iMAP: Implicit Mapping and Positioning in Real-Time (ICCV 2021)

DAVIAN Vision Seminar / 2022.04.11 / 배광탁

Task

<u>RGB-D SLAM</u> with continual learning

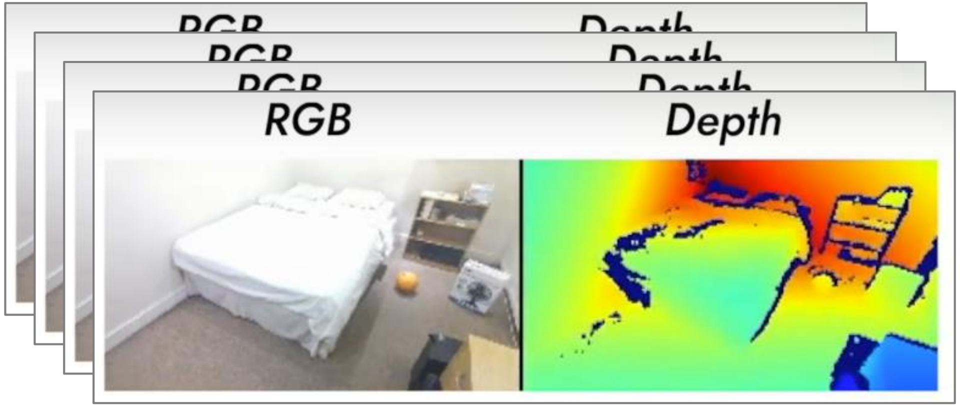
(SLAM : Simultaneous Localization and Mapping)

Azure Kinect

iPhone 12 Pro

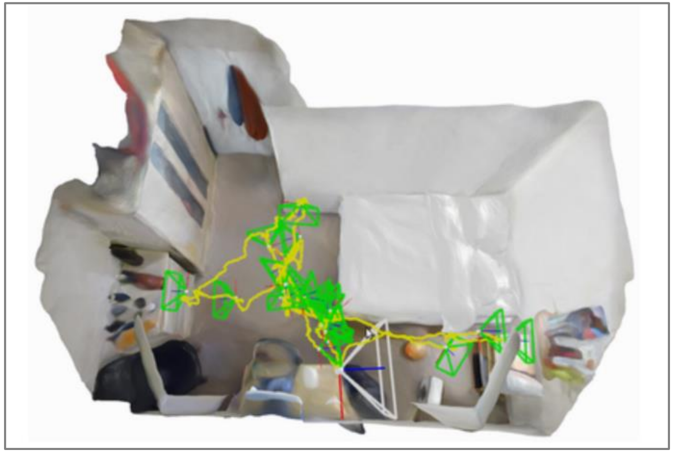
Input

<u>RGB-D Image Sequences</u>



Output

<u>Camera Poses and 3D Scene Structure</u>



Contribution

1. [Joint Optimisation]

: *Jointly* optimising *a full 3D map and camera poses* by using *implicit neural scene representation*

→ ability of INR, memory efficient scene modeling

2. [Active Sampling]

: *Incrementaly* training an implicit scene network in *real-time*

→ pratical techniques of INR

Method 1. Joint Optimisation

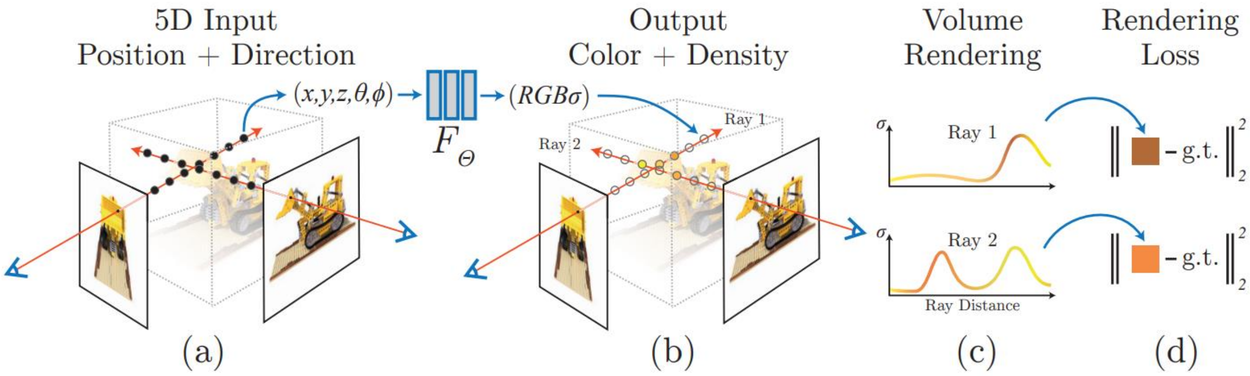
Preliminaries : NeRF[1]

Task : novel view synthesis

Input : 3D point coordinates and viewing direction

Output : color and volume density

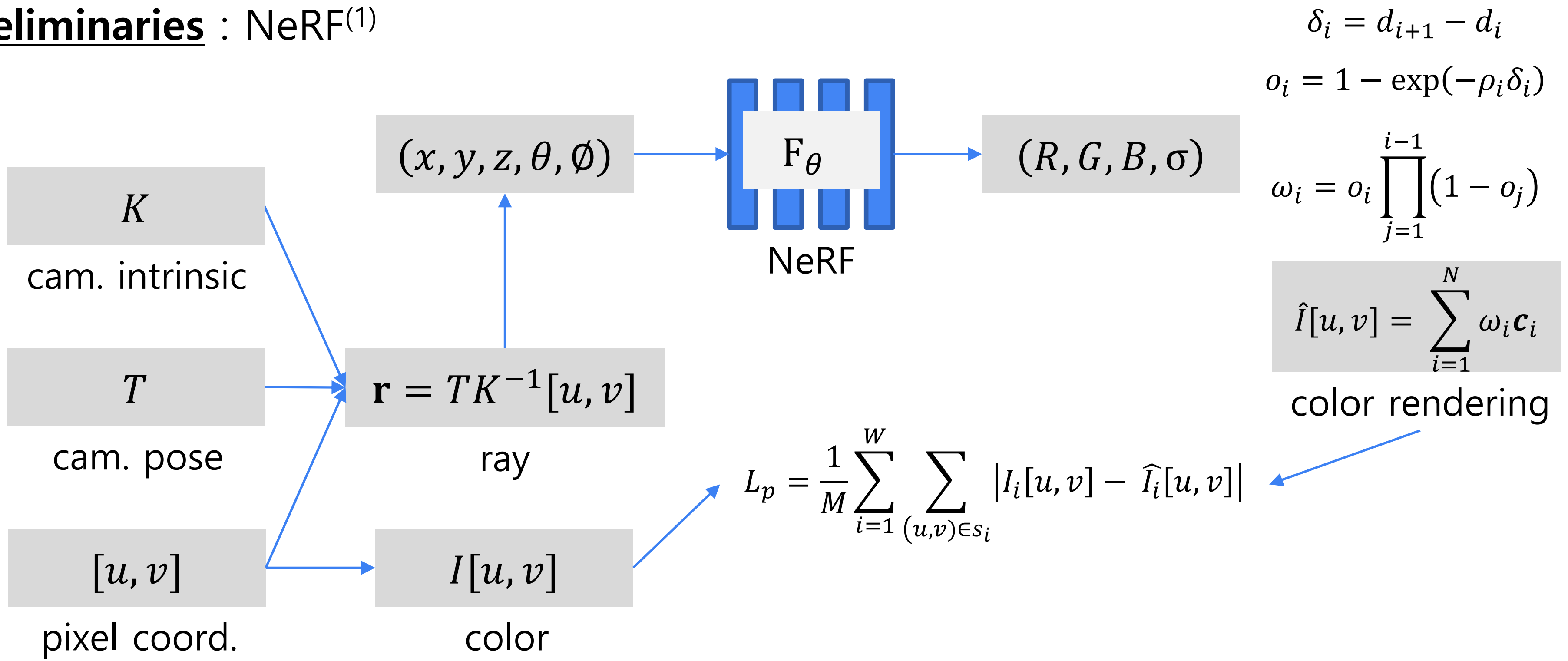
Method : training MLP with images from sparse set of views



(1) Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *European conference on computer vision*. Springer, Cham, 2020.

Method 1. Joint Optimisation

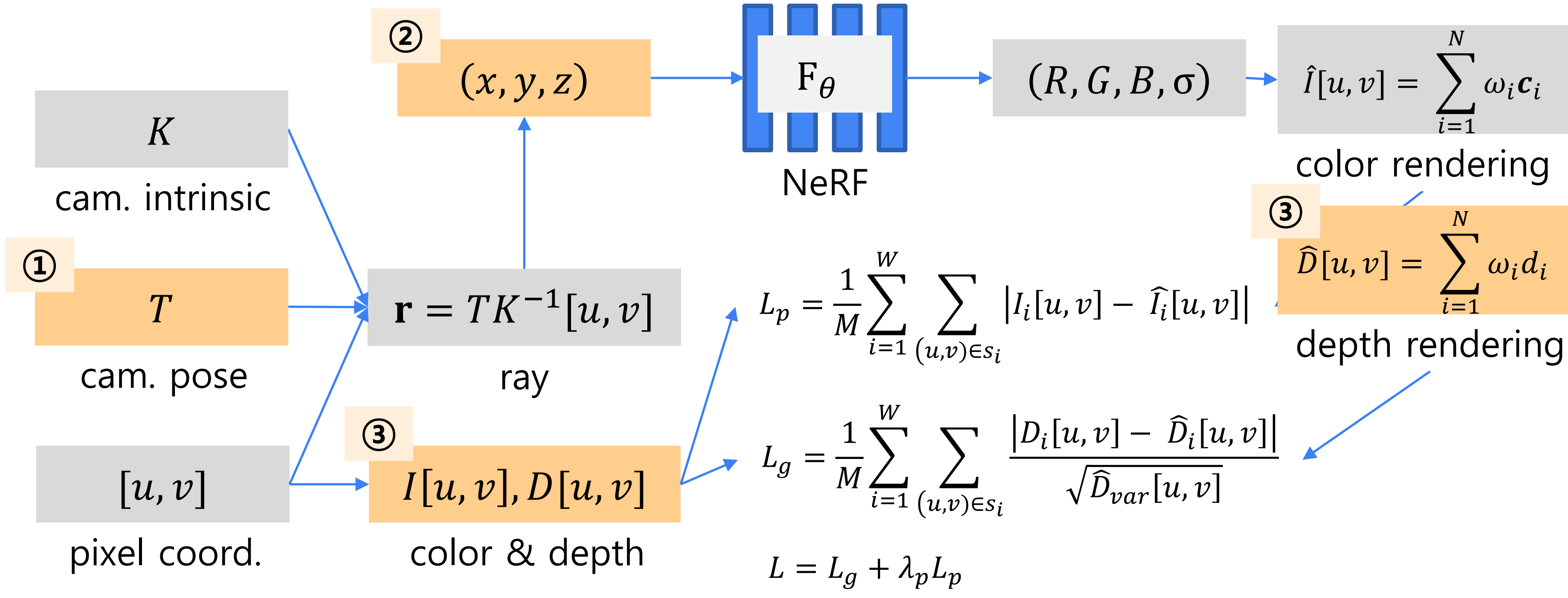
Preliminaries : NeRF[1]



(1) Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *European conference on computer vision*. Springer, Cham, 2020.

Method 1. Joint Optimisation

iMAP



Method 1. Joint Optimisation

iMAP

(x, y, z) → F_θ → (R, G, B, σ) →

NeRF

$$\hat{I}[u,v] = \sum_{i=1}^{N} \omega_i \boldsymbol{c}_i$$

color rendering

$$\hat{D}[u,v] = \sum_{i=1}^{N} \omega_i d_i$$

depth rendering

$$L_g = \frac{1}{M} \sum_{i=1}^{W} \sum_{(u,v) \in s_i} \frac{\left| D_i[u,v] - \hat{D}_i[u,v] \right|}{\sqrt{\hat{D}_{var}[u,v]}}$$

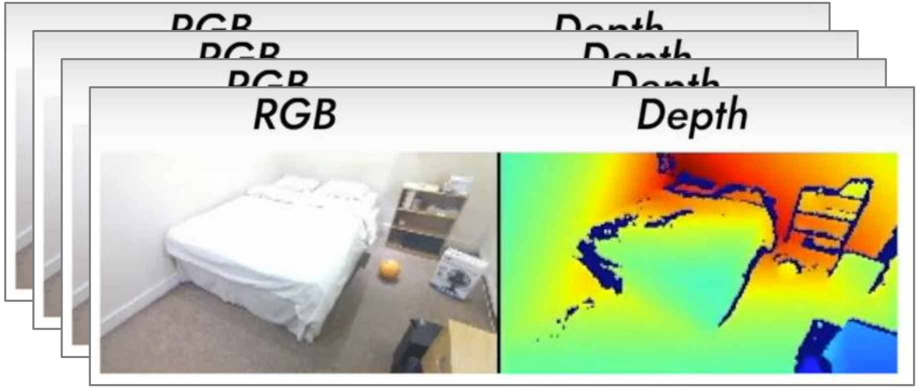
$$\hat{D}_{var}[u,v] = \sum_{i=1}^{N} \omega_i \left(\hat{D}[u,v] - d_i \right)^2$$

Method 2. Active Sampling

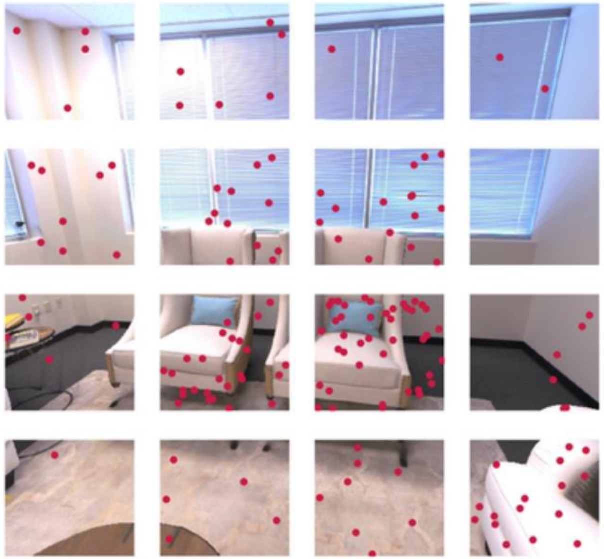
Keyframe Selection → continual learning

Input

RGB-D Image Sequences



Active Sampling → practical INR



Method 2. Active Sampling

<u>Keyframe Selection</u> → continual learning

<u>RGB-D Image Sequences</u> → *<u>Registered keyframes</u>*

T=0

T=1

⋮

T=N-1

T=N

$$P = \frac{1}{|s|} \sum_{(u,v)\in s} \mathbb{1}\left(\frac{\left|D[u,v] - \hat{D}[u,v]\right|}{D[u,v]} < t_D\right)$$

$$P < t_P$$

<u>T=0</u>

T=3

⋮

T=N-7

<u>T=N</u>

Method 2. Active Sampling

<u>**Image Active Sampling**</u> → practical INR

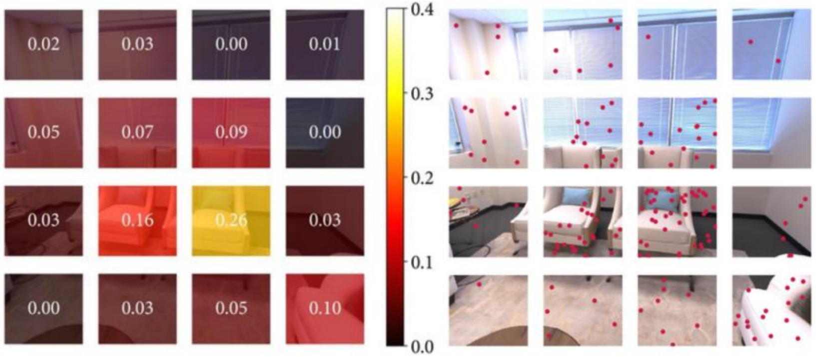
uniform sampling → loss statistics → probability distribution → re-sample

$$L_i[j] = \frac{1}{|r_j|} \sum_{(u,v)\in r_j} \left|I_i[u,v] - \widehat{I_i}[u,v]\right| + \left|D_i[u,v] - \widehat{D_i}[u,v]\right|$$

$$f_i[j] = \frac{L_i[j]}{\sum_{m=1}^{64} L_i[m]}$$

$$n_i \cdot f_i[j]$$

uniform samples per region



Method 2. Active Sampling

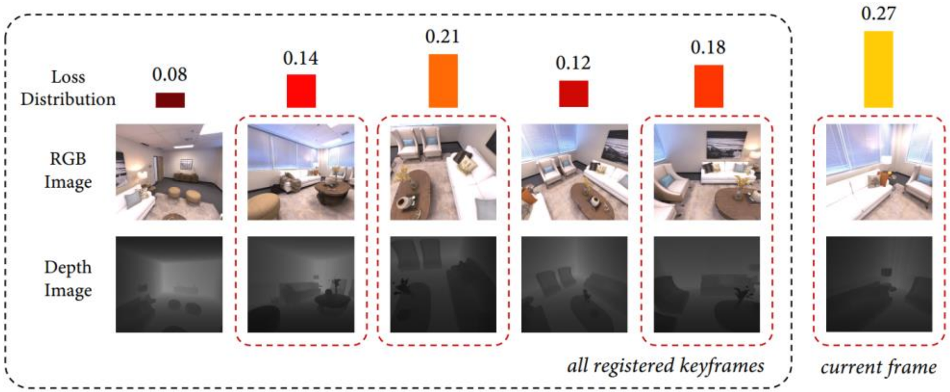
Keyframe Active Sampling → newly explored, highly detailed, or started to forget

Registered keyframes

T=0		n_0
T=3		n_3
⋮	⋮	⋮
T=N-7		n_{N-7}
T=N		n_N

loss distribution across keyframes

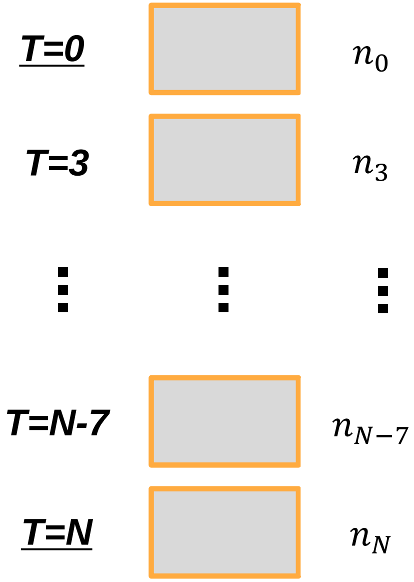
→ different number of samples n_i for each keyframes,



Method 2. Active Sampling

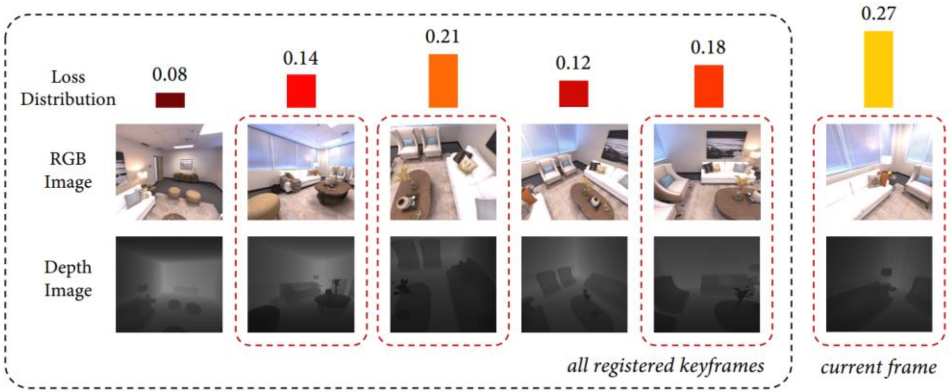
Bounded Keyframe Selection → bound joint optimisation computation

Registered keyframes



bounded window with constantly changing frames

→ *W-2* randomly sampled + *1* last keyframe + *1* current live frame



Experimental Results

Dataset : Replica dataset(simulated), Azure Kinect RGB-D, TUM RGB-D dataset

Metric : Accuracy, Completion, Completion Ratio, ATE RMSE

Quantitative Results :

		room-0	room-1	room-2	office-0	office-1	office-2	office-3	office-4	Avg.
iMAP	**# Keyframes**	11	12	12	10	11	10	14	11	13.37
	Acc. [cm]	3.58	3.69	4.68	5.87	3.71	4.81	4.27	4.83	4.43
	Comp. [cm]	5.06	4.87	5.51	6.11	5.26	5.65	5.45	6.59	**5.56**
	Comp. Ratio [< 5cm %]	83.91	83.45	75.53	77.71	79.64	77.22	77.34	77.63	**79.06**
TSDF Fusion	**Acc.** [cm]	4.21	3.08	2.88	2.70	2.66	4.27	4.07	3.70	**3.45**
	Comp. [cm]	5.04	4.35	5.40	10.47	10.29	6.43	6.26	4.78	6.63
	Comp. Ratio [< 5cm %]	76.90	79.87	77.79	79.60	71.93	71.66	65.87	77.11	75.09

Table 1: Reconstruction results for 8 indoor Replica scenes. We report the highest reached completion ratio in each scene along with the corresponding accuracy and completion values at that point.

iMAP [MB]	Width = 128 0.26	Width = 256 1.04	Width = 512 4.19
TSDF Fusion [MB]	Res. = 128 8.38	Res. = 256 67.10	Res. = 512 536.87

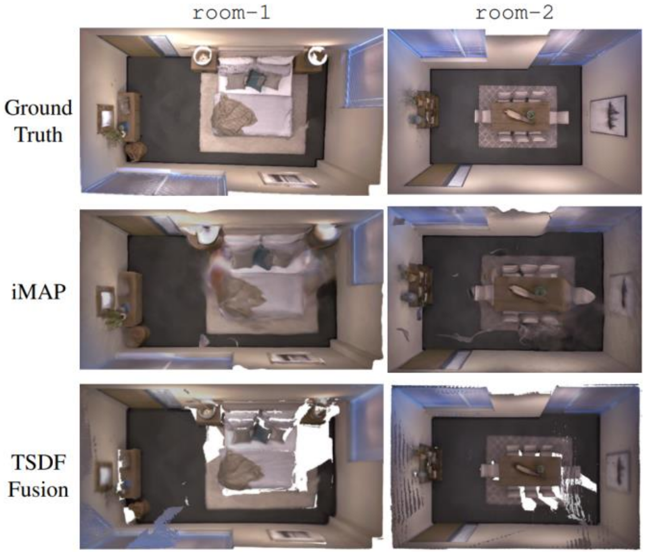
Table 2: Memory consumption: for iMAP as a function of network size, and for TSDF fusion of voxel resolution.

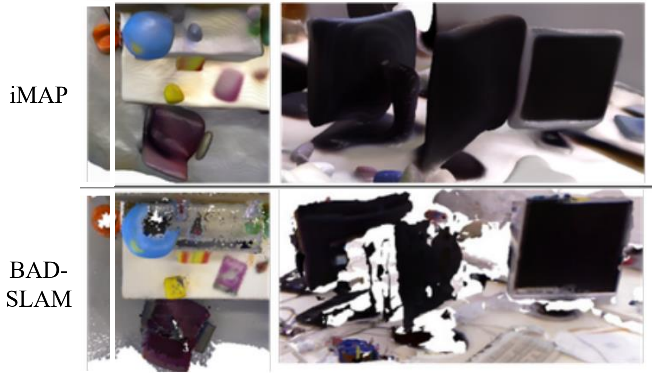
	fr1/desk (cm)	fr2/xyz (cm)	fr3/office (cm)
iMAP	4.9	2.0	5.8
BAD-SLAM	1.7	1.1	1.73
Kintinuous	3.7	2.9	3.0
ORB-SLAM2	1.6	0.4	1.0

Table 3: ATE RMSE in cm on TUM RGB-D dataset.

Experimental Results

Qualtative Results :





(a) Chair (b) Back of Objects (c) Small Objects (d) Black Chair

Experimental Results

Ablative Analysis :

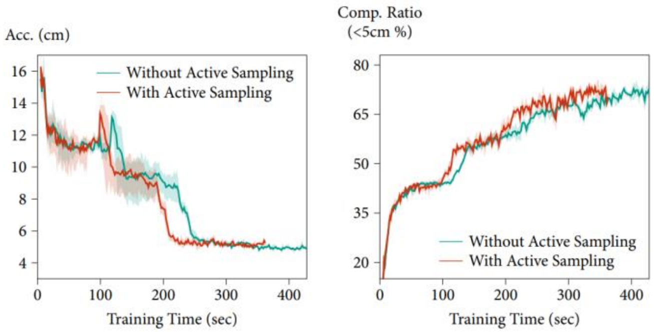


Figure 12: Active sampling obtains better completion with faster accuracy convergence than pure random sampling.

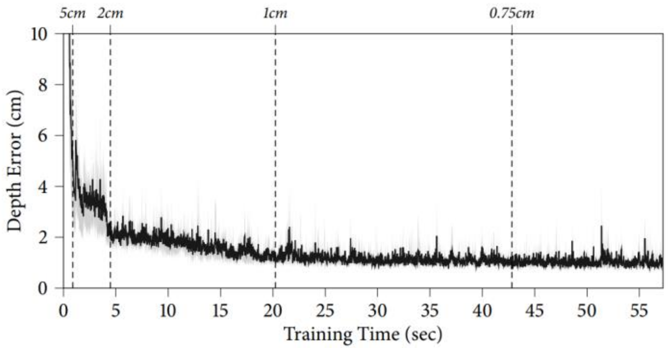


Figure 13: Reaching 5cm, 2cm, 1cm and 0.75cm depth error requires around 1, 4, 20, 43 seconds respectively.

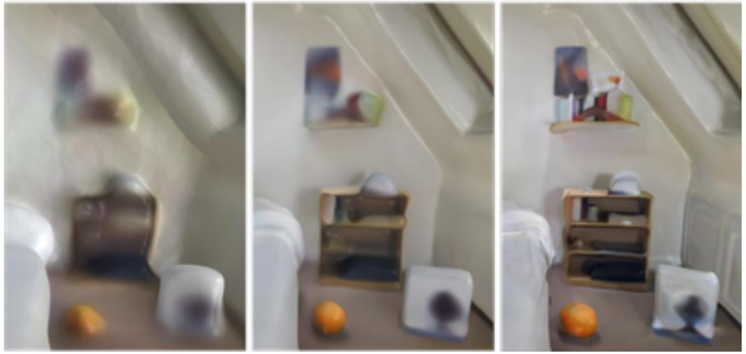


Figure 14: Evolution of reconstruction detail.