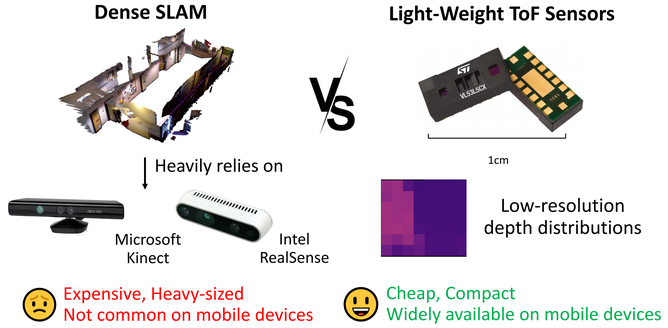
Multi-Modal Neural Radiance Field for Monocular Dense SLAM with a Light-Weight ToF Sensor

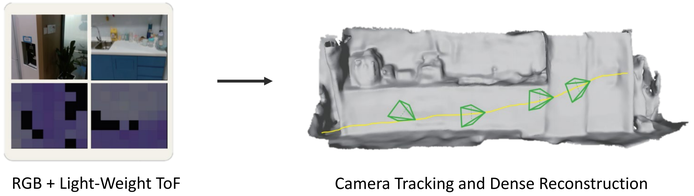
Xinyang Liu, Yijin Li, Yanbin Teng, Hujun Bao, Guofeng Zhang, Yinda Zhang, Zhaopeng Cui*

ICCV23 PARIS

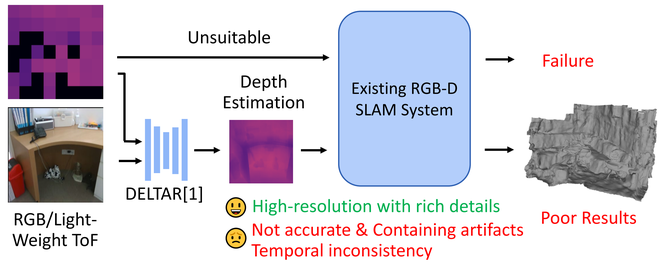
Motivation



Can we build a dense SLAM system with light-weight ToF sensors?



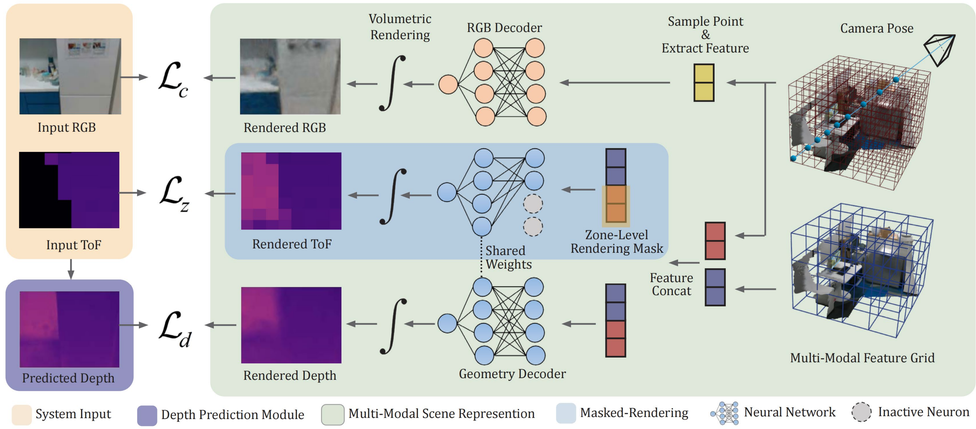
Challenges



- Simply using light-weight ToF signals as low-resolution depths in current RGB-D SLAM systems leads to failure.
- The estimated depth contains artifacts and lacks temporal consistency.

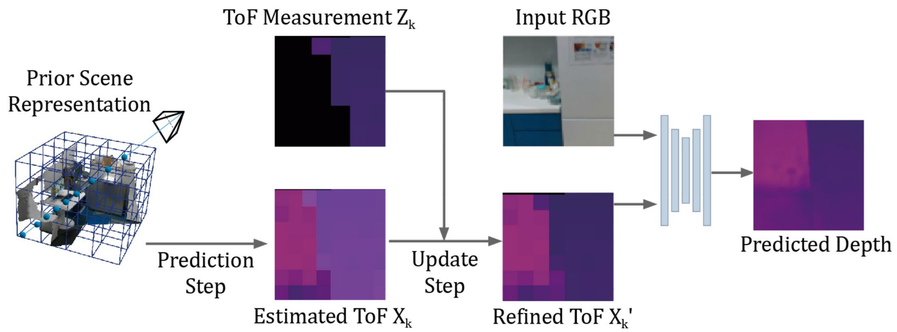
Method

Multi-Modal Implicit Scene Representation



- Our multi-modal implicit scene representation supports rendering both the zone-level signals of light-weight ToF sensors and pixel-wise RGB/depth images.
- We use a rendering mask to render zone-level ToF signals.

Temporal Filtering Module



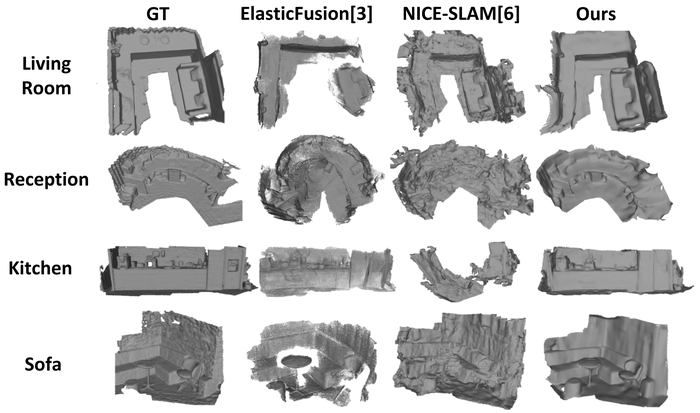
- When ToF signals are missing or noisy, the depth predictions may contain severe artifacts.
- We develop an explicit temporal filtering module to enhance the ToF signals.

Experiments

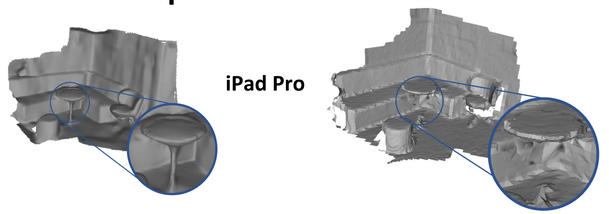
Quantitative Results

Scene Name		Kitchen	Sofa	Office	Reception	Living room	Office2	Sofa2	Avg.
KinectFusion[2]	Acc.↓	-	0.190	0.211	0.261	-	0.267	0.135	0.213
	Comp.↓	-	0.048	0.046	0.064	-	**0.078**	0.064	**0.060**
	F-score	-	0.278	0.288	0.285	-	0.274	0.381	0.301
ElasticFusion[3]	Acc.↓	0.092	0.135	0.084	0.297	0.151	**0.096**	0.122	0.140
	Comp.↓	**0.065**	0.048	0.082	0.305	0.216	0.147	0.047	0.130
	F-score	0.553	0.420	0.529	0.274	0.382	0.416	0.481	0.436
BundleFusion[4]	Acc.↓	0.170	0.100	0.103	0.122	-	0.121	0.123	0.123
	Comp.↓	0.088	**0.030**	**0.038**	**0.057**	-	0.214	0.034	0.077
	F-score	0.373	0.571	0.474	0.470	-	0.442	0.527	0.476
iMAP[5]	Acc.↓	-	0.135	0.229	0.365	0.225	0.233	0.139	0.221
	Comp.↓	-	0.054	0.103	0.245	0.291	0.139	0.069	0.150
	F-score	-	0.445	0.315	0.238	0.170	0.255	0.416	0.307
NICE-SLAM[6]	Acc.↓	0.303	0.119	0.116	0.216	0.103	0.156	0.464	0.211
	Comp.↓	0.456	0.042	0.070	0.199	**0.089**	0.163	0.045	0.152
	F-score	0.221	0.554	0.411	0.402	0.400	0.273	0.401	0.380
Ours	Acc.↓	**0.081**	**0.068**	**0.067**	**0.079**	**0.078**	0.113	**0.121**	**0.087**
	Comp.↓	0.071	0.041	0.045	0.062	0.122	0.085	**0.033**	0.066
	F-score	**0.559**	**0.661**	**0.646**	**0.643**	**0.496**	**0.557**	**0.656**	**0.604**

Qualitative Results



Comparison with iPad Pro



References [1] Li Yijin, et al. DELTAR: Depth Estimation from a Light-Weight ToF Sensor and RGB Image. ECCV 2022. [2] Shahram Izadi, et al. KinectFusion: real-time 3d reconstruction and interaction using a moving depth camera. UIST 2011. [3] Thomas Whelan, et al. ElasticFusion: Realtime dense SLAM and light source estimation. IJRR 2016. [4] Angela Dai, et al. BundleFusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. ToG 2017. [5] Edgar Sucar, et al. iMAP: Implicit mapping and positioning in real-time. ICCV 2021. [6] Zhu Zihan, et al. NICE-SLAM: Neural implicit scalable encoding for SLAM. CVPR 2022.