iSDF: Real-Time Neural Signed Distance Fields for Robot Perception

Joseph Ortiz[1,2] Alexander Clegg[2] Jing Dong[3] Edgar Sucar[1]
David Novotny[2] Michael Zollhoefer[3] Mustafa Mukadam[2]

***Abstract*— We present iSDF, a continual learning system for real-time signed distance field (SDF) reconstruction. Given a stream of posed depth images from a moving camera, it trains a randomly initialised neural network to map input 3D coordinate to approximate signed distance. The model is self-supervised by minimising a loss that bounds the predicted signed distance using the distance to the closest sampled point in a batch of query points that are actively sampled. In contrast to prior work based on voxel grids, our neural method is able to provide adaptive levels of detail with plausible filling in of partially observed regions and denoising of observations, all while having a more compact representation. In evaluations against alternative methods on real and synthetic datasets of indoor environments, we find that iSDF produces more accurate reconstructions, and better approximations of collision costs and gradients useful for downstream planners in domains from navigation to manipulation. Code and video results can be found at our project page: https://joeaortiz.github.io/iSDF/.**

I. Introduction

Signed distance fields are a common map representations in both robotics and vision that associate each point in space with the signed distance to the closest surface, encoding the surface as the zero level set. In robotics, SDFs are a commonly used for collision avoidance in motion planning [24], [20], [17], [4], [12], however they are usually assumed as given a priori or too expensive to compute in real-time. On the other hand, in real-time vision systems, depth fusion is commonly used to produce truncated signed distance fields (TSDF) [14], [13], but there is little work on reconstructing non-truncated SDFs in room scale environments. In order to close this gap, in this work, we focus on the problem of real-time full SDF reconstruction.

Prior work on real-time reconstruction of non-truncated SDFs [15], [7] (most notably Voxblox [15]) operates in two stages, first reconstructing surface geometry and then transforming it to the SDF using wavefront propagation algorithms. These methods are based on voxel grids and are limited to a resolution of 5cm by real-time constraints due to the cost of wavefront propagation.

Neural fields provide an alternative paradigm to voxel grids for modelling scenes [16], [10], [11] and can be optimised from scratch to accurately fit a specific scene. Building on recent advances, we present iSDF (*incremental Signed Distance Fields*), a system for real-time SDF reconstruction based on a neural signed distance field that maps a 3D coordinate $\mathbf{x}$ to the signed distance value s, $f(\mathbf{x};\theta) = s$. Specifically we build two advances; iMAP [18] which demonstrated that neural radiance fields can be trained in real-time as part of a SLAM system and the body of work using neural SDFs for object modelling [6], [23], [8], [22], [21], [9], [1].

[1]Imperial College London, [2]Meta AI, [3]Reality Labs Research.
joeaortiz16@gmail.com

II. iSDF - Real-time SDF reconstruction

iSDF is a system for real-time SDF reconstruction that takes as input a stream of posed depth images captured by a moving camera and, during online operation, optimises a neural SDF of the environment. We assume an external tracking module and focus on mapping.

The SDF is modelled by a randomly initialised MLP with 4 hidden layer and the "off-axis" positional embedding [2]. At each iteration, we select a subset of frames, i.e. a few posed depth images. Given the continual learning setting, we follow iMAP [18] and select frames via active sampling from a sparse set of representative keyframes to mitigate catastrophic forgetting. For each frame, we sample 200 random pixels and then, along each backprojected ray, generate 20 stratified samples, 8 Gaussian samples around the measured depth and one sample at the depth. These points for all sampled pixels form the batch of points that are fed to the MLP.

To supervise the predicted signed distances, we propose to use the distance to the closest surface point in the batch. This distance b will be larger than the true signed distance and so we use it to bound the prediction via the loss:

$$\mathscr{L}_{\text{sdf}}(f(\mathbf{x};\theta), b) = \max\left(0,\ e^{-\beta f(\mathbf{x};\theta)} - 1,\ f(\mathbf{x};\theta) - b\right) . \quad (1)$$

Close to the surface, we expect the bound to be very tight and so we instead directly supervise the SDF prediction to take the bound value. The closest surface point in the batch can also provide signal to supervise the spatial gradient of the predicted signed distance and we introduce a loss $\mathscr{L}_{\text{grad}}$ penalising the cosine distance between the predicted signed distance and the vector from the closest surface point to the query point. Following Gropp et al. [6] we also apply Eikonal regularisation: $\mathscr{L}_{\text{eik}}(f(\mathbf{x};\theta)) = |\|\nabla_{\mathbf{x}} f(\mathbf{x};\theta)\| - 1|$. The network parameters are optimised to minimise the loss:

$$l(\theta) = \mathscr{L}_{\text{sdf}} + 0.02\,\mathscr{L}_{\text{grad}} + 0.25\,\mathscr{L}_{\text{eik}} . \quad (2)$$

III. Results

Experimental setup. We experiment with sequences from the ScanNet dataset [3] and the synthetic ReplicaCAD dataset [19] (in which sequences are generated by simulating the measurements by a camera mounted on a mobile manipulator). To evaluate the SDF reconstruction, we precompute a voxel grid (resolution 1cm) of ground truth SDF

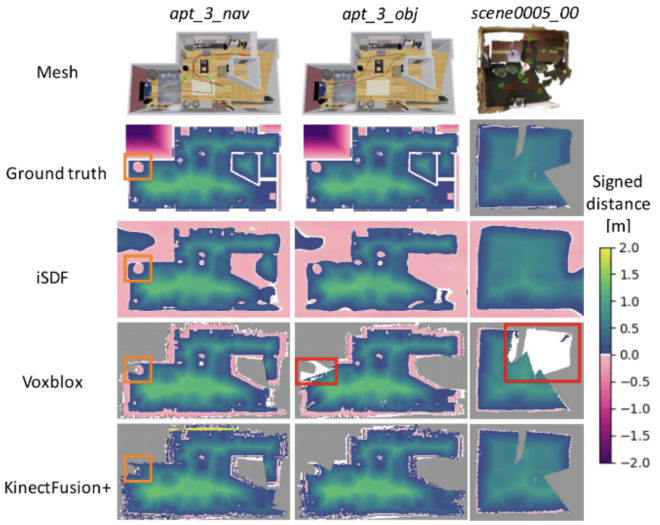

Fig. 1. Slices at constant height of the reconstructed SDF at the end of the sequence. The meshes are shown for reference with the camera trajectory and keyframes selected by iSDF overlaid. For Voxblox and KinectFusion+, the slices are greyed out in the non-visible region as neither method makes predictions in this region. The ground truth ScanNet [3] slices are greyed out in the non-visible regions as we only have ground truth SDF values in visible regions. White regions in the Voxblox and KinectFusion+ slices are regions that are visible but unmapped.

values by voxelizing the mesh into an occupancy grid and then computing the euclidean distance transform (EDT). We compare iSDF against two voxel-based methods for real-time SDF reconstruction: (1) Voxblox [15] and (2) KinectFusion+. Both methods operate in two stages, first reconstructing the surface before computing the SDF via wavefront propagation. Voxblox is a CPU only method and to the best of our knowledge there are no GPU methods for real-time SDF reconstruction. Consequently, KinectFusion+ is our own GPU implementation in C++ with custom CUDA kernels for fusion and the efficient EDT [5].

Qualitative results. To visualise the reconstructed SDFs, we compare 2D slices of the SDFs at constant heights in Fig. 1. Unlike Voxblox and KinectFusion+ which only map the visible region, iSDF is predictive and reconstructs a plausible field in the full environment. For example, in the interiors of the beanbag (highlighted by the orange boxes in Fig. 1) or the interior of a wall, iSDF is the only method that returns predictions. As Voxblox and KinectFusion+ employ fusion in the first stage, there are holes in the reconstructions in regions that no rays reach, shown in white in the slices. These holes are problematic for downstream planners as they must be marked as unnavigable and given a high cost.

Quantitative results. In Fig. 2 we show that iSDF outperforms prior work in terms of SDF accuracy on both synthetic and real datasets in indoor environments. We also evaluate collision costs and gradients derived from signed distance and find that iSDF outperforms alternatives on these metrics, demonstrating its utility for downstream planners in domains from navigation to manipulation. The difference in accuracy is primarily due to the large voxel size used by Voxblox and KinectFusion+ for real-time performance.

Additional results. We provide further qualitative and

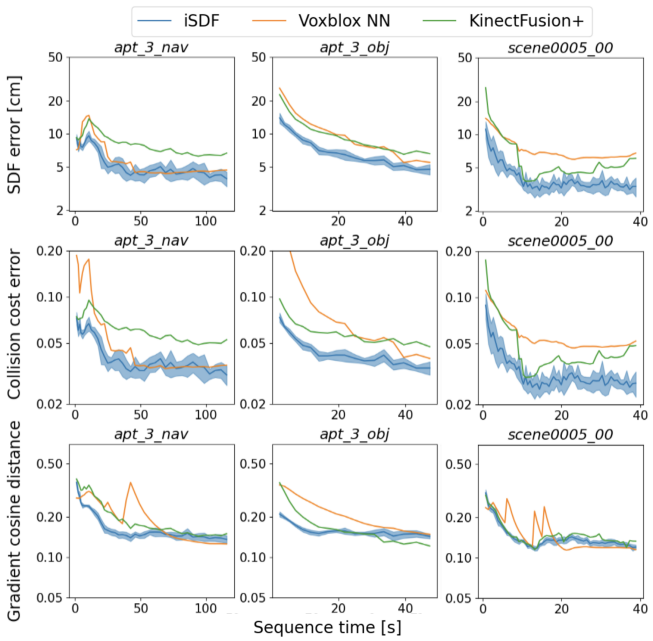

Fig. 2. We compare iSDF, Voxblox and KinectFusion+ along our three evaluation metrics: SDF error, collision cost error (using the cost function from CHOMP [24]) and cosine distance between the true and predicted spatial gradient of the SDF. The metrics are evaluated at regular fixed intervals during the sequences with 200k points sampled in the visible region. Results for iSDF are averaged over 10 runs with different network initialisations; Voxblox and KinectFusion+ are deterministic.

quantitative results for more sequences on our website (`https://joeaortiz.github.io/iSDF/`). We show that iSDF inherits the positive characteristics of neural fields. In contrast to prior work based on voxel grids, iSDF is: (i) efficient - it can seamlessly allocate memory capacity to model different parts of an environment with different levels of detail, and (ii) predictive - it can denoise and consolidate noisy measurements and sensibly fill in gaps in partially observed regions; all while requiring only 1MB of memory for the network weights. Results demonstrating these properties can be found on our project website.

IV. Limitations and Conclusions

There remain many important challenges in training neural fields in real-time, for example: handling dynamic environments, using local models to reduce replay, and designing more flexible positional embeddings. In addition, for downstream planning, it would be useful to have a measure of confidence associated with the iSDF predictions.

To conclude, we have presented iSDF, a continual learning system for real-time signed distance field reconstruction. Key to our approach is a novel loss that bounds the predicted signed distance using the distance to the closest batch surface point to enable learning signed distances away from surfaces. In evaluations against alternative methods on real and synthetic datasets, we show that iSDF produces more accurate reconstructions as well as better collision costs and gradients for downstream planners. We hope that the generality and differentiability of iSDF means it can easily be integrated into downstream robotics applications.

REFERENCES

[1] M. Atzmon and Y. Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2565–2574, 2020.

[2] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. *arXiv*, 2021.

[3] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scene. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] J. Dong, M. Mukadam, F. Dellaert, and B. Boots. Motion Planning as Probabilistic Inference using Gaussian Processes and Factor Graphs. In *Robotics: Science and Systems*, volume 12, page 4, 2016.

[5] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. Technical report, Cornell Computing and Information Science, 2004.

[6] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman. Implicit Geometric Regularization for Learning Shapes. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3569–3579, 2020.

[7] L. Han, F. Gao, B. Zhou, and S. Shen. Fiesta: Fast incremental euclidean distance fields for online motion planning of aerial robots. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, pages 4423–4430. IEEE, 2019.

[8] Y. Jiang, D. Ji, Z. Han, and M. Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1251–1261, 2020.

[9] S. Liu, Y. Zhang, S. Peng, B. Shi, M. Pollefeys, and Z. Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2019–2028, 2020.

[10] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019.

[11] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

[12] M. Mukadam, J. Dong, X. Yan, F. Dellaert, and B. Boots. Continuous-time Gaussian process motion planning via probabilistic inference. *The International Journal of Robotics Research*, 37(11):1319–1340, 2018.

[13] R. A. Newcombe, D. Fox, and S. M. Seitz. DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[14] 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 *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011.

[15] H. Oleynikova, Z. Taylor, M. Fehr, R. Siegwart, and J. Nieto. Voxblox: Incremental 3D Euclidean Signed Distance Fields for On-Board MAV Planning. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, 2017.

[16] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019.

[17] J. Schulman, Y. Duan, J. Ho, A. Lee, I. Awwal, H. Bradlow, J. Pan, S. Patil, K. Goldberg, and P. Abbeel. Motion planning with sequential convex optimization and convex collision checking. *The International Journal of Robotics Research*, 33(9):1251–1270, 2014.

[18] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison. iMAP: Implicit Mapping and Positioning in Real-Time. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[19] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra. Habitat 2.0: Training Home Assistants to Rearrange their Habitat. *arXiv preprint arXiv:2106.14405*, 2021.

[20] M. Toussaint. Robot trajectory optimization using approximate inference. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1049–1056, 2009.

[21] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

[22] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, R. Basri, and Y. Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Neural Information Processing Systems (NIPS)*, 2021.

[23] J. Zhang, Y. Yao, and L. Quan. Learning signed distance field for multi-view surface reconstruction. *arXiv preprint arXiv:2108.09964*, 2021.

[24] M. Zucker, N. Ratliff, A. D. Dragan, M. Pivtoraiko, M. Klingensmith, C. M. Dellin, J. A. Bagnell, and S. S. Srinivasa. Chomp: Covariant hamiltonian optimization for motion planning. *The International Journal of Robotics Research*, 32(9-10):1164–1193, 2013.