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

Xinyang Liu[1], Yijin Li[1], Yanbin Teng[1], Hujun Bao[1], Guofeng Zhang[1], Yinda Zhang[2], Zhaopeng Cui[1*]
[1]State Key Lab of CAD&CG, Zhejiang University [2]Google

In this supplementary document, we provide more implementation details in Sec. A, describe more details about the light-weight ToF sensor we used in Sec. B, more details on the evaluation in Sec. C, and show more qualitative results in Sec. D. Finally, we discuss our limitations in Sec. E. More qualitative results can be found in our supplementary video.

A. More Implementation Details

A.1. Multi-Modal Implicit Scene Representation

As described in Sec. 3.2 of our main paper, we propose a new multi-model implicit scene representation with multi-level grid features. In the multi-model feature grids, the dimension of features for the geometry grid at each level is 4. Since we only encode color at the finest level, color features are encoded with a dimension of 6. We use the dense grid from tiny-cuda-nn [2] library to implement the multi-level feature grids for acceleration. For both geometry and color features, we use small MLPs with two hidden layers consisting of 32 neurons as decoders.

The input dimension of the geometry decoder is 16, corresponding to the total geometry feature size. When decoding pixel-level depth, all the input neurons are active; while decoding zone-level depth, only 8 of the 16 neurons corresponding to the zone-level features are active. The input dimension of the RGB decoder is 9, including 6 for the color features extracted from the feature grid and 3 for the view direction vector.

A.2. Loss Function

In this section, we provide more details of the SDF supervision and the SDF regularization terms used in the mapping process.

SDF Supervision Term. Apart from supervising the rendered depth, we also supervise the intermediate SDF prediction. Following [3, 7, 1], we approximate the ground-truth SDF supervision $b(\mathbf{x}) = \widetilde{D}[u,v] - d$ based on our depth predictions. It is noticeable that this approximation is an upper bound of the ground-truth SDF value. For near-surface points, the differences between the ground-truth SDF value and the approximation are expected to be smaller. As a result, we apply the following near-surface loss for the near-surface points ($|\widetilde{D}[u,v] - d| <= t$) as in [3]:

$$\ell_{sdf}(\mathbf{x}) = |\phi_{pix}(\mathbf{x}) - b(\mathbf{x})|, \tag{1}$$

where ϕ_{pix} represents the pixel-level SDF value of point $\mathbf{x}$ as mentioned in Sec. 3.2 of our main paper. The truncation threshold t is a hyper-parameter and we set it to 16cm. For points far from the surface ($|\widetilde{D}[u,v] - d| > t$), we apply the following loss to encourage free space prediction as [3]:

$$\ell_{fs}(\mathbf{x}) = \max\left(0, e^{-\beta\phi_{pix}(\mathbf{x})} - 1, \phi_{pix}(\mathbf{x}) - b(\mathbf{x})\right), \tag{2}$$

where β is a hyper-parameter controlling the exponential penalty term when the SDF prediction is negative, and we set $\beta = 5$ in our experiment.

SDF Regularization Term. To alleviate the ill-posed problem in under-constrained regions, we employ two additional regularization terms on the SDF prediction: Eikonal regularization ℓ_{eik} and smoothness regularization ℓ_{smooth}.

Eikonal regularization is widely used in previous works [8, 3] , encouraging the prediction to approximate a signed distance function. This regularization can help propagate the SDF field from the near-surface regions to free space. For a given point $\mathbf{x}$, the Eikonal regularization is applied via the loss ℓ_{eik}:

$$\ell_{eik}(\mathbf{x}) = (1 - ||\nabla\phi_{pix}(\mathbf{x})||)^2. \tag{3}$$

Since we do not have a high-quality depth map as input like [10, 5], we also add a smoothness regularization that encourages nearby points to have similar normal direction as in [7]:

$$\ell_{smooth}(\mathbf{x}) = ||\nabla\phi_{pix}(\mathbf{x}) - \nabla\phi_{pix}(\mathbf{x} + \epsilon)||^2, \tag{4}$$

where ϵ is a small perturbation with a random direction and length of δ_s. We set δ_s to 4mm empirically.

*Corresponding author.

	VL53L5CX(ours)	Apple LiDAR
Cost	$2-3	∼$20
Resolution	8×8	256×192
Power	0.2W	3-4W
Main Usage	Autofocus	AR & VR

Table A. **Comparison with Apple LiDAR.**

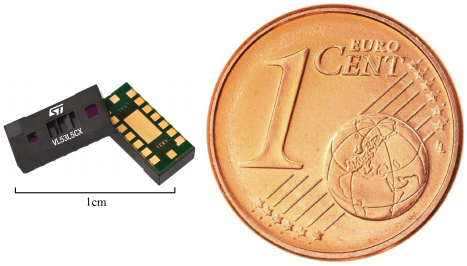

Figure A. **VL53L5CX ToF Sensor.** The sensor is extremely small in size (6.4 × 3.0 × 1.5 mm) and runs at a fairly low power consumption (about 200mW). We compare the size of the VL53L5CX sensor with a one cent euro coin in the figure.

Experimental Settings. We use the first 60 frames to do initialization. During the initialization process, the frames are sequentially added to the optimization process every $N_a = 100$ iterations, and after all frames are added, an extra $N_e = 300$ iterations of optimization are performed. We set the number of tracking iterations N_t to 50, the number of mapping iterations N_m to 150, the number of sampling pixels M to 5000, the number of sampling zones Z to 500, the number of neighbor frames N_n to 60, the window size W_s to 30 and implement a hierarchical sampling strategy similar to NeuS [8] to obtain the sampling points along the ray. We first sample $N_c = 96$ coarse samples and add 12 samples at each step based on weights computed from the previously sampled points.

B. Light-Weight ToF Sensor Details

In this paper, we use ST VL53L5CX [4] (shown in Fig. A, denoted as L5) as a typical representative of light-weight ToF sensors. Comparing with common commodity level depth sensors (e.g. Intel RealSense, apple LiDAR, etc.), light-weight ToF sensors are a magnitude lower in power consumption and price. We compare the basic information between L5 and apple LiDAR in Table A. L5 outputs depth in the resolution of 8×8, and each measures the depth distribution in a large zone. Its diagonal field-of-view (FoV) is 63°, similar to common monocular cameras. Unlike the conventional ToF sensor that provides per-pixel high-resolution (usually higher than 0.03 megapixels) depth map, L5 provides depth distribution measurements within zones in an extremely low resolution.

L5 emits infrared rays and measures depth based on the time taken for the wave to bounce back to the emitter. However, due to the light-weight electronic design, L5 is only able to give statistical information about the depth in a large zone. The depth distribution is initially obtained by counting the number of photons returned in each discretized range of time and then fitted with a Gaussian distribution model to compress the raw information due to its tight bandwidth limit.

For each zone, L5 also returns a status code to show whether the measurement in that zone is valid. If the number of photons received in a zone is too small or the measurements are unstable, the corresponding zone will be marked as invalid. More details can be found on STMicroelectronics's webpage[1].

C. Evaluation Details

C.1. Mesh Culling

Implicit methods can usually complete the scene geometry for unseen regions. For a fair comparison between implicit and explicit methods, we cull surfaces that are not observed inside any camera frustums or occluded by other objects.

C.2. Depth Metrics

C.3. Reconstruction Metrics

We evaluate the quality of the scene reconstruction using the following standard metrics where p and p^* are the vertices in generated mesh P and GT mesh P^* respectively:

- Accuracy (Acc.):

$$\frac{1}{|P|} \sum_{p \in P} \min_{p* \in P^*} \|p - p^*\|. \tag{5}$$

- Completeness (Comp.):

$$\frac{1}{|P^*|} \sum_{p* \in P^*} \min_{p \in P} \|p - p^*\|. \tag{6}$$

- Precision (Prec.):

$$\frac{1}{|P|} \sum_{p \in P} \min_{p* \in P^*} \|p - p^*\| < 0.05. \tag{7}$$

- Recall (Recal.):

$$\frac{1}{|P^*|} \sum_{p* \in P^*} \min_{p \in P} \|p - p^*\| < 0.05. \tag{8}$$

- F-score:

$$\frac{2 \times \text{Prec.} \times \text{Recal.}}{\text{Prec.} + \text{Recal.}}. \tag{9}$$

[1] https://www.st.com/content/st_com/en/premium-content/premium-content-time-of-flight.html

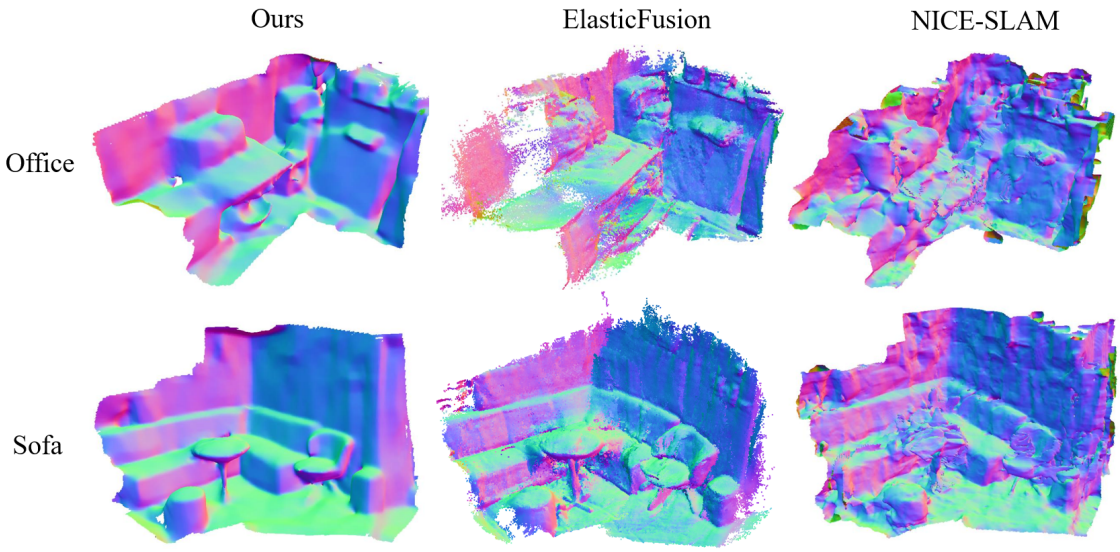

Figure B. **More Reconstruction Results.** We show the final mesh generated by NICE-SLAM [10], ElasticFusion [9] and our method. The mesh is visualized with the vertex normal. Compared to other methods, our method recover cleaner and sharper scene geometry.

In general, F-score is considered as the most proper metric to evaluate the quality of the scene reconstruction [6] since both the accuracy and completeness of the reconstruction are considered.

We evaluate the performance of the depth prediction using the following standard metrics where $\hat{d}_i$ represents predicted depth, d_i represents ground truth depth, and N is the number of valid ground truth values:

- Threshold Accuracy (δ_i with i=1,2,3):

$$\frac{\sum_{i=1}^{N}[max(\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i}) < 1.25^i]}{N}, \tag{10}$$

  where [] denotes Iverson brackets.

- Mean Absolute Relative Error (REL):

$$\frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{d}_i - d_i|}{d_i}. \tag{11}$$

- Root Mean Square Error (RMSE):

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{d}_i - d_i)^2}. \tag{12}$$

In the ablation study about the temporal filtering technique for depth prediction, we do the evaluation separately for normal and hard cases. Here, we give a detailed definition of the division criteria. For the normal case, the raw L5 signal is of normal quality, leading to relatively accurate depth prediction. While for the hard case, the raw L5 signal is noisy or has large amounts of missing data. We classify a prediction as a normal case if its RMSE error is less than 0.4; otherwise, we regard it as a hard case.

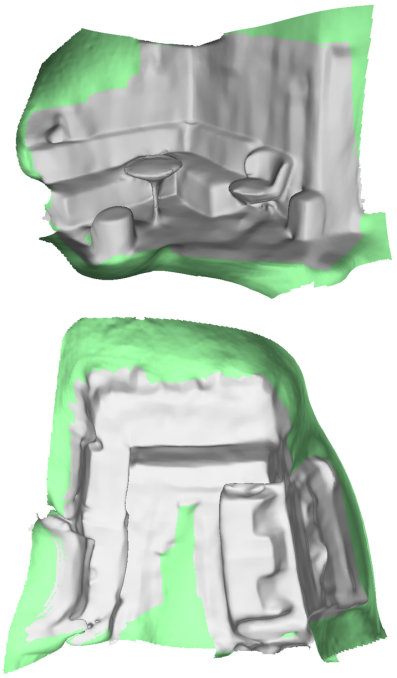

Figure C. **Geometry Forecast.** The white-colored area is the region with observations, while the green-colored area represents the unseen but forecasted region. It is noticeable that our method can generate reasonable mesh even in the unseen region.

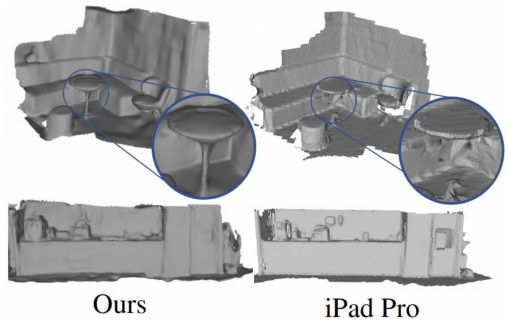

Figure D. **Comparison to a Smartphone Scan.**

D. More Qualitative Results

Mesh Visualization with Vertex Normal. To better highlight the differences in reconstruction quality, we provide additional visualizations in Fig. B using the vertex normal to color the mesh. It is noticeable that our approach outperforms the others and produces high-quality scene reconstruction results.

Geometry Forecast. Our method is able to make reasonable predictions in unseen regions thanks to the multi-modal implicit scene representation. As shown in Fig. C, the hole in the floor is well filled, and the walls are correctly expanded to unobserved regions.

Comparison to a Smartphone Scan. We used an iPad Pro (with "3D Scanner App") to re-scan two testing scenes under as closely identical conditions as possible to ours. As shown in Fig. D, our method can achieve reconstruction results comparable to the iPad Pro using a much cheaper light-weight ToF and outperforms iPad Pro on thin objects thanks to the multi-modal implicit scene representation.

E. Limitations

Firstly, the range of a light-weight ToF sensor is usually limited to several meters and the sunlight has a strong interference on the sensor, so our method currently focuses on indoor scenes. We plan to further improve the system to overcome the limitation of ToF sensors in outdoor scenarios. Secondly, the computational overhead of the proposed method is still relatively high for mobile platforms. In future work, we plan to further reduce the computational burden and make it efficient enough to run on mobile robots. At last, we also plan to add semantic information into our system for high-level scene understanding.

References

[1] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6290–6301, 2022. 1

[2] Thomas Müller. tiny-cuda-nn. https://github.com/NVlabs/tiny-cuda-nn, 4 2021. 1

[3] Joseph Ortiz, Alexander Clegg, Jing Dong, Edgar Sucar, David Novotny, Michael Zollhoefer, and Mustafa Mukadam. isdf: Real-time neural signed distance fields for robot perception. In *Proceedings of the Robotics: Science and Systems*, 2022. 1

[4] STMicroelectronics. Time-of-Flight 8x8 multizone ranging sensor with wide field of view. https://www.st.com/en/imaging-and-photonics-solutions/vl53l5cx.html. Accessed 19-Jul-2022. 2

[5] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021. 1

[6] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15593–15602, Nashville, TN, USA, 2021. IEEE. 3

[7] Jingwen Wang, Tymoteusz Bleja, and Lourdes Agapito. Go-surf: Neural feature grid optimization for fast, high-fidelity rgb-d surface reconstruction. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 433–442. IEEE, 2022. 1

[8] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021. 1, 2

[9] Thomas Whelan, Renato F Salas-Moreno, Ben Glocker, Andrew J Davison, and Stefan Leutenegger. ElasticFusion: Real-time dense SLAM and light source estimation. *The International Journal of Robotics Research*, 35(14):1697–1716, 2016. 3

[10] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 1, 3