Unified Neural Scene Representation For Robotics

Yue Pan
Center for Robotics, University of Bonn

I. Motivation and Related Work

The ability to perceive and understand the environment is fundamental for robots operating in complex real-world scenarios. At the core of this capability lies the construction of maps — building a digital twin of the robot's workspace based on sensor observations. Such scene representations serve as the foundation for the robot's spatial awareness, enabling accurate state estimation and safe interaction with its surroundings [15].

Throughout the years, researchers have proposed various scene representations optimized for specific applications. My research aims to develop a unified scene representation for diverse robotic tasks, eliminating redundant task-specific maps.

Over the past decades, explicit scene representations have been widely used in robotics for tasks such as localization [33], planning [3], and exploration [31]. These methods represent scenes using landmark feature points [4], dense point clouds [38], surfels [2], meshes [34], or volumetric grids storing occupancy [5] or signed distance functions [19]. While point-based methods enable localization, volumetric representations better support planning and obstacle avoidance. Research has focused on improving the scalability of volumetric representations [6, 20, 35] and enabling incremental mapping [21, 24]. Beyond geometric mapping, recent researches have extended volumetric maps to capture semantics [27, 30], topological structure [8], and temporal dynamics [29] of the scene. However, most of explicit volumetric representations face two key limitations: their fixed spatial resolution creates a fidelity-scalability trade-off, and their rigid grid structure prevents deformation during online pose adjustments caused by loop closures or multi-agent interactions.

Recently, implicit neural representations have been proposed to leverage neural networks to implicitly encode either geometry [16, 26], appearance [17], or semantics [39] into compact latent spaces, enabling perception systems to learn directly from raw sensor data and fine-grained scene details continuously. These continuous representations offer advantages like compact storage, and better handling of regions with sparse observations, while supporting conversion to explicit representations for downstream tasks. To improve scalability and enable incremental mapping, recent methods [18, 37] employ hybrid architectures that combine explicitly stored local features (in regular grids or sparse point sets) with shared MLP decoders to map these features to various scene properties like volume density, distance field, or color. Point-based implicit neural representations [1, 28, 37] store optimizable features in a neural point cloud, which has advantages over grid-based alternatives through its flexible layout and inherent elasticity under transformations for example caused by loop closures.

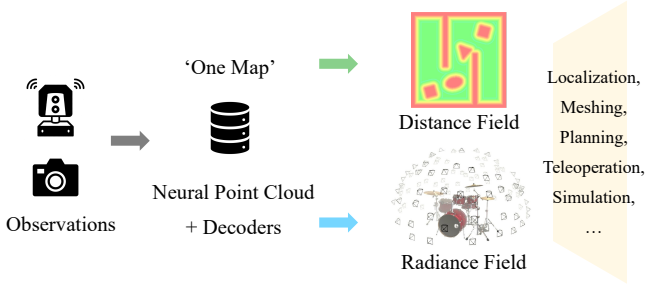

Fig. 1: Our unified neural scene representation for robotics to achieve 'one map to rule them all'.

My research leverages point-based implicit neural maps, also known as neural point map, to develop a unified scene representation that enables diverse downstream tasks such as localization, mesh reconstruction, navigation, teleoperation, and simulation for robotics learning, as shown in Fig. 1. Using this representation, I develop SLAM systems that incrementally build globally consistent neural point maps while achieving accurate state estimation and scene reconstruction. Compared to concurrent works [28], my research focuses on large-scale scenes, improved computational efficiency, and maintaining global consistency during long-term missions. My contributions are summarized as follows: (i) A neural point-based distance field that enables accurate and globally consistent LiDAR SLAM with loop closure corrections [25], demonstrating extensibility through integration of appearance and semantic information for metric-semantic mapping. (ii) An extension of [25] by unifying distance fields and Gaussian splatting radiance fields within a neural point map with mutual geometric consistency, implemented in a LiDAR-visual SLAM system that achieves superior accuracy in localization, reconstruction, and novel view rendering [23]. (iii) A method to handle object-level submaps with pretrained shape priors, enabling pose estimation and shape completion. This is demonstrated through panoptic mapping in greenhouse environments [22], improving fruit harvesting performance [14].

II. Preliminary Results

A. LiDAR SLAM with Point-based Neural Distance Field

In [25], we presented a novel LiDAR SLAM system that achieves large-scale globally consistent mapping using a compact point-based implicit neural scene representation. Our approach alternates between online incremental learning of a local implicit map modeling the distance field using LiDAR observations [40] and odometry estimation via correspondence-free point-to-implicit distance field registration [36]. By using

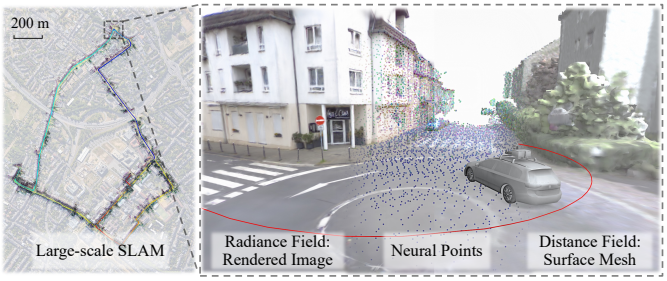

Fig. 2: Unified neural point map built by the proposed SLAM system. Taking image and LiDAR data as input, the system incrementally builds a compact and globally consistent map modeling the radiance field and distance field, supporting various downstream tasks.

sparse neural points as local feature embeddings that are inherently elastic and deformable, we can effectively maintain global consistency of both the neural points and underlying distance field through loop closure corrections. Extensive evaluations on various datasets demonstrate that our SLAM system achieves superior or comparable localization accuracy to state-of-the-art methods while building more consistent and compact implicit maps that enable more accurate and complete mesh reconstruction. The system runs at sensor frame rate on a moderate GPU. We further demonstrate the versatility of our map representation through extensions to metric-semantic mapping by encoding semantic labels into the neural points. Additional works showcase its applicability to modeling 4D dynamic scenes [41] and enabling efficient 2D MCL [12].

B. Unifying Distance and Radiance Fields in Neural Maps

Though the distance field representation in [25] enables accurate localization, surface reconstruction and obstacle avoidance, it falls short of providing photorealistic novel view rendering, which is crucial for applications requiring dense photometric information. Inspired by Scaffold-GS [13], in the paper [23], we develop a novel point-based model that additionally represents a Gaussian splatting radiance field which enables real-time rendering. Each neural point spawn multiple Gaussian surfel primitives with its latent geometric and appearance features and globally shared MLP decoders. By enforcing geometric consistency between the distance field and the radiance field, we achieve mutual improvements: the distance field provides geometric structure to guide radiance field optimization, while the radiance field's dense photometric cues and multi-view consistency enhance the distance field's accuracy in regions with sparse LiDAR measurements. We further develop a LiDAR-visual SLAM system using such a unified scene representation, as shown in Figure 2. Experimental results on challenging datasets demonstrate that our method incrementally constructs globally consistent maps that outperform baseline methods in novel view rendering fidelity, surface reconstruction quality, odometry estimation accuracy, and map memory efficiency. The resulting Gaussian splatting radiance field enables effective active scene reconstruction [10].

C. Object-level Submap using Pre-trained Shape Priors

While the aforementioned works treat the scene as a whole without leveraging pre-trained priors, real-world environments often contain multiple objects-of-interest with characteristic shapes and structures. For many robotic tasks like manipulation and navigation, accurately estimating the pose and complete shape of these objects is crucial. We represent each object-of-interest as a separate submap with its own latent shape feature and pre-trained MLP decoder modeling the shape priors. This enables joint optimization of object pose and shape features using accumulated observations. We demonstrate this approach through panoptic mapping of fruits in greenhouse environments for agricultural robotics [22]. Using a multi-resolution representation [30], we model fruits at high resolution while representing background vegetation at lower resolution. Our method leverages high-precision 3D scans of fruits to learn generalizable fruit shape priors offline, which are combined with an occlusion-aware differentiable rendering pipeline during online inference. This enables accurate completion of partial fruit observations and estimation of 7-DoF fruit poses within the map. The resulting scene representation has been applied to robotic fruit grasping tasks [14] with a better fruit harvesting success rate than previous methods.

III. Future Work

I plan to continue working on the following topics towards an unified neural scene representation used by more robust and scalable spatial perception systems for long-term autonomy.

High-level Semantics and Physical Properties. Beyond geometric and appearance information, I aim to extend the neural point representation to encode high-level scene semantics and physical properties. While my previous works [22, 25] integrated pre-defined semantic labels, I plan to leverage large visual-language models for open-vocabulary scene understanding [7]. This would enable language-based spatial reasoning with the optimized features of neural points, advancing towards more sophisticated embodied intelligence. By incorporating multi-modal sensing [32] (e.g., tactile and multi-spectral data) and visual inference with diffusion priors [11], one can encode physical properties like materials and affordances. These rich scene representations would enhance planning and manipulation for online tasks while helping close the sim-to-real gap for robotics learning. As shown in [23], enforcing consistency among different modalities can further improve the overall map quality through their complementary nature.

Neural Point Hierarchy. Currently, the single-resolution neural point map requires dense point allocation in freespace for path planning and obstacle avoidance, resulting in high memory usage. I plan to develop a multi-resolution hierarchy where freespace can be efficiently represented using sparse neural points at Voronoi graph nodes. This hierarchy could also encode higher-level spatial abstractions, enabling more compact storage and improved spatial reasoning for planning [9].

Long-term Persistence. To enable long-term autonomy, the scene representation must be both persistent and continuously updated, capable of accommodating environmental changes across multiple timescales with temporal consistency. I plan to use the object-level submaps for efficient long-term change detection, object retrieval and map update like in [29, 42].

References

[1] J. Abou-Chakra, F. Dayoub, and N. Sünderhauf. ParticleNeRF: A Particle-Based Encoding for Online Neural Radiance Fields. In *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2024.

[2] 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.

[3] A. Bircher, M. Kamel, K. Alexis, H. Oleynikova, and R. Siegwart. Receding Horizon "Next-Best-View" Planner for 3D Exploration. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2016.

[4] C. Campos, R. Elvira, J.J.G. Rodríguez, J.M. Montiel, and J.D. Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Trans. on Robotics (TRO)*, 37(6):1874–1890, 2021.

[5] G. Grisetti, C. Stachniss, and W. Burgard. Improved Techniques for Grid Mapping with Rao-Blackwellized Particle Filters. *IEEE Trans. on Robotics (TRO)*, 23(1):34–46, 2007.

[6] A. Hornung, K. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard. OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees. *Autonomous Robots*, 34(3):189–206, 2013.

[7] C. Huang, O. Mees, A. Zeng, and W. Burgard. Visual Language Maps for Robot Navigation. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2023.

[8] N. Hughes, Y. Chang, and L. Carlone. Hydra: A Real-time Spatial Perception System for 3D Scene Graph Construction and Optimization. In *Proc. of Robotics: Science and Systems (RSS)*, 2022.

[9] N. Hughes, Y. Chang, S. Hu, R. Talak, R. Abdulhai, J. Strader, and L. Carlone. Foundations of spatial perception for robotics: Hierarchical representations and real-time systems. *Intl. Journal of Robotics Research (IJRR)*, 43(10):1457–1505, 2024.

[10] L. Jin, X. Zhong, Y. Pan, J. Behley, C. Stachniss, and M. Popovic. ActiveGS: Active Scene Reconstruction using Gaussian Splatting. *IEEE Robotics and Automation Letters (RA-L)*, 10(5):4866–4873, 2025.

[11] B. Ke, K. Qu, T. Wang, N. Metzger, S. Huang, B. Li, A. Obukhov, and K. Schindler. Marigold: Affordable Adaptation of Diffusion-Based Image Generators for Image Analysis. *arXiv preprint*, arXiv:2505.09358, 2025.

[12] H. Kuang, Y. Pan, X. Zhong, L. Wiesmann, J. Behley, and C. Stachniss. Improving Indoor Localization Accuracy by Using an Efficient Implicit Neural Map Representation. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2025.

[13] T. Lu, M. Yu, L. Xu, Y. Xiangli, L. Wang, D. Lin, and B. Dai. Scaffold-GS: Structured 3d gaussians for view-adaptive rendering. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[14] F. Magistri, Y. Pan, J. Bartels, J. Behley, C. Stachniss, and C. Lehnert. Improving Robotic Fruit Harvesting Within Cluttered Environments Through 3D Shape Completion. *IEEE Robotics and Automation Letters (RA-L)*, 9(8):7357–7364, 2024.

[15] R. Mascaro and M. Chli. Scene Representations for Robotic Spatial Perception. *Annual Review of Control, Robotics, and Autonomous Systems*, 8(5):1–27, 2024.

[16] 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.

[17] 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.

[18] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. on Graphics*, 41(4):102:1–102:15, 2022.

[19] 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.

[20] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-Time 3D Reconstruction at Scale Using Voxel Hashing. *ACM Trans. on Graphics (TOG)*, 32(6):1–11, 2013.

[21] H. Oleynikova, Z. Taylor, M. Fehr, R. Siegwar, and J. Nieto. Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2017.

[22] Y. Pan, F. Magistri, T. Läbe, E. Marks, C. Smitt, C. McCool, J. Behley, and C. Stachniss. Panoptic Mapping with Fruit Completion and Pose Estimation for Horticultural Robots. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2023.

[23] Y. Pan, X. Zhong, L. Jin, L. Wiesmann, M. Popović, J. Behley, and C. Stachniss. PINGS: Gaussian Splatting Meets Distance Fields within a Point-Based Implicit Neural Map. In *Proc. of Robotics: Science and Systems (RSS)*, 2025.

[24] 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.

[25] Y. Pan, X. Zhong, L. Wiesmann, T. Posewsky, J. Behley, and C. Stachniss. PIN-SLAM: LiDAR SLAM Using a Point-Based Implicit Neural Representation for Achieving Global Map Consistency. *IEEE Trans. on Robotics (TRO)*, 40:4045–4064, 2024.

[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] A. Rosinol, M. Abate, Y. Chang, and L. Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2020.

[28] E. Sandström, Y. Li, L. Van Gool, and M. R. Oswald. Point-SLAM: Dense Neural Point Cloud-based SLAM. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2023.

[29] L. Schmid, M. Abate, Y. Chang, and L. Carlone. Khronos: A Unified Approach for Spatio-Temporal Metric-Semantic SLAM in Dynamic Environments. In *Proc. of Robotics: Science and Systems (RSS)*, 2024.

[30] L. Schmid, J. Delmerico, J. Schönberger, J. Nieto, M. Pollefeys, R. Siegwart, and C. Cadena. Panoptic multi-tsdfs: a flexible representation for online multi-resolution volumetric mapping and long-term dynamic scene consistency. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2022.

[31] C. Stachniss, G. Grisetti, and W. Burgard. Information Gain-based Exploration Using Rao-Blackwellized Particle Filters. In *Proc. of Robotics: Science and Systems (RSS)*, 2005.

[32] S. Suresh, H. Qi, T. Wu, T. Fan, L. Pineda, M. Lambeta, J. Malik, M. Kalakrishnan, R. Calandra, M. Kaess, J. Ortiz, and M. Mukadam. NeuralFeels with Neural Fields: Visuotactile Perception for In-hand Manipulation. *Science Robotics*, 9(96), 2024.

[33] S. Thrun, D. Fox, W. Burgard, and F. Dellaert. Robust Monte Carlo Localization for Mobile Robots. *Artificial Intelligence*, 128(1-2), 2001.

[34] 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.

[35] I. Vizzo, T. Guadagnino, J. Behley, and C. Stachniss. VDBFusion: Flexible and Efficient TSDF Integration of Range Sensor Data. *Sensors*, 22(3):1296, 2022.

[36] L. Wiesmann, T. Guadagnino, I. Vizzo, N. Zimmerman, Y. Pan, H. Kuang, J. Behley, and C. Stachniss. LocNDF: Neural Distance Field Mapping for Robot Localization. *IEEE Robotics and Automation Letters (RA-L)*, 8(8):4999–5006, 2023.

[37] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann. Point-NeRF: Point-Based Neural Radiance Fields. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[38] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang. FAST-LIO2: Fast Direct LiDAR-Inertial Odometry. *IEEE Trans. on Robotics (TRO)*, 38(4):2053–2073, 2022.

[39] S. Zhi, T. Laidlow, S. Leutenegger, and A. Davison. In-Place Scene Labelling and Understanding with Implicit Scene Representation. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2021.

[40] X. Zhong, Y. Pan, J. Behley, and C. Stachniss. SHINE-Mapping: Large-Scale 3D Mapping Using Sparse Hierarchical Implicit Neural Representations. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2023.

[41] X. Zhong, Y. Pan, C. Stachniss, and J. Behley. 3D LiDAR Mapping in Dynamic Environments using a 4D Implicit Neural Representation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[42] L. Zhu, S. Huang, K. Schindler, and I. Armeni. Living Scenes: Multi-object Relocalization and Reconstruction in Changing 3D Environments. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.