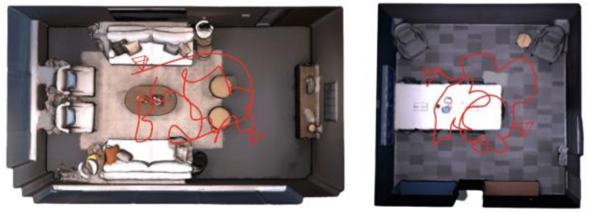
IBD-SLAM: Learning Image-Based Depth Fusion for Generalizable SLAM

Minghao Yin[1] Shangzhe Wu[2] Kai Han[1]

[1]Visual AI Lab, The University of Hong Kong [2]Visual Geometry Group, University of Oxford

Background & Contribution

We address the problem of visual SLAM. We aim to develop a generalizable visual SLAM system that does not require network optimization during the mapping process.

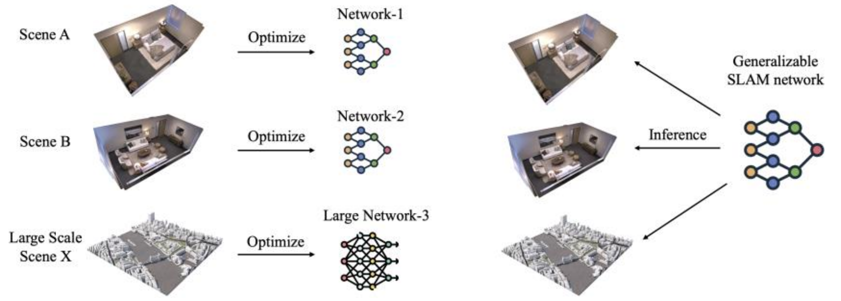


Key contributions:

- We design IBD-SLAM, which can generalize to novel scenes without the need to retrain the model for scene-specific representation.
- We propose to utilize *xyz*-map for high-quality depth fusion.
- Our method runs 10× faster than the previous state-of-the-art methods during the mapping stage.

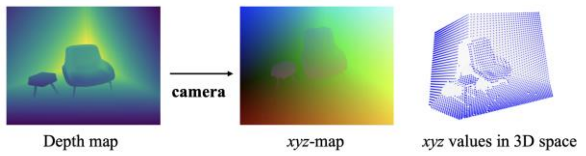
Motivation

- Existing SLAM methods require high time consumption and increased memory usage as the scene scales up.
- Existing SLAM systems based on implicit representations require per-scene optimization for mapping, which limits the generalizability to unseen scenes.

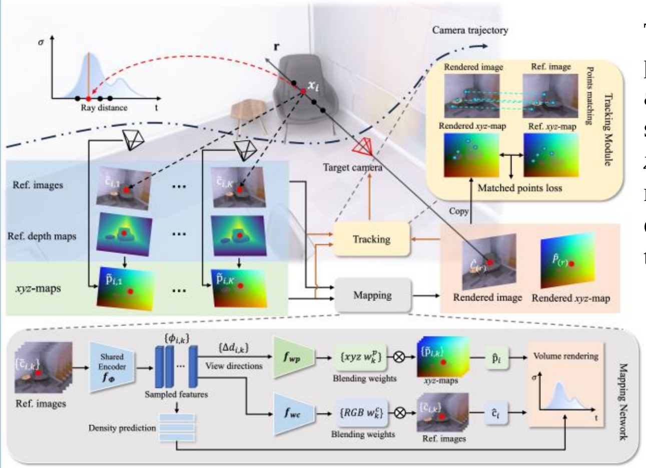


Method

- IBD-SLAM predicts the target view RGB image & *xyz*-map by fusing multi-view inputs from previous frames.



Depth maps suffer from multi-view inconsistency in camera coordinates. We propose to utilize *xyz*-map in world coordinates for fusing.



Tracking: Matched points on the reference and rendered images should share the same *xyz* values. By minimizing their *xyz* differences, we optimize the target camera pose.

Mapping: The final colors and *xyz* values are obtained as a weighted sum of their correspondences in reference views.

Quantitative Results

➢Reconstruction results & Time consumption

	Depth L1 ↓	Acc. ↓	Comp.↓	Comp Ratio ↑
Orb-SLAM2	4.49/3.35†	3.97/3.36†	4.05/3.60†	82.4/86.3†
NICE-SLAM	3.55/2.49†	2.87/2.42†	3.13/2.65†	87.1/90.3†
ESLAM	**2.30/1.29**†	2.82/2.34†	2.97/2.14†	89.5/**94.7**†
Co-SLAM	2.59/1.60†	2.66/2.21†	3.21/2.36†	88.9/92.7†
Ours	2.41/1.53†	**2.25/1.83**†	**2.93/2.02**†	**90.9**/93.8†

	Track ↓ [ms x it]	Map ↓ [ms x it]	#param ↓
NICE-SLAM	7.8x10	82.5x60	17.4M
ESLAM	**6.9x8**	18.4x15	9.29M
Co-SLAM	5.8x10	9.8x10	0.26M
Ours	5.4x20	**12.3x1**	**0.04M**

➢Tracking results on TUM-RGBD and Scannet datasets

	fr1/desk	fr2/xyz	fr3/office	Avg
iMAP[35]	7.2	2.1	9.0	6.1
DI-Fusion[15]	4.4	2.1	15.6	7.4
NICE-SLAM[53]	2.7	**1.8**	3.0	2.5
Ours	**1.8**	**1.8**	**2.7**	**2.2**

	0000	0059	0106	0169	Avg
iMAP[35]	55.95	32.06	17.50	70.51	44.00
DI-Fusion[15]	66.99	128.00	18.50	75.80	72.32
NICE-SLAM[53]	8.64	12.25	8.09	10.28	9.89
Ours	**7.96**	**9.19**	**7.13**	**7.98**	**9.44**

➢Ablation study

	Depth L1 ↓	Acc. ↓	Comp. ↓	Comp. Ratio ↑
w/o novel	1.80	2.66	2.91	91.2
Shared-net	2.35	3.02	3.89	87.7
Ours	**1.53**	**1.83**	**2.02**	**93.8**

(a) Ablation study of model design

In table (b), † denotes results without regularization losses. And ‡ denotes Poisson surface reconstruction results.

	Depth L1 ↓	Acc. ↓	Comp. ↓	Comp. Ratio ↑
iMAP	4.39	4.77	5.02	75.5
iMAP‡	5.17	6.19	6.87	61.4
NICE-SLAM	2.49	2.42	2.62	90.3
NICE-SLAM‡	3.13	3.05	3.16	87.2
Ours†	1.72	2.05	2.30	92.2
Ours	**1.53**	**1.83**	**2.02**	**93.8**

(b) Effects of mesh generation methods

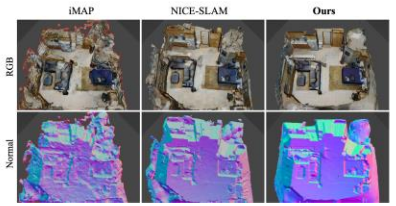
	Depth L1 ↓	Acc. ↓	Comp. ↓	Comp. Ratio ↑
w/o $\mathcal{L}_{xyz}$	3.02	2.81	3.23	88.2
w/o $\mathcal{L}_{depth}$	2.09	2.23	2.45	92.0
w/o $\mathcal{L}_{rgb}$	2.17	2.35	2.64	91.6
w/o $\mathcal{L}_{normal}$	1.65	1.98	2.20	92.7
w/o $\mathcal{L}_{wrap}$	1.62	1.92	2.13	93.3
w/o $\mathcal{L}_{flow}$	1.57	1.86	2.20	93.1
Ours	**1.53**	**1.83**	**2.02**	**93.8**

(c) Ablation study of pretraining loss functions

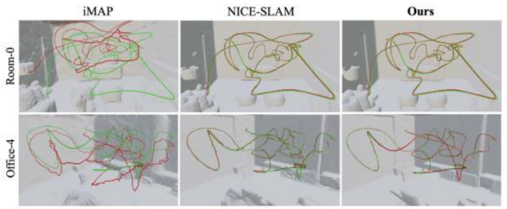
Qualitative results

➢Reconstruction: Replica

➢Reconstruction: Scannet



➢Tracking: Replica



➢Reconstruction results with different fusion methods