Query Quantized Neural SLAM

Sijia Jiang, Jing Hua*, **Zhizhong Han**

Department of Computer Science, Wayne State University, Detroit, MI, USA
sijiajiang@wayne.edu, jinghua@wayne.edu, h312h@wayne.edu

Abstract

Neural implicit representations have shown remarkable abilities in jointly modeling geometry, color, and camera poses in simultaneous localization and mapping (SLAM). Current methods use coordinates, positional encodings, or other geometry features as input to query neural implicit functions for signed distances and color which produce rendering errors to drive the optimization in overfitting image observations. However, due to the run time efficiency requirement in SLAM systems, we are merely allowed to conduct optimization on each frame in few iterations, which is far from enough for neural networks to overfit these queries. The underfitting usually results in severe drifts in camera tracking and artifacts in reconstruction. To resolve this issue, we propose query quantized neural SLAM which uses quantized queries to reduce variations of input for much easier and faster overfitting a frame. To this end, we quantize a query into a discrete representation with a set of codes, only allow neural networks to observe a finite number of variations. This makes neural networks get more and more familiar to these codes after overfitting more and more previous frames. Moreover, we also introduce novel initialization, losses, and argumentation to stabilize the optimization with huge uncertainty in the early optimization, constrain the optimization space, and estimate camera poses more accurately. We justify the effectiveness of each design and report visual and numerical comparisons on widely used benchmarks to show our superiority over the latest methods in both reconstruction and camera tracking.

Introduction

Neural implicit representations have made huge progress in simultaneous localization and mapping (SLAM) (Zhu et al. 2022, 2023a; Wang, Wang, and Agapito 2023; Sucar et al. 2021; Stier et al. 2023). These methods represent geometry and color as continuous functions to reconstruct smooth surfaces and render plausible novel views, which shows advantages over point clouds in classic SLAM systems (Koestler et al. 2022). Current methods learn neural implicits in a scene through rendering them into RGBD images through volume rendering and minimizing the rendering errors to ground truth observations. To render a color (Wang et al. 2021), depth (Yu et al. 2022), or normal (Wang et al. 2022) at a pixel, we query neural implicit representations for signed distances or occupancy labels and color at many locations along a ray, which are integrated based on volume rendering equations.

We usually use coordinates, positional encodings, or other features as the input of neural implicit representations (Park et al. 2019; Songyou Peng 2020; Müller et al. 2022; Rosu and Behnke 2023; Li et al. 2023b), which we call a query. Queries are continuous vectors which allows neural networks to generalize well on unseen queries that are similar to the ones seen before. Continuity is good for generalization but also brings huge variations for neural networks to overfit. Neural networks need to see these queries or similar ones lots of times so that they can infer and remember attributes like geometry and color at these queries, which takes time. However, this runtime efficiency does not meet the requirement of SLAM systems, what is more critical, we are only allowed to conduct optimization on current frame in merely few iterations, no beyond frames are observable. Underfitting on these queries results in huge drifts in camera tracking and artifacts in reconstruction. Therefore, how to query neural implicit representations to make overfitting more efficiently in SLAM is still a challenge.

To overcome this challenge, we introduce query quantized neural SLAM to jointly model geometry, color, and camera poses from RGBD images. We learn a neural singed distance function (SDF) to represent geometry in a scene through rendering the SDF with a color function to overfit image observations. We propose to quantize a query into a discrete representation with a set of codes, and use the discrete query as the input of neural SDF, which significantly reduces the variations of queries and improves the performance of reconstruction and camera tracking. Our idea is to make neural networks get more and more familiar to these quantized queries after overfitting more and more previous frames, which leads to faster and easier coveragence at each frame. We provide a thorough solution to discretize queries like coordinates, positional encodings, or other geometry features for overfitting each frame more effectively. Moreover, to support our quantized queries, we also introduce novel initialization, losses, and augmentation to stabilize the optimization with huge uncertainty in the very beginning, constrain the optimization space, and estimate camera poses more accurately. We evaluate our methods on widely used

*Corresponding author: Jing Hua
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

benchmarks containing synthetic data and real scans. Our numerical and visual comparisons justify the effectiveness of our modules, and show superiorities over the latest methods in terms of accuracy in scene reconstruction and camera tracking. Our contributions are listed below.

1. We present query quantized neural SLAM for joint scene reconstruction and camera tracking from RGBD images. We justify the idea of improving SLAM performance by reducing query variations through quantization.
2. We present novel initialization, losses, and augmentation to stabilize the optimization. We show that the stabilization is the key to make quantized queries work in SLAM.
3. We report state-of-the-art performance in scene reconstruction and camera tracking in SLAM.

Related Work

Neural implicit representations achieve impressive results across various applications (Wang et al. 2022; Guo et al. 2022; Rosu and Behnke 2023; Li et al. 2023b; Müller et al. 2022). With supervision from 3D annotations (Liu et al. 2021; Tang et al. 2021), point clouds (Atzmon and Lipman 2020; Zhao et al. 2020; Atzmon and Lipman 2021; Chen, Liu, and Han 2022), or multi-view images (Fu et al. 2022; Yu et al. 2022; Wang et al. 2022; Guo et al. 2022), neural SDFs or occupancy functions can be estimated using additional constraints or volume rendering.

Multi-view Reconstruction. Classic multi-view stereo (MVS) (Schönberger and Frahm 2016; Schönberger et al. 2016) uses photo consistency to estimate depth maps but struggles with large viewpoint variations and complex lighting. Space carving (Laurentini 1994) reconstructs 3D structures as voxel grids without relying on color.

Recent methods leverage neural networks to predict depth maps using depth supervision (Yao et al. 2018; Koestler et al. 2022) or multi-view photo consistency (Zhou et al. 2017).

Neural implicit representations have gained popularity for learning 3D geometry from multi-view images. Early works compared rendered outputs to masked input segments using differentiable surface renderers (Jiang et al. 2020; Niemeyer et al. 2020; Sun et al. 2021). DVR (Niemeyer et al. 2020) and IDR (Yariv et al. 2020) model radiance near surfaces for rendering.

NeRF (Mildenhall et al. 2020) and its variants (Park et al. 2021; Müller et al. 2022; Sun et al. 2021) combine geometry and color via volume rendering, excelling in novel view synthesis without masks. UNISURF (Oechsle, Peng, and Geiger 2021) and NeuS (Wang et al. 2021) improve on this by rendering occupancy functions and SDFs. Further advancements integrate depth (Yu et al. 2022; Azinović et al. 2022; Zhu et al. 2022), normals (Wang et al. 2022; Guo et al. 2022), and multi-view consistency (Fu et al. 2022) to enhance accuracy. Depth images play a key role by guiding ray sampling (Yu et al. 2022) or providing rendering supervision (Yu et al. 2022; Lee et al. 2023), enabling more precise surface estimation.

Neural SLAM. Early work employed neural networks to learn policies for exploring 3D environments. More recent methods (Zhang et al. 2023; Xinyang et al. 2023; Teigen et al. 2023; Sandström et al. 2023) learn neural implicit representations from RGBD images. iMAP (Sucar et al. 2021) uses an MLP as the only scene representation in a realtime SLAM system. NICE-SLAM (Zhu et al. 2022) presents a hierarchical scene representation to reconstruct large scenes with more details. Its following work NICER-SLAM (Zhu et al. 2023b) uses monocular geometric cues instead of depth images as supervision. Co-SLAM (Wang, Wang, and Agapito 2023) jointly uses coordinate and sparse parametric encodings to learn neural implicit functions. Segmentation priors (Kong et al. 2023; Haghighi et al. 2023) also show their ability to improve the performance of SLAM. Also, vMAP (Kong et al. 2023) represents each object in the scene as a neural implicit in a SLAM system. Depth fusion is also integrated with neural SDF as a prior for more accurate geometry modeling in SLAM (Hu and Han 2023).

Neural Representations with Vector Quantization. Vector quantization, first introduced in VQ-VAE (Oord, Vinyals, and Kavukcuoglu 2017) for image generation, has since been applied to binary neural networks (Gordon et al. 2023), data augmentation (Wu et al. 2022), compression (Dupont et al. 2022), novel view synthesis (Yang et al. 2023b), point cloud completion (Fei et al. 2022), image synthesis (Gu et al. 2022), and 3D reconstruction/generation using Transformers or diffusion models (Corona-Figueroa et al. 2023; Li et al. 2023a). Unlike these approaches, we quantize input queries to approximate continuous representations for SLAM systems, addressing runtime efficiency and visibility constraints during optimization. Unlike Gaussian splatting-based SLAM methods (Keetha et al. 2024; Matsuki et al. 2024; Huang et al. 2024b,a; Yu, Sattler, and Geiger 2024), our approach focuses on recovering high-fidelity SDFs.

Method

Overview. Following previous methods (Wang, Wang, and Agapito 2023; Zhu et al. 2022; Hu and Han 2023), our neural SLAM jointly estimates geometry, color and camera poses from J frames of RGBD images I and D. Our SLAM estimates camera poses O_j for each frame j, and infers an SDF f_s and a color function f_c which predict a signed distance $s = f_s(\tilde{q})$ and a color $c = f_c(\tilde{q})$ for an quantized query $\tilde{q}$. $\tilde{q}$ is quantized from its continuous representation q, which is not limited to a coordinate p but also includes its positional encoding $h(p)$, geometry feature $g(p)$, and interpolation from fused depth prior $t(p)$.

Fig. 1 illustrates our framework. Starting from a continuous query q, we first quantize it into a quantized query $\tilde{q}$, which is the input to our neural implicit representations including SDF f_s and color function f_c, predicting a signed distance s and a color c. We accumulate signed distances and colors at queries sampled along a ray into a rendered color and a depth through volume rendering. We tune f_s, f_c, and $\{O_j\}$ by minimizing rendering errors. After the optimization, we extract the zero-level set of f_s as the surface of the scene using the marching cubes algorithm (Lorensen and Cline 1987).

Coordinate Quantization. For the coordinate p of a query q, we directly discretize p as its nearest vertex on an extremely high resolution 3D grid, such as 12800^3, which be-

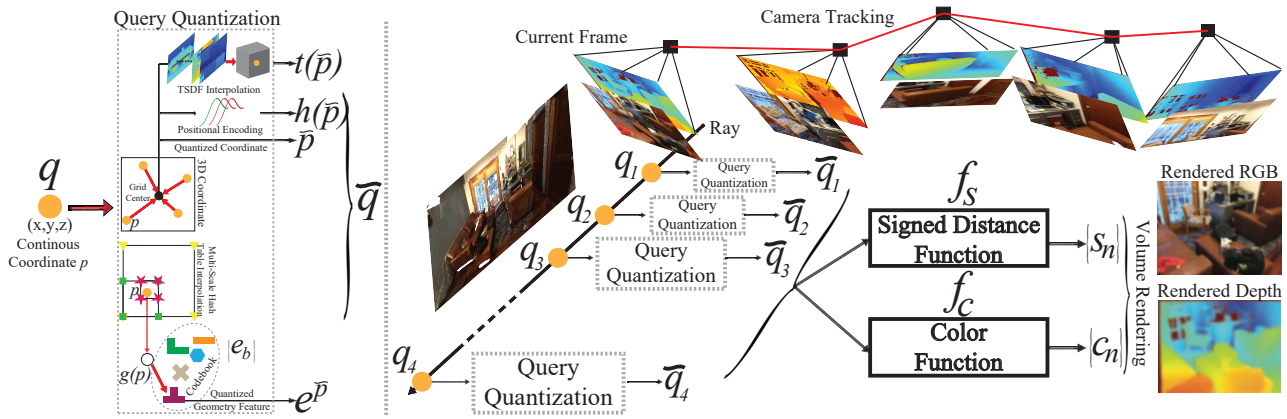

Figure 1: Overview of our method.

comes a quantized coordinate $\tilde{p}$. We use coordinate quantization (Jiang, Hua, and Han 2023) to reduce the coordinate variations, preserve contentiousness, and stabilize the optimization with high frequency postional encodings. Moreover, we use one-blob encoding (Müller et al. 2019) along with the quantized coordinates $\tilde{p}$ as the positional encoding $h(\tilde{p})$. We denote $h(\tilde{p})$ as $h^{\tilde{p}}$ for simplicity in the following.

Geometry Feature Quantization. We follow Instant-NGP (Müller et al. 2022) to build up a multi-resolution hash-based feature grid θ_g as geometry features in the scene. We put learnable features at vertices on the multi-resolution grid, and use the trilinear interpolation to get a geometry feature $g(p)$ at the location p of query q. We normalize the length of $g(p)$ to be 1 to balance the importance of different features that are used to update the same discrete code.

Following VQ-VAE (Oord, Vinyals, and Kavukcuoglu 2017), we maintain a set of B learnable codes $\{e_b\}_{b=1}^{b=B}$, and quantize each geometry feature $g(p)$ into one of the codes by the nearest neighbor search using a L_2 norm as a metric,

$$e^{\tilde{p}} =_{\{e_b\}} ||e_b - g(p)||_2^2, \tag{1}$$

where we denote the nearest code to $g(p)$ in the codebook as $e^{\tilde{p}}$. After each iteration, we normalize the length of each code to be 1 to make codes comparable with each other.

Codebook Initialization. Our preliminary results show that the initialization of B codes really matters. Different from using relatively clean point clouds as supervision (Yang et al. 2023a), random initialization using uniform or Gaussian distributions for each code brings more uncertainties when there are already lots uncertainties in SDF f_s, color function f_c, and the estimated camera poses in the very beginning. These uncertainties cause unstable optimization which results in large drifts in camera tracking that is hard to be corrected in the following optimization iterations. We found that using Bernoulli distribution to initialize entries to either 0 or 1 in each code can significantly constrain changes on these codes, and stabilize the optimization,

$$e_b \sim Bernoulli(0.5). \tag{2}$$

Quantization of Additional Geometry Priors. It has shown that using additional geometry priors as a part of input can improve the reconstruction accuracy in SLAM Hu and Han (2023). It uses a signed distance $t(p)$ at location p as a part of input. $t(p)$ is a scalar interpolated from a TSDF grid θ_t which is incrementally fused from input depth images. We simply quantize $t(p)$ by using quantized coordinates $\tilde{p}$ for the interpolation from θ_t. We denote the quantized signed distance interpolation as $t^{\tilde{p}}$.

Quantized Queries. In sum up, for a continuous query q formed by coordinate p, positional encoding $h(p)$, geometry feature $g(p)$, and TSDF interpolation $t(p)$, we quantize q into a discrete representation,

$$\tilde{q} = [\tilde{p}, h^{\tilde{p}}, e^{\tilde{p}}, t^{\tilde{p}}]. \tag{3}$$

Volume Rendering. We follow NeRF to do volume rendering at current frame j, we render a RGB image $\bar{I}_j$ and a depth image $\bar{D}_j$. This produces rendering errors in terms of RGB color and depth to the input I_j and D_j, which drives the optimization to minimize.

With the estimated camera poses O_j, we shoot a ray R_k at a randomly sampled pixel on view I_j. R_k starts from the camera origin o and points a direction of r. We sample N points along the ray R_k using stratified sampling and uniformly sample near the depth, where each point is sampled at $p_n = o + d_n r$ and d_n corresponds to the depth value of p_n on the ray, where each location p_n indicates a query q_n. We quantize each query q_n into $\tilde{q}_n$ using Eq. 3. Then, the SDF f_s and the color function f_c predict a signed distance $s_n = f_s(\tilde{q})$ and a color $c_n = f_c(\tilde{q})$. Following Neural-RGBD (Azinović et al. 2022), we use a simple bell-shaped function formed by the product of two Sigmoid functions δ to transform signed distances s_n into volume density w_n,

$$w_n = \delta(s_n/t)\delta(-s_n/t), \tag{4}$$

where t is the truncation distance. With w_n, we render RGB $\bar{I}_j$ and depth $\bar{D}_j$ images through alpha blending,

$$\begin{aligned} \bar{I}_j(k) &= \frac{1}{\sum_{n'=1}^{N} w_{n'}} \sum_{n'=1}^{N} w_{n'} c_{n'}, \\ \bar{D}_j(k) &= \frac{1}{\sum_{n'=1}^{N} w_{n'}} \sum_{n'=1}^{N} w_{n'} d_{n'}. \end{aligned} \tag{5}$$

Loss Function. With estimated camera poses, we evaluate the rendering errors at K rays on the rendered $\bar{I}_j$ and $\bar{D}_j$,

$$L_I = \frac{1}{JK} \sum\nolimits_{j,k=1}^{J,K} ||I_j(k) - \bar{I}_j(k)||_2^2,$$
$$L_D = \frac{1}{JK} \sum\nolimits_{j,k=1}^{J,K} ||D_j(k) - \bar{D}_j(k)||_2^2. \quad (6)$$

With the input depth D_j, we can also impose two constraints on the predicted signed distances in the free space between the camera and the surface and area near the surface. We use a threshold tr of signed distances to set up a bandwidth around a surface. For queries outside the bandwidth, we truncate their signed distances into either 1 or -1. Thus, we use an empty space loss $L_{s'}$ to supervise the predicted signed distances $L_{s'} = \sum_{n,k,j} ||s_n - t_r||_2^2$. Moreover, we approximate the signed distances at queries q_n within the bandwidth as $d_n - d'_n$, where d_n is the depth observation at the pixel on D_j and d'_n is the depth at query q_n. Thus, we use $L_s = \sum_{n,k,j} ||s_n - (d_n - d'_n)||_2^2$ to supervise the predicted signed distances s_n.

To learn the B codes $\{e_b\}$, we impose two constraints. One is that we push the code $e^{\tilde{p}}$ that the geometry feature $g(p)$ matches in Eq. 1 to be similar to g(p). We use a MSE,

$$L_g = ||sg[e^{\tilde{p}}] - g(p)||_2 + \lambda ||e^{\tilde{p}} - sg[g(p)]||_2, \quad (7)$$

where sg stands for the stop gradient (Oord, Vinyals, and Kavukcuoglu 2017) operator. The key idea behind stop gradient is to decouple the training of the SDF f_s, color function f_c from the training of B codes. We use $\lambda = 0.1$ in all our experiments.The other is that we diversify the B codes $\{e_b\}$ to prevent them from going to the same point in the feature space using a diverse loss $L_e = \sum_b^B \sum_{b'}^B ||e_b - e_{b'}||_2$.

Our loss function L includes all loss terms above. We jointly minimize all loss terms with balance weights α, β, γ, ζ and η below,

$$\min_{f_s, f_c, \{e_b\}, \theta_g} L_I + \alpha L_D + \beta L_g - \gamma L_e + \zeta\, L_s + \eta L_{s'}. \quad (8)$$

Details in SLAM. With RGBD input, we jointly estimate camera poses for each frame and infer the SDF f_s to model geometry. For camera tracking, we first initialize the pose of current frame using a constant speed assumption, which provides us a coarse pose estimation according to poses of previous frames. We use the coarse pose estimation to shoot rays and render RGB and depth images. We minimize the same loss function in Eq. 8 by only refining the camera poses and keeping other parameters fixed. We refine camera poses and other parameters at the same time in a bundle adjustment procedure every 5 frames, where we also add the pose of current frame as one additional parameter in Eq. 8.

For reconstruction, we render rays from the current view and key frames in each batch. Instead of key frame images, we follow Co-SLAM (Wang, Wang, and Agapito 2023) to store rays randomly sampled 5% of all pixels from each key frame in a key frame ray list. This allows us to insert new key frames more frequently and maintain a larger key frame coverage. We select a key frame very 55 frames.

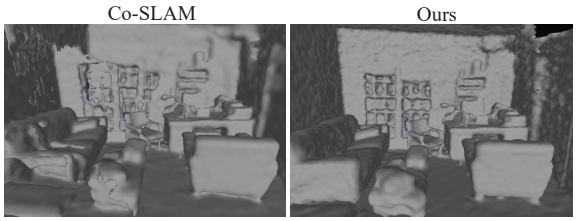

Figure 2: Visual comparison in reconstruction on ScanNet.

With the estimated camera poses, we incrementally fuse input depth images D_j into a TSDF grid θ_t in a resolution of 256. We do trilinear interpolation on θ_t to obtain the prior interpolation t(p) at a query q.

Augmentation of Geometry Priors. Although depth fusion priors (Hu and Han 2023) show that the TSDF θ_t can improve the reconstruction accuracy in SLAM, we found that the interpolation $t^{\tilde{p}}$ of geometry priors significantly degenerate the performance in our preliminary results. Our analysis shows that the neural networks learn a shortcut from the input to the output, which directly maps the geometry prior $t^{\tilde{p}}$ as the predicted signed distance at most queries, ignoring any geometry constraints like camera poses. The reason why it works well with depth fusion priors (Hu and Han 2023) is that it predicts occupancy probabilities but not signed distances, which differentiates the input from the output.

To resolve this problem, we introduce a simple augmentation to manipulate the geometry prior interpolation $t^{\tilde{p}}$ through a linear transformation. We use $t^{\tilde{p}} \leftarrow tanh(t^{\tilde{p}})$ to make signed distances still comparable to each other in the range of $[-1, 1]$ but also shift away from the original TSDF interpolations.

Implementation Details. We run Query Quantized Neural SLAM on an NVIDIA RTX 3090ti GPU, achieving 34 FPS on the Replica dataset with default settings. For query sampling, we sample $N = 43$ points per ray, including 32 uniformly sampled and 11 near-surface queries. We use $B = 128$ codes for vector quantization and a 256^3 TSDF resolution with a truncated threshold $t_r = 10$ voxel size near surfaces. Following DP Prior (Hu and Han 2023), we incrementally fuse a TSDF using coarsely estimated camera poses. Rays are sampled for volume rendering, and depth fusing is redone with refined poses for the next frame. Loss parameters are set as $t = 0.1, \alpha = 0.02, \beta = 0.06, \gamma = 0.0001, \zeta = 200, \eta = 2$.

Experiments and Analysis

Datasets. We evaluate the query quantized neural SLAM on real-world indoor scenes from 4 datasets and 8 synthetic Replica (Straub et al. 2019) scenes following Co-SLAM. Additionally, we assess reconstruction quality on 7 noisy SyntheticRGBD (Rajpal et al. 2023) scenes and compare our reconstruction and camera tracking accuracy to SOTAs on 6 scenes from NICE-SLAM (Zhu et al. 2022) with BundleFusion ground truth poses. Camera tracking is also reported on 3 scenes from TUM RGB-D (Sturm et al. 2012).

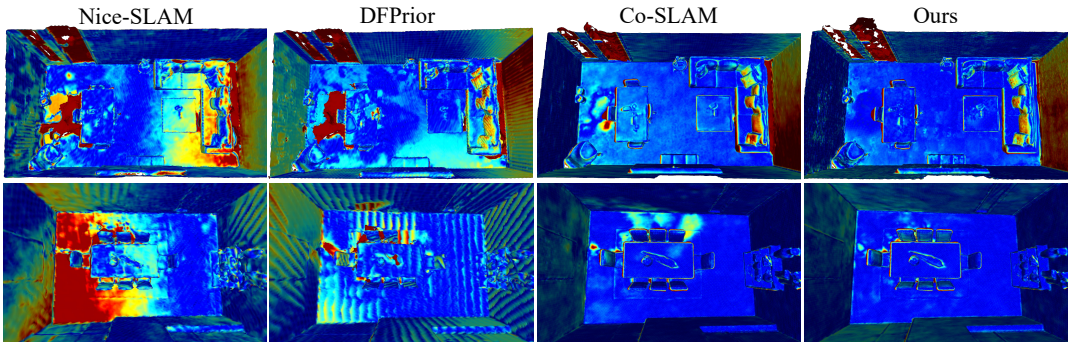

Figure 3: Visual comparisons on Replica.

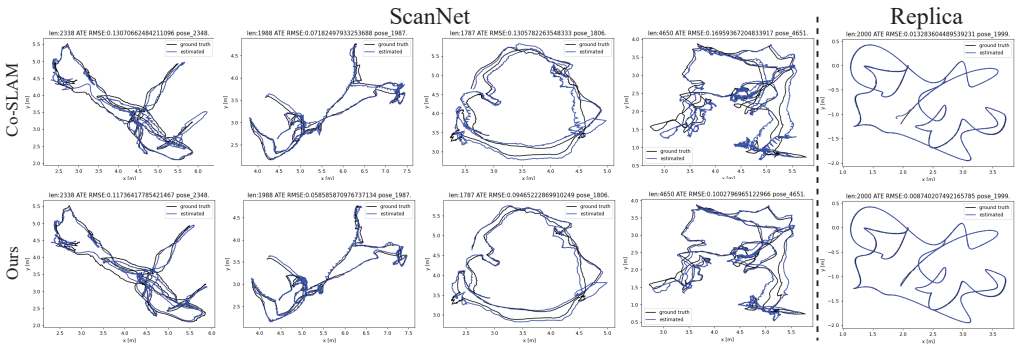

Figure 4: Visual comparisons in camera tracking on ScanNet and Replica.

		room0	room1	room2	office0	office1	office2	office3	office4	Avg.
iMAP	Depth L1[cm]↓	5.08	3.44	5.78	3.79	3.76	3.97	5.61	5.71	4.64
	Acc.[cm]↓	4.01	3.04	3.84	3.34	2.10	4.06	4.20	4.34	3.62
	Comp.[cm]↓	5.84	4.40	5.07	3.62	3.62	4.73	5.49	6.65	4.93
	Comp. Ratio↑	78.34	85.85	79.40	83.59	88.45	79.73	73.90	74.77	80.50
NICE	Depth L1[cm]↓	1.79	1.33	2.20	1.43	1.58	2.70	2.10	2.06	1.90
	Acc.[cm]↓	2.44	2.10	2.17	1.85	1.56	3.28	3.01	2.54	2.37
	Comp.[cm]↓	2.60	2.19	2.73	1.84	1.82	3.11	3.16	3.61	2.63
	Comp. Ratio ↑	91.81	93.56	91.48	94.93	94.11	88.27	87.68	87.23	91.13
DF Prior	Depth L1[cm]↓	1.44	1.90	2.75	1.43	2.03	7.73	4.81	1.99	3.01
	Acc.[cm]↓	2.54	2.70	2.25	2.14	2.80	3.58	3.46	2.68	2.77
	Comp.[cm]↓	2.41	2.26	2.46	1.76	1.94	2.56	2.93	3.27	2.45
	Comp. Ratio ↑	93.22	94.75	93.02	96.04	94.77	91.89	90.17	88.46	92.79
Co-SLAM	Depth L1[cm]↓	1.05	0.85	2.37	1.24	1.48	1.86	1.66	1.54	1.51
	Acc.[cm]↓	**2.11**	**1.68**	**1.99**	1.57	**1.31**	**2.84**	**3.06**	2.23	**2.10**
	Comp.[cm]↓	2.02	1.81	1.96	**1.56**	1.59	2.43	2.72	2.52	2.08
	Comp. Ratio ↑	95.26	95.19	93.58	96.09	94.65	91.63	90.72	90.44	93.44
Ours	Depth L1[cm]↓	1.09	**0.69**	2.48	**1.18**	**0.99**	**1.76**	**1.54**	1.68	**1.42**
	Acc.[cm]↓	2.38	2.62	2.0	**1.55**	1.37	3.43	3.94	**2.16**	2.43
	Comp.[cm]↓	**1.76**	**1.77**	**1.82**	**1.57**	**1.39**	**2.14**	**2.55**	**2.46**	**1.93**
	Comp. Ratio ↑	**96.39**	**95.49**	**94.28**	**96.10**	**95.4**	**94.07**	**91.78**	**91.53**	**94.38**

Table 1: Numerical comparison in each scene on Replica.

Metrics. We adopt Co-SLAM's culling strategy and evaluate reconstruction accuracy using Depth L1 (cm), Accuracy (cm), Completion (cm), and Completion Ratio ($< 5\text{cm}\%$). For camera tracking, we report ATE RMSE (Sturm et al. 2012) (cm). Our main baselines include iMAP (Sucar et al. 2021), NICE-SLAM (Zhu et al. 2022), NICER-SLAM (Zhu et al. 2023b), DF Prior (Hu and Han 2023), Co-SLAM, and Go-Surf (Wang, Bleja, and Agapito 2022), ensuring fair comparison with Co-SLAM's mesh culling.

Evaluations

Results on Replica. We evaluate our method on 8 Replica scenes, comparing reconstruction accuracy with iMAP, NICE-SLAM, NICER-SLAM, Co-SLAM, and DF Prior under the same conditions. Tab. 1 shows our method significantly improves surface completion and completion ratios, with visual comparisons in Fig. 3. Our superior reconstruction is due to more accurate camera tracking, as reported in Tab. 6 and visually compared with Co-SLAM in Fig. 4.

Figure 5: Reconstruction comparisons on SyntheticRGBD.

	Acc.[cm]↓	Comp.[cm]↓	Comp. Ratio ↑
NICE-SLAM	**21.46**	7.39	60.89
DF Prior	22.91	8.26	52.08
Co-SLAM	36.89	5.75	68.46
Ours	39.67	**5.09**	**69.89**

Table 2: Reconstruction comparisons on ScanNet 6 scenes.

Results on SyntheticRGBD. Tab. 7 shows numerical comparisons with iMAP, NICE-SLAM, Co-SLAM, and DF Prior on the SyntheticRGBD (Rajpal et al. 2023) dataset. Our method achieves higher accuracy, particularly in completeness and completion ratio metrics. Fig. 5 highlights our superior geometric detail, such as window frames and floors

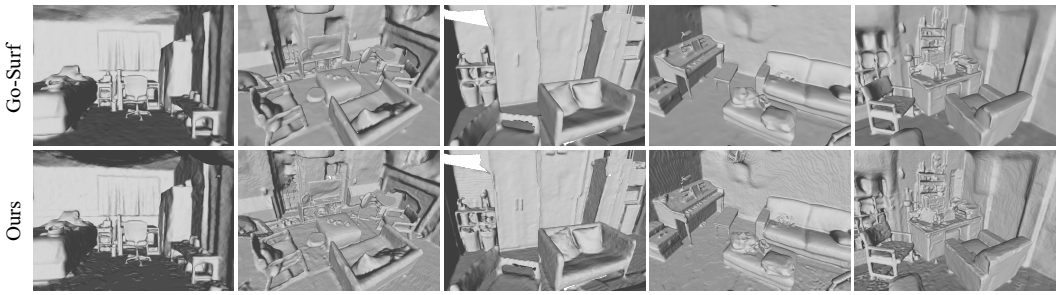

Figure 6: Visual comparisons with Go-Surf on ScanNet.

	fr1/desk (cm)	fr2/xyz (cm)	fr3/office (cm)
iMAP	4.9	2.0	5.8
NICE-SLAM	2.7	1.8	3.0
Co-SLAM	2.7	1.9	**2.67**
Ours	**2.61**	**1.7**	2.70

Table 3: ATE RMSE(cm) in tracking on TUMRGBD.

		0000	0002	0005	0050	Avg.
Go-Surf	Acc.[cm]↓	3.18	**3.48**	14.79	29.36	12.70
	Comp.[cm]↓	2.37	2.91	2.04	2.87	2.55
	Comp. Ratio [<5cm %] ↑	94.04	84.94	92.78	88.31	90.01
Ours	Acc.[cm]↓	**3.17**	4.08	**13.63**	**23.45**	**11.08**
	Comp.[cm]↓	**2.33**	**2.83**	**1.97**	**2.81**	**2.48**
	Comp. Ratio [<5cm %] ↑	**94.3**	**85.48**	**94.01**	**88.38**	**90.54**

Table 4: Numerical comparison in each scene on Scannet.

in front of sofas. Query quantized neural SLAM reconstructs smoother, more complete surfaces with enhanced detail.

Results on ScanNet. We evaluate our method on real ScanNet scans. Tab. 2 shows our method outperforms NICE-SLAM, Co-SLAM, and DF Prior numerically, while Fig. 2 highlights sharper, more compact surfaces. Tab. 5 and Fig. 4 demonstrate improved camera tracking, particularly on complex real scans, thanks to our quantized queries.

Results on TUMRGBD. We follow Co-SLAM to report our tracking performance on TUMRGBD. The numerical comparisons in Tab. 3 show that our quantized queries also make networks estimate camera poses more accurately.

Application in Multi-View Reconstruction. We evaluate our quantized queries for multi-view reconstruction using Go-Surf's neural implicit function. Tab. 4 shows our approach consistently outperforms Go-Surf on 4 ScanNet scenes in Accuracy (cm), Completion (cm), and Completion ratio ($<5cm\%$). Fig. 6 demonstrates more compact surfaces and enhanced geometric details achieved through better convergence with our quantized queries.

Scene ID	0000	0059	0106	0169	0181	0207	Avg.
iMAP	55.95	32.06	17.50	70.51	32.10	11.91	36.67
NICE-SLAM	8.64	12.25	**8.09**	10.28	**12.93**	**5.59**	9.63
Co-SLAM	7.18	12.29	9.57	6.62	13.43	7.13	9.37
Ours	**6.99**	**9.47**	8.82	**6.48**	13.30	5.86	**8.49**

Table 5: ATE RMSE(cm) comparisons on ScanNet.

	rm-0	rm-1	rm-2	off-0	off-1	off-2	off-3	off-4	Avg.
NICE	1.69	2.04	1.55	0.99	0.90	**1.39**	3.97	3.08	1.95
NICER	1.36	1.60	1.14	2.12	3.23	2.12	1.42	2.01	1.88
DF Prior	1.39	1.55	2.60	1.09	1.23	1.61	3.61	1.42	1.81
Co-SLAM	0.72	1.32	1.27	0.62	0.52	2.07	1.47	0.84	1.10
Ours	**0.58**	**1.16**	**0.87**	**0.52**	**0.48**	1.74	**1.22**	**0.73**	**0.91**

Table 6: ATE RMSE(cm) comparisons on Replica.

		BR	CK	GR	GWR	MA	TG	WR	Avg.
iMAP	Depth L1[cm]↓	24.03	63.59	26.22	21.32	61.29	29.16	81.71	47.22
	Acc.[cm]↓	10.56	25.16	13.01	11.90	29.62	12.98	24.82	18.29
	Comp.[cm]↓	11.27	31.09	19.17	20.39	49.22	21.07	32.63	26.41
	Comp. Ratio ↑	46.91	12.96	21.78	20.48	10.72	19.17	13.07	20.73
NICE	Depth L1[cm]↓	3.66	12.08	10.88	2.57	1.72	7.74	5.59	6.32
	Acc.[cm]↓	3.44	10.92	5.34	2.63	6.55	3.57	9.22	5.95
	Comp.[cm]↓	3.69	12.00	4.94	3.15	3.13	5.28	4.89	5.30
	Comp. Ratio↑	87.69	55.41	82.78	87.72	85.04	72.05	71.56	77.46
Co-SLAM	Depth L1[cm]↓	3.51	5.62	1.95	1.25	1.41	4.66	2.74	3.02
	Acc.[cm]↓	**1.97**	**4.68**	2.10	**1.89**	1.60	3.38	**5.03**	2.95
	Comp.[cm]↓	1.93	**4.94**	2.96	2.16	2.67	2.74	3.34	2.96
	Comp. Ratio ↑	94.75	68.91	90.80	95.04	86.98	86.74	84.94	86.88
Ours	Depth L1[cm]↓	**3.45**	5.63	**1.09**	1.46	**1.28**	**4.18**	**2.16**	**2.75**
	Acc.[cm]↓	2.04	7.16	**1.83**	2.07	**1.56**	**1.63**	5.25	3.07
	Comp.[cm]↓	**1.84**	5.17	**2.53**	**2.01**	**2.66**	**2.61**	**3.01**	**2.83**
	Comp. Ratio ↑	**95.88**	**69.10**	**92.44**	**95.77**	**88.01**	**87.75**	**88.14**	**88.16**

Table 7: Numerical comparison in each scene on Synthetic.

Analysis

Why Quantized Queries Work. For time-sensitive task SLAM, the network can merely get updated in few iterations like 10 in our method at each frame. Thus, the convergence efficiency is vital to inference accuracy. Our quantized queries significantly reduce the variations of input, which makes neural network always see queries that have been observed at previous frames, leading to fast overfitting on the current frame. We record the iteration when our neural network converges at each frame, and visualize the integral of converge iteration at each frame in Fig. 10. The comparison with Co-SLAM which needs continuous queries show that quantized queries need much fewer iterations than Co-SLAM to converge at a frame. We determine if the optimization converges according to the RGB rendering loss L_I with a threshold of 0.0002.

Fig. 8 (b) shows the merits of quantized codes in camera tracking. We compare ours with Co-SLAM in terms of accuracy in different iterations. We see that tracking accuracy is relatively stable and does not get larger as Co-SLAM once quantized codes converge after seeing about 500 frames.

Figure 7: Visual comparisons in ablation studies.

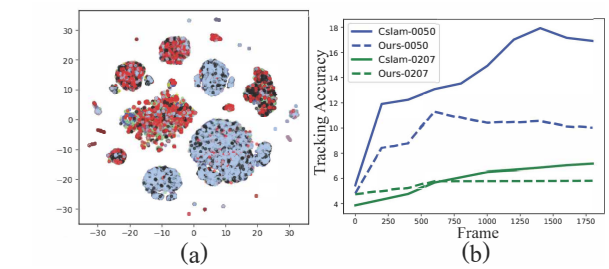

Figure 8: (a) Codebook visualization with TSNE (Color indicates segmentation labels. Sofa:Red, Wall:Blue.). (b) Comparisons of tracking accuracy (the lower the better) with Co-SLAM during optimization.

Code Distribution. For each vertex on the reconstructed meshes, we query its ID of codes and visualize the ID of codes as color on mesh in Fig. 9. We can see that different codes can be used to generate different geometry, and one code can be used to generate similar structures. We also visualize the nearest codes at all vertices with TSNE in Fig. 8 (a) and colorize these codes using the GT segmentation labels of vertices. We can see that codes show some patterns, and several codes which group together may generate the same semantic object like sofa (red) and wall (blue).

Ablation Studies

Merits of Quantization. We report merits of quantization on queries in Tab. 8.

The degenerated results with either continuous coordinate "w/o Gridcor", or continuous geometry features "w/o Codebook", or no geometry prior "w/o TSDF", or continuous geometry prior "w/o TSDF1" show the advantages of quantized queries. Fig. 7 shows the visual comparisons. We can see that the reconstruction gets decreased without the geometry prior TSDF, and we can not estimate accurate zero level set if no codebook is used.

	0050	0059	0106	0207	Avg.
w/o Gridcor	11.34	9.96	9.01	6.29	9.15
w/o Codebook	14.76	11.24	9.37	6.58	10.49
w/o TSDF	13.16	10.53	9.34	6.46	9.87
w/o TSDF1	11.19	9.87	9.14	6.17	9.09
w/o tanh	14.23	11.39	9.41	6.89	10.48
w/o Bernoulli	12.78	10.46	9.17	6.80	9.80
64 codes	15.15	11.50	9.18	6.30	10.53
256 codes	14.98	10.48	9.37	6.21	10.26
L_I	139.00	117.31	201.85	137.53	148.9
$L_I + L_D$	101.76	102.71	221.87	107.22	133.39
$L_I + L_D + L_g - L_e$	83.97	99.83	84.66	74.71	85.79
$L_I + L_D + L_g - L_e + L_{s'}$	13.76	11.01	**8.78**	6.13	9.92
Full Model	**10.02**	**9.47**	8.82	**5.86**	**8.54**

Table 8: Abalation study on 4 scenes on ScanNet. ATE RMSE(cm) comparisons in tracking.

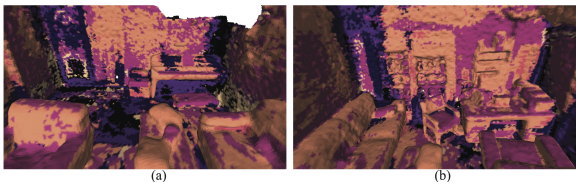

Figure 9: Code ID at vertices on the reconstructed mesh.

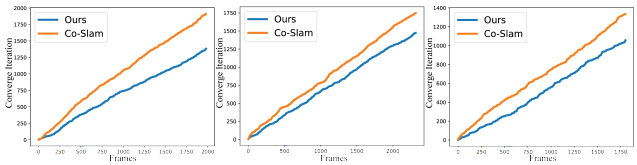

Figure 10: Merits of quantized queries in convergence on scene 0059, 0106, and 0207 from ScanNet.

Code Initialization. We conduct experiments to highlight the importance of codebook initialization. We try to use other distribution like uniform distribution to replace the Bernoulli distribution in the code initialization. The result "w/o Bernoulli" in Tab. 8 and Fig. 7 (a) shows that Bernoulli distribution can significantly stabilize the optimization by constraining the optimization space with uncertainties since very beginning. This initialization is a key to our success.

Effectiveness of Loss Terms. We justify the effectiveness of each loss term in Tab. 8. We incrementally add one loss term each time. Each loss can improve the tracking performance.

Effect of Code Number. We also explore the effect of code number in Tab. 8. We try different numbers including $B = \{64, 256\}$. We see that either fewer or more codes degenerate the performance. This is because too few codes are not enough to represent diverse geometries while too many codes make it hard to learn patterns well as codes in the overfitting on a scene.

Effect of TSDF Augmentation. Using singed distances interpolated from TSDF fusion as a geometry prior also improves reconstruction accuracy. We report results "w/o tanh" without signed distance interpolations in Tab. 8. The degenerated results justify its effectiveness and the importance of making the input different to the output.

Conclusion

We present query quantized neural SLAM for joint camera pose estimation and scene reconstruction. By quantizing queries—coordinates, positional encodings, geometry features, or priors—we reduce query variations, enabling faster neural network convergence per frame. Our novel initialization, losses, and augmentations stabilize optimization, making quantized coordinates effective for neural SLAM. Extensive evaluations on standard benchmarks show our method outperforms existing approaches in both camera tracking and reconstruction quality.

References

Atzmon, M.; and Lipman, Y. 2020. SAL: Sign Agnostic Learning of Shapes From Raw Data. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Atzmon, M.; and Lipman, y. 2021. SALD: Sign Agnostic Learning with Derivatives. In *International Conference on Learning Representations*.

Azinović, D.; Martin-Brualla, R.; Goldman, D. B.; Nießner, M.; and Thies, J. 2022. Neural RGB-D Surface Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6290–6301.

Chen, C.; Liu, Y.-S.; and Han, Z. 2022. Latent Partition Implicit with Surface Codes for 3D Representation. In *European Conference on Computer Vision*.

Corona-Figueroa, A.; Bond-Taylor, S.; Bhowmik, N.; Gaus, Y. F. A.; Breckon, T. P.; Shum, H. P.; and Willcocks, C. G. 2023. Unaligned 2D to 3D Translation with Conditional Vector-Quantized Code Diffusion using Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14585–14594.

Dupont, E.; Loya, H.; Alizadeh, M.; Goliński, A.; Teh, Y. W.; and Doucet, A. 2022. COIN++: Neural compression across modalities. *arXiv preprint arXiv:2201.12904*.

Fei, B.; Yang, W.; Chen, W.-M.; and Ma, L. 2022. VQ-DcTr: Vector-quantized autoencoder with dual-channel transformer points splitting for 3D point cloud completion. In *Proceedings of the 30th ACM international conference on multimedia*, 4769–4778.

Fu, Q.; Xu, Q.; Ong, Y.-S.; and Tao, W. 2022. Geo-Neus: Geometry-Consistent Neural Implicit Surfaces Learning for Multi-view Reconstruction. In *Advances in Neural Information Processing Systems*.

Gordon, C.; Chng, S.-F.; MacDonald, L.; and Lucey, S. 2023. On Quantizing Implicit Neural Representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 341–350.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10696–10706.

Guo, H.; Peng, S.; Lin, H.; Wang, Q.; Zhang, G.; Bao, H.; and Zhou, X. 2022. Neural 3D Scene Reconstruction with the Manhattan-world Assumption. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Haghighi, Y.; Kumar, S.; Thiran, J.-P.; and Gool, L. V. 2023. Neural Implicit Dense Semantic SLAM. arXiv:2304.14560.

Hu, P.; and Han, Z. 2023. Learning Neural Implicit through Volume Rendering with Attentive Depth Fusion Priors. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Huang, B.; Yu, Z.; Chen, A.; Geiger, A.; and Gao, S. 2024a. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24*, SIGGRAPH '24. ACM.

Huang, H.; Li, L.; Hui, C.; and Yeung, S.-K. 2024b. Photo-SLAM: Real-time Simultaneous Localization and Photorealistic Mapping for Monocular, Stereo, and RGB-D Cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jiang, S.; Hua, J.; and Han, Z. 2023. Coordinate Quantized Neural Implicit Representations for Multi-view 3D Reconstruction. In *IEEE International Conference on Computer Vision*.

Jiang, Y.; Ji, D.; Han, Z.; and Zwicker, M. 2020. SDFDiff: Differentiable Rendering of Signed Distance Fields for 3D Shape Optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Keetha, N.; Karhade, J.; Jatavallabhula, K. M.; Yang, G.; Scherer, S.; Ramanan, D.; and Luiten, J. 2024. SplaTAM: Splat, Track & Map 3D Gaussians for Dense RGB-D SLAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Koestler, L.; Yang, N.; Zeller, N.; and Cremers, D. 2022. Tandem: Tracking and dense mapping in real-time using deep multi-view stereo. In *Conference on Robot Learning*, 34–45. PMLR.

Kong, X.; Liu, S.; Taher, M.; and Davison, A. J. 2023. vMAP: Vectorised Object Mapping for Neural Field SLAM. *arXiv preprint arXiv:2302.01838*.

Laurentini, A. 1994. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2): 150–162.

Lee, S.; Park, G.; Son, H.; Ryu, J.; and Chae, H. J. 2023. FastSurf: Fast Neural RGB-D Surface Reconstruction using Per-Frame Intrinsic Refinement and TSDF Fusion Prior Learning. *arXiv preprint arXiv:2303.04508*.

Li, Y.; Dou, Y.; Chen, X.; Ni, B.; Sun, Y.; Liu, Y.; and Wang, F. 2023a. Generalized Deep 3D Shape Prior via Part-Discretized Diffusion Process. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16784–16794.

Li, Z.; Müller, T.; Evans, A.; Taylor, R. H.; Unberath, M.; Liu, M.-Y.; and Lin, C.-H. 2023b. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, S.-L.; Guo, H.-X.; Pan, H.; Wang, P.; Tong, X.; and Liu, Y. 2021. Deep Implicit Moving Least-Squares Functions for 3D Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Lorensen, W. E.; and Cline, H. E. 1987. Marching cubes: A high resolution 3D surface construction algorithm. *Computer Graphics*, 21(4): 163–169.

Matsuki, H.; Murai, R.; Kelly, P. H. J.; and Davison, A. J. 2024. Gaussian Splatting SLAM.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision*.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *arXiv:2201.05989*.

Müller, T.; McWilliams, B.; Rousselle, F.; Gross, M.; and Novák, J. 2019. Neural importance sampling. *ACM Transactions on Graphics (ToG)*, 38(5): 1–19.

Niemeyer, M.; Mescheder, L.; Oechsle, M.; and Geiger, A. 2020. Differentiable Volumetric Rendering: Learning Implicit 3D Representations without 3D Supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Oechsle, M.; Peng, S.; and Geiger, A. 2021. UNISURF: Unifying Neural Implicit Surfaces and Radiance Fields for Multi-View Reconstruction. In *International Conference on Computer Vision*.

Oord, A. v. d.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Park, K.; Sinha, U.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Seitz, S. M.; and Martin-Brualla, R. 2021. Nerfies: Deformable Neural Radiance Fields. *IEEE International Conference on Computer Vision*.

Rajpal, A.; Cheema, N.; Illgner-Fehns, K.; Slusallek, P.; and Jaiswal, S. 2023. High-Resolution Synthetic RGB-D Datasets for Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1188–1198.

Rosu, R. A.; and Behnke, S. 2023. PermutoSDF: Fast Multi-View Reconstruction with Implicit Surfaces using Permutohedral Lattices. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sandström, E.; Ta, K.; Gool, L. V.; and Oswald, M. R. 2023. Uncle-SLAM: Uncertainty Learning for Dense Neural SLAM. In *International Conference on Computer Vision Workshops (ICCVW)*.

Schönberger, J. L.; and Frahm, J.-M. 2016. Structure-from-Motion Revisited. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Schönberger, J. L.; Zheng, E.; Pollefeys, M.; and Frahm, J.-M. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision*.

Songyou Peng, L. M. M. P. A. G., Michael Niemeyer. 2020. Convolutional Occupancy Networks. In *European Conference on Computer Vision*.

Stier, N.; Ranjan, A.; Colburn, A.; Yan, Y.; Yang, L.; Ma, F.; and Angles, B. 2023. FineRecon: Depth-aware Feed-forward Network for Detailed 3D Reconstruction. *arXiv preprint*.

Straub, J.; Whelan, T.; Ma, L.; Chen, Y.; Wijmans, E.; Green, S.; Engel, J. J.; Mur-Artal, R.; Ren, C.; Verma, S.; Clarkson, A.; Yan, M.; Budge, B.; Yan, Y.; Pan, X.; Yon, J.; Zou, Y.; Leon, K.; Carter, N.; Briales, J.; Gillingham, T.; Mueggler, E.; Pesqueira, L.; Savva, M.; Batra, D.; Strasdat, H. M.; Nardi, R. D.; Goesele, M.; Lovegrove, S.; and Newcombe, R. 2019. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv preprint arXiv:1906.05797*.

Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; and Cremers, D. 2012. A Benchmark for the Evaluation of RGB-D SLAM Systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*.

Sucar, E.; Liu, S.; Ortiz, J.; and Davison, A. J. 2021. iMAP: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6229–6238.

Sun, J.; Xie, Y.; Chen, L.; Zhou, X.; and Bao, H. 2021. NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video. *IEEE Conference on Computer Vision and Pattern Recognition*.

Tang, J.; Lei, J.; Xu, D.; Ma, F.; Jia, K.; and Zhang, L. 2021. SA-ConvONet: Sign-Agnostic Optimization of Convolutional Occupancy Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Teigen, A. L.; Park, Y.; Stahl, A.; and Mester, R. 2023. RGB-D Mapping and Tracking in a Plenoxel Radiance Field. *arXiv preprint arXiv:2307.03404*.

Wang, H.; Wang, J.; and Agapito, L. 2023. Co-SLAM: Joint Coordinate and Sparse Parametric Encodings for Neural Real-Time SLAM. arXiv:2304.14377.

Wang, J.; Bleja, T.; and Agapito, L. 2022. GO-Surf: Neural Feature Grid Optimization for Fast, High-Fidelity RGB-D Surface Reconstruction. In *International Conference on 3D Vision*.

Wang, J.; Wang, P.; Long, X.; Theobalt, C.; Komura, T.; Liu, L.; and Wang, W. 2022. NeuRIS: Neural Reconstruction of Indoor Scenes Using Normal Priors. In *European Conference on Computer Vision*.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In *Advances in Neural Information Processing Systems*, 27171–27183.

Wu, H.; Lei, C.; Sun, X.; Wang, P.-S.; Chen, Q.; Cheng, K.-T.; Lin, S.; and Wu, Z. 2022. Randomized Quantization for Data Agnostic Representation Learning. *arXiv preprint arXiv:2212.08663*.

Xinyang, L.; Yijin, L.; Yanbin, T.; Hujun, B.; Guofeng, Z.; Yinda, Z.; and Zhaopeng, C. 2023. Multi-Modal Neural Radiance Field for Monocular Dense SLAM with a Light-Weight ToF Sensor. In *International Conference on Computer Vision (ICCV)*.

Yang, X.; Lin, G.; Chen, Z.; and Zhou, L. 2023a. Neural Vector Fields: Implicit Representation by Explicit Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16727–16738.

Yang, Y.; Liu, W.; Yin, F.; Chen, X.; Yu, G.; Fan, J.; and Chen, T. 2023b. VQ-NeRF: Vector Quantization Enhances Implicit Neural Representations. *arXiv preprint arXiv:2310.14487*.

Yao, Y.; Luo, Z.; Li, S.; Fang, T.; and Quan, L. 2018. MVSNet: Depth Inference for Unstructured Multi-view Stereo. *European Conference on Computer Vision*.

Yariv, L.; Kasten, Y.; Moran, D.; Galun, M.; Atzmon, M.; Ronen, B.; and Lipman, Y. 2020. Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance. *Advances in Neural Information Processing Systems*, 33.

Yu, Z.; Peng, S.; Niemeyer, M.; Sattler, T.; and Geiger, A. 2022. MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction. *ArXiv*, abs/2022.00665.

Yu, Z.; Sattler, T.; and Geiger, A. 2024. Gaussian Opacity Fields: Efficient and Compact Surface Reconstruction in Unbounded Scenes. arXiv:2404.10772.

Zhang, Y.; Tosi, F.; Mattoccia, S.; and Poggi, M. 2023. GO-SLAM: Global Optimization for Consistent 3D Instant Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhao, W.; Lei, J.; Wen, Y.; Zhang, J.; and Jia, K. 2020. Sign-Agnostic Implicit Learning of Surface Self-Similarities for Shape Modeling and Reconstruction from Raw Point Clouds. *CoRR*, abs/2012.07498.

Zhou, T.; Brown, M.; Snavely, N.; and Lowe, D. G. 2017. Unsupervised Learning of Depth and Ego-Motion from Video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6612–6619.

Zhu, Z.; Peng, S.; Larsson, V.; Cui, Z.; Oswald, M. R.; Geiger, A.; and Pollefeys, M. 2023a. NICER-SLAM: Neural Implicit Scene Encoding for RGB SLAM. *arXiv preprint arXiv:2302.03594*.

Zhu, Z.; Peng, S.; Larsson, V.; Cui, Z.; Oswald, M. R.; Geiger, A.; and Pollefeys, M. 2023b. NICER-SLAM: Neural Implicit Scene Encoding for RGB SLAM. *CoRR*, abs/2302.03594.

Zhu, Z.; Peng, S.; Larsson, V.; Xu, W.; Bao, H.; Cui, Z.; Oswald, M. R.; and Pollefeys, M. 2022. NICE-SLAM: Neural Implicit Scalable Encoding for SLAM. In *IEEE Conference on Computer Vision and Pattern Recognition*.