

SHINE-Mapping: Large-Scale 3D Mapping Using Sparse Hierarchical Implicit Neural Representations

Xingguang Zhong*

Yue Pan*

Jens Behley

Cyrill Stachniss

Abstract—Accurate mapping of large-scale environments is an essential building block of most outdoor autonomous systems. Challenges of traditional mapping methods include the balance between memory consumption and mapping accuracy. This paper addresses the problem of achieving large-scale 3D reconstruction using implicit representations built from 3D LiDAR measurements. We learn and store implicit features through an octree-based, hierarchical structure, which is sparse and extensible. The implicit features can be turned into signed distance values through a shallow neural network. We leverage binary cross entropy loss to optimize the local features with the 3D measurements as supervision. Based on our implicit representation, we design an incremental mapping system with regularization to tackle the issue of forgetting in continual learning. Our experiments show that our 3D reconstructions are more accurate, complete, and memory-efficient than current state-of-the-art 3D mapping methods.

I. INTRODUCTION

Localization and navigation in large-scale outdoor scenes is a common task of mobile robots, especially for self-driving cars. An accurate and dense 3D map of the environment plays a relevant role in achieving these tasks, and most systems use or maintain a model of their surroundings. Usually, for outdoors, dense 3D maps are built based on range sensors such as 3D LiDAR [3], [18]. Due to the limited memory of most mobile robots performing all computations onboard, maps should be compact but, at the same time, accurate enough to represent the environment in sufficient detail.

Current large-scale mapping methods often use spatial grids or various tree structures as map representation [7], [10], [15], [25], [36]. For these models, it can be hard to simultaneously satisfy both desires, accurate and detailed 3D information but not requiring substantial memory resources. Additionally, these methods often do not perform well in areas that have been only sparsely covered with sensor data. In such areas, they usually cannot reconstruct a map at a high level of completeness.

Recently, neural network-based representations [14], [16], [26] attracted significant interest in the computer vision and robotics communities. By storing information about the environment in the neural network implicitly, these approaches can achieve remarkable accuracy and high-fidelity reconstructions using a comparably compact representation.

* Equal contribution.

All authors are with the University of Bonn, Germany. Cyrill Stachniss is additionally with the Department of Engineering Science at the University of Oxford, UK, and with the Lamarr Institute for Machine Learning and Artificial Intelligence, Germany.

This work has partially been funded by the European Union under the grant agreements No 101070405 (DigiForest) and No 101017008 (Harmony).

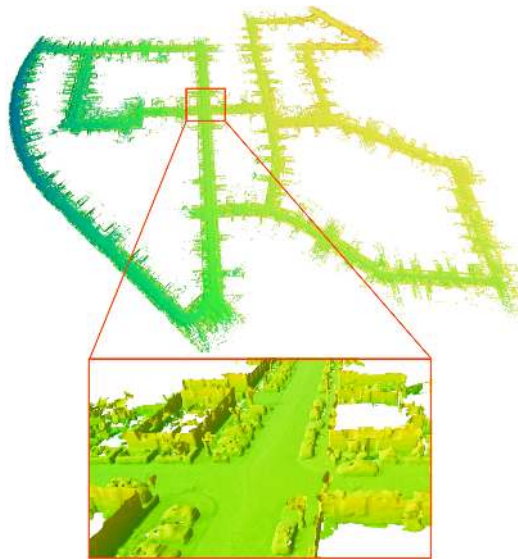


Fig. 1: The incremental reconstruction result of the proposed approach on KITTI odometry sequence 00.

Due to such advantages, several recent works have used implicit representation for 3D scene reconstruction built from images data or RGB-D frames [2], [31], [33], [37], [40]. Comparably, little has been done in the context of LiDAR data. Furthermore, most of these methods only work in relatively small indoor scenes, which is difficult to use in robot applications for large-scale outdoor environments.

In this paper, we address this challenge and investigate effective means of using an implicit neural network-based map representation for outdoor robots using range sensors such as LiDARs. We took inspiration from the recent work by Takikawa et al. [34], which represents surfaces with a sparse octree-based feature volume and can adaptively fit 3D shapes with multiple levels of detail.

The main contribution of this paper is a novel approach called SHINE-Mapping that enables large-scale incremental 3D mapping and allows for accurate and complete reconstructions exploiting sparse hierarchical implicit representation. An example illustration of an incremental mapping result generated by our approach on the KITTI dataset [6] is given in Fig. 1. Our approach exploits an octree-based sparse data structure to store incrementally learnable feature vectors and a shared shallow multilayer perceptron (MLP) as the decoder to transfer local features to signed distance values. We design a binary cross entropy-based loss function for efficient and robust local feature optimization. By interpolating among local features, our representation enables us to query geometric information at any resolution. We use

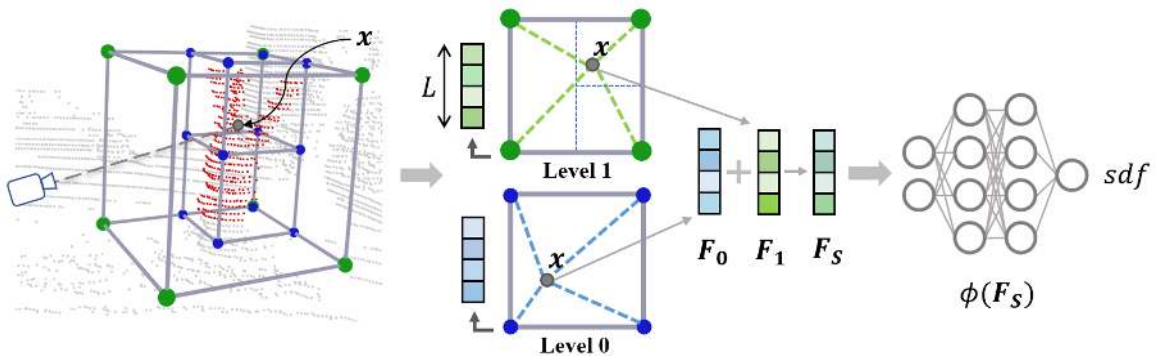


Fig. 2: Overview of querying for SDF value in our map representation. For querying a SDF value, we first determine the grids in each level with features where the query point x belongs to and determine the interpolated feature at the location of the query point by trilinear interpolation of the vertex features. We add the interpolated features of every level as F_s to regress the SDF value using the MLP ϕ .

point clouds as observation to optimize local features online and achieve incremental mapping through the extension of the octree structure. Our approach yields an incremental mapping system based on our implicit representation and uses the feature update regularization to tackle the issue of catastrophic forgetting [13]. As the experiments illustrate, our mapping approach (i) is more accurate in the regions with dense observations than TSDF fusion-based methods [21], [36] and volume rendering-based implicit neural mapping method [40]; (ii) is more complete in the region with sparse observations than the non-learning-based mapping methods; (iii) generates more compact map than TSDF fusion-based methods.

The open-source implementation is available at: https://github.com/PRBonn/SHINE_mapping.

II. RELATED WORK

The usage of LiDAR data for large-scale environment mapping is often a key component to enable localization and navigation but also enables applications involving visualization, BIMs, or augmented reality. Besides surfel-based representations [3], triangle meshes [4], [12], [29], [35], and histograms [30], octree-based occupancy representations [7] are a popular choice for representing the environment. The seminal work of Newcombe et al. [18] that enabled real-time reconstruction using truncated signed distance function (TSDF) [5] popularized the use of volumetric integration methods [10], [21], [24], [36], [38]. The abovementioned representations result in an explicit geometrical representation of the environment that can be used for localization [4] and navigation [22].

While such explicit geometric representations enable detailed reconstructions in large-scale environments [19], [36], [38], the recent emergence of neural representations, like NeRF [16] for novel view synthesis inspired researchers to leverage their capabilities for mapping and reconstruction [2], [8], [17], [27], [31], [32], [33], [34], [40]. Such implicit representations represent the environment via MLPs that estimate a density that can be used to volumetrically render novel views [16] or scene geometry [17], [31], [40].

For the implicit neural mapping with sequential data, iMap [31], iSDF [23] and Neural RGBD [2] follow NeRF to use a single MLP to represent the entire indoor scene.

Extending these methods to large-scale mapping is impractical due to the limited model capacity. By combining a shallow MLP with optimizable local feature grids, NICE-SLAM [40] and Go-SURF [37] can achieve more accurate and faster surface reconstruction in larger scenes such as multiple rooms. However, the most notable shortcoming of these methods is the enormous memory cost of dense voxel structures. Our method leverages the octree-based sparse feature grid to reduce memory consumption significantly.

Additionally, incremental mapping with implicit representation can be regarded as a continual learning problem, which faces the challenge of the so-called catastrophic forgetting [13]. To solve this problem, iMap [31] and iSDF [23] replay keyframes from historical data and train the network together with current observations. Nevertheless, for large-scale incremental 3D mapping, such replay-based methods will inevitably store more and more data and need complex resampling mechanisms to keep it memory-efficient. By storing features in local grids and using a fixed MLP decoder, NICE-SLAM [40] claims it does not suffer too much from the forgetting issue. But as the feature grid size increases, the impact of the forgetting problem becomes more severe, which would be illustrate in detail in Sec. III. In our approach, we achieve incremental mapping with limited memory by leveraging feature update regularization.

III. OUR APPROACH – SHINE-MAPPING

We propose a framework for large-scale 3D mapping based on an implicit neural representation taking point clouds from a range sensor such as LiDAR with known poses as input. Our implicit map representation uses learnable octree-based hierarchical feature grids and a globally shared MLP decoder to represent the continuous signed distance field (SDF) of the environment. As illustrated in Fig. 2, we optimize local feature vectors online to capture the local geometry by using direct measurements to supervise the output signed distance value from the MLP. Using this learned implicit map representation, we can generate an explicit geometric representation in the form of a triangle mesh by marching cubes [11] for visualization and evaluation.

A. Implicit Neural Map Representation

Our implicit map representation has to store spatially located features in the 3D world. The SDF values will then

be inferred from these features through the neural network. Our network will not only use features from one spatial resolution but combine features at H different resolutions. The H hierarchical levels would always double the spatial extent in x, y, z -direction from one level to the next. We found that H equals 3 or 4 is sufficient for good results.

Spatial Model. We first have to specify our data structure. To enable large-scale mapping with implicit representation, one could choose an octree-based map representation similar to the one proposed in NGLOD [34]. In this case, one would store features per octree node corner and consider several levels of the tree when combining features to feed the network. However, NGLOD is designed to model a single object, and the authors do not target the incremental extension of the mapped area. Extending the map is vital for the online robotic operation as the robot’s workspace is typically not known beforehand, especially when navigating through large outdoor scenes.

Based on the NGLOD, we build the octree from the point cloud directly and take a different approach to store our features using hash tables, one table for each level, such that we maintain H hash tables during mapping. For addressing the hash tables, while still being able to find features of upper levels quickly, we can exploit unique Morton codes. Morton codes, also known as Z-order or Lebesgue curve, map multidimensional data to one dimension, i.e., the spatial hash code, while preserving the locality of the data points. This setup allows us to easily extend the map without allocating memory beforehand while still being able to maintain the H most-detailed levels. Thus, this provides an elegant and efficient, i.e., average case $\mathcal{O}(1)$ access to all features.

Features. In our representation, we store the feature, a one-dimensional vector of length L , in each corner of the tree nodes. The MLP will take the same length vector as the input and compute the SDF values. For that operation, the feature vectors from at most H levels of our representation are combined to form the network’s input. For these features stored in corners, we randomly initialize their values when created and then optimize them until convergence by using the training pairs sampled along the rays.

Fig. 2 illustrates the overall process of training our implicit map representation, and the right-hand side of the image illustrates the combination of features from different levels. For clearer visualization, we depict only two levels, green and blue. For any query location, we start from the lowest level (level 0) and compute a trilinear interpolation for the position \mathbf{x} . This yields the feature vector at level 0. Then, we move up the hierarchy to level 1 to repeat the process. We combine features through summation for up to H levels.

Next, we feed the summed-up feature vector into a shared shallow MLP with M hidden layers to obtain the SDF value. As the whole process is differentiable, we can optimize the feature vectors and the MLP’s parameters jointly through backpropagation. To ensure our shared MLP generalizes well at various scales, we do not stack the query point coordinates under the final feature vector as done by Takikawa et al. [34].

Our MLP does not need to be pre-trained if mapping

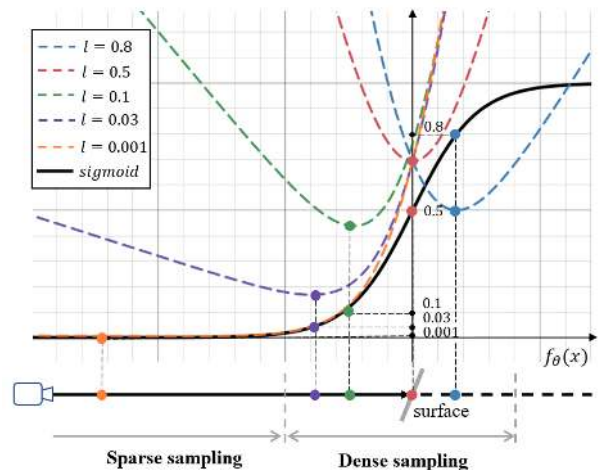


Fig. 3: The dashed lines show the L_{bce} curves for different labels l with network output $f_{\theta}(\mathbf{x})$ as horizontal coordinates. l is calculated from different sample point (colorful dot) along the ray. As can be seen, for the point close to the surface, its loss is more sensitive to changes in network output. Moreover, if the network mispredicts the sign’s symbol for any sampling point, there will be a significant loss, which is what we expect.

operates in batch mode, i.e., all range measurements taken at known poses are available. In case we map incrementally, we, in contrast, use a pre-trained MLP and keep it fixed to minimize the effects of catastrophic forgetting.

B. Training Pairs and Loss Function

Range sensors such as LiDARs typically provide accurate range measurements. Thus, instead of using the volume rendering depth as the supervision as done by vision-based approaches [2], [20], [37], [40], we can obtain more or less directly from the range readings. We obtain training pairs by sampling points along the ray and directly use the signed distance from sampled point to the beam endpoint as the supervision signal. This signal is often referred to as the projected signed distance along the ray.

For SDF-based mapping, the regions of interest are the values close to zero as they define the surfaces. Therefore, sampled points closer to the endpoint should have a higher impact as the precise SDF value far from a surface has very little impact. Thus, instead of using an L2 loss, as, for example, used by Ortiz et al. [23], we map the projected signed distance to $[0, 1]$ before feeding to the actual loss through the sigmoid function: $S(x) = 1/(1 + e^{x/\sigma})$, where the σ is a hyperparameter to control the function’s flatness and indicates the magnitude of the measurement noise. Given the sampled data point $\mathbf{x}_i \in \mathbb{R}^3$, we calculate its projected signed distance to the surface d_i , then use the value after sigmoid mapping $l_i = S(d_i)$ as supervision label and apply the binary cross entropy (BCE) as our loss function:

$$L_{bce} = l_i \cdot \log(o_i) + (1 - l_i) \cdot \log(1 - o_i), \quad (1)$$

where $o_i = S(f_{\theta}(\mathbf{x}_i))$ represents the signed distance output of our model $f_{\theta}(\mathbf{x}_i)$ after the sigmoid mapping, which can be regarded as an indicator for the occupancy probability. The effect of BCE loss is visualized in Fig. 3. Additionally, the

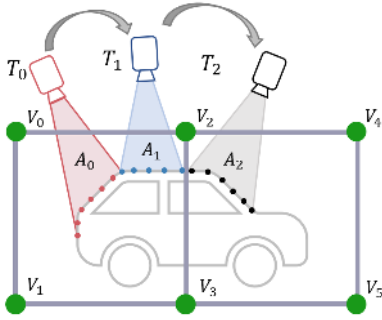


Fig. 4: An example for the catastrophic forgetting in feature grid based incremental mapping.

sigmoid function also realizes soft truncation of signed distance, and the truncation range can be adjusted by changing the σ in the sigmoid function. In order to improve efficiency, we uniformly sample N_f points in the free space and N_s points inside the truncation band $\pm 3\sigma$ around the surface.

Since our network output is the signed distance value, we add an Eikonal term to the loss function to encourage accurate, signed distance values like [23]. Therefore, for batch-based feature optimization training, i.e., non-online mapping, our loss function is as follows:

$$L_{\text{batch}} = L_{\text{bce}} + \lambda_e \underbrace{\left(\left\| \frac{\partial f_{\theta}(\mathbf{x}_i)}{\partial \mathbf{x}_i} \right\| - 1 \right)^2}_{\text{Eikonal loss}}, \quad (2)$$

where λ_e is a hyperparameter representing the weight for the Eikonal loss, and the gradient of the input sample can be efficiently calculated through automatic differentiation.

C. Incremental Mapping Without Forgetting

As we explained in Sec. II, it is not a good choice to use the replay-based method in large-scale implicit incremental mapping. Thus, in this section, we focus on only using the current observation to train our feature field incrementally.

Firstly, we describe the reason catastrophic forgetting happens in feature grid-based implicit incremental mapping. As shown in the Fig. 4, we can only obtain partial observations of the environment at each frame (T_0 , T_1 and T_2). During the incremental mapping, at frame T_0 , we use the data capture from area A_0 to optimize the feature V_0 , V_1 , V_2 , and V_3 . After the training converge, V_0 , V_1 , V_2 , and V_3 will have an accurate geometric representation of A_0 . However, if we move forward and use the data from frame T_1 to train and update V_0 , V_1 , V_2 , and V_3 , the network will only focus on reducing the loss generated in A_1 and does not care about the performance in A_0 anymore. This may lead to a decline in the reconstruction accuracy in A_0 , which is why catastrophic forgetting happens. In addition, when we continue to use the observation at frame T_2 for training, the feature V_2 , V_3 will be updated again, and we cannot guarantee that the feature update caused by the data from T_2 would not deteriorate the mapping accuracy in A_0 and A_1 . Fig. 8 shows the impact of this forgetting problem during incremental mapping, which will become more severe as the grid size increases.

An intuitive idea to solve this problem is to limit the update direction of the local feature vector to make current

update not affect the previously modeled area excessively. Inspired by the regularization-based methods in the continual learning field [1], [9], [39], we add a regularization term to the loss function:

$$L_r = \sum_{i \in A} \Omega_i (\theta_i^t - \theta_i^*)^2, \quad (3)$$

where A refers to the set of local feature parameters that are used in this iteration, θ_i^t represent the current parameter values, and θ_i^* represent parameter values, which have converged in the training of previous scans. The term Ω_i act as the importance weights for different parameters.

In our case, we regard these importance weights as the sensitivity of the loss of previous data to a parameter change, suggested by Zenke et al. [39]. After the convergence of each scan's training, we recalculate the importance weights by:

$$\Omega_i = \min \left(\Omega_i^* + \sum_{k=1}^N \left\| \frac{\partial L_{\text{bce}}(\mathbf{x}_k, l_k)}{\partial \theta_i} \right\|, \Omega_m \right), \quad (4)$$

where Ω_i^* represents the previous importance value, and Ω_m is used to limit the weight to prevent gradient explosion. The iteration from $k = 1$ to $k = N$ means that all samples are considered in the training of this scan.

Intuitively, the gradient of the loss L_{bce} to θ_i indicates the change of the loss on previous data by adjusting the parameter θ_i . In the following training, we prefer changing the parameters with small importance weights to avoid a significant impact on previous losses. Therefore, for incremental mapping, our complete loss function is:

$$L_{\text{incr}} = L_{\text{bce}} + \lambda_e L_{\text{eikonal}} + \lambda_r L_r, \quad (5)$$

where λ_r is a hyperparameter used to adjust the degree of keeping historical knowledge. As our experiments will show, this approach reduces the risk of catastrophic forgetting.

IV. EXPERIMENTAL EVALUATION

The main focus of this work is an incremental and scalable 3D mapping system using a sparse hierarchical feature grid and a neural decoder. In this section, we analyze the capabilities of our approach and assess its performance.

A. Experimental Setup

Our model is first evaluated qualitatively and quantitatively on two publicly available outdoor LiDAR datasets that come with (near) ground truth mesh information. One is the synthetic MaiCity dataset [35], which consists of a sequence of 64 beam noise-free simulated LiDAR scans of an urban scenario. The other is the more challenging, non-simulated Newer College dataset [28], recorded at Oxford University using handheld LiDAR with cm-level measurement noise and substantial motion distortion. Near ground truth data is available from a terrestrial scanner. On these two datasets, we evaluate the mapping accuracy, completeness, and memory

TABLE I: Parameter settings

| H | L | M | N_s | N_f | $\sigma [m]$ |
|-----|-----|-----|-------|-------|--------------|
| 4 | 8 | 2 | 5 | 5 | 0.05 |

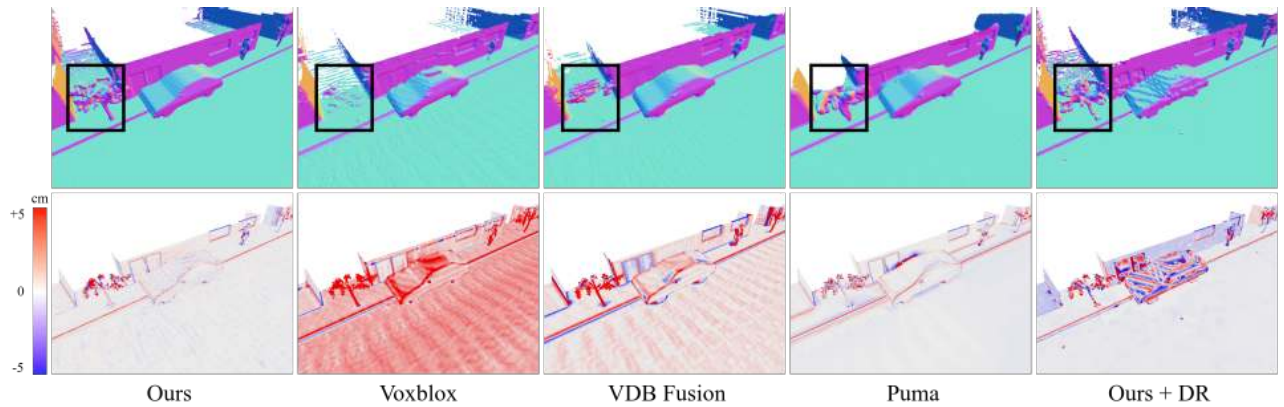


Fig. 5: A comparison of different methods on the *MaiCity* dataset. The first row shows the reconstructed mesh and a tree is highlighted in the black box. The second row shows the error map of the reconstruction overlaid on the ground truth mesh, where the blue to red colormap indicates the signed reconstruction error from -5 cm to +5 cm.

TABLE II: Quantitative results of the reconstruction quality on the *MaiCity* dataset. We report the distance error metrics, namely completion, accuracy and Chamfer-L1 in cm. Additionally, we show the completion ratio and F-score in % with a 10 cm error threshold.

| Method | Comp. ↓ | Acc. ↓ | C-l1 ↓ | Comp.Ratio ↑ | F-score ↑ |
|------------|------------|------------|------------|--------------|-------------|
| Voxblox | 7.1 | 1.8 | 4.8 | 84.0 | 90.9 |
| VDB Fusion | 6.9 | 1.3 | 4.5 | 90.2 | 94.1 |
| Puma | 32.0 | 1.2 | 16.9 | 78.8 | 87.3 |
| Ours + DR | 3.3 | 1.5 | 3.7 | 94.0 | 90.7 |
| Ours | 3.2 | 1.1 | 2.9 | 95.2 | 95.9 |

efficiency of our approach and compare the results to those of previous methods. Second, we use the KITTI odometry dataset [6] to validate the scalability for incremental mapping. Finally, we showcase that our method can also achieve high-fidelity 3D reconstruction indoors. In all our experiments, we set the parameters as Tab. I.

B. Mapping Quality

This first experiment evaluates the mapping quality in terms of accuracy and completeness on the *MaiCity* and the *Newer College* dataset. We compare our approach against three other mapping systems: two state-of-the-art TSDF fusion-based methods, namely Voxblox [21] and VDB Fusion [36] as well as against the Possion surface-based reconstruction system Puma [35]. All three methods are scalable to outdoor LiDAR mapping, and source code is available. With our data structure, we additionally implemented differentiable rendering that is used in recent neural mapping systems [2], [37], [40] to supervise our map representation. We denote it as Ours + DR in the experiment results.

Although our approach can directly infer the SDF at an arbitrary position, we reconstruct the 3D mesh by marching cubes on the same fixed-size grid to have a fair comparison against the previous methods relying on marching cubes. In other words, we regard the 3D reconstruction quality of the resulting mesh as the mapping quality. For the quantitative assessment, we use the commonly used reconstruction metrics [14] calculated from the ground truth and predicted mesh, namely accuracy, completion, Chamfer-L1 distance, completion ratio, and F-score. Instead of the unfair accuracy metric used to calculate the Chamfer-L1 distance, we report

TABLE III: Quantitative results of the reconstruction quality on the *Newer College* dataset. We report completion, accuracy and Chamfer-L1 in cm. Additionally, we show the completion ratio and F-score in % calculated with a 20 cm error threshold.

| Method | Comp. ↓ | Acc. ↓ | C-l1 ↓ | Comp.Ratio ↑ | F-score ↑ |
|------------|-------------|------------|------------|--------------|-------------|
| Voxblox | 14.9 | 9.3 | 12.1 | 87.8 | 87.9 |
| VDB Fusion | 12.0 | 6.9 | 9.4 | 91.3 | 92.6 |
| Puma | 15.4 | 7.7 | 11.5 | 89.9 | 91.9 |
| Ours + DR | 11.4 | 11.1 | 11.2 | 92.5 | 86.1 |
| Ours | 10.0 | 6.7 | 8.4 | 93.6 | 93.7 |

the reconstruction accuracy calculated using the ground truth mesh masked by the intersection area of the 3D reconstruction from all the compared methods.

Tab. II lists the obtained results. When fixing the voxel size or feature grid size, here to a size of 10 cm, our method outperforms all baselines, the non-learning as well as the learnable neural rendering-based ones. The superiority of our method is also visible in the results depicted in Fig. 5. As can be seen in the first row, our method has the most complete and smoothest reconstruction, visible, for example, when inspecting the tree or the pedestrian. The error map depicted in the second row also backs up the claims that our method can achieve better mapping accuracy in areas with dense measurements and higher completeness in areas with sparse observations compared to the state-of-the-art approaches.

As shown in Tab. III and Fig. 6, we get better performance compared with the other methods on the more noisy *Newer College* dataset with the same 10 cm voxel size. The results indicate that our method can handle measurement noise well. Besides, our approach is able to build the map, eliminating the dynamic objects such as the moving pedestrians, while not deleting the sparse static objects such as the trees. This is difficult to achieve for the space carving in Voxblox and VDB Fusion and the density thresholding in Puma.

C. Memory Efficiency

Fig. 7 depicts the memory usage in relation to the mapping quality. The same set of mapping voxel size settings from 10 to 100 cm are used for the methods. The results indicate that our method can create map with smaller memory consistently for all settings. Meanwhile, our mapping quality hardly gets

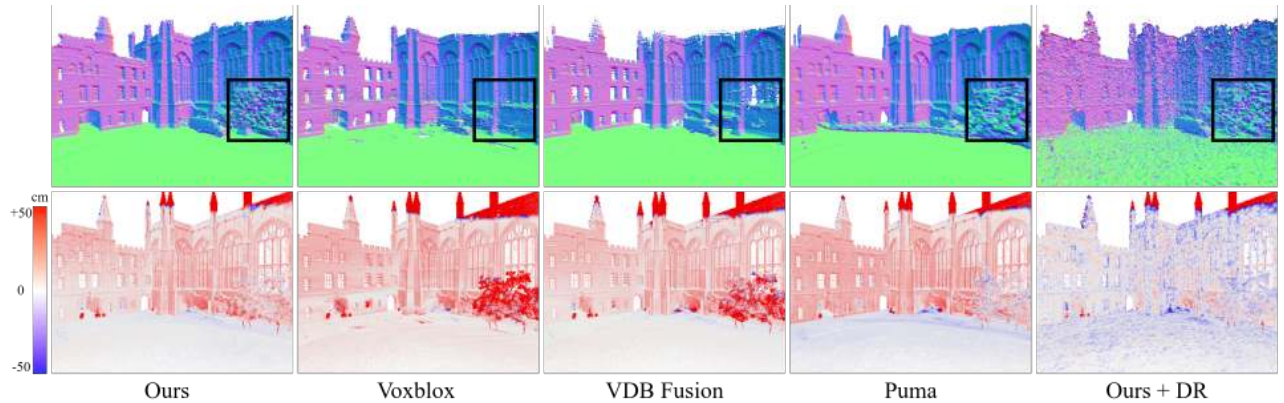


Fig. 6: A comparison of different methods on the *Newer College* dataset. The first row shows the reconstructed mesh and a tree is highlighted in the black box. The second row shows the error map of the reconstruction overlaid on the ground truth mesh, where the blue to red colormap indicates the signed reconstruction error from -50 cm to +50 cm.

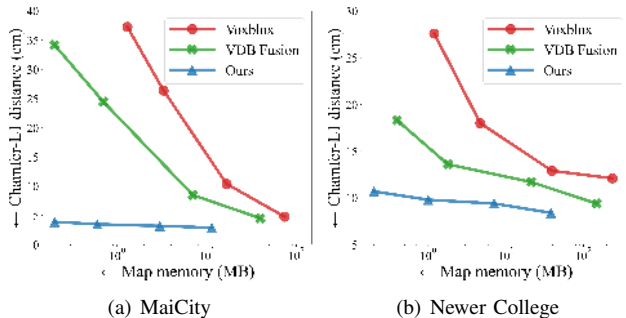


Fig. 7: Comparison of map memory efficiency versus the reconstruction Chamfer-L1 distance for different voxel sizes (10 cm, 20 cm, 50 cm and 100 cm). For our SHINE-Mapping, the voxel size represents the octree’s leaf node grid resolution.

worse with a lower feature grid resolution, while the mapping error of Voxblox and VDB Fusion increases significantly.

Our representation using hash tables storing features allows for efficient memory usage. Although the sparse data structures are also used in Voxblox and VDB Fusion, they need to allocate memory with the same voxel size for the truncation band close to the surface and even the free space covered by the measurement rays with the space carving.

D. Scalable Incremental Mapping

The next experiment showcases exemplarily the ability of SHINE-Mapping to scale to larger environments, even when performing the incremental mapping. For this, we use the KITTI dataset. As shown in Fig. 1, our method reconstructs a driving sequence over a distance of about 4 km with the incremental update of the hierarchical feature grids using the regularization-based continual learning strategy.

Additionally, we provide a qualitative comparison between our incremental mapping with and without the feature update regularization in Fig. 8. The regularized approach manages to clearly reconstruct the structures that may vanish or distort as the consequences of forgetting during the incremental mapping without regularization.

E. Indoor Mapping and Filling Occluded Areas

Lastly, we provide an example for indoor mapping by our approach. Fig. 9 shows our lab in Bonn, reconstructed from sequential point clouds. We used a 3 cm leaf node resolution

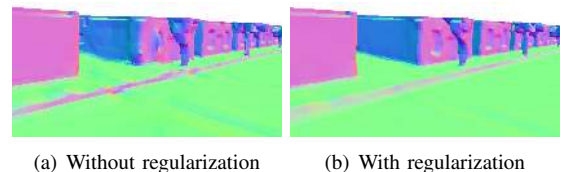


Fig. 8: A comparison of the incremental mapping results with and without regularization. The resolution of feature leaf nodes is 50 cm.



Fig. 9: Indoor mapping of the IPB office dataset. SHINE-Mapping’s reconstruction of the whole floor is shown on the left. A close-up view of the reconstructed mesh and the input point cloud of one room are shown on the right. Our approach manages to conduct reasonable scene completion in the highlighted box.

instead of 10 cm used outdoors to cover the details of indoor environments better. The map memory here is only 84 MB. Our method can achieve not only the detailed reconstruction of the furnitures or the approx. 1 cm thick whiteboard on the wall, but also a reasonable scene completion in the occluded areas where no observation are available.

V. CONCLUSION

This paper presents a novel approach to large-scale 3D SDF mapping using range sensors. Our model does not explicitly store signed distance values in a voxel grid. Instead, it uses an octree-based implicit representation consisting of features stored in hash tables and which can, through a neural network, be turned into SDF values. The network and the features can be learned end-to-end from range data. We evaluated our approach on both simulated and real-world datasets and show our reconstruction approach has advantages over current state-of-the-art mapping systems. The experiments suggest that our method achieves more accurate and complete 3D reconstruction with lower map memory than the compared methods while operating online incrementally. Furthermore, our approach can provide a reasonable guess about the structure for regions not covered by the sensor, for example, due to occlusions.

REFERENCES

- [1] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018.
- [2] D. Azinović, R. Martin-Brualla, D.B. Goldman, M. Nießner, and J. Thies. Neural rgb-d surface reconstruction. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [3] J. Behley and C. Stachniss. Efficient Surfel-Based SLAM using 3D Laser Range Data in Urban Environments. In *Proc. of Robotics: Science and Systems (RSS)*, 2018.
- [4] X. Chen, I. Vizzo, T. Läbe, J. Behley, and C. Stachniss. Range Image-based LiDAR Localization for Autonomous Vehicles. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2021.
- [5] B. Curless and M. Levoy. A Volumetric Method for Building Complex Models from Range Images. In *Proc. of the Intl. Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 303–312. ACM, 1996.
- [6] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012.
- [7] A. Hornung, K. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard. OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees. *Autonomous Robots*, 34:189–206, 2013.
- [8] J. Huang, S.S. Huang, H. Song, and S.M. Hu. Di-fusion: Online implicit 3d reconstruction with deep priors. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [9] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A.A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
- [10] T. Kühner and J. Kümmerle. Large-Scale Volumetric Scene Reconstruction using LiDAR. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2020.
- [11] W. Lorensen and H. Cline. Marching Cubes: a High Resolution 3D Surface Construction Algorithm. In *Proc. of the Intl. Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 163–169, 1987.
- [12] Z. Marton, R. Rusu, and M. Beetz. On fast surface reconstruction methods for large and noisy point clouds. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2009.
- [13] M. McCloskey and N.J. Cohen. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.
- [14] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [15] A. Milan, T. Pham, V. G. D. Morrison, A. Tow, L. Liu, J. Erskine, R. Grinover, A. Gurman, T. Hunn, N. Kelly-Boxall, D. Lee, M. McTaggart, G. Rallos, A. Razjigaev, T. Rowntree, T. Shen, R. Smith, S. Wade-McCue, Z. Zhuang, C. Lehnert, G. Lin, I. Reid, P. Corke, and J. Leitner. Semantic segmentation from limited training data. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2018.
- [16] B. Mildenhall, P. Srinivasan, M. Tancik, J. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2020.
- [17] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):102:1–102:15, 2022.
- [18] R.A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A.J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *Proc. of the Intl. Symposium on Mixed and Augmented Reality (ISMAR)*, 2011.
- [19] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3D Reconstruction at Scale using Voxel Hashing. *Proc. of the SIGGRAPH Asia*, 32(6), 2013.
- [20] M. Oechsle, S. Peng, and A. Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2021.
- [21] H. Oleynikova, Z. Taylor, M. Fehr, R. Siegwar, and J. Nieto. Voxblox: Incremental 3d euclidean signed distance fields for on-board map planning. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 1366–1373, 2017.
- [22] H. Oleynikova, Z. Taylor, R. Siegwart, and J. Nieto. Safe local exploration for replanning in cluttered and unknown environments for micro-aerial vehicles. *IEEE Robotics and Automation Letters (RA-L)*, 2018.
- [23] J. Ortiz, A. Clegg, J. Dong, E. Sucar, D. Novotny, M. Zollhoefer, and M. Mukadam. isdf: Real-time neural signed distance fields for robot perception. In *Proc. of Robotics: Science and Systems (RSS)*, 2022.
- [24] E. Palazzolo, J. Behley, P. Lottes, P. Giguere, and C. Stachniss. ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019.
- [25] Y. Pan, Y. Kompis, L. Bartolomei, R. Mascaro, C. Stachniss, and M. Chli. Voxfield: Non-projective signed distance fields for online planning and 3d reconstruction. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2022.
- [26] J.J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [27] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger. Convolutional occupancy networks. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*. Springer, 2020.
- [28] M. Ramezani, Y. Wang, M. Camurri, D. Wisth, M. Mattamala, and M. Fallon. The newer college dataset: Handheld lidar, inertial and vision with ground truth. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2020.
- [29] F. Ruetz, E. Hernández, M. Pfeiffer, H. Oleynikova, M. Cox, T. Lowe, and P. Borges. Oypc mesh: 3d free-space representation for local ground vehicle navigation. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2019.
- [30] C. Stachniss and W. Burgard. Mapping and Exploration with Mobile Robots using Coverage Maps. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2003.
- [31] E. Sucar, S. Liu, J. Ortiz, and A.J. Davison. imap: Implicit mapping and positioning in real-time. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2021.
- [32] J. Sun, X. Chen, Q. Wang, Z. Li, H. Averbuch-Elor, X. Zhou, and N. Snavely. Neural 3D reconstruction in the wild. In *Proc. of the Intl. Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2022.
- [33] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [34] T. Takikawa, J. Litalien, K. Yin, K. Kreis, C. Loop, D. Nowrouzezahrai, A. Jacobson, M. McGuire, and S. Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [35] I. Vizzo, X. Chen, N. Chebrolu, J. Behley, and C. Stachniss. Poisson Surface Reconstruction for LiDAR Odometry and Mapping. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2021.
- [36] I. Vizzo, T. Guadagnino, J. Behley, and C. Stachniss. Vdbfusion: Flexible and efficient tsdf integration of range sensor data. *Sensors*, 22(3), 2022.
- [37] J. Wang, T. Bleja, and L. Agapito. Go-surf: Neural feature grid optimization for fast, high-fidelity rgb-d surface reconstruction. In *Proc. of the Intl. Conf. on 3D Vision (3DV)*, 2022.
- [38] T. Whelan, M. Kaess, H. Johannsson, M. Fallon, J.J. Leonard, and J. McDonald. Real-time large scale dense RGB-D SLAM with volumetric fusion. *Intl. Journal of Robotics Research (IJRR)*, 34(4-5):598–626, 2014.
- [39] F. Zenke, B. Poole, and S. Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995, 2017.
- [40] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M.R. Oswald, and M. Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.