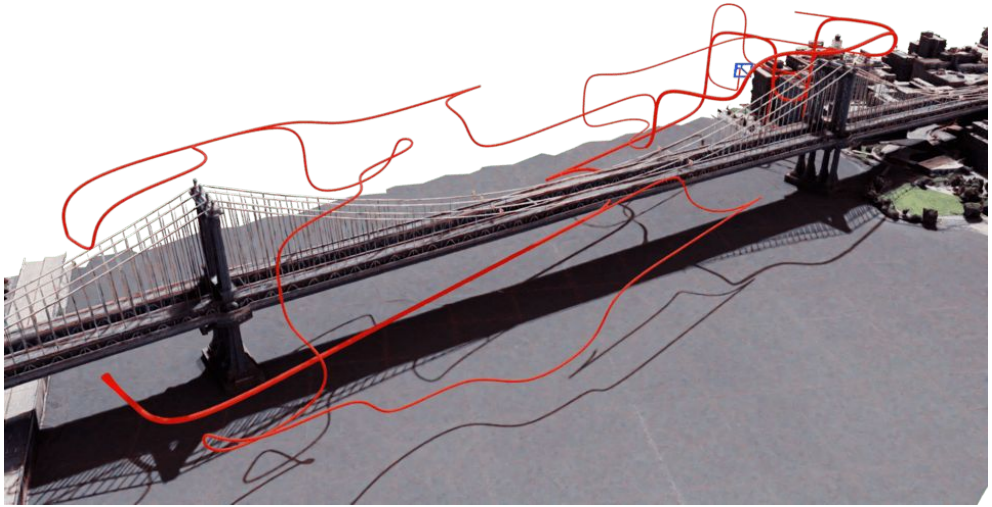
Active View Planning in 3D scenes: Qualitative Results

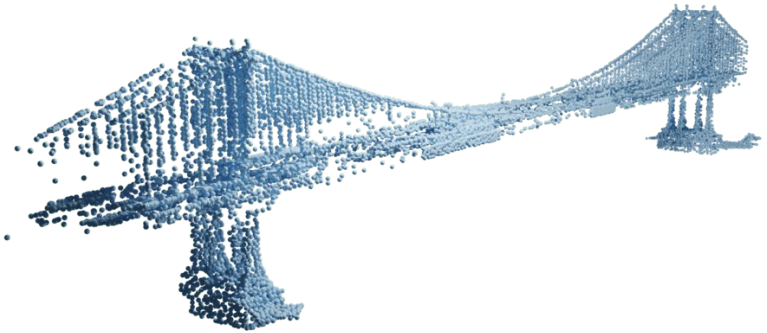




Alhambra Palace

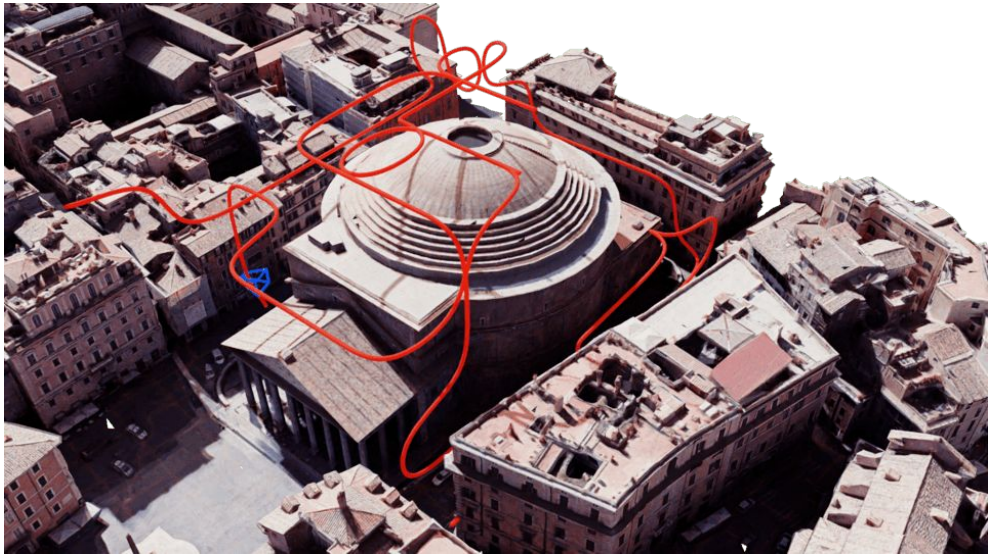
Active View Planning in 3D scenes: Qualitative Results





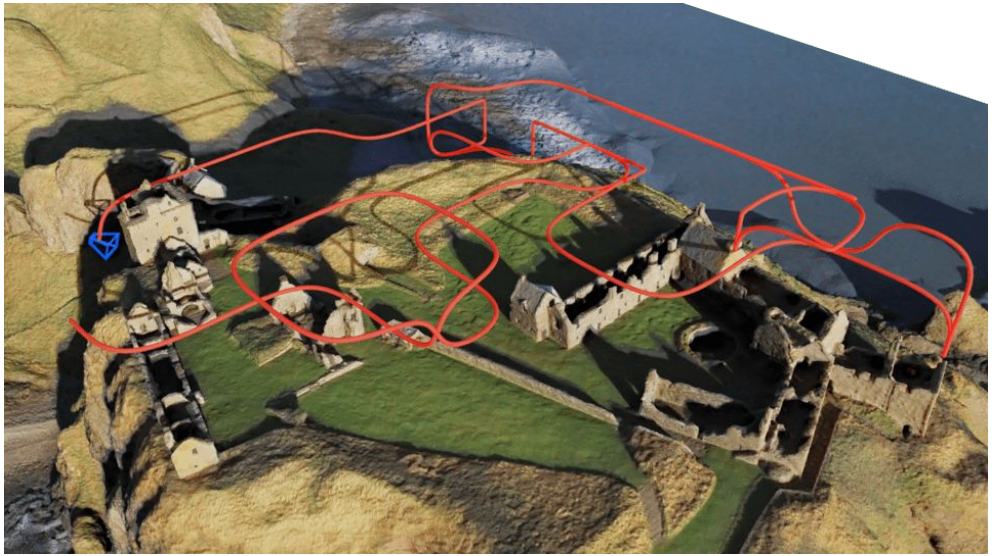
Manhattan Bridge

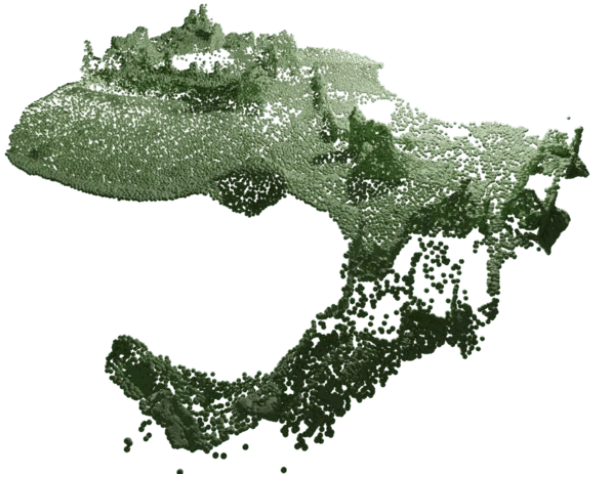
Active View Planning in 3D scenes: Qualitative Results



Pantheon (Rome, not Paris hehe)

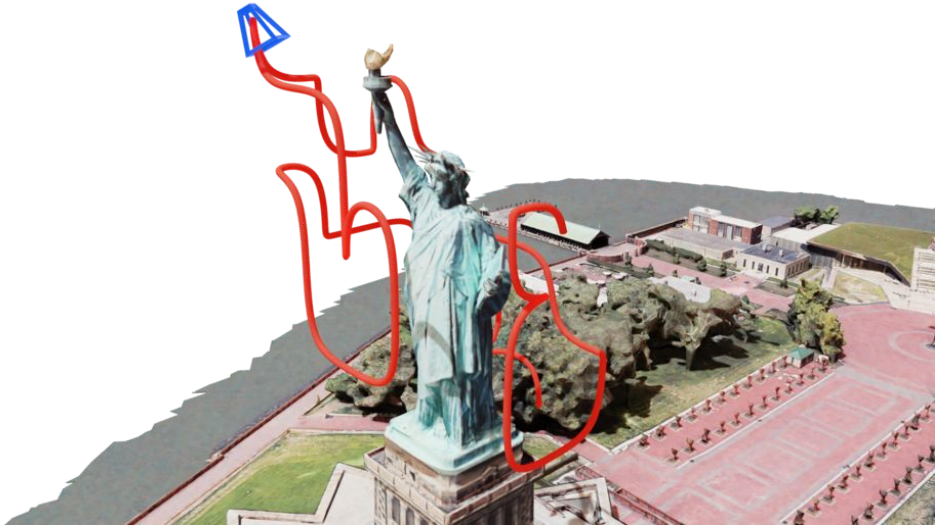
Active View Planning in 3D scenes: Qualitative Results





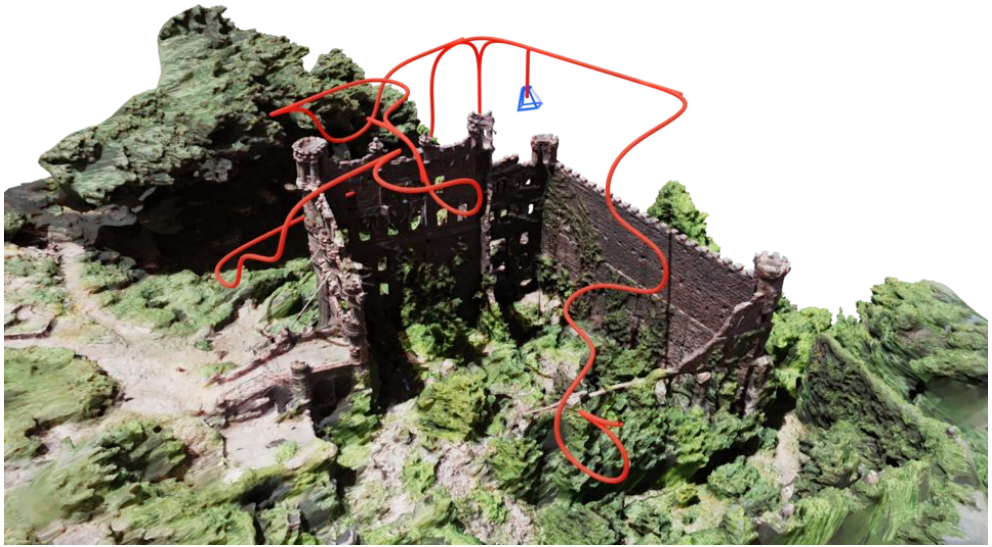
Dunnotar Castle

Active View Planning in 3D scenes: Qualitative Results



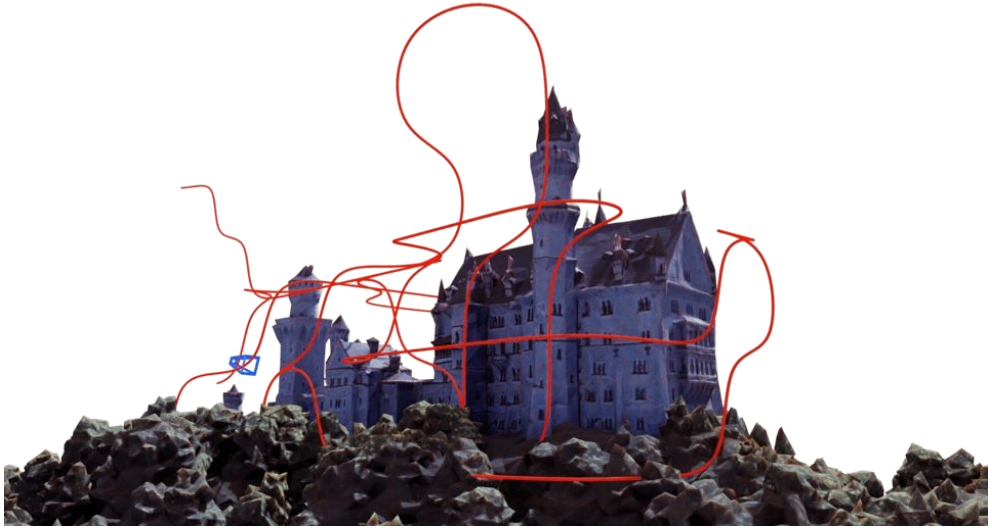
Statue of Liberty

Active View Planning in 3D scenes: Qualitative Results



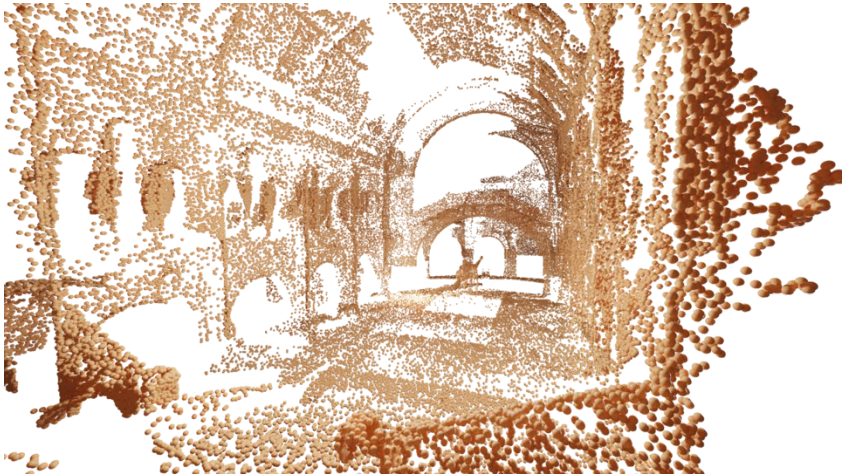
Bannerman Castle

Active View Planning in 3D scenes: Qualitative Results



Neuschwanstein Castle

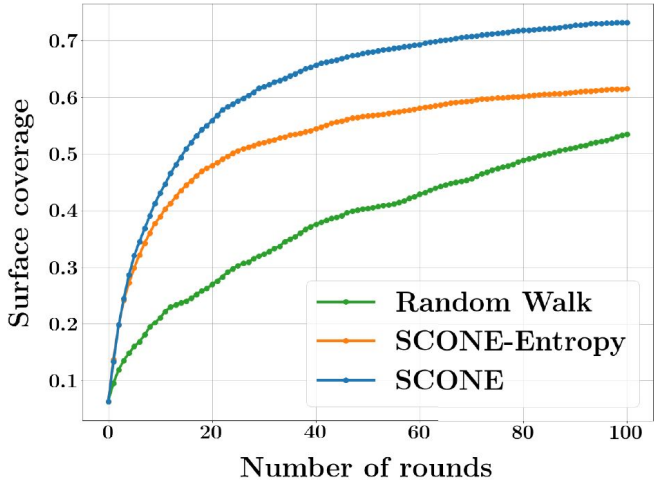
Active View Planning in 3D scenes: Qualitative Results



Natural History Museum, London

Active View Planning in 3D scenes: Quantitative Results

	Method		
3D scene	Random Walk	SCONE-Entropy	SCONE
Dunnottar Castle	0.355 ± 0.106	0.456 ± 0.041	**0.739** ± 0.050
Manhattan Bridge	0.405 ± 0.089	0.361 ± 0.065	**0.685** ± 0.034
Alhambra Palace	0.384 ± 0.086	0.437 ± 0.047	**0.567** ± 0.031
Leaning Tower	0.286 ± 0.122	0.415 ± 0.023	**0.542** ± 0.026
Neuschwanstein Castle	0.403 ± 0.032	0.538 ± 0.040	**0.653** ± 0.025
Colosseum	0.308 ± 0.061	0.512 ± 0.024	**0.571** ± 0.024
Eiffel Tower	0.495 ± 0.062	0.741 ± 0.017	**0.762** ± 0.020
Fushimi Castle	0.584 ± 0.078	0.802 ± 0.022	**0.841** ± 0.027
Pantheon	0.175 ± 0.065	0.351 ± 0.020	**0.396** ± 0.036
Bannerman Castle	0.321 ± 0.121	**0.667** ± 0.023	0.642 ± 0.047
Christ the Redeemer	0.600 ± 0.146	0.839 ± 0.038	**0.859** ± 0.022
Statue of Liberty	0.469 ± 0.075	0.681 ± 0.018	**0.693** ± 0.032
Natural History Museum	0.147 ± 0.024	0.080 ± 0.010	**0.177** ± 0.031
Mean	0.380	0.529	**0.625**



Thank you!