Neural Graph Mapping for Dense SLAM with Efficient Loop Closure

Leonard Bruns[1] Jun Zhang[2] Patric Jensfelt[1]
[1]KTH Royal Institute of Technology, Stockholm, Sweden
[2]TU Graz, Graz, Austria
{leonardb, patric}@kth.se, jun.zhang@tugraz.at

Abstract

Existing neural field-based SLAM methods typically employ a single monolithic field as their scene representation. This prevents efficient incorporation of loop closure constraints and limits scalability. To address these shortcomings, we propose a neural mapping framework which anchors lightweight neural fields to the pose graph of a sparse visual SLAM system. Our approach shows the ability to integrate large-scale loop closures, while limiting necessary reintegration. Furthermore, we verify the scalability of our approach by demonstrating successful building-scale mapping taking multiple loop closures into account during the optimization, and show that our method outperforms existing state-of-the-art approaches on large scenes in terms of quality and runtime. Our code is available open-source at https://github.com/roym899/neural_graph_mapping.

1. Introduction

Simultaneous localization and mapping (SLAM) using cameras, often referred to as visual SLAM, has been a long standing problem in computer vision [7, 9, 14]. In particular, dense visual SLAM aims to construct a detailed geometric representation of the environment enabling various applications, such as, occlusion handling in augmented reality [26], planning and collision checking in robotics [1], or camera localization [21]. Often volumetric scene representations are employed as they are well-suited for online data integration. Traditional volumetric representations use grid-based structures [25], often incorporating acceleration approaches like octrees [10, 18] or voxel hashing [27] in practice. However, incorporating loop closure constraints in volumetric maps is difficult [6, 30, 41] compared to sparse keypoint-based maps, which can easily be deformed.

Recently, neural fields have emerged as a memory-efficient volumetric scene representation due to their amenability to differentiable rendering-based optimization [34]. Building on the first neural field-based SLAM method iMAP [33], various adaptations of this method have been proposed. These adaptations primarily focused on enhancing optimization speed by altering network architecture, sampling scheme, and rendering formulation. Nevertheless, the majority of existing neural field-based SLAM methods [12, 33, 39, 43, 44] remain constrained by their monolithic data structure (i.e., a single fixed-size architecture). In that regard, neural fields share the same limitation as other volumetric scene representations, that is, after data has been integrated, the volumetric scene cannot easily be deformed to take loop closure constraints into account.

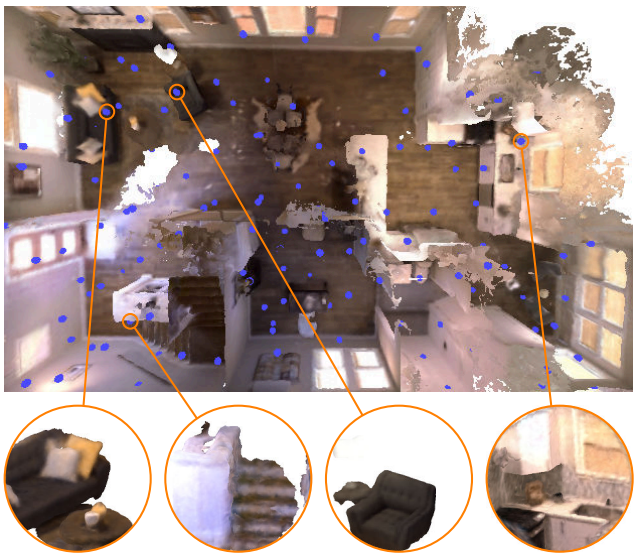


Figure 1. We propose to represent a scene by a set of neural fields (centers indicated by blue spheres) anchored to keyframes in a pose graph with each field capturing the scene within a ball surrounding it. This allows to dynamically extend the scene while also incorporating loop closure deformations into the volumetric scene representation without requiring full reintegration.

To address this issue, we propose to represent the scene by an extendable set of lightweight neural fields (see Fig. 1). These fields are anchored to keyframes within a pose graph, collectively forming the volumetric map while maintaining the ability to deform as the keyframe poses change upon

loop closure. Our design allows to combine the benefits of sparse pose graph-based SLAM methods while maintaining a volumetric scene representation that remains consistent with the pose graph without requiring costly full reintegration of previous sensor data. Furthermore, our method eliminates the necessity for fixed scene boundaries common among existing neural-field-based SLAM methods. It achieves this by allocating additional fields dynamically as new parts of the scene come into view.

To summarize, our contributions are

- a novel SLAM framework that combines the robust, accurate tracking and efficient loop closure handling of sparse visual SLAM with the differentiable-rendering-based dense mapping of neural scene representations,
- a multi-field scene representation in which neural fields are anchored relative to a pose graph allowing for large-scale map deformations while limiting necessary reintegration, and
- a thorough comparison to multiple state-of-the-art methods on scenes of varying scales; including a novel set of sequences for the larger Replica scenes that allow to benchmark robustness and scalability.

2. Related Work

2.1. Traditional Volumetric Loop Closure

The difficulty of incorporating loop closure constraints in volumetric scene representations is not limited to neural representations. Traditional volumetric representations such as sparse grids and octrees also require solutions to efficiently adapt to pose graph changes. Notable examples include: Kintinuous [41] in which the volumetric map is only used for local fusion and globally a mesh is deformed based on a deformation graph; BundleFusion [6] in which frames whose pose has changed are removed and reintegrated prioritized based on the amount of change; and VoxGraph [30] in which submaps are aligned and resulting constraints are included in the pose graph optimization. In this work, we aim to enable efficient loop closure specifically for neural field representations in which data reintegration is expensive.

2.2. Neural Field-Based Scene Representations

Neural fields are parametrized differentiable functions that map coordinates to quantities at that point in space. Initially, neural fields were trained using 3D supervision and were used to represent shape as occupancy [4, 19] or signed distance fields (SDFs) [28]. Shortly after, neural radiance fields (NeRFs) were introduced optimized using only 2D images through differentiable volume rendering [20].

Focusing on surface reconstruction, multiple works proposed SDF-based representations replacing the density-based approach in NeRFs. [2] use a heuristic rendering model that allows to optimize SDFs yielding improved surface reconstructions; however, their model is not occlusion-aware. An occlusion-aware approach for unbiased SDF-based volume rendering was proposed in [40].

Focusing on training and inference speed, [23] showed significant speed-ups by employing a multi-resolution hash encoding in which learnable features in a hash table are used instead of a fixed Fourier encoding to lift the low-dimensional, low-frequency input point to a high-dimensional, high-frequency embedding. [31] further proposed to replace the multi-resolution voxel grid with a multi-resolution permutohedral lattice due to its better scalability to higher dimensions. We adopt this approach as the positional encoding for each field.

Fewer works have investigated the use of multiple neural fields. In Block-NeRF [35] multiple fields are combined to represent whole neighborhoods. However, the field positions are predefined. In Nerflets [42] fields of varying size are composed to achieve an editable scene representation. NeRFuser [8] proposes methods for registering and blending multiple NeRFs. Finally, vMAP [15] uses multiple fields to represent individual objects highlighting the potential of vectorizing neural field evaluations. The latter is also employed in this work.

2.3. Neural Field-Based SLAM

iMAP [33] was the first SLAM method using a multilayer perceptron (MLP) as the underlying scene representation. During optimization previous keyframes have to be continuously reintegrated to avoid forgetting. Subsequently, various modifications of this approach have been proposed aiming to alleviate this forgetting issue and improve run-time.

NICE-SLAM [44] addresses the forgetting issue by employing a hierarchical feature grid combined with a fixed, pretrained decoder MLP. By adopting an SDF-based field combined with efficient learnable positional encodings, ESLAM [12] and Co-SLAM [39] demonstrate significantly improved optimization times. Specifically, ESLAM uses a tri-plane [3] encoding, whereas Co-SLAM uses the aforementioned multi-resolution voxel hash encoding [23]. Co-SLAM further complements the hash encoding with one-blob encodings [22] to improve scene completion. However, none of the above approaches support integration of loop closure constraints, limiting their applicability to larger scenes, in which accumulation of drift is unavoidable.

Matsuki et al. [17] propose a similar idea to ours, also using multiple fields posed relative to a pose graph. However, they employ space warping in which each field in principle covers the entire scene. Synthesized views from multiple fields are merged through alpha compositing. This approach is well-suited for novel view synthesis; however, generalizes poorly to mesh extraction and other geometric queries. Instead, in our work, each field covers a Euclidean

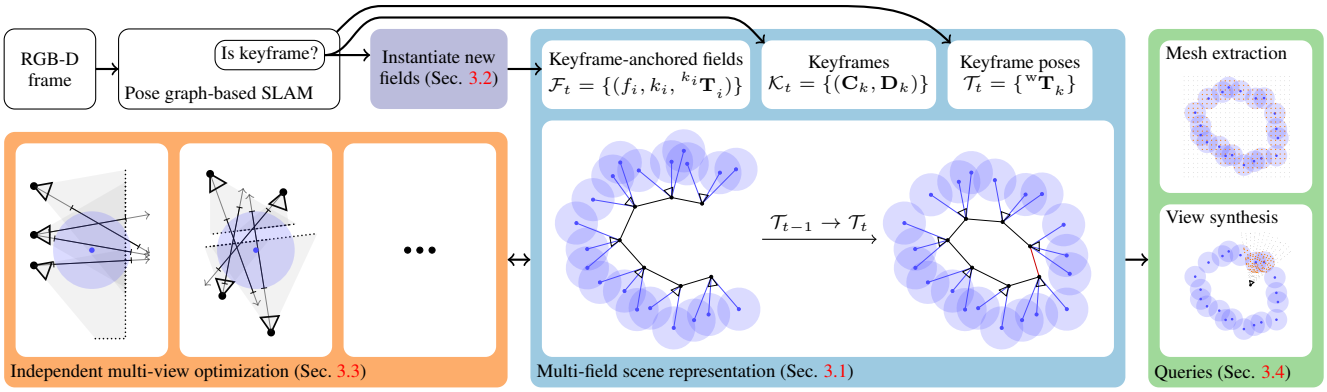


Figure 2. Neural graph map framework overview. Our framework can be used with any pose graph-based SLAM system and maintains a set of neural fields anchored relative to the keyframes of the SLAM system's pose graph. Each field captures the scene in a sphere surrounding it in local coordinates. The set of such fields represents a deformable volumetric scene representation that can easily adapt to and stay in sync with the pose graph. To enable efficient optimization and scalability the fields are optimized independently and in parallel. This design allows to average the output of overlapping fields in 3D leading to well-defined queries (e.g., mesh extraction and novel view synthesis).

ball around it, which allows general geometric queries. In MIPS-Fusion [36] global bundle adjustment at the level of fields is performed. Notably, both MIPS-Fusion and NEWTON are limited to a small number ($\sim$10) of fields, where each field can be interpreted as a submap. In contrast, our formulation allows training and maintenance of hundreds of fields and can therefore incorporate local and global pose graph deformations in a more fine-grained manner.

Finally, GO-SLAM [43] also incorporates loop closure constraints through extending DROID-SLAM [38] and biasing the optimization towards keyframes whose pose changed the most. However, it still uses a single neural field, which requires reoptimization after loop closure. Instead we aim to achieve instant map deformation upon loop closure by anchoring fields to keyframes.

3. Method

We consider the problem of online mapping with an RGB-D camera without ground-truth poses. That is, given a stream of RGB-D frames $(\mathbf{C}_t, \mathbf{D}_t)$ composed of color images $\mathbf{C}_t \in \mathbb{R}^{H \times W \times 3}$ and depth maps $\mathbf{D}_t \in \mathbb{R}^{H \times W}$, our goal is to build a dense scene representation at each step $t \in \{1, ..., T\} = \mathcal{N}_t$ taking into account only the data from the current and previous steps $\mathcal{N}_{\leq t}$.

Figure 2 gives an overview of the proposed framework. Similar to other recent neural field-based mapping approaches [15, 17, 43], our approach relies on an off-the-shelf keyframe-based SLAM system to provide a set of posed keyframes as well as the pose of the current frame.

3.1. Multi-Field Scene Representation

Let $\mathcal{K}_t = \{(\mathbf{C}_k, \mathbf{D}_k) \mid k \in \mathcal{N}_k \subseteq \mathcal{N}_{\leq t}\}$ and $\mathcal{T}_t = \{{}^{\mathrm{w}}\mathbf{T}_k \in \mathrm{SE}(3) \mid k \in \mathcal{N}_k\}$ denote the set of keyframes and keyframe poses at time step t, respectively.

We propose to represent the scene by an extendable set of posed neural fields

$$\mathcal{F}_t = \{(f_i, k_i, {}^{k_i}\mathbf{T}_i) \mid i \in \{1, ..., F_t\} = \mathcal{N}_f\}, \quad (1)$$

where each field

$$\begin{aligned} f_i : \mathbb{R}^3 &\to \mathbb{R}^3 \times \mathbb{R} \\ \boldsymbol{x} &\mapsto (\boldsymbol{c}, s) \end{aligned} \quad (2)$$

maps a 3D point $\boldsymbol{x}$ to a color $\boldsymbol{c}$ and truncated signed distance s at that point (see Sec. 3.3 for further details). Each field is rigidly attached to a parent keyframe with index $k_i \in \mathcal{I}(\mathcal{K}_t)$ (with $\mathcal{I}$ denoting the set of indices) by a transform ${}^{k_i}\mathbf{T}_i \in \mathrm{SE}(3)$, which can be decomposed into a translation ${}^{k_i}\boldsymbol{t}_i \in \mathbb{R}^3$ and rotation ${}^{k_i}\mathbf{R}_i \in \mathrm{SO}(3)$. To limit the complexity that each field stores, each field only captures the scene within a sphere of fixed radius r around its center.

In the following sections the proposed strategies to instantiate and anchor fields to the pose graph (Sec. 3.2), optimize the fields (Sec. 3.3), and query (e.g., for mesh extraction and view synthesis) the full scene representation (Sec. 3.4) are described. Finally, further architecture details are provided (Sec. 3.5).

3.2. Field Instantiation

Whenever a new keyframe k is added to the pose graph new fields are instantiated such that all observed 3D points ${}^{\mathrm{w}}\mathcal{X}_k$ in the keyframe are covered by at least one field of radius r. An approximate two-stage algorithm is used (see Fig. 3). First, the uncovered 3D points $\mathcal{X}_{\mathrm{unc}} = \{\boldsymbol{x} \in {}^{\mathrm{w}}\mathcal{X}_k \mid \|\boldsymbol{x} - {}^{\mathrm{w}}\mathbf{T}_{k_i}\, {}^{k_i}\boldsymbol{t}_i\| > r \,\forall\, i \in \mathcal{N}_f\}$ are found. Second, the space is divided into voxel cells with a side length of $g = 2r/\sqrt{3}$, such that a cell is fully covered by a field of radius r when the center of the field is at the center of the

cell. New fields are instantiated in the center of all cells that contain a point from $\mathcal{X}_{\text{unc}}$ and no field center. This scheme ensures that a minimum distance of $g/2$ to existing fields is maintained; however, points in $\mathcal{X}_{\text{unc}}$ might remain uncovered when fields move away from the cell centers through pose graph deformations. To alleviate the chance of uncovered points over time, the voxel grid is randomly shifted for every added keyframe.

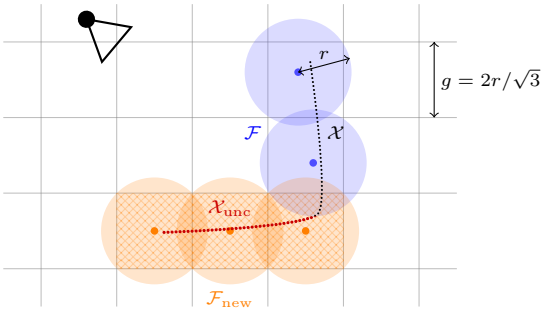


Figure 3. Grid-based instantiation scheme. A new keyframe (top-left) observes world points $\mathcal{X}$ (black & red dots). The points $\mathcal{X}_{\text{unc}}$ (red dots) that are not covered by any of the existing fields $\mathcal{F}$ (blue circles) are determined. Cells that contain uncovered points and no existing field center (hatched orange) are used to instantiate new fields $\mathcal{F}_{\text{new}}$ (orange circles). The new fields are positioned in the center of the uncovered cells and anchored to the keyframe.

3.3. TSDF-Based Optimization

For each new frame coming at time step t, a fixed number of optimization steps N_{it} are performed on up to N_{f} fields. During the optimization the fields f_i are trained independently, allowing for efficient parallelization [15]. Therefore, we drop the field index i in the following.

Due to its well-defined isosurface for mesh extraction, we model the geometry as a truncated signed distance field (TSDF) adopting the TSDF-based rendering model from [2] and modify it to be occlusion-aware. Specifically, instead of directly converting signed distances to sample weights as done in [2, 39], we first convert them to occupancy probabilities. That is, given N_{s} samples along a ray $\boldsymbol{x}_i = \boldsymbol{o} + l_i\boldsymbol{d}, i = 1, ..., N_{\text{s}}$ with origin $\boldsymbol{o}$ and direction $\boldsymbol{d}$, the colors and signed distances $(\boldsymbol{c}_i, s_i) = f(\boldsymbol{x}_i)$ along the ray are computed. The signed distances are converted to occupancy probabilities using

$$o_i = 4\sigma\left(\frac{\eta s_i}{\tau}\right)\sigma\left(-\frac{\eta s_i}{\tau}\right), \tag{3}$$

where σ denotes the sigmoid function, τ is the truncation distance, and η is a parameter that determines how quickly occupancy probability decays around the surface. Note that this definition ensures $o_i = 1.0$ for $s_i = 0.0$. The ray's rendered color and depth are then computed via the weight $w_i = o_i \prod_{j=1}^{i-1}(1 - o_j)$ as

$$\boldsymbol{c} = \sum_{i=1}^{N_{\text{s}}} w_i \boldsymbol{c}_i \quad d = \sum_{i=1}^{N_{\text{s}}} w_i l_i \tag{4}$$

assuming the ray's direction $\boldsymbol{d}$ was scaled appropriately.

Sampling Strategy Each optimization iteration follows a three-stage sampling procedure. Specifically, we sample: (1) N_{f} fields that will be optimized in the next iteration; (2) N_{r} ray segments (with associated observed color $\mathbf{c}$ and distance d) for each sampled field; (3) N_{s} query points along each sampled ray segment. The sampling strategy is visualized in Fig. 4 (see supplementary material for further details).

Loss Our loss is a weighted sum of four terms: color, depth, TSDF, and free-space. The weighing has been found experimentally to achieve a good trade-off between appearance and geometry quality.

The color loss is computed as the mean L1-norm of the difference between observed and estimated color

$$l_{\text{color}} = \frac{1}{|\mathcal{C}|}\sum_{(\boldsymbol{c},\tilde{\boldsymbol{c}})\in\mathcal{C}} \|\boldsymbol{c} - \tilde{\boldsymbol{c}}\|_1 \tag{5}$$

and the depth loss is computed as the mean Huber loss [11] between observed and estimated depth with $\delta = 5\,\text{cm}$

$$l_{\text{depth}} = \frac{1}{|\mathcal{D}|}\sum_{(d,\tilde{d})\in\mathcal{D}} \text{Huber}_\delta(d, \tilde{d}), \tag{6}$$

where $\mathcal{C}$ and $\mathcal{D}$ are sets containing tuples of the observed and estimated color and depth, respectively.

Finally, the set of all query points $\mathcal{X}$ is filtered into two sets $\mathcal{X}_{\text{tsdf}}$ and $\mathcal{X}_{\text{fs}}$ based on whether a query point's projected depth l_i is within the truncation distance τ of the corresponding observed depth $l_{\text{obs},i}$ or is more than τ in front of it, respectively. A query point's associated estimated signed distance $\tilde{s}_i$ and its signed distance to the observed depth $s_i = l_{\text{obs},i} - l_i$ are then used to compute the TSDF loss as

$$l_{\text{tsdf}} = \frac{1}{|\mathcal{X}_{\text{tsdf}}|}\sum_{i\in\mathcal{I}(\mathcal{X}_{\text{tsdf}})} (\tilde{s}_i - s_i)^2 \tag{7}$$

and the free-space loss as

$$l_{\text{fs}} = \frac{1}{|\mathcal{X}_{\text{fs}}|}\sum_{i\in\mathcal{I}(\mathcal{X}_{\text{fs}})} (\tilde{s}_i - \tau)^2. \tag{8}$$

Table 1. Quantitative mesh reconstruction quality on the synthetic dataset (**best** ●, second best ●).

		Replica	NRGBD									Replica-Big		
		Avg.	br	ck	gr	gwr	ki	ma	sc	tg	wa	apt0	apt1	apt2
NICE-S. [44]	Acc (cm)	**2.38** ●	2.46	10.76	2.33	2.71	9.18	1.70 ●	4.55	8.37	7.37	15.98	17.56	23.39
	Acc. R. (%)	**93.17** ●	92.41 ●	65.34	93.87	93.63	57.29	95.06 ●	71.69	56.22	77.38	30.67	41.48	28.20
	Comp. (cm)	2.84 ●	4.82	14.21	3.91	3.19	12.82	3.36	10.69	8.02	5.39	12.73	11.59	19.77
	Comp. R. (%)	90.11 ●	86.18	53.58	86.66	87.64	50.04	84.06 ●	58.79	59.24	69.04	31.06	42.21	26.80
Co-S. [39]	Acc (cm)	4.29	**2.21** ●	4.73	**1.89** ●	**2.02** ●	7.40	1.74	3.30 ●	2.07	6.24	37.69	12.89	30.19
	Acc. R. (%)	87.02	**93.24** ●	75.16	**95.17** ●	**94.84** ●	77.72	93.98	78.01 ●	92.05 ●	84.92 ●	18.85	62.11	34.71
	Comp. (cm)	4.09	**2.06** ●	**8.76** ●	2.93 ●	**2.41** ●	5.14 ●	**2.75** ●	**4.29** ●	**2.83** ●	**3.85** ●	17.91	7.07	17.75
	Comp. R. (%)	86.39	93.49 ●	63.17 ●	91.32 ●	**93.96** ●	**78.19** ●	**86.41** ●	70.90 ●	**86.62** ●	**81.70** ●	21.93	66.37	36.75
GO-S. [43]	Acc (cm)	2.89	3.89	4.08 ●	2.50	2.87	3.28 ●	**1.54** ●	6.46	**1.48** ●	5.46 ●	14.20 ●	8.92 ●	11.82 ●
	Acc. R. (%)	89.31	77.64	81.94 ●	91.46	86.87	84.74 ●	**97.44** ●	66.62	**96.32** ●	73.90	**57.93** ●	74.38 ●	**80.22** ●
	Comp. (cm)	6.18	9.25	29.60	9.50	4.50	**5.11** ●	4.60	12.35	7.26	12.57	**7.98** ●	5.19 ●	5.87 ●
	Comp. R. (%)	74.20	64.29	54.56	71.31	75.12	72.58	75.53	54.24	70.78	57.33	**59.21** ●	73.90 ●	81.06 ●
Ours	Acc (cm)	2.88 ●	2.26 ●	**3.21** ●	1.90 ●	2.19 ●	**2.19** ●	1.89	**2.60** ●	2.05 ●	**4.00** ●	**10.14** ●	**5.26** ●	**11.45** ●
	Acc.-Ratio (%)	89.50 ●	90.58	**87.81** ●	94.60 ●	93.78 ●	**96.18** ●	93.01	**93.45** ●	91.50	**85.97** ●	56.09 ●	**91.64** ●	72.35 ●
	Comp. (cm)	**2.36** ●	2.42 ●	9.03 ●	**2.71** ●	2.44 ●	30.37	3.11 ●	4.67 ●	3.66 ●	4.64 ●	8.71 ●	**2.48** ●	**3.94** ●
	Comp. R. (%)	**92.11** ●	**94.01** ●	**77.83** ●	**91.62** ●	92.35 ●	73.03 ●	83.77	**87.78** ●	81.76 ●	81.07 ●	58.52 ●	**96.26** ●	**82.70** ●

* Acknowledging the small differences on most Replica sequences, we report the average here and include the full results in the supplementary material.

† We evaluated Co-SLAM without ground-truth initialization for fair comparison (see supplementary material for further discussion.)

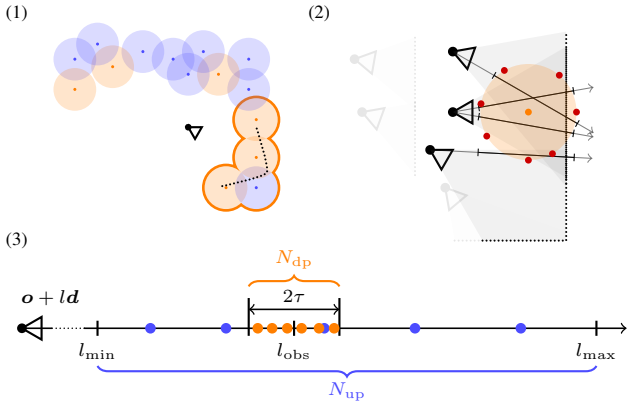


Figure 4. Multi-view sampling procedure. (1) A subset of fields (orange circles) to optimize in the next iteration is sampled with half the fields stemming from the currently observed fields (orange outline). (2) For each field, the observing keyframes (opaque frustums) are approximated based on whether samples on the field boundary (red dots) fall into the observed keyframe region. Ray segments to supervise the field are then sampled from the observing keyframes. (3) Each sampled ray segment $[l_{\text{min}}, l_{\text{max}}]$ is approximated by N_{up} uniformly sampled query points that cover the whole field and N_{dp} depth-guided query points distributed within the truncation distance τ around the observed distance l_{obs}.

3.4. Novel View Synthesis and Mesh Extraction

Due to the independent optimization described in Sec. 3.3 all fields will be supervised at least within a sphere of radius r. Therefore, overlapping regions will contain different colors and distances. To reduce transition artifacts at the boundary of fields, it is possible to query the k nearest fields and average the outputs based on the field distances (see supplementary material for further details). For query points outside of all fields' spheres empty space is assumed, that is, $s = 1$.

Novel View Synthesis Since no depth observation is available at inference time, a fixed far distance l_{far} is used. Query points are uniformly distributed within the ray interval $[0, l_{\text{far}}]$ at the same density as the depth-guided points, that is, $N_{\text{p,inf}} = l_{\text{far}} N_{\text{dp}} / (2\tau)$ samples are used.

Mesh Extraction To extract a mesh from our scene representation the marching cubes algorithm [16] with an unbiased isosurface level of 0 is used.

3.5. Architecture Details

ORB-SLAM2 is used as the keyframe-based SLAM system in this work [24]. However, our proposed method is agnostic to the exact SLAM framework as long as a pose graph as described in Sec. 3.1 is available. We precomputed the ORB-SLAM results and will publish them in conjunction with our code to improve reproducibility.

Each field f_i is parameterized by a permutohedral hash encoding [31] followed by a linear layer with ReLU activation and another linear layer mapping to the output color and distance. We also experimented with other positional encodings, such as Fourier encoding [34], triplane encoding [3, 12], and voxel hash encoding [23]; however, the best results given the same time budget were achieved with the permutohedral hash encoding closely followed by the voxel hash encoding.

To efficiently evaluate and optimize multiple neural fields in parallel, optimization and evaluation is vectorized following [15]. One limitation imposed by PyTorch's [29] vectorization framework is that the batch size has to be the same for all evaluated networks. The design choices de-

scribed in Sec. 3.3 are tailored to this limitation to allow sampling the same number of rays per supervised field.

4. Experiments

To verify the proposed approach, we report results following the reconstruction-based evaluation protocol of Wang et al. [39].

4.1. Experimental Setup

Datasets We evaluate our method quantitatively on three synthetic datasets of varying difficulty: the *Replica* [32] sequences from iMAP [33], the *NRGBD* dataset by [2], and three novel sequences on the larger Replica scenes (*Replica-Big*). The latter contains challenging trajectories with larger loops spanning multiple rooms. Further, we present qualitative results on two real-world datasets: ScanNet [5] and Kintinuous [41]. The latter contains a large-scale loop, whereas the tested ScanNet sequences typically contain one or two smaller loops.

Metrics We adopt the reconstruction metrics reported by [39], but use completion ratio instead of depth L1, which is skewed by holes in the mesh and therefore less intuitive than the point set-based metrics. Completion ratio, defined akin to accuracy ratio, naturally completes their point set-based metrics. Therefore, we report the methods' accuracy, completion, accuracy ratio, and completion ratio at a $5\,\mathrm{cm}$ threshold. We further use F1-score to summarize reconstruction quality in a single metric [37]. We do not evaluate tracking metrics, since we use ORB-SLAM2 as the tracking backend and our focus is on dense mapping, while efficiently supporting map deformations due to loop closures.

Baselines We compare our method to three state-of-the-art neural field-based SLAM methods: NICE-SLAM [44], Co-SLAM [39], and GO-SLAM [43]. Co-SLAM is, to the best of our knowledge, the best available method in terms of quality and run-time at the time of writing. However, neither NICE-SLAM nor Co-SLAM can integrate loop closure constraints. GO-SLAM is, to our knowledge, the only other published neural field-based method that also incorporates loop closure constraints.

Implementation Details For all experiments we use the same parameters, most notably, $r = 1\,\mathrm{m}$, $N_{\mathrm{it}} = 5$, $N_{\mathrm{f}} = 32$, $N_{\mathrm{r}} = 512$, $N_{\mathrm{up}} = 8$, $N_{\mathrm{dp}} = 16$ and $k = 2$. We adjust the truncation distance τ from $0.1\,\mathrm{m}$ for synthetic data to $0.2\,\mathrm{m}$ for real data to account for the increased depth noise. For ORB-SLAM2 we use the default parameters in the RGB-D setup increasing the number of extracted features in challenging scenes with low texture. All remaining parameters are provided in the supplementary material and as part of the released code.

Table 2. Run-time and model size comparison of all methods. Our method compares favorably in terms of processing time; however, is less efficient in terms of model size.

	FPS (Hz)			Model Size (MB)		
	`br`	`ck`	`apt0`	`br`	`ck`	`apt0`
NICE-SLAM	2.74	2.70	2.25	**5.3**	21.6	57.1
Co-SLAM	12.70	12.84	10.75	7.0	**7.0**	**7.0**
GO-SLAM	17.35	12.06	6.36	66.5	66.5	66.5
Ours	**20.52**	**20.17**	**20.15**	27.0	82.1	222.3

4.2. Reconstruction Quality Evaluation

Table 1 reports quantitative results on all datasets. Our method achieves state-of-the-art results on the smaller scenes (Replica, NRGBD), while outperforming existing methods on larger scenes (Replica-Big). The qualitative results on Replica-Big shown in Figure 5 further illustrates the differences between the methods. Both NICE-SLAM and Co-SLAM drift off and cannot correct for the drift upon returning to previous frames leading to poor results. GO-SLAM also benefiting from loop closures successfully maps large parts; however, fails in the last section of `apt0` and generally exhibits worse reconstruction quality. While GO-SLAM achieves quantitatively similar results as our method on `apt0`, we note that qualitatively our method shows better scene completion and robustness.

In Fig. 15 we further show results on two scenes of the ScanNet dataset also including a comparison to BundleFusion [6]. The benefit of neural scene representations over classic representations in terms of scene completion can easily be observed. While not as much drift accumulates on these scenes, our method still achieves on par results to the otherwise best performing method Co-SLAM, while achieving slightly more globally consistent results (see, e.g., bottom-right corner of carpet).

Figure 7 shows results on the Kintinuous sequence before and after loop closure without additional optimization. Specifically, our framework allows immediate map deformation incorporating the large-scale loop closure by only adjusting the field poses connecting the corridors. GO-SLAM did not close the loop and Co-SLAM failed in the second corner.

4.3. Run-Time and Model Size Analysis

In Tab. 2 we compare run-time[1] and model size of our method to the baselines on three representative scenes of different sizes. Despite the overhead of parallelized training of multiple fields, our approach compares favorably in

[1] We report time per frame by benchmarking the execution time for the whole sequence and dividing it by the number of frames. Mesh extraction and other logging has been deactivated for all methods for fair comparison. All evaluations were run with an Intel Core i9-13900KF and NVIDIA GeForce RTX 4090.

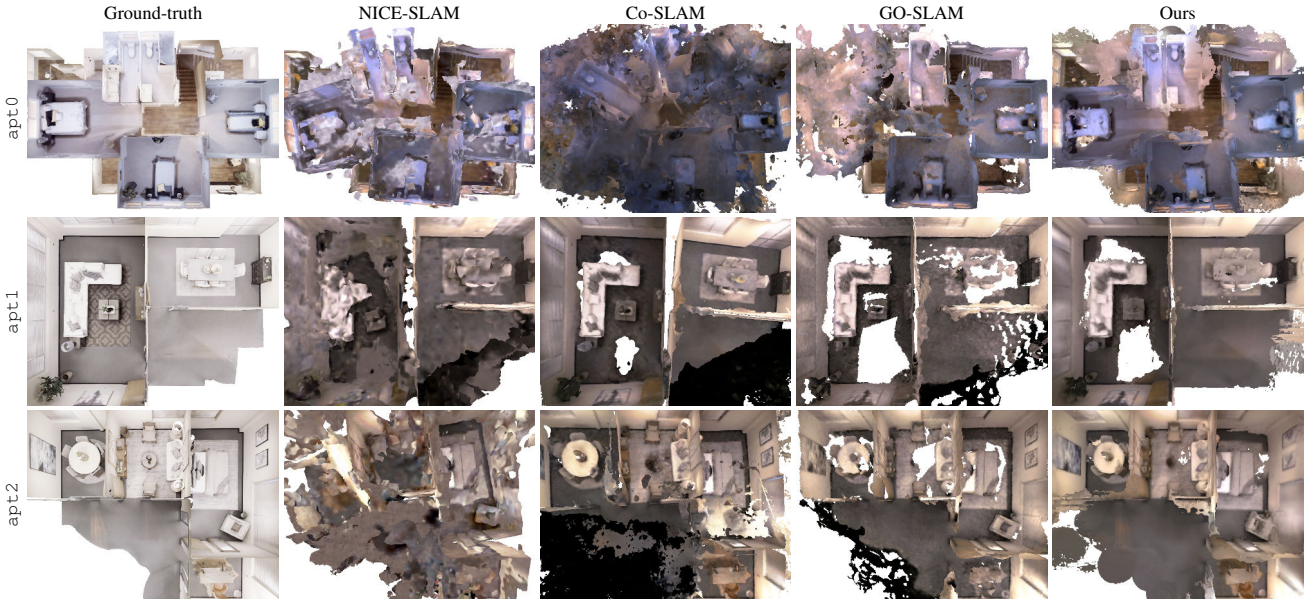


Figure 5. Qualitative comparison of final reconstruction on Replica-Big. NICE-SLAM and Co-SLAM do not achieve globally consistent results. GO-SLAM and ours achieve significantly better results due to supporting loop closure. Particularly in terms of scene completion our method outperforms GO-SLAM.

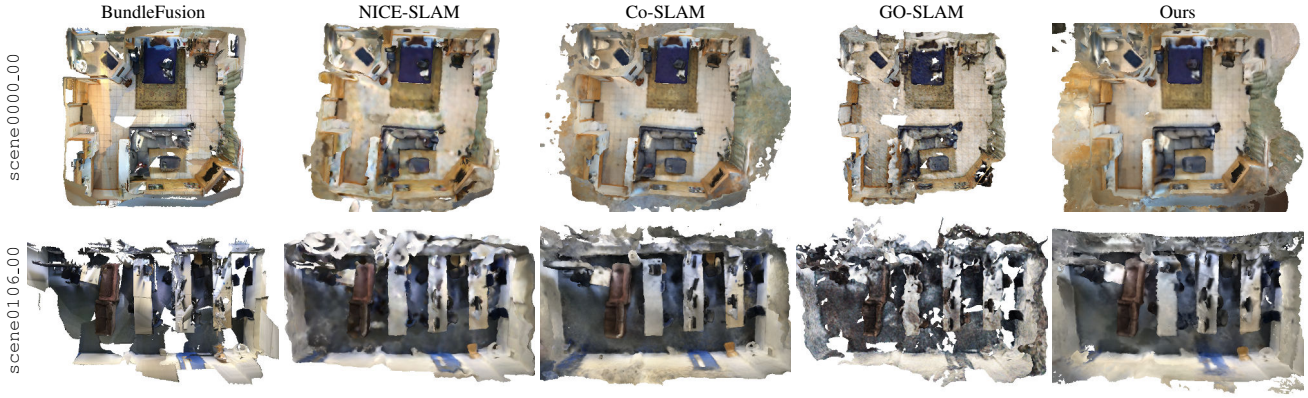
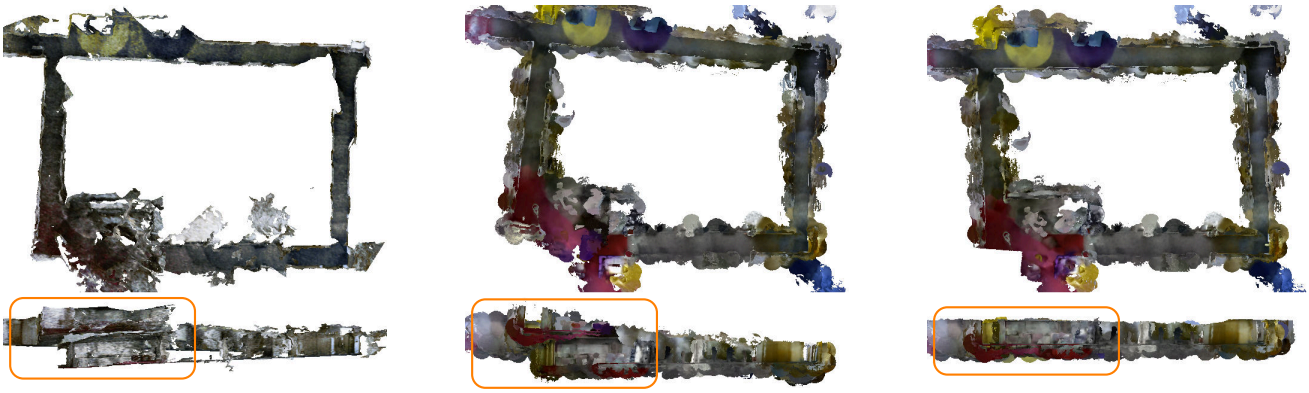


Figure 6. Qualitative comparison of final reconstruction on ScanNet. BundleFusion achieves globally consistent results; however, leaves many regions uncompleted. Our method achieves overall the best alignment, while maintaining the advantages in terms of scene completion of neural scene representations.



(a) Go-SLAM end of sequence (frame 4014) (b) Ours before loop closure (frame 3855) (c) Ours after loop closure (frame 3856)

Figure 7. Qualitative results on the Kintinuous sequence (first row shows top view; second row shows side view). No additional optimization has been performed between extracting the two meshes for our method; only the field poses have changed successfully connecting the corridors (note the duplicate floors in the highlighted regions in (a) and (b)).

terms of run-time. We note that GO-SLAM significantly slows down for longer sequences, whereas our methods maintains a consistent frame rate. Our proposed multi-field representation requires more parameters than the monolithic baselines especially for larger scenes. We suspect this is due to the less efficient hash table use, which depending on the field's content will be unnecessarily large. Sharing hash tables among multiple fields might be a viable future direction to further reduce the model size.

Field Radius and Hash Map Size In Fig. 8 the memory requirement and F1-score for different field radii r and hash table sizes T is reported. In general, larger fields require larger hash maps to maintain the same quality. Our approach works across a wide range of r, but the ability to adapt to loop closures will be limited for larger r.

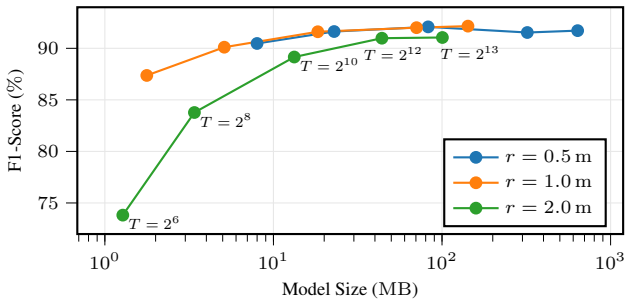


Figure 8. Impact of field radius r and hash table size T on model size and reconstruction quality. Larger radii lead to fewer fields and smaller model size. However, larger radii require larger hash tables to achieve similar quality.

Batch Size and Iterations Increasing the number of iterations per frame or the batch size will directly influence run-time and quality. In Fig. 9 the processing time per frame and F1-score is reported for various settings. Notably, our approach maintains high reconstruction quality even at frame rates of over 50 Hz.

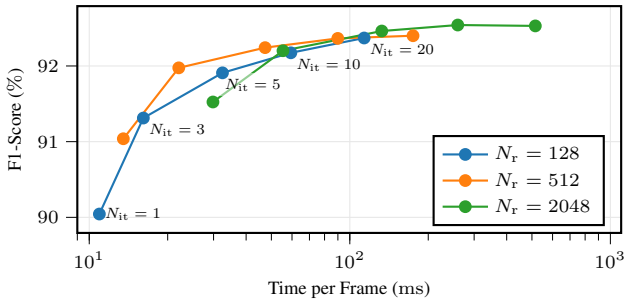


Figure 9. Impact of number of iterations per frame N_{it} and number of rays per field N_{r} (i.e., batch size per field) on run-time and reconstruction quality.

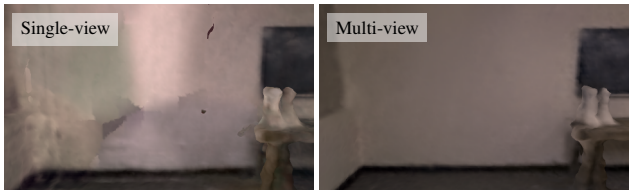


Figure 10. Single-view supervision leads to artifacts when the opposite side of a wall is observed. Multi-view supervision significantly reduces this local forgetting effect.

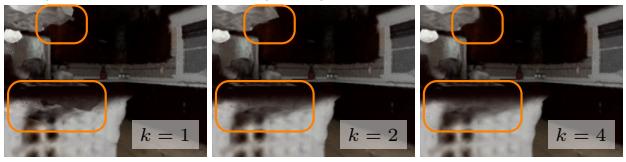


Figure 11. Comparison of rendered views with different number of neighbors k. By increasing the number of nearest neighbors taken into account, the number of visible field transitions decreases. These transitions are most visible in unobserved regions, such as the underside of the countertop shown here (see highlighted regions).

4.4. Ablation Study

Multi-View Supervision In Fig. 10 we compare optimization with targets sampled from all keyframes as described in Sec. 3.3 to targets sampled from a single view only. In the latter case, optimization alternates between a random previous keyframe and the latest keyframe. The figure shows that single-view supervision exhibits local forgetting effects, particularly notable when a previously seen wall is observed from the other side. Multi-view optimization avoids this local forgetting effect by combining previous and current observations in each optimization step.

K-Nearest Neighbor Averaging Figure 11 shows renderings of our model with varying values for k (see Sec. 3.4). For higher k visible transitions between fields are reduced and smoothed out. Note that due to our independent training scheme, all fields are trained in overlapping regions making averaging at the query point a viable strategy. While higher values for k lead to improved results, it also multiplies the number of queries required for rendering and mesh extraction (note that optimization time is unaffected by k).

5. Conclusion

In this work, we presented a novel approach to volumetric mapping that anchors lightweight neural fields to the pose graph of a sparse visual SLAM system. This framework allows to incorporate loop closures into the volumetric map at near-zero cost improving robustness and scalability. Compared to existing neural mapping approaches that support loop closure, our approach deforms the map reducing the need for reintegration while still allowing well-defined geometric queries (such as mesh extraction) without relying on

any space warping or image-space fusion.

Limitations While our approach enables efficient incorporation of loop closure constraints, it is not without drawbacks. In particular the multi-field representation is less memory-efficient compared to monolithic neural field representations, since those can use the available network capacity relatively unconstrained, whereas this adaptiveness is constrained to the local spheres in our approach. Furthermore, the neural scene representation is not used to improve the SLAM result and a tigher integration of dense mapping and sparse tracking could lead to improved robustness.

References

[1] Michal Adamkiewicz, Timothy Chen, Adam Caccavale, Rachel Gardner, Preston Culbertson, Jeannette Bohg, and Mac Schwager. Vision-only robot navigation in a neural radiance world. *RAL*, 7(2):4606–4613, 2022. 1

[2] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *CVPR*, pages 6290–6301, 2022. 2, 4, 6

[3] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, pages 16123–16133, 2022. 2, 5

[4] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, pages 5939–5948, 2019. 2

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 6, 4

[6] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. BundleFusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM TOG*, 36(4):1, 2017. 1, 2, 6

[7] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. MonoSLAM: Real-time single camera SLAM. *IEEE TPAMI*, 29(6):1052–1067, 2007. 1

[8] Jiading Fang, Shengjie Lin, Igor Vasiljevic, Vitor Guizilini, Rares Ambrus, Adrien Gaidon, Gregory Shakhnarovich, and Matthew R Walter. NeRFuser: Large-scale scene representation by NeRF fusion. *arXiv preprint arXiv:2305.13307*, 2023. 2

[9] Christopher G Harris and JM Pike. 3D positional integration from image sequences. *Image and Vision Computing*, 6(2): 87–90, 1988. 1

[10] Armin Hornung, Kai M Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. Octomap: An efficient probabilistic 3d mapping framework based on octrees. *Autonomous Robots*, 34:189–206, 2013. 1

[11] Peter J Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101, 1964. 4

[12] Mohammad Mahdi Johari, Camilla Carta, and François Fleuret. ESLAM: Efficient dense SLAM system based on hybrid representation of signed distance fields. In *CVPR*, pages 17408–17419, 2023. 1, 2, 5

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 2

[14] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *ISMAR*, pages 225–234, 2007. 1

[15] Xin Kong, Shikun Liu, Marwan Taher, and Andrew J Davison. vMAP: Vectorised object mapping for neural field SLAM. In *CVPR*, pages 952–961, 2023. 2, 3, 4, 5

[16] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *ACM SIGGRAPH*, pages 163–169, 1987. 5

[17] Hidenobu Matsuki, Keisuke Tateno, Michael Niemeyer, and Federic Tombari. NEWTON: Neural view-centric mapping for on-the-fly large-scale SLAM. *arXiv preprint arXiv:2303.13654*, 2023. 2, 3

[18] Donald Meagher. Geometric modeling using octree encoding. *Computer Graphics and Image Processing*, 19(2):129–147, 1982. 1

[19] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *CVPR*, pages 4460–4470, 2019. 2

[20] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421. Springer, 2020. 2

[21] Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. LENS: Localization enhanced by NeRF synthesis. In *CoRL*, pages 1347–1356, 2022. 1

[22] Thomas Müller, Brian McWilliams, Fabrice Rousselle, Markus Gross, and Jan Novák. Neural importance sampling. *ACM TOG*, 38(5):1–19, 2019. 2

[23] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 41(4), 2022. 2, 5

[24] Raul Mur-Artal and Juan D Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *TRO*, 33(5):1255–1262, 2017. 5

[25] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *ISMAR*, pages 127–136, 2011. 1

[26] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. DTAM: Dense tracking and mapping in real-time. In *ICCV*, pages 2320–2327, 2011. 1

[27] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3D reconstruction at scale using voxel hashing. *ACM TOG*, 32(6):1–11, 2013. 1

[28] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, pages 165–174, 2019. 2

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8026–8037, 2019. 5

[30] Victor Reijgwart, Alexander Millane, Helen Oleynikova, Roland Siegwart, Cesar Cadena, and Juan Nieto. Voxgraph: Globally consistent, volumetric mapping using signed distance function submaps. *RAL*, 5(1):227–234, 2019. 1, 2

[31] Radu Alexandru Rosu and Sven Behnke. PermutoSDF: Fast multi-view reconstruction with implicit surfaces using permutohedral lattices. In *CVPR*, pages 8466–8475, 2023. 2, 5

[32] Julian Straub et al. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 6, 5

[33] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. iMAP: Implicit mapping and positioning in real-time. In *ICCV*, pages 6229–6238, 2021. 1, 2, 6, 5

[34] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 33:7537–7547, 2020. 1, 5

[35] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-NeRF: Scalable large scene neural view synthesis. In *CVPR*, pages 8248–8258, 2022. 2

[36] Yijie Tang, Jiazhao Zhang, Zhinan Yu, He Wang, and Kai Xu. MIPS-Fusion: Multi-implicit-submaps for scalable and robust online neural RGB-D reconstruction. *arXiv preprint arXiv:2308.08741*, 2023. 3

[37] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, pages 3405–3414, 2019. 6

[38] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual slam for monocular, stereo, and RGB-D cameras. *NeurIPS*, 34:16558–16569, 2021. 3

[39] Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Co-SLAM: Joint coordinate and sparse parametric encodings for neural real-time SLAM. In *CVPR*, pages 13293–13302, 2023. 1, 2, 4, 5, 6, 3

[40] 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. *NeurIPS*, 34, 2021. 2

[41] Thomas Whelan, Michael Kaess, Hordur Johannsson, Maurice Fallon, John J Leonard, and John McDonald. Real-time large-scale dense RGB-D SLAM with volumetric fusion. *IJRR*, 34(4-5):598–626, 2014. 1, 2, 6, 3, 4

[42] Xiaoshuai Zhang, Abhijit Kundu, Thomas Funkhouser, Leonidas Guibas, Hao Su, and Kyle Genova. Nerflets: Local radiance fields for efficient structure-aware 3D scene representation from 2D supervision. In *CVPR*, pages 8274–8284, 2023. 2

[43] Youmin Zhang, Fabio Tosi, Stefano Mattoccia, and Matteo Poggi. GO-SLAM: Global optimization for consistent 3D instant reconstruction. In *ICCV*, pages 3727–3737, 2023. 1, 3, 5, 6, 4

[44] 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 *CVPR*, 2022. 1, 2, 5, 6, 3, 4

Neural Graph Mapping for Dense SLAM with Efficient Loop Closure

Supplementary Material

6. Method Details

6.1. Sampling Strategy

As described in Sec. 3.3 and visualized in Fig. 4, a three-stage sampling procedure is used. First, a subset of fields is sampled, then rays are sampled for each field, and finally points are sampled along each ray.

Fields Especially with larger scenes, it is important that recently added fields and those currently observed are optimized with a higher rate than out-of-view fields that have already been optimized before. To achieve this, the currently observed fields $\mathcal{F}_t^{\text{obs}} \subseteq \mathcal{F}_t$ are determined and sampled with a higher probability. Specifically, a total of N_f fields are sampled; half from the currently observed fields and the remaining ones from all fields $\mathcal{F}_t$ discarding duplicates.

Rays To sample supervision targets, each sampled field i is approximated by a set of points $\boldsymbol{q}_j^i, j = 1, ..., N_{\text{approx}}$ sampled uniformly on the field's sphere of radius r. These points are projected into all keyframes. A field is considered visible in a keyframe, if at least one of the field's points $\boldsymbol{q}_j^i$ is inside the keyframe's frustum and the projected depth of $\boldsymbol{q}_j^i$ is smaller than the observed depth at the projected 2D point. This yields a set of keyframes $\mathcal{K}_t^i \subseteq \mathcal{K}_t$. N_r rays per field (i.e., target rays) are then sampled via the 2D bounding boxes of the projected points $\boldsymbol{q}_j^i$ in the keyframes $\mathcal{K}_t^i$. For each target ray $(\boldsymbol{o}, \boldsymbol{d})$ the closest point to the field center is computed as $\boldsymbol{o} + l_\text{c}\boldsymbol{d}$ and only a ray segment $[l_\text{c} - r, l_\text{c} + r]$ covering the sphere will be considered for optimization.

Points Given a ray segment $[l_{\text{min}}, l_{\text{max}}]$, N_{up} points are uniformly sampled across the segment, and N_{dp} points are uniformly sampled in the truncation interval τ around the observed depth, that is, in the interval $[l_{\text{obs}} - \tau, l_{\text{obs}} + \tau]$; the full ray interval is used, if there is no depth measurement or $l_{\text{obs}} \notin [l_{\text{min}}, l_{\text{max}}]$. This yields a total of $N_\text{p} = N_{\text{up}} + N_{\text{dp}}$ query points per ray segment during optimization.

In total, each optimization iteration will contain a maximum of $N_\text{f} N_\text{r} N_\text{p}$ query points.

6.2. K-Nearest Neighbors Queries

We compute the color and signed distance at a query point as the weighted average of the k nearest fields. Specifically, let $\boldsymbol{x} \in \mathbb{R}^3$ denote the query point. Let $\boldsymbol{c}_i, s_i, d_i, i = 1, ..., k$ denote the returned color, signed distance, and distance to the field center for the k nearest fields for the query point $\boldsymbol{x}$.

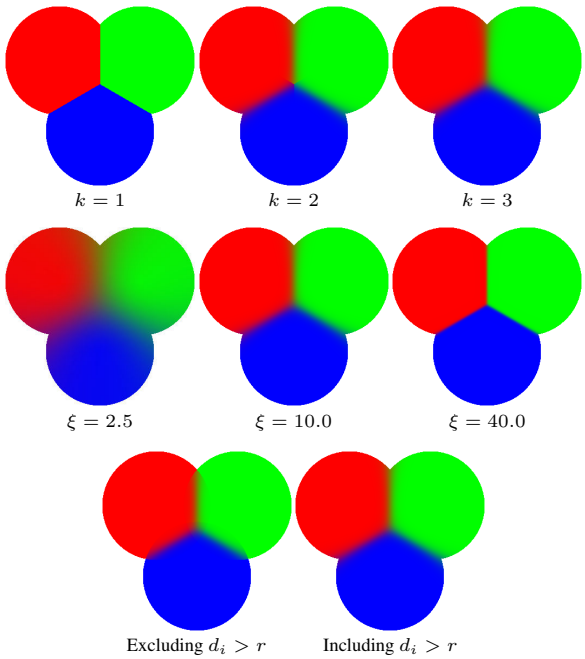


Figure 12. Visualization of k-nearest neighbor distance-based averaging. The top row shows the effect of varying k. The second row shows the effect of varying ξ. The last row shows the effect of excluding fields with distances d_i greater than the field radius r from the averaging.

We compute weights based on the softmax of the negative distances, that is,

$$u_i = \frac{e^{-\xi d_i}}{\sum_{j=1}^{k} e^{-\xi d_j}}, \tag{9}$$

where ξ determines the transition speed. The combined color and signed distance are then computed as a weighted sum, that is, $\boldsymbol{c} = \sum_{i=1}^{k} u_i \boldsymbol{c}_i$ and $s = \sum_{i=1}^{k} u_i s_i$.

We always use the k nearest fields even when only the closest field is within radius r. However, we set ξ sufficiently high such that the transition region becomes small and fields with $d_i \gg r$ will have no significant contribution to the final value. This is a feasible strategy, since fields are optimized for all ray segments intersecting them even when the segment is terminating outside the sphere. Hence, each field will in practice capture a region larger than a sphere with radius r.

Figure 12 illustrates the effect of this weighted averaging for different values of ξ, k, and with and without $u_i = 0$ for $d_i > r$ on a 2D toy example with three fields of fixed color. Note that the distance-weighted averaging leads to smooth transitions in the overlapping regions. When forc-

Table 3. Overview of parameters.

Parameter	Value	Description
r	$1\,\mathrm{m}$	Field radius
N_f	2	Number of fields optimized in parallel in each iterations
N_r	512	Number of ray segments sampled per field during optimization
N_up	8	Number of uniformly sampled points distributed along each ray segment during optimization
N_dp	16	Number of depth-guided points distributed along each ray segment during optimization
τ	$0.1\,\mathrm{m}$ or $0.2\,\mathrm{m}^*$	Truncation distance; used for scaling depth-guided sampling and for capping the supervision range (i.e., dividing samples into free-space samples and TSDF samples)
η	20.0	Determines how fast occupancy probability decays around surfaces
k	2	Number of nearest neighbors used during evaluation
ξ	10.0	Distance weighing determining transition speed between two fields
L	1	Number of MLP layers following the permutohedral encoding; excluding the final linear layer
T	2^{12}	Hash table size for permutohedral encoding
N_levels	16	Number of resolution levels for permutohedral encoding
N_fpl	16	Number of features per level for permutohedral encoding
r_coarse	1.0	Coarsest resolution for permutohedral encoding
r_fine	0.0001	Finest resolution for permutohedral encoding
λ_color	1.0	Weight of color loss
λ_depth	1.0	Weight of depth loss
λ_fs	40.0	Weight of free-space loss
λ_tsdf	50.0	Weight of TSDF loss
δ	$5\,\mathrm{cm}$	Huber loss threshold
γ	1×10^{-3}	Learning rate used for Adam optimizer [13]
λ	1×10^{-5}	Weight decay used for Adam optimizer [13]

* The truncation distance is increased for real-world datasets to account for the increase in depth noise.

ing $u_i = 0$ for $d_i > r$ transitions on the boundaries are unavoidable, hence we opt for the strategy described in the previous paragraph. For the experiments we use $k = 2$ and $\xi = 10$.

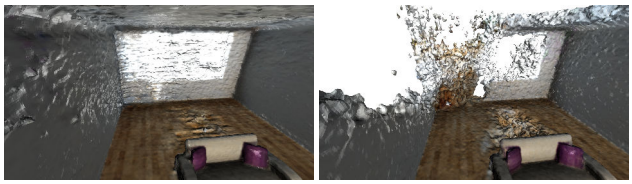
6.3. Parameters

In Tab. 3 we provide a full list of parameters, the used value to achieve the experimental results, and a brief description. Parameters were tuned manually and the same setting is used for all experiments (with the exception of τ, which is increased for the real-world datasets).

7. Experiment Details

Baseline Setup All baselines are evaluated using the parameters published as part of the published code. For additional datasets for which no parameters were provided, the most similar dataset's parameters were adopted (i.e., for Replica-Big the provided setup for Replica is used; for Kintinuous the setup for ScanNet). Scene boundaries were manually adjusted to cover the observed area with extra margin to account for errors in positioning.

In our experiments, we noticed that Co-SLAM [39] uses the ground-truth pose of the first frame to initialize the SLAM system, which leads to axis-aligned planes. We found that planes (such as walls, floors, and ceilings) which are axis-aligned are significantly better completed using the one-blob encoding [22] than generally-oriented planes. Therefore, for a fair comparison, we modified Co-SLAM's implementation to start from a random orientation instead. We note that this mainly reduces qualitative scene completion, however, on one of the Replica scenes it leads to tracking issues and hence poor reconstruction results. Figure 13 shows an example of the scene completion capability of Co-SLAM with and without ground-truth initialization (i.e., with and without axis-aligned planes).



(a) With axis-alignment (b) Without axis-alignment

Figure 13. Co-SLAM result with and without ground-truth initialization. Ground-truth initialization leads to axis-aligned walls and floors, which in turn leads to significantly better scene completion.

7.1. Evaluation Protocol

Let ${}^\mathrm{w}\mathcal{M}_\mathrm{gt}$ and ${}^{\tilde{\mathrm{w}}}\mathcal{M}_\mathrm{est}$ denote the ground-truth mesh and estimated mesh respectively. We assume that ${}^{\tilde{\mathrm{w}}}\mathcal{M}_\mathrm{est}$ has already been globally aligned with ${}^\mathrm{w}\mathcal{M}_\mathrm{gt}$, which is typically achieved by either aligning the first frame in the sequence or by aligning the trajectories using the Umeyama algorithm. Starting from ${}^\mathrm{w}\mathcal{M}_\mathrm{gt}$ and ${}^{\tilde{\mathrm{w}}}\mathcal{M}_\mathrm{est}$ further preprocessing steps are performed before the evaluation metrics are computed.

1. Unobserved parts of the ground-truth mesh ${}^\mathrm{w}\mathcal{M}_\mathrm{gt}$ are removed. In particular, we apply two removal steps. First, vertices falling more than $2\,\mathrm{cm}$ outside the scene bounding box are removed. The scene bounding box is

Table 4. Comparison of mesh quality on Replica (**best** ●, second best ●).

		Replica								
		room0	room1	room2	offi0	offi1	offi2	offi3	offi4	Avg.
NICE-SLAM [44]	Acc (cm)	2.47	2.21	2.17	1.90	1.61	3.13	2.92	2.60	**2.38**
	Acc.-Ratio (%)	93.38	94.92	**93.75**	94.87	95.30	89.78	90.17	93.21	**93.17**
	Comp. (cm)	2.93	2.31	2.77	2.37	2.15	2.89	3.42	3.91	2.84
	Comp. Ratio (%)	90.90	93.57	90.97	92.58	92.15	88.78	86.20	85.72	90.11
Co-SLAM [39]	Acc (cm)	**1.99**	19.90	**1.92**	**1.55**	**1.33**	**2.76**	**2.61**	**2.22**	4.29
	Acc.-Ratio (%)	**95.37**	38.68	93.51	96.15	96.75	**90.92**	**92.04**	**92.70**	87.02
	Comp. (cm)	2.37	17.47	**2.08**	**1.54**	1.68	2.39	2.73	**2.47**	4.09
	Comp. Ratio (%)	**93.43**	40.03	**93.16**	**96.04**	94.57	**91.99**	**90.92**	**90.96**	86.39
GO-SLAM [43]	Acc (cm)	3.45	**2.15**	2.90	1.88	1.72	3.51	3.97	3.56	2.89
	Acc.-Ratio (%)	84.59	**96.04**	89.50	**96.31**	**97.55**	84.49	79.64	86.34	89.31
	Comp. (cm)	6.83	4.26	8.46	3.12	4.17	6.74	7.43	8.39	6.18
	Comp. Ratio (%)	69.44	82.89	72.65	85.21	82.89	69.40	63.81	67.33	74.20
Ours	Acc (cm)	2.63	2.25	2.86	1.88	2.07	3.45	4.92	2.98	2.88
	Acc.-Ratio (%)	90.80	93.01	86.99	93.83	92.43	87.73	83.01	88.20	89.50
	Comp. (cm)	**2.25**	**1.86**	3.57	1.67	1.79	**2.34**	**2.69**	2.67	**2.36**
	Comp. Ratio (%)	93.23	**94.98**	89.62	95.59	93.34	91.35	89.40	89.34	**92.11**

computed as the intersection of a manually-set bounding box (used to exclude outliers in the depth map present in some scenes) and an automatically computed bounding box (based on the ground-truth trajectory and depth maps). Second, vertices that are not in front or up to 3 cm behind any rendered depth map are removed. These depth maps are rendered from the ground-truth trajectory and from additional virtual views manually placed to improve the evaluation of scene completion (same as in Co-SLAM [39]). This yields a culled ground-truth mesh used for evaluation ${}^{\mathrm{w}}\mathcal{M}^{*}_{\mathrm{gt}}$.

2. To further equalize slight differences in alignment between different methods, we perform another alignment step using point-to-plane-based iterative closest point from ${}^{\tilde{\mathrm{w}}}\mathcal{M}_{\mathrm{est}}$'s vertices to ${}^{\mathrm{w}}\mathcal{M}_{\mathrm{gt}}$'s vertices yielding an aligned estimated mesh ${}^{\mathrm{w}}\mathcal{M}_{\mathrm{est}}$.
3. The aligned estimated mesh ${}^{\mathrm{w}}\mathcal{M}_{\mathrm{est}}$ follows the same removal process as the ground-truth mesh (see step 1 above) yielding the culled estimated mesh used for evaluation ${}^{\mathrm{w}}\mathcal{M}^{*}_{\mathrm{est}}$

For the evaluation, $N_{\mathrm{samples}} = 200\,000$ points are uniformly sampled on both meshes yielding the point sets $\mathcal{G} = \{\boldsymbol{x}_i \sim \mathcal{U}({}^{\mathrm{w}}\mathcal{M}^{*}_{\mathrm{gt}}) \mid i = 1, ..., N_{\mathrm{samples}}\}$ and $\mathcal{E} = \{\boldsymbol{y}_i \sim \mathcal{U}({}^{\mathrm{w}}\mathcal{M}^{*}_{\mathrm{est}}) \mid i = 1, ..., N_{\mathrm{samples}}\}$, where $\mathcal{U}(\cdot)$ denotes the uniform distribution. The point sets are used to compute accuracy, completion, accuracy ratio, and completion ratio as

$$\mathrm{Acc}(\mathcal{G}, \mathcal{E}) = \frac{1}{|\mathcal{E}|} \sum_{\boldsymbol{y} \in \mathcal{E}} \min_{\boldsymbol{x} \in \mathcal{G}} \|\boldsymbol{y} - \boldsymbol{x}\| \tag{10}$$

$$\mathrm{Comp}(\mathcal{G}, \mathcal{E}) = \frac{1}{|\mathcal{G}|} \sum_{\boldsymbol{x} \in \mathcal{G}} \min_{\boldsymbol{y} \in \mathcal{E}} \|\boldsymbol{x} - \boldsymbol{y}\| \tag{11}$$

$$\mathrm{AR}(\mathcal{G}, \mathcal{E}) = \frac{1}{|\mathcal{E}|} \sum_{\boldsymbol{y} \in \mathcal{E}} \left[\min_{\boldsymbol{x} \in \mathcal{G}} \|\boldsymbol{y} - \boldsymbol{x}\| < \Delta \right] \tag{12}$$

$$\mathrm{CR}(\mathcal{G}, \mathcal{E}) = \frac{1}{|\mathcal{G}|} \sum_{\boldsymbol{x} \in \mathcal{G}} \left[\min_{\boldsymbol{y} \in \mathcal{E}} \|\boldsymbol{x} - \boldsymbol{y}\| < \Delta \right], \tag{13}$$

where $[\cdot]$ denotes the Iverson bracket and $\Delta = 5\,\mathrm{cm}$ in our experiments. Since accuracy ratio and completion ratio can be interpreted as precision and recall of the reconstruction, we further use the F1-score

$$F_1(\mathcal{G}, \mathcal{E}) = \frac{2}{\mathrm{AR}(\mathcal{G}, \mathcal{E})^{-1} + \mathrm{CR}(\mathcal{G}, \mathcal{E})^{-1}} \tag{14}$$

to summarize reconstruction performance in one metric.

8. Additional Results

Detailed Replica Results Table 4 shows per-scene results on the Replica dataset. On most scenes Co-SLAM achieves the best result; however, it fails on `room1` leading to worse average results. We note that our contributions do not aim to improve scene reconstruction on small scenes and this experiment merely serves to highlight overall competitive results. That is, our method scales to large scale scenes by incorporating loop closure constraints, without significantly hurting small scene performance.

Additional Qualitative Results Figure 14 shows an extended comparison at the end of the Kintinuous [41] sequence. Our method is the only one that successfully closes the loop. Co-SLAM drifts off significantly in the second corner failing to map subsequent parts successfully.

In Fig. 15, Fig. 16, and Fig. 17 qualitative results on the ScanNet, Replica, and NRGBD datasets are shown, respectively. In most scenes our approach performs close to the best performing method Co-SLAM, while outperforming NICE-SLAM and GO-SLAM. Co-SLAM achieves slightly more detailed and smoother results, which might be due to their more effective hash table use and the additional smoothness loss.

Figure 14. Qualitative comparison on the Kintinuous dataset [41] at the end of the sequence. Three views are shown. Note that our method is the only one closing the loop successfully. The last row shows the corner in which Co-SLAM is failing (mapping a 90 degree corner as a U-turn).

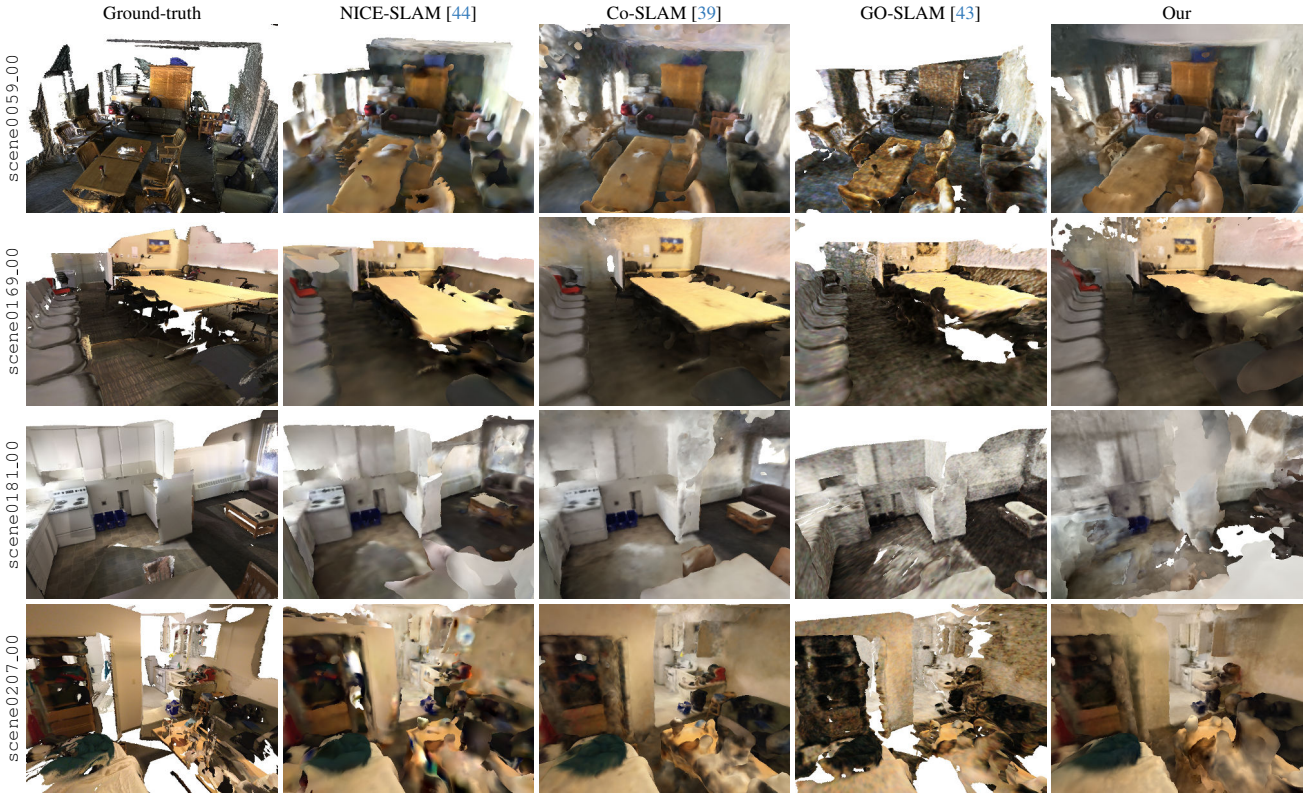


Figure 15. Qualitative comparison of final meshes extracted by all methods on the ScanNet dataset [5]. Our method fails on `scene0181_00` due to tracking issues in the underlying SLAM system in a feature-less region.

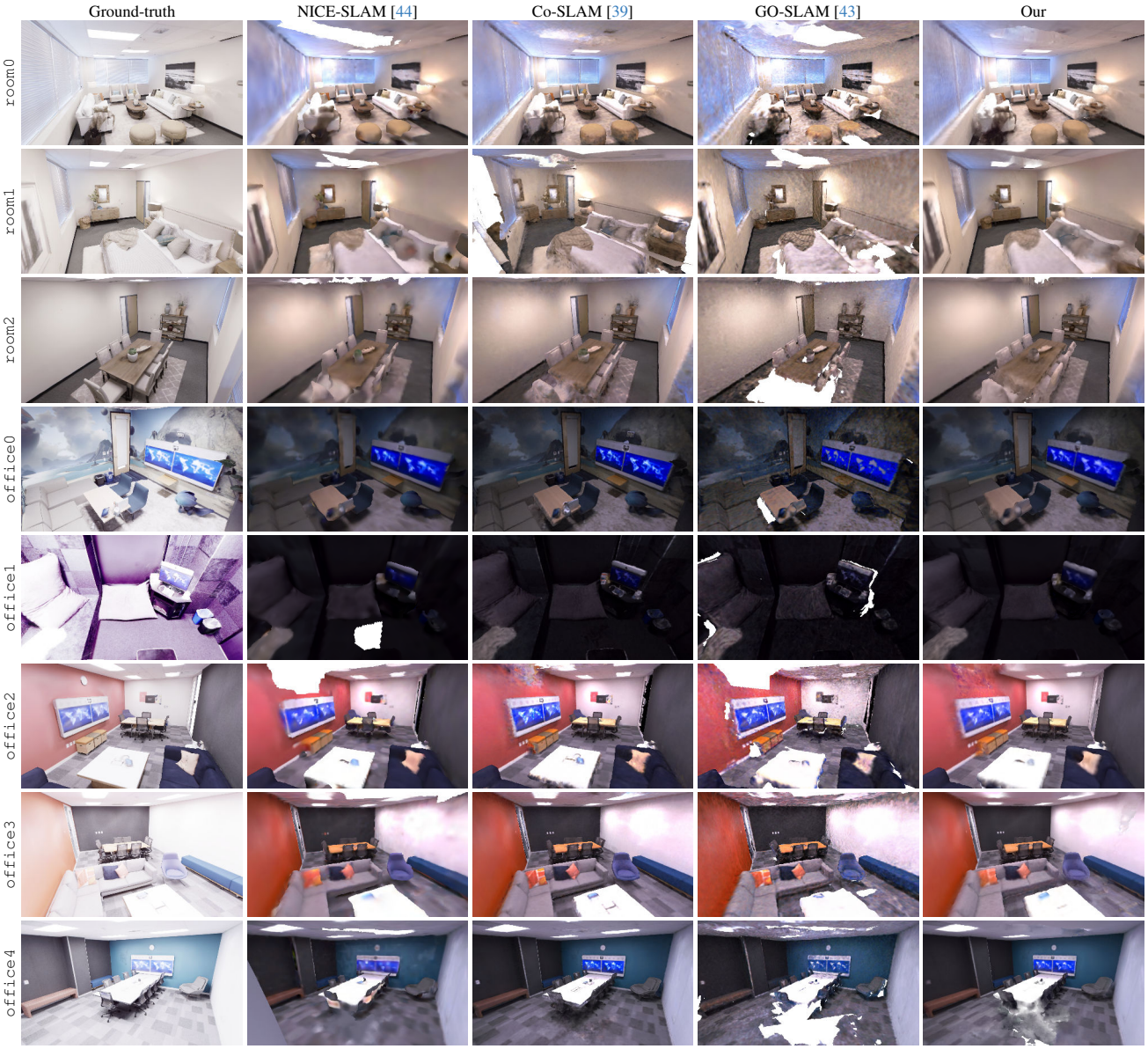


Figure 16. Qualitative comparison of final meshes extracted by all methods on the Replica dataset [32, 33].

Figure 17. Qualitative comparison of final meshes extracted by all evaluated methods on the NRGBD dataset [2].