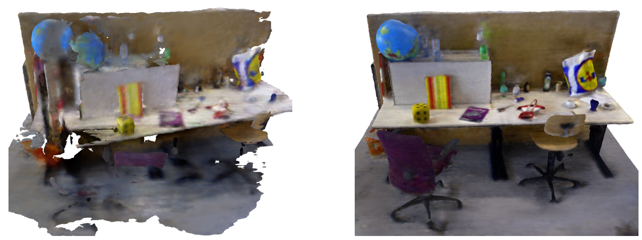
Uni-SLAM: Uncertainty-Aware Neural Implicit SLAM for Real-Time Dense Indoor Scene Reconstruction

Shaoxiang Wang, Yaxu Xie, Chun-Peng Chang, Christen Millerdurai, Alain Pagani, Didier Stricker
German Research Center for Artificial Intelligence
firstname.lastname@dfki.de

Abstract

Neural implicit fields have recently emerged as a powerful representation method for multi-view surface reconstruction due to their simplicity and state-of-the-art performance. However, reconstructing thin structures of indoor scenes while ensuring real-time performance remains a challenge for dense visual SLAM systems. Previous methods do not consider varying quality of input RGB-D data and employ fixed-frequency mapping process to reconstruct the scene, which could result in the loss of valuable information in some frames.

In this paper, we propose Uni-SLAM, a decoupled 3D spatial representation based on hash grids for indoor reconstruction. We introduce a novel defined predictive uncertainty to reweight the loss function, along with strategic local-to-global bundle adjustment. Experiments on synthetic and real-world datasets demonstrate that our system achieves state-of-the-art tracking and mapping accuracy while maintaining real-time performance. It significantly improves over current methods with a 25% reduction in depth L1 error and a 66.86% completion rate within 1 cm on the Replica dataset, reflecting a more accurate reconstruction of thin structures. Project page: https://shaoxiang777.github.io/project/uni-slam/



(a) (b)

Figure 1. The reconstructed 3D mesh on the TUM RGB-D dataset [53], generated using our proposed method without uncertainty-guided reweighting and strategy, is illustrated in Fig. 1a. Conversely, Fig. 1b demonstrates the 3D mesh produced by our method after the incorporation of the uncertainty-aware strategy.

1. Introduction

Dense visual Simultaneous Localization and Mapping (SLAM) aims at reconstructing a dense 3D map of an unknown environment while simultaneously estimating the accurate camera pose. Traditional SLAM algorithms [12, 13, 37, 39] focus on localization accuracy for real-time large-scale applications, whereas Neural Radiance Fields (NeRFs) [35] significantly enhance dense 3D reconstruction and novel view synthesis, spurring the development of NeRF-based dense visual SLAM techniques.

As pioneering efforts, iMAP [55] and Nice-SLAM [76] utilize neural representations for both tracking and mapping, but slow convergence limits their low-latency mapping capabilities. SDF-based methods [11, 19, 21, 61] offer faster convergence and higher rendering accuracy. But they treat all data even with varying quality equally, alternating tracking and mapping at a constant frequency (every n frames). However, in dense NeRF-SLAM, the quality of RGB-D input data varies throughout the sequence (such as invalid depth), significantly impacting both camera pose estimation and scene reconstruction. Furthermore, constant mapping, this simple approach may lead to missing potentially effective information in frames where no mapping process occurs. Therefore, treating all data uniformly in dense NeRF-SLAM systems is suboptimal, leading to overconfidence in poor-quality data and inefficient use of valuable information.

Our dense SLAM method, Uni-SLAM, tackles these challenges by: 1) Differentiating data quality through pixel-level uncertainty analysis and loss reweighting to identify outliers; 2) Using image-level uncertainty to guide local-to-global bundle adjustment for comprehensive reconstruction; and 3) Employing decoupled hash grids to separately represent geometry and appearance, enabling real-time capture of high-frequency details in indoor scenes. **Contributions** of our method are summarized as follows:

- We introduce a novel form of uncertainty, termed *predictive uncertainty*, which enables pixel-level loss reweighting without the need for additional training. By leveraging this uncertainty, our method dynamically identifies and prioritizes valuable regions in the input data, enhancing the performance of mapping and tracking processes. This approach proves particularly effective when dealing with varying levels of input data quality, ensuring more robust and accurate outcomes.

- Image-level uncertainty dynamically activates mapping with strategic local-to-global bundle adjustment, preserving valuable image information and enhancing global stability while capturing local color and geometry.
- We propose an efficient scene representation using hash grids to decouple the scene's geometry and appearance. This approach enhances spatial representation of high-frequency signals while maintaining real-time performance. Our method achieves state-of-the-art results on the Replica [52], ScanNet [9], and TUM RGB-D [53] datasets.

2. Related Work

The proposed method encompasses SLAM, implicit spatial representation and uncertainty modeling. Therefore, we focus the discussion of related work on these specific methods to better highlight our contributions.

Dense Visual SLAM. Early dense visual SLAM approaches, like PTAM [25] and DTAM [39], used feature-based methods, separating tracking and mapping tasks for efficiency. ORB-SLAM [37] further refined this with a feature-based approach for camera trajectory and 3D map construction. DROID-SLAM [60] introduced optical flow for precise real-time visual odometry and dense mapping. Learning-based methods [28, 48, 72] improved feature extraction and robustness. Recent works [7, 26, 32, 44, 74] combine ORB-SLAM for robust tracking with NeRF-based mapping. Others [19, 29, 46, 55, 59, 61, 75, 76] integrate tracking and mapping in an interactive process. This paper explores uncertainty's impact in joint optimization scenarios.

Scene Representation. Most common scene representation for dense mapping are grid-based (including voxel grids [8, 38, 56], octrees [58, 70], voxel hashing [36, 40]), surfel clouds [6, 64, 67] and multi layer perceptron (MLP)-based [2, 44, 71]. Grid-based methods offer the advantages of easy neighborhood finding and fast tri-linear interpolation. However, they require manual grid resolution specification and waste memory in empty regions [70, 75, 76]. Point-based methods avoid pre-specified resolutions but have complex neighborhood searches and low convergence speeds, which hinder real-time reconstruction. Additionally, they cannot fill in empty holes or make reasonable guesses for unscanned areas [23, 34, 46, 67, 69]. MLP-based methods suffer from slow convergence and catastrophic forgetting in large scenes [55, 59], as updating all weights during optimization can cause forgetting issues.

Uncertainty Modeling in Scene Reconstruction. The computer vision community has increasingly recognized the importance of uncertainty estimation across fields such as next-best-view (NBV) selection [27, 42, 57], segmentation [16, 22, 31], depth estimation [17, 18], and SLAM [3, 5, 47]. Uncertainty assessment enhances model interpretability and reduces critical errors. Kendall et al. [24] identify two types of uncertainty in Bayesian deep learning: *aleatoric* (due to inherent data ambiguity) and *epistemic* (arising from limited data) [1, 20, 65].

In NeRF-based novel view synthesis with *known camera pose*, integrating uncertainty has led to improvements in handling blur, dynamic objects, and confidence visualization [14, 33, 50, 51, 68]. However, its application in dense NeRF-SLAM with *unknown camera pose* remains underexplored. Sandström *et al*. [47] introduce a SLAM system that estimates aleatoric depth uncertainty, while Rosinol *et al*. [45] propose fast uncertainty propagation for cleaner 3D meshes. To our knowledge, we are the first to use novel-defined predictive uncertainty, caused by limited unobserved data, to reweight dense implicit SLAM and guide local-to-global BA.

3. Method

Our overall pipeline is illustrated in Fig. 2. The input consists of a sequence of RGB-D images and known camera intrinsic parameters. Through a decoupled scene representation, we estimate the camera pose, the implicit truncated signed distance function (TSDF), depth, color and uncertainty. In Sec. 3.1, our efficient independent scene representation using two hash grids is described. In Sec. 3.2, we present our novel uncertainty model and explain how it reweights the loss function in Sec. 3.3. Finally, Sec. 3.4 presents the uncertainty-guided strategic BA and keyframe selection.

3.1. Neural Scene Representation

All existing implicit NeRF-based SLAM systems exhibit various issues in scene representation: ***a)*** *MLP-based [55] forgetting problem and insufficient spatial representation capability when using tri-planes [4, 21].* ***b)*** *Coupled geometry and appearance information [11, 61, 76] increases training difficulty, resulting in poor reconstruction quality.* ***c)*** *Coarse-to-fine dense grids [76] rely on heuristic resolution selection and require longer training times and high memory usage, failing to meet real-time requirements.* In our method, the hypothesis is that geometry and color

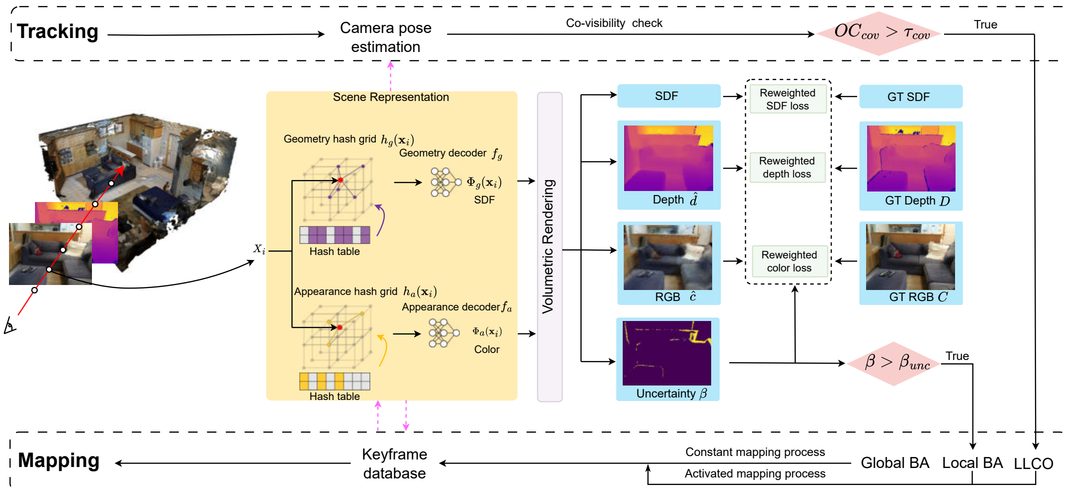


Figure 2. **Uni-SLAM Architecture Overview.** Our framework consists of two threads, tracking and mapping. While tracking is performed every frame for RGB-D stream, besides constant mapping is performed every n frame constantly with global BA, activated additional mapping process is executed to capture local scene information based on uncertainty and co-visibility check with local BA and local loop closure optimization (LLCO). Our proposed pixel-level uncertainty method adaptively filters outlier pixels and reweights the loss function, enabling more precise localization during tracking and the reconstruction of color and geometric information in mapping.

information should not be sampled at the same frequency. To verify this, we opt for a decoupled representation, using multiresolution hash grids [36] model for each of them. We show in our experiments that this decoupled hash grid representation favors speed, hole-filling ability, and low memory footprint while not sacrificing accuracy. The raw SDF $\Phi_g(\mathbf{x}_i)$ and the raw color $\Phi_a(\mathbf{x}_i)$ are decoded via tiny MLPs geometry decoder f_g and appearance decoder f_a:

$$\Phi_g(\mathbf{x}_i) = f_g\left(h_g(\mathbf{x}_i)\right) \quad \text{and} \quad \Phi_a(\mathbf{x}_i) = f_a\left(h_a(\mathbf{x}_i)\right) \tag{1}$$

where $h_g(\mathbf{x}_i)$ and $h_a(\mathbf{x}_i)$ represent multiresolution geometry hash grids and appearance hash grids respectively in Fig. 2. We set the multiresolution level to $L = 16$, and only visualize one resolution level hash grid here for clarity. The decoupled representation effectively reduces the network's confusion when faced with appearance and geometry information of varying complexity. For more implementation details of hash grid, we refer readers to the supplementary Sec. A.1, B.1 and [36].

Depth and Color Volume Rendering. We follow [76] to render depth and color via integration along the sampling rays as $\hat{\boldsymbol{c}} = \sum_{i=1}^{N} w_i \boldsymbol{\phi_a}(\mathbf{x}_i)$ and $\hat{\boldsymbol{d}} = \sum_{i=1}^{N} w_i d_i$, where d_i represents the distance from camera center to the current sample point $\mathbf{x}_i$ along this ray. $\mathbf{x}_i$ is sampled and guided by depth image as [61]. w_i is the weight of the current sampling point, which can be converted from the density $\sigma(\mathbf{x}_i)$ as

$$w_i = \exp\left(-\sum_{j=1}^{i-1} \boldsymbol{\sigma}(\mathbf{x}_j)\right)\left(1 - \exp\left(-\boldsymbol{\sigma}(\mathbf{x}_i)\right)\right) \tag{2}$$

where $\sigma(\mathbf{x}_i) = \frac{1}{\alpha} \cdot \mathrm{Sigmoid}\left(\frac{-\phi_{\boldsymbol{g}}(\mathbf{x}_i)}{\alpha}\right)$ is the 3D volumetric density that can be converted from the SDF $\Phi_g(\mathbf{x}_i)$ [41], α is a learnable parameter which controls the sharpness of the model. This method of conversion through density, compared to direct conversion [61, 70] and surface-based conversion [62, 74], offers better interpretability, aligning closely with the original volumetric rendering in NeRF [35]. Moreover, we leverage this representation to derive our definition of uncertainty, which will be discussed in the following section.

3.2. Uncertainty Modeling

Our primary objective is to derive an uncertainty measure that can indicate the quality of the color and depth images, allowing us to reweight the loss functions during tracking and mapping. However, to our knowledge, no NeRF-based dense SLAM system has yet addressed predictive uncertainty, which reflects the model's confidence explicitly in its predictions for each view.

Specifically, inspired by the vanilla NeRF formulation [35] (Eq. 3), we utilize the **termination probability** concept from the volume rendering equation.

$$w_i = \overbrace{\underbrace{\exp\left(-\sum_{j=1}^{i-1} \boldsymbol{\sigma}(\mathbf{x}_j)\right)}_{\text{transmittance } T_i} \underbrace{\left(1 - \exp\left(-\boldsymbol{\sigma}(\mathbf{x}_i)\right)\right)}_{\text{occupancy } o_i}}^{\text{termination probability at sample } i} = T_i \cdot o_i \tag{3}$$

where T_i describes *transmittance* at sample point t_i along the ray from t_0 to t_{i-1} without hitting any other particle, *occupancy* o_i represents the probability that the ray collides with a particle at position t_i independently of the previously

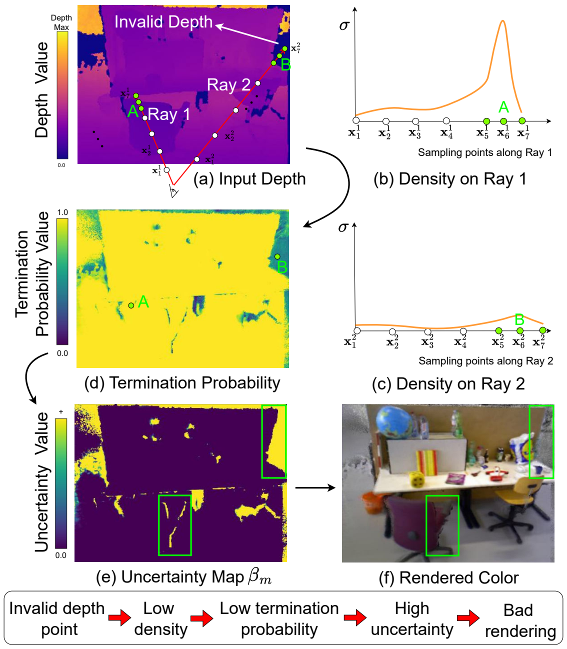


Figure 3. **Termination Probability and Uncertainty.** This figure illustrates the termination probability and uncertainty during ray sampling. For pixel A with valid depth (sampling by Ray 1), the sampling density is high along this ray, leading to a high termination probability and lower uncertainty. In contrast, for pixel B with invalid depth (sampling by Ray 2), the sampling density is low along this ray, resulting in a lower termination probability and higher uncertainty, as seen in the uncertainty map (e). This leads to degraded rendering quality in regions with high uncertainty, as shown in (f). Back-projected points A and B correspond to the surfaces of the hit objects in 3D space. For point B with invalid depth, we can estimate an approximate depth value based on the model in its current state.

light path. The product of the two $w_i = T_i \cdot o_i$ represents the termination probability, i.e. the probability that the light can reach the spatial location t_i.

We define the accumulated termination probability of N sampling points along a current sampling ray r as

$$p(r) = \sum_{i=1}^{N} w_i = 1 - \exp\left(-\sum_{i=1}^{N} \sigma\left(\mathbf{x}_i\right)\right) \tag{4}$$

$p(r) = 1$ ideally when the rendering is perfect (camera tracking is accurate and the region has been already observed before). Conversely, in never unobserved regions the NeRF model will estimate a low termination probability $p(r) \approx 0$ along the current ray r. The value is bounded by $(0, 1)$. We validate this in our experiments and visualize the termination probability in Fig. 3 (d), and the mathematical proof is included in the supplementary material Sec. A.2.

In [57] Sünderhauf et al. define uncertainty based on deep ensembles. However, full deep ensembles require training multiple models with different initializations, and are unsuitable for real-time systems like SLAM due to the high computational cost of maintaining several models. For a given image with an estimated pose, a pixel with index m is associated to a corresponding ray r_m. Inspired by [57], based on the probability $p(r_m)$, we defined the pixel-level predictive uncertainty as

$$\beta_m = (1 - p(r_m))^2 \tag{5}$$

As shown in Fig. 3(a), pixel B with invalid depth, we can only estimate an approximate depth value based on the model in its current state. Using this estimated depth for ray sampling results in a rendering with low accumulated termination probability in Fig. 3(d), indicating higher uncertainty as seen in Fig. 3(e) the uncertainty map.

For a rendered image associated with M sampled rays, we introduce a novel image-level predictive uncertainty β defined as

$$\beta = \frac{1}{M} \sum_{m=1}^{M} \beta_m \tag{6}$$

This image-level uncertainty β indicates the model's confidence in its current position estimate. A low β value suggests that the model is familiar with the area because of the accurate estimated camera position and sufficient sampling rays. Conversely, a high β value indicates that the model is less familiar with the area, suggesting that it should be more cautious and attentive in this region.

This predictive uncertainty, reflecting the model's knowledge limitations on the current camera pose, can be reduced by gathering more data, such as by taking data slowly to avoid drastic changes in motion state. How to use the defined uncertainty in the loss function and keyframe selection will be discussed in Sec. 3.3 and Sec. 3.4.

We also compared our model-free uncertainty approach with a learnable uncertainty model, based on Gaussian assumptions, as in BayesRays [15]. Our experiments show that this idea not only brings undesirable increased model complexity, making the model much slower, but also leads to poorer results in terms of reconstruction quality. Details can be found in the supplementary material Sec. B.3.

3.3. Uncertainty-guided Loss Function

Our mapping and tracking processes are carried out by minimizing our objective functions with respect to the network parameters θ and the camera parameters $\{R_i|t_i\}$ as [61]. We hypothesize that pixels with invalid depth or motion blur, caused by sensor issues or sudden motion changes, should exhibit high uncertainty, while well-observed regions should display low uncertainty. This premise enables us to effectively incorporate predicted uncertainty into the objective function, with the goal of progressively filtering out outliers to enhance localization accuracy and rendering quality. Inspired by the definition of SSIM loss in NeRF on-the-go [43] and the masked uncertainty learning in DebSDF [66], we define pixel-level binary confidence function as

$$CF_m = \mathbb{1}\left(1-\beta_m\right) = \begin{cases} 1 & \text{if } \beta_m \leq \beta_{unc_m} \\ 0 & \text{if } \beta_m > \beta_{unc_m} \end{cases} \tag{7}$$

where β_{unc_m} is a threshold for pixel-level uncertainty.

Near the surface we set hyperparameter truncation distance τ_{trunc} and approximate the ground truth SDF of sampling point $\mathbf{x}_i$ by $b\left(\mathbf{x}_i\right) = D_m - D_{m,i}^{ray}$, where D_m is current ray depth, $D_{m,i}^{ray}$ is the distance from camera center w.r.t. sampling point. The points that lie within the truncation distance $[-\tau_{\text{trunc}}, \tau_{\text{trunc}}]$, $i.e.|b\left(\mathbf{x}_i\right)| < \tau_{\text{trunc}}$ form the set X^{tr}.

The loss associated to the points belonging to X^{tr} is

$$\mathcal{L}^{tr}(X^{tr}) = \frac{1}{M^*}\sum_{m=1}^{M}\frac{CF_m}{|X^{tr}|}\sum_{\mathbf{x}_i \in X^{tr}}\left(\Phi_g(\mathbf{x}_i)\tau_{\text{trunc}} - b\left(\mathbf{x}_i\right)\right)^2 \tag{8}$$

where M is the number of sampled points, M^* is the number of valid sampled points after reweighting by Eq. (7).

We further refine the set of sampling points inside the truncation distance in two subgroups. Assuming accurate valid depth ground truth, we assign greater weights to sample points at the *center* (closer to the surface) $X_c^{tr} = \{\mathbf{x}_i \mid \; |b\left(\mathbf{x}_i\right)| \leq 0.4\tau_{\text{trunc}}\}$ to accelerate convergence and achieve more accurate geometry, while points at the *tail* of the truncation region constitute X_t^{tr}, and associate different losses to these two groups as follows:

$$\mathcal{L}_c^{tr} = \mathcal{L}^{tr}(X_c^{tr}) \quad \text{and} \quad \mathcal{L}_t^{tr} = \mathcal{L}^{tr}(X_t^{tr}) \tag{9}$$

Considering the points outside the truncation distance as the free space set X^{fs}, which are far from the surface $|b\left(\mathbf{x}_i\right)| > \tau_{\text{trunc}}$. In this area the loss function encourages $\Phi_g(\mathbf{x}_i)$ to have the value equal to one as

$$\mathcal{L}^{fs} = \frac{1}{M^*}\sum_{m=1}^{M}\frac{CF_m}{|X^{fs}|}\sum_{\mathbf{x}_i \in X^{fs}}\left(\Phi_g(\mathbf{x}_i) - 1\right)^2 \tag{10}$$

The color and depth losses are defined as follows:

$$\mathcal{L}_{rgb}^{track} = \frac{1}{M^*}\sum_{m=1}^{M}\left(C[u,v] - \hat{\boldsymbol{c}}_m\right)^2 \cdot CF_m \tag{11}$$

$$\mathcal{L}_{rgb}^{map} = \frac{1}{M}\sum_{m=1}^{M}\left(C[u,v] - \hat{\boldsymbol{c}}_m\right)^2 \tag{12}$$

$$\mathcal{L}_{dep} = \frac{1}{M^*}\sum_{m=1}^{M}\left(D[u,v] - \hat{\boldsymbol{d}}_m\right)^2 \cdot CF_m \tag{13}$$

where $C[u,v]$ and $D[u,v]$ are the ground-truth values for color and depth respectively. Note the reweighting confidence function CF_m is not applied to color loss in the mapping process.

Tracking Loss Function. The loss function for the tracking process is achieved by the following weighting scheme:

$$\mathcal{L}_t = \lambda_{rgb}\mathcal{L}_{rgb}^{track} + \lambda_{dep}\mathcal{L}_{dep} + \mathcal{L}_{sdf} \tag{14}$$

where $\mathcal{L}_{sdf} = \lambda_c^{tr}\mathcal{L}_c^{tr} + \lambda_t^{tr}\mathcal{L}_t^{tr} + \lambda_{fs}\mathcal{L}_{fs}$.

During tracking, the scene representation remains unchanged and only the camera pose is optimized (as shown by the magenta dashed line in Fig. 2). CF_m helps us select the most confidently estimated data for optimal optimization. If certain pixels are already predicted incorrectly, continuing to assign them a high weight is not beneficial. Therefore, when applying the tracking loss function, it is crucial to focus on pixels that are correctly estimated with high confidence. This means that the loss for pixels which are misestimated with high uncertainty can be neglected.

Mapping Loss Function. The total loss function for mapping loss is defined as:

$$\mathcal{L}_m = \lambda_{rgb}\mathcal{L}_{rgb}^{map} + \lambda_{dep}\mathcal{L}_{dep} + \mathcal{L}_{sdf} \tag{15}$$

Unlike tracking, the mapping process relies more on RGB information to compensate for invalid depth, requiring a distinct treatment of $\mathcal{L}_{rgb}^{tracking}$ and $\mathcal{L}_{rgb}^{map}$. Additionally, since scene representation is optimized only during mapping, we do not reweight $\mathcal{L}_{rgb}^{map}$ with the confidence function CF_m in Eq. (12).

3.4. Strategic Bundle Adjustment

In bundle adjustment (BA), keyframes are selected first, followed by joint optimization of camera poses and scenes. Traditional dense SLAM techniques require storing keyframe images for dense pixel-level loss calculation. Recent NeRF-based SLAM methods like iMap [55] and Nice-SLAM [75] use local BA, selecting a small fraction of keyframes and points through a sliding window. In [19, 61], global BA optimizes all keyframes. However, none of these NeRF-based SLAM methods incorporate uncertainty management in keyframe selection or BA. Performing mapping process every n frames is unreasonable due to the random motion states and varying quality of depth and color images, which provide different information to the scene representation. Any misestimation (e.g., outlier pose) will have a global impact and might cause false reconstruction. Therefore, corrective and remedial strategies are needed. To better balance efficiency and accuracy, we propose an uncertainty-guided local-to-global bundle adjustment, as depicted in Fig. 4. Tracking operations are executed for every frame, while mapping with global BA occurs every n frames constantly. In order to capture local information, our Uni-SLAM system can activate **additional mapping processes** with local BA based on image-level uncertainty β if $\beta > \beta_{unc}$, where β_{unc} is the threshold for image-level

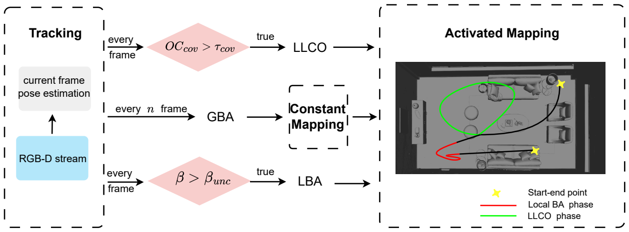


Figure 4. **Strategic BA.** While the tracking process is performed at every frame, we perform a constant mapping with global bundle adjustment (GBA) at a fixed frequency. Thus, the pose and map are optimized using all keyframes from the start to the end of the frame sequence. If an outlier frame is detected based on its uncertainty, a local bundle adjustment (LBA) is performed, as shown in red. If a loop closure is detected, a local loop closure optimization (LLCO) is performed, as shown in green in the figure.

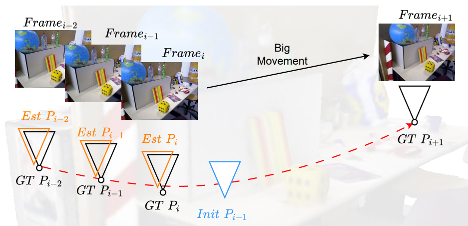


Figure 5. **Activated additional local BA**. From position P_i to P_{i+1}, sudden large movements lead to difficulties in pose estimation and increased uncertainty due to unseen areas. The initialization of $Init\ P_{i+1}$ based on the constant speed assumption is hard to optimize. Therefore, besides constant global BA, we activate additional local BA based on image-level uncertainty to optimize local information. This simulates slowing down the movement. Its effectiveness can be found in Fig. 8 and Tab. 7.

uncertainty. In local BA, we use only keyframes that visually overlap with the current frame, mitigating the impact of outlier frames. Fig. 5 illustrates this necessity.

For local BA keyframe selection, we first initialize spatial sample points in 3D space using the current frame's camera pose. These points are then back-projected onto previous keyframes to check how many fall within image boundaries, determining overlap. Prioritizing local over global information, this method enables efficient local map updates with a limited number of M sample points and informs our co-visibility check. Eq. (16) defines the overlapping coefficient of co-visibility $OC_{cov}(i, c)$ between i-keyframe I_i and current frame I_c, $I_i \in$ *Keyframe Database* $\{I_1, I_2, \ldots, I_n\}$

$$OC_{cov}(i, c) = \frac{|I_i \cap I_c|}{|I_c|} \tag{16}$$

At the end of the tracking process for every frame, we calculate the co-visibility with negligible computational overhead. If the co-visibility is larger than threshold τ_{cov} (set at 0.95), it indicates a loop closure. In this case, the **additional mapping process** with local loop closure optimization (LLCO) is performed immediately. This process optimizes only the keyframes from the current frame to the loop closure point, as shown in green circle in Fig. 4. This approach enables efficient use of M sample points and improves system stability.

4. Experiments and Results

4.1. Experimental Setup

Datasets. We evaluate Uni-SLAM using diverse benchmarks, including the synthetic Replica dataset [52] with 8 high-quality indoor scene reconstructions, as well as the realistic ScanNet [9] and TUM RGB-D datasets [53].
Metrics. We assess the quality of our reconstruction from multiple perspectives. For tracking accuracy, we adopt *ATE RMSE [cm]* [54]. We analyze the reconstruction quality using 3D and 2D metrics. For 3D metrics, the meshes produced by marchingcubes [30] are evaluated by *Depth L1 [cm], Accuracy [cm], Reconstruction completion [cm]*, and *Completion ratio [1cm]%*. Those meshes are culled following [2] before evaluation. For 2D rendering, we provide the peak signal-to-noise ratio (PSNR), SSIM [63], and LPIPS [73]. The rendering metrics are evaluated every 5 frames on full-resolution images.
Baselines and Implementation. We primarily compare our method to existing state-of-the-art dense implicit RGB-D SLAM systems such as Nice-SLAM [76], Co-SLAM [61], ESLAM [21], and BSLAM [19]. For BSLAM we produce results with their novel proposed hybrid model. We reproduce their results using the open-source code and report the middle value after 5 runs. The results of iMAP* [55] are adopted from Nice-SLAM. For a fair comparison, we extract mesh at $1cm$ resolution. In our pipeline implementation, we set the hash grid level to 16 for both geometry and appearance grids. We randomly select 4,000 sampling points for the mapping process and 2,000 for the tracking process. The truncation distance is set to 6 cm. Additional details can be found in Supp. Sec. A.1.

Method	Rm 0	Rm 1	Rm 2	Off 0	Off 1	Off 2	Off 3	Off 4	Ave.
iMAP* [55]	5.23	3.09	2.58	2.4	1.17	5.67	5.08	2.23	3.24
Nice-SLAM [76]	0.97	1.31	1.07	0.88	1.00	1.06	1.10	1.13	1.06
MIPS-Fusion [59]	1.10	1.20	1.10	0.70	0.80	1.30	2020	1.10	1.19
Co-SLAM [61]	0.66	2.25	1.07	0.65	0.53	2.12	1.32	0.85	1.18
ESLAM [21]	0.69	0.70	0.52	0.57	0.55	0.58	0.72	0.63	0.63
BSLAM [19]	0.71	0.88	1.5	0.61	0.49	2.14	1.63	1.66	1.19
Ours	**0.49**	**0.48**	**0.40**	**0.37**	**0.36**	**0.48**	**0.56**	**0.44**	**0.45**

Table 1. **Tracking performance on Replica [52]**(RMSE $\downarrow$ [cm]).

4.2. Qualitative and Quantitative Evaluation

Reconstruction & Rendering. Fig. 6 compares the mesh reconstructions of Co-SLAM [61], BSLAM [19] and ours to ground truth mesh on Replica. Our method can achieve more accurate thin geometric details and high-fidelity colors, such as captured chair legs and thin tables. Quantitatively, Tab. 2 compares reconstruction and

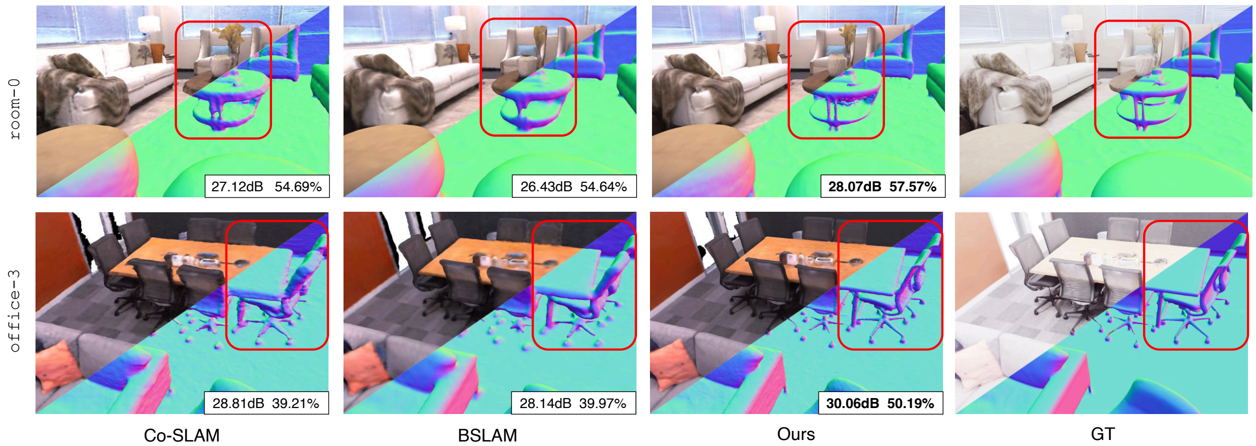


Figure 6. **Mesh Evaluation on Replica [52].** Our method outstands with its thin geometry details and higher texture fidelity compared to Co-SLAM [61] and BSLAM [19]. For example, the table and vase in `room-0`; the thinner office desk, chair backrest, and detailed reconstructed chair legs in `office-3`. In the lower right corner, we note rendering quality in PSNR$[dB]$ $\uparrow$ and geometric evaluation in completion ratio $[< 1cm\%]$ $\uparrow$. Please zoom-in for more details.

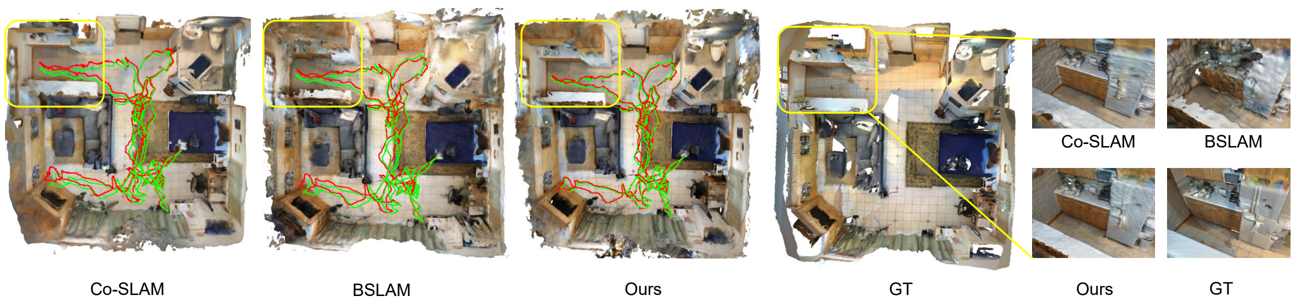


Figure 7. **Mesh Evaluation on ScanNet [9].** The estimated pose is shown in red, and the ground truth camera pose is shown in green. Our method stands out with its more accurate trajectory and higher quality reconstruction, such as the corners of the kitchen.

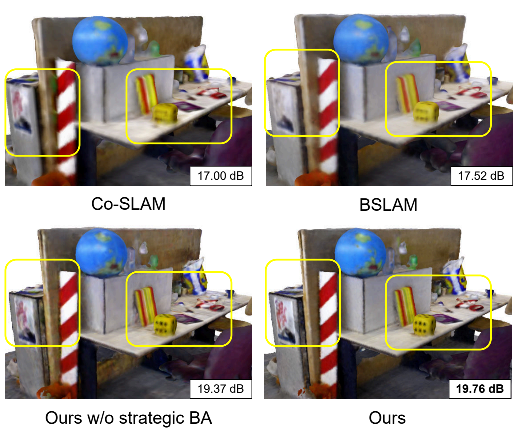


Figure 8. **Mesh Evaluation on TUM RGB-D [53].** Our method stands out with its geometry details and higher texture fidelity. Without strategic BA (only with global BA), the performance can be suboptimal due to missing local information.

Method	Reconstruction $[cm]$ Depth L1↓	Acc.↓	Comp.↓	Comp. Ratio $[\%]$ ↑	Rendering PSNR$[dB]$ ↑	SSIM↑	LPIPS↓
iMAP* [55]	8.23	7.16	5.83	20.33	17.32	0.6535	0.3425
Nice-SLAM [76]	3.18	1.90	1.53	36.93	24.42	0.8091	0.2335
Co-SLAM [61]	2.15	1.16	1.12	55.94	30.27	0.9396	0.2468
ESLAM [21]	1.18	0.97	1.05	63.99	30.19	0.9421	0.2433
BSLAM [19]	2.52	1.12	1.10	57.18	29.55	0.9335	0.2361
Ours	**0.89**	**0.92**	**0.92**	**66.86**	**31.62**	**0.9584**	**0.1853**

Table 2. **Reconstruction and Rendering Performance on Replica** [52]**.** To reflect the ability to reconstruct geometric details, we report completion ratio $[< 1cm\%]$. For the details of the evaluations for each scene, refer to the supplementary material.

Method		Sc.00	Sc.59	Sc.106	Sc.169	Sc181	Sc.207	Ave.
iMAP* [55]	ICCV 21	42.7	17.8	15.0	39.1	24.7	20.1	26.6
Nice-SLAM [76]	CVPR 22	12.0	14.0	7.9	10.9	13.4	6.2	10.7
MIPS-Fusion [59]	SA 23	7.9	10.7	9.7	9.7	14.2	7.8	10.0
ESLAM [21]	CVPR 23	7.3	8.5	7.5	6.5	**9.0**	5.7	7.4
Co-SLAM [61]	CVPR 23	7.2	12.3	9.6	6.6	13.4	7.1	9.4
BSLAM [19]	CVPR 24	7.29	12.2	9.0	8.8	13.4	6.65	9.56
Ours		**6.12**	**7.77**	**7.41**	**5.82**	9.77	**5.21**	**7.01**

Table 3. **Tracking Performance on ScanNet** [9](RMSE [cm]) **.** On average, our method achieved the best results.

rendering performance on the Replica dataset and shows best among 3D metrics and 2D metrics, beating all implicit dense SLAM. In Fig. 7 we show that our method can achieve more accurate localization and finer realistic details on ScanNet. We attribute this to our sufficient model capability and online uncertainty-aware activated additional mapping process, which can capture more details locally. The reconstructed mesh on TUM RGB-D is shown in Fig. 8. The results show that our reconstruction quality benefits from strategic BA.

Tracking. Tab. 1 compares our methods to state-of-the-art implicit dense RGB-D neural SLAM system on 8 scenes of Replica datasets [52] in tracking performance. We outperform on all scenes and achieve an average improvement of 62%, 29% and 62% on RMSE over Co-SLAM, and ESLAM and BSLAM respectively. The tracking performance on ScanNet and TUM RGB-D is shown in Tab. 3 and Tab. 4 respectively. We primarily attribute this to the uncertainty reweighted loss function, where only the most reliable information is emphasized. Although classic methods are still showing state-of-the-art accurate tracking on TUM RGB-D, our method outperforms neural methods on average and bridges the gap between those two categories.

	Method	fr1/ desk	fr2/ xyz	fr3/ office	Ave.
NeRF-Based	iMAP* [55]	5.15	2.39	5.76	4.43
	Nice-SLAM [76]	5.00	3.17	5.05	4.41
	MIPS-Fusion [59]	3.00	1.40	4.6	3.0
	Co-SLAM [61]	3.05	1.88	2.85	2.59
	ESLAM [21]	2.54	**1.13**	2.75	2.14
	BSLAM [19]	2.87	1.38	2.95	2.39
	Ours	**2.37**	1.17	**2.62**	**2.05**
Classic	ORB-SLAM2 [37]	**1.6**	**0.4**	**1.0**	**1.0**
	BundleFusion [10]	1.6	1.1	2.2	1.63
	BAD-SLAM [49]	1.7	1.1	1.7	1.5

Table 4. **Tracking Performance on TUM RGB-D** [53] (RMSE [cm]) **.**

4.3. Analysis on Design Choices

Runtime and Memory Analysis In Tab. 5, we compare runtime and memory usage, benchmarking all methods on NVIDIA GeForce RTX 4090 GPU using `room0` of Replica [52]. We report tracking and mapping times per iteration and compare iteration steps to show convergence speed. Our model achieves real-time performance on par with SOTA results at speeds exceeding 8 FPS.

Method	Tracking [ms x it.] ↓	Mapping [ms x it.] ↓	FPS↑	Time Mins↓	Params.↓
Nice-SLAM [76]	6.5 x 10	29.3x60	1.8	18.51	12.13M
Co-SLAM [61]	**4.6 x 10**	**6.6 x 10**	**9.07**	**3.67**	**1.72M**
ESLAM [21]	7.9 x 8	18.8 x 15	5.55	6.01	<u>6.78M</u>
BSLAM [46]	11 x 20	15 x 20	1.66	20.3	17.38M
Ours	<u>7.0 x 8</u>	<u>8.1 x 13</u>	<u>8.37</u>	<u>4.02</u>	12.69M

Table 5. Runtime and Memory Usage Comparison.

Ablation of Model Design. We encoded geometry and appearance using different structures and validated our design choices on the Replica dataset [52], as shown in Tab. 6. By ablating various combinations of hash grids [36] and tri-planes [4], we found that using two hash grids without a third learnable uncertainty grid (h-h-n) produced the best results. Introducing a third learnable uncertainty grid (h-h-u) under the Gaussian assumption made training and convergence more complex. Further details can be found in Supplementary Sec. B.3.

Method	Reconstruction $[cm]$				Rendering/Tracking/Time		
	Depth L1↓	Acc.↓	Comp.↓	Comp. Ratio $[\%]$ ↑	PSNR$[dB]$ ↑	RMSE$[cm]$ ↓	Mins↓
h-h-u	3.75	1.79	1.65	31.52	27.33	1.51	6.53
h-t-n	0.93	1.01	1.15	64.69	30.98	0.47	4.79
t-h-n	0.97	1.17	1.09	63.82	31.32	0.50	4.65
Ours(h-h-n)	**0.89**	**0.92**	**0.92**	**66.86**	**31.62**	**0.45**	**3.97**

Table 6. Ablation of model design.

Ablation on Reweighting. In Fig. 9, we present a quantitative analysis of the application of model uncertainty to various loss terms on TUM RGB-D [53]. Configuration (d) achieves the highest localization accuracy and rendering quality. During tracking, reweighting all terms to focus on only low-uncertainty information improves localization. In mapping, color information can compensate for invalid depth values, so reweighting is not applied to the color term. This strategy enhances reconstruction quality in both geometry (lower depth L1) and appearance (higher PSNR) compared to configuration (e).

Method		Reweighting Term			Tracking/Rendering		
		SDF	Depth	Color	RMSE [cm] ↓	PSNR [dB] ↑	Depth L1[m]↓
a)	Tracking	✗	✗	✗	7.18	16.76	0.347
	Mapping	✗	✗	✗			
b)	Tracking	✗	✗	✗	2.32	19.82	0.111
	Mapping	✓	✓	✗			
c)	Tracking	✓	✓	✓	6.57	17.25	0.281
	Mapping	✗	✗	✗			
d)	Tracking	✓	✓	✓	**2.05**	**21.23**	**0.099**
	Mapping	✓	✓	✗			
e)	Tracking	✓	✓	✓	2.21	20.17	0.115
	Mapping	✓	✓	✓			

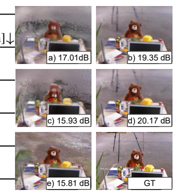


Figure 9. Ablation on loss term reweighting

Ablation of strategic BA. Tab. 7 shows localization accuracy and rendering quality under different BA strategies on 6 ScanNet scenes. Experimental results demonstrate that our uncertainty-guided strategic BA method achieves optimal performance by dynamically activating the mapping process and selecting keyframes. Fig. 8 ablates the reconstructed mesh without strategic BA.

Method	Keyframe Selection			Camera pose	ATE [cm]		PSNR↑
	Local	Global	LC		RMSE↓	Mean↓	
w/o BA				✗	17.58	15.15	17.63
LBA	✓			✓	8.77	7.23	20.62
GBA		✓		✓	8.35	7.17	21.52
LBA + GBA	✓	✓		✓	7.23	6.56	21.59
LBA + GBA + LLCO	✓	✓	✓	✓	**7.01**	**6.15**	**21.77**

Table 7. Ablation of strategic BA: LBA selects 20 local keyframes, GBA includes all keyframes, and LLCO focuses on keyframes in loop closure.

5. Conclusion

We present Uni-SLAM, a novel uncertainty-guided dense implicit SLAM approach. In decoupled scene representation, we propose utilizing model-free predictive uncertainty to reweight the loss function at the pixel level to capture effective information, achieving high-frequency geometric reconstruction. By leveraging image-level uncertainty, we strategically perform bundle adjustment to balance local-to-global information. Overall, our method achieves state-of-the-art high-fidelity mapping and accurate tracking in real-time among dense SLAM.

We accept a trade-off in efficiency through random sampling in a real-time required SLAM system. However, active sampling based on uncertainty should further improve efficiency and yield finer edge structures. We leave this for future work.

Acknowledgements: This research has been partially funded by the EU projects CORTEX2 (GA Nr 101070192) and FLUENTLY (GA Nr 101058680).

References

[1] Shervin Ardeshir and Navid Azizan. Uncertainty in contrastive learning: On the predictability of downstream performance. *arXiv preprint arXiv:2207.09336*, 2022. 2

[2] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6290–6301, 2022. 2, 6

[3] Henry Carrillo, Ian Reid, and José A Castellanos. On the comparison of uncertainty criteria for active slam. In *2012 IEEE International Conference on Robotics and Automation*, pages 2080–2087. IEEE, 2012. 2

[4] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 2, 8

[5] Hao-Wei Chen, Ting-Hsuan Liao, Hsuan-Kung Yang, and Chun-Yi Lee. Pixel-wise prediction based visual odometry via uncertainty estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2518–2528, 2023. 2

[6] Hae Min Cho, HyungGi Jo, and Euntai Kim. Sp-slam: Surfel-point simultaneous localization and mapping. *IEEE/ASME Transactions on Mechatronics*, 27(5):2568–2579, 2021. 2

[7] Chi-Ming Chung, Yang-Che Tseng, Ya-Ching Hsu, Xiang-Qian Shi, Yun-Hung Hua, Jia-Fong Yeh, Wen-Chin Chen, Yi-Ting Chen, and Winston H Hsu. Orbeez-slam: A real-time monocular visual slam with orb features and nerf-realized mapping. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9400–9406. IEEE, 2023. 2

[8] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 2

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 6, 7

[10] 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 Transactions on Graphics (ToG)*, 36(4):1, 2017. 8

[11] Tianchen Deng, Guole Shen, Tong Qin, Jianyu Wang, Wentao Zhao, Jingchuan Wang, Danwei Wang, and Weidong Chen. Plgslam: Progressive neural scene represenation with local to global bundle adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19657–19666, 2024. 1, 2

[12] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014. 1

[13] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 15–22. IEEE, 2014. 1

[14] Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' rays: Uncertainty quantification for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20061–20070, 2024. 2

[15] Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' rays: Uncertainty quantification for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20061–20070, 2024. 4

[16] Christopher J Holder and Muhammad Shafique. Efficient uncertainty estimation in semantic segmentation via distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3087–3094, 2021. 2

[17] Julia Hornauer and Vasileios Belagiannis. Gradient-based uncertainty for monocular depth estimation. In *European Conference on Computer Vision*, pages 613–630. Springer, 2022. 2

[18] Julia Hornauer, Adrian Holzbock, and Vasileios Belagiannis. Out-of-distribution detection for monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1911–1921, 2023. 2

[19] Tongyan Hua and Lin Wang. Benchmarking implicit neural representation and geometric rendering in real-time rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21346–21356, 2024. 1, 2, 5, 6, 7, 8

[20] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021. 2

[21] Mohammad Mahdi Johari, Camilla Carta, and François Fleuret. Eslam: Efficient dense slam system based on hybrid representation of signed distance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17408–17419, 2023. 1, 2, 6, 7, 8

[22] Kim-Celine Kahl, Carsten T Lüth, Maximilian Zenk, Klaus Maier-Hein, and Paul F Jaeger. Values: A framework for systematic validation of uncertainty estimation in semantic segmentation. *arXiv preprint arXiv:2401.08501*, 2024. 2

[23] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21357–21366, 2024. 2

[24] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 2

[25] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM international symposium on mixed and augmented reality*, pages 225–234. IEEE, 2007. 2

[26] Xin Kong, Shikun Liu, Marwan Taher, and Andrew J Davison. vmap: Vectorised object mapping for neural field slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 952–961, 2023. 2

[27] Soomin Lee, Le Chen, Jiahao Wang, Alexander Liniger, Suryansh Kumar, and Fisher Yu. Uncertainty guided policy for active robotic 3d reconstruction using neural radiance fields. *IEEE Robotics and Automation Letters*, 7(4):12070–12077, 2022. 2

[28] Ruihao Li, Sen Wang, and Dongbing Gu. Deepslam: A robust monocular slam system with unsupervised deep learning. *IEEE Transactions on Industrial Electronics*, 68(4):3577–3587, 2020. 2

[29] Lorenzo Liso, Erik Sandström, Vladimir Yugay, Luc Van Gool, and Martin R Oswald. Loopy-slam: Dense neural slam with loop closures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20363–20373, 2024. 2

[30] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, 21(4):163–169, 1987. 6

[31] Kira Maag and Asja Fischer. Uncertainty-weighted loss functions for improved adversarial attacks on semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3906–3914, 2024. 2

[32] Yunxuan Mao, Xuan Yu, Kai Wang, Yue Wang, Rong Xiong, and Yiyi Liao. Ngel-slam: Neural implicit representation-based global consistent low-latency slam system. *arXiv preprint arXiv:2311.09525*, 2023. 2

[33] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7210–7219, 2021. 2

[34] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18039–18048, 2024. 2

[35] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1, 3

[36] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2, 3, 8

[37] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017. 1, 2, 8

[38] 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 *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. Ieee, 2011. 2

[39] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *2011 international conference on computer vision*, pages 2320–2327. IEEE, 2011. 1, 2

[40] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32(6):1–11, 2013. 2

[41] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 3

[42] Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. Activenerf: Learning where to see with uncertainty estimation. In *European Conference on Computer Vision*, pages 230–246. Springer, 2022. 2

[43] Weining Ren, Zihan Zhu, Boyang Sun, Jiaqi Chen, Marc Pollefeys, and Songyou Peng. Nerf on-the-go: Exploiting uncertainty for distractor-free nerfs in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8931–8940, 2024. 5

[44] Antoni Rosinol, John J Leonard, and Luca Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3437–3444. IEEE, 2023. 2

[45] Antoni Rosinol, John J Leonard, and Luca Carlone. Probabilistic volumetric fusion for dense monocular slam. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3097–3105, 2023. 2

[46] Erik Sandström, Yue Li, Luc Van Gool, and Martin R Oswald. Point-slam: Dense neural point cloud-based slam. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18433–18444, 2023. 2, 8

[47] Erik Sandström, Kevin Ta, Luc Van Gool, and Martin R Oswald. Uncle-slam: Uncertainty learning for dense neural slam. In *Proceedings of the International Conference on Computer Vision (ICCV) Workshops*. IEEE, 2023. Available as arXiv preprint arXiv:2306.11048. 2

[48] Muhamad Risqi U Saputra, Pedro PB De Gusmao, Chris Xiaoxuan Lu, Yasin Almalioglu, Stefano Rosa, Changhao Chen, Johan Wahlström, Wei Wang, Andrew Markham, and Niki Trigoni. Deeptio: A deep thermal-inertial odometry with visual hallucination. *IEEE Robotics and Automation Letters*, 5(2):1672–1679, 2020. 2

[49] Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 134–144, 2019. 8

[50] Jianxiong Shen, Antonio Agudo, Francesc Moreno-Noguer, and Adria Ruiz. Conditional-flow nerf: Accurate 3d modelling with reliable uncertainty quantification. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 2

[51] Jianxiong Shen, Adria Ruiz, Antonio Agudo, and Francesc Moreno-Noguer. Stochastic neural radiance fields: Quantifying uncertainty in implicit 3d representations. In *2021 International Conference on 3D Vision (3DV)*, pages 972–981. IEEE, 2021. 2

[52] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 6, 7, 8

[53] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012. 1, 2, 6, 7, 8

[54] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012. 6

[55] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021. 1, 2, 5, 6, 7, 8

[56] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 2

[57] Niko Sünderhauf, Jad Abou-Chakra, and Dimity Miller. Density-aware nerf ensembles: Quantifying predictive uncertainty in neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9370–9376. IEEE, 2023. 2, 4

[58] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11367, 2021. 2

[59] 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. *ACM Transactions on Graphics (TOG)*, 42(6):1–16, 2023. 2, 6, 7, 8

[60] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 2

[61] Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13293–13302, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[62] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in neural information processing systems*, 2021. 3

[63] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[64] Thomas Whelan, Stefan Leutenegger, Renato Salas-Moreno, Ben Glocker, and Andrew Davison. Elasticfusion: Dense slam without a pose graph. In *Proceedings of the Robotics: Science and Systems*. RSS, 2015. 2

[65] Kira Wursthorn, Markus Hillemann, and Markus Ulrich. Comparison of uncertainty quantification methods for cnn-based regression. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:721–728, 2022. 2

[66] Yuting Xiao, Jingwei Xu, Zehao Yu, and Shenghua Gao. Debsdf: Delving into the details and bias of neural indoor scene reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 5

[67] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 2

[68] Shangjie Xue, Jesse Dill, Pranay Mathur, Frank Dellaert, Panagiotis Tsiotra, and Danfei Xu. Neural visibility field for uncertainty-driven active mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18122–18132, 2024. 2

[69] Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19595–19604, 2024. 2

[70] Xingrui Yang, Hai Li, Hongjia Zhai, Yuhang Ming, Yuqian Liu, and Guofeng Zhang. Vox-fusion: Dense tracking and mapping with voxel-based neural implicit representation. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 499–507. IEEE, 2022. 2, 3

[71] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 2

[72] Masashi Yokozuka, Shuji Oishi, Simon Thompson, and Atsuhiko Banno. Vitamin-e: Visual tracking and mapping with extremely dense feature points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2019. 2

[73] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE*

conference on computer vision and pattern recognition, pages 586–595, 2018. 6

[74] Youmin Zhang, Fabio Tosi, Stefano Mattoccia, and Matteo Poggi. Go-slam: Global optimization for consistent 3d instant reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3727–3737, 2023. 2, 3

[75] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R Oswald, Andreas Geiger, and Marc Pollefeys. Nicer-slam: Neural implicit scene encoding for rgb slam. In *2024 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2024. 2, 5

[76] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 1, 2, 3, 6, 7, 8